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 BEAST 2 program. Gamma category count was set 4. For the substitution model I used Blossum62. For the Markov Chain Monte Carlo parameters, number of generations was set to 1000000 and sampling tree every 200 generation. For the model, I used Yule model. For the birth rate I used Gamma.

Structure of Zaire Ebola sGP was doanloaded from PDB (ID 5KEM), for other proteins there aren't structures available in the PDB database, I used Alphafold 2 to predict the structure of

the protein. Then the proteins were docked to 70SGP2A aptamer. For energy of binding analysis, I used FoldX program.

Results:

Bayesian phylogenetic tree was generated and the energy of binding for each protein to the aptamer was calculated. The likelihood of tree is -5000.3988.

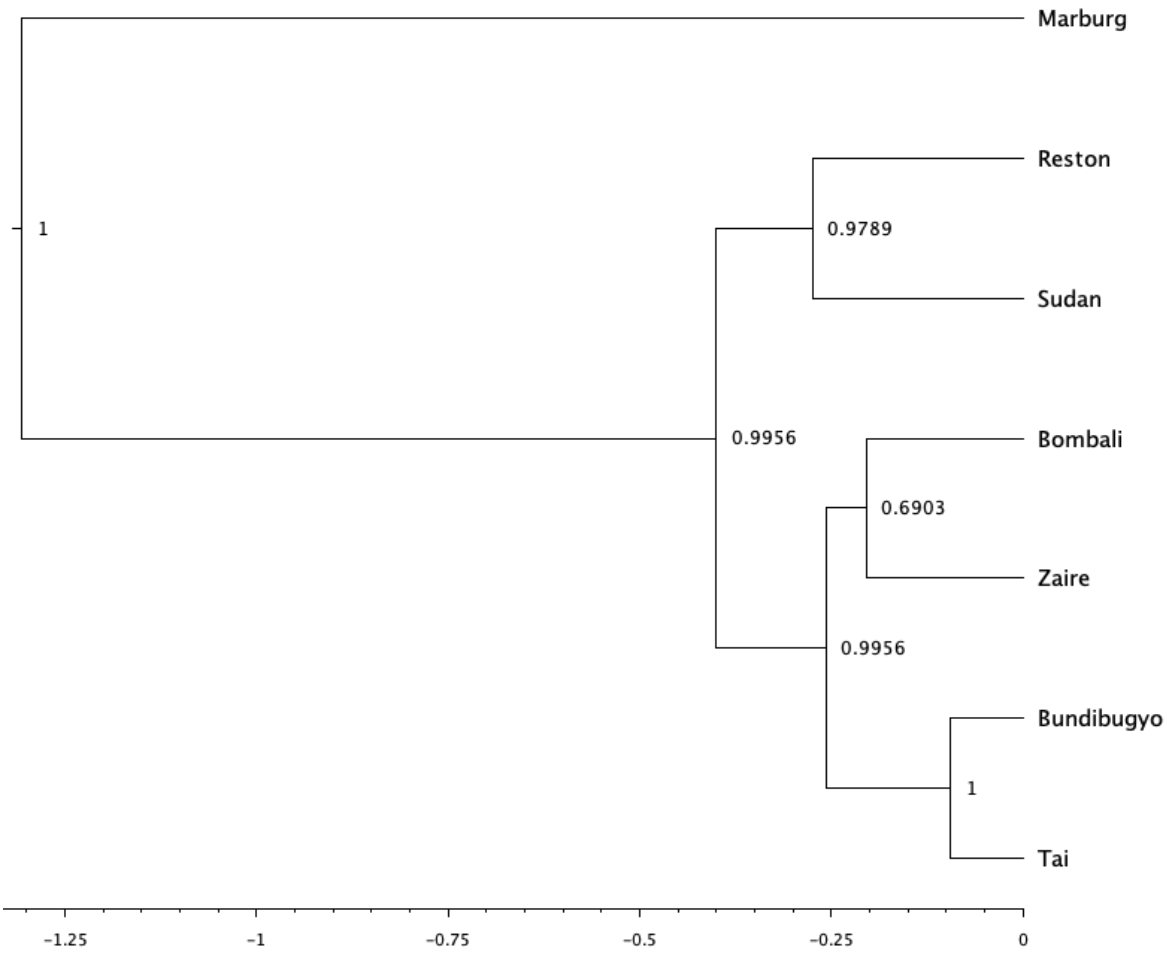

Figure 1. Bayesian phylogenetic tree of EBOV sGP generated using BEAST 2 program. Marburg GP is the outgroup. Posterior probabilities of each clade is indicated at the node.

