Decoding HIV's Family Tree

Imani Lee

Principles of Bioinformatics

November 4, 2025

Introduction

The HIV-1 envelope glycoprotein is an important part of the virus that helps it enter human cells. It's made up of two subunits called gp120 and gp41, which work together to bind to immune cells and allow the virus to infect them. Understanding how different HIV-1 envelope sequences are related to each other is useful for tracking how the virus spreads and for developing better treatments. In this study, I analyzed HIV-1 envelope gene sequences to figure out their evolutionary relationships and determine which mathematical model best describes how these sequences change over time.

Materials and Methods

I started by aligning 133 HIV-1 envelope gene sequences using a program called Clustal Omega with the default settings. These sequences came from GenBank and were mostly from HIV-1 isolates collected in the United States. My main sequence of interest was AY156744.1, which is an HIV-1 clone from the USA. To figure out which evolutionary model best fit my data, I used MEGA12 software to test 24 different models. The program calculated scores called AICc and BIC for each model, and the model with the lowest scores is considered the best fit. After selecting the best model, I used it to build a phylogenetic tree using Maximum Likelihood analysis. This tree shows how all 133 sequences are related to each other based on their genetic similarities and differences. I also did a BLAST search on sequence AY156744.1 and read some papers about HIV-1 envelope genes to better understand what my results mean.

Results

The model testing showed that GTR+G was the best fit for my data.