100% -0 0 [Bombali]
100% -0 0 [Bundibugyo]
100% 0 0 [Marburg]
100% -0 0 [Reston]
100% -0 0 [Sudan]
100% -0 0 [Tai]
100% -0 0 [Zaire]
100% 0.06 0.15 [Bundibugyo,Tai]
100% 0.77 2.04 [Bombali,Bundibugyo,Marburg,Reston,Sudan,Tai,Zaire]
99.56% 0.19 0.38 [Bombali,Bundibugyo,Tai,Zaire]
99.56% 0.31 0.56 [Bombali,Bundibugyo,Reston,Sudan,Tai,Zaire]
98.22% 0.17 0.47 [Reston,Sudan]
70.44% 0.14 0.3 [Bombali,Zaire]
22.44% 0.16 0.38 [Bombali,Bundibugyo,Tai]
6.89% 0.15 0.32 [Bundibugyo,Tai,Zaire]
1.11% 0.31 0.71 [Bombali,Bundibugyo,Sudan,Tai,Zaire]

Figure 2. Clades and their posterior probability

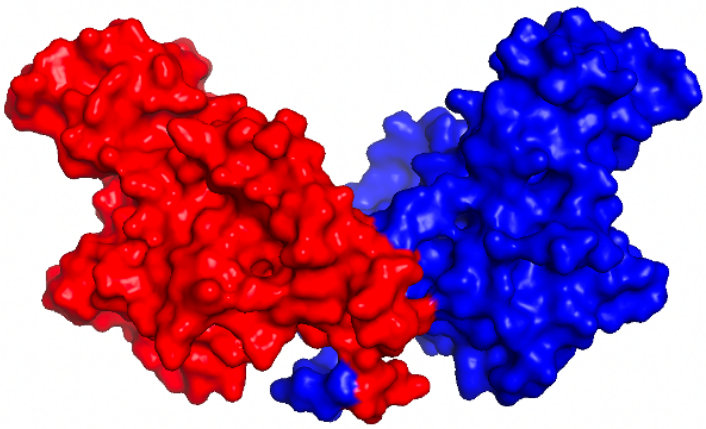

Figure 3. Structure of Zaire EBOV downloaded from PDB database.

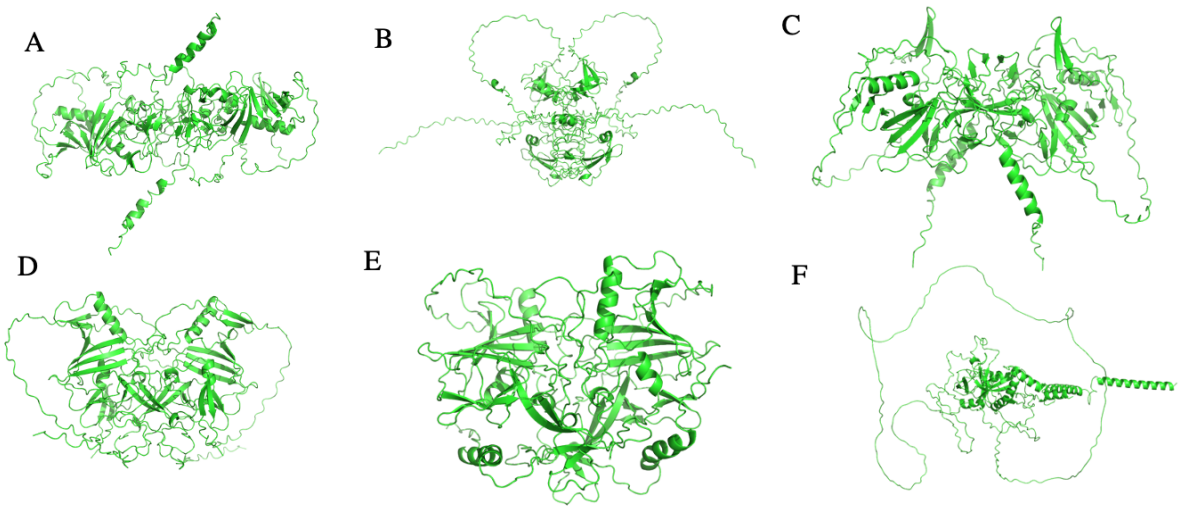

Figure 4. Structure of different strains of EBOV sGP and Marburg GP predicted using Alphafold2. A. Sudan EBOV sGP, B. Tai Forest, C. Bundibugyo, D. Bombali, E. Reston, F. Marburg GP.

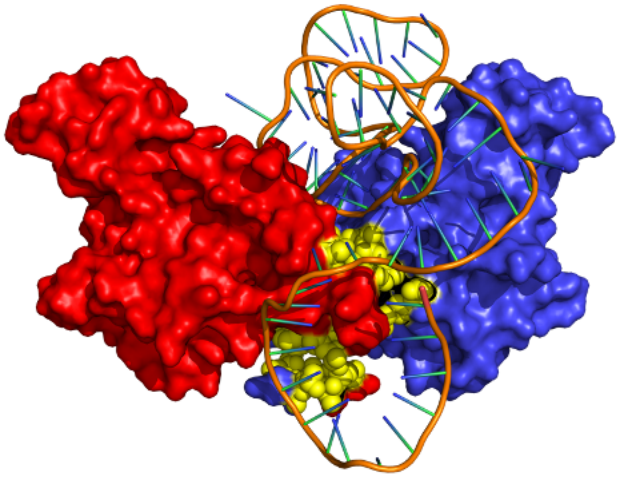

Figure 5. Zaire sGP docked to EBOV aptamer

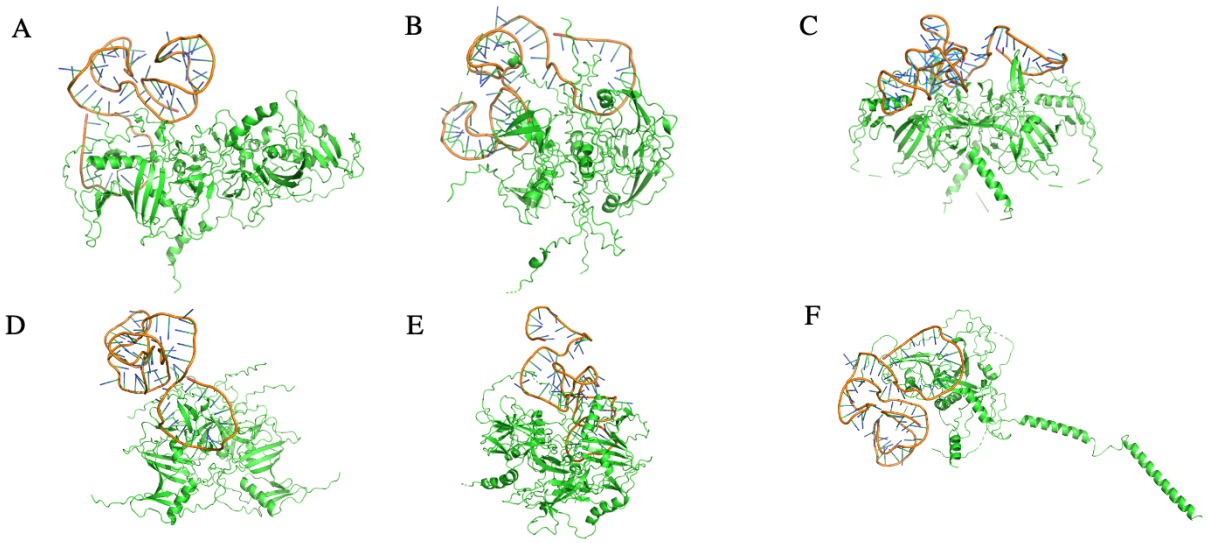

Figure 6. sGP and GP structures docked to 70SGP2A aptamer. A. Sudan EBOV sGP, B. Tai Forest, C. Bundibugyo, D. Bombali, E. Reston, F. Marburg GP

Virus name and protein	Energy of binding (kcal/mol)
Bombali EBOV sGP	70.07
Bundibugyo EBOV sGP	90.6
Reston EOV sGP	43.31
Sudan EBOV sGP	24.25
Tai EBOV sGP	56.1
Zaire EBOV sGP	31.19
Marburg virus GP	53.75

Table 1. Binding energy values of EBOV sGP and Marburg GP to 70SGP2A aptamer.

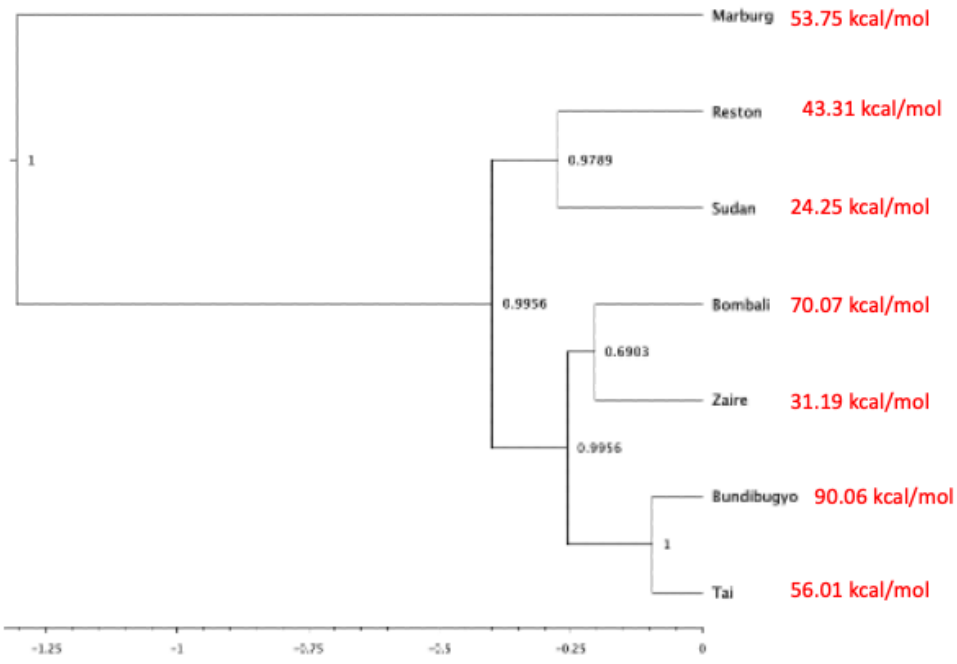

Figure 7. Phylogenetic tree with binding energy values of viral sGP or GP to EBOV aptamer.

Discussion:

I have constructed a Bayesian phylogenetic tree. The phylogenetic tree shows the Marburg GP as the outgroup correctly. In my previous attempt I used MrBayes program to build the phylogenetic tree and using some of the substitution models the Marburg virus was grouped with EBOV strains which was not correct. I decided to not include that data and do my analysis with BEAST 2 and Blossum62 substitution model. In the tree Bombali and Zaire are grouped together with a 0.6903 posterior probability. Bombali is a newly discovered EBOV (2018) and it is probably not human pathogenic, Zaire on the other had is the deadliest strain of EBOV. Bundibugyo and Tai forest are making one clade while Bundibugyo is human pathogenic and Tai forest is non-human pathogenic. Reston and Sudan strains make one strains. In this tree, in each clade there is one non-human pathogenic virus and one human pathogenic virus. Which could suggest that the two viruses shared a common ancestor and as they evolved in different animals they gained different pathogenic abilities.

The 70SGP2A aptamer that we have in the lab is selected against Zaire EBOV sGP but it also binds to Sudan EBOV sGP. I was interested to see if this aptamer is going to bind to other strains of EBOV and the values of binding energies is related to if the viruses are in one clade.

This aptamer was experimentally shown that will bind to Sudan and Zaire EBOV and the predicted binding energy values from FoldX show that the energy of binding to these strains is less than others. Marburg GP has a binding energy of 53.75 which can suggest that this aptamer might bind to this virus as well. Bundibugyo EBOV sGP has the highest binding energy which can show that this aptamer might not bind to this protein while this strain is one of the human pathogenic strains.

Figure 7 shows the binding energy values on the phylogenetic tree. My hypothesis was that the strains that are grouped together were going to have not the same but close binding energy value, and the Marburg virus (outgroup) is going to have the highest binding energy value. But the phylogenetic tree shows that the binding energy values are not related to the position of a strain in the tree. This tree is built based on amino acid sequences not the structures. This could suggest that some of these strains could have similar sequences, but they might be structurally different.

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.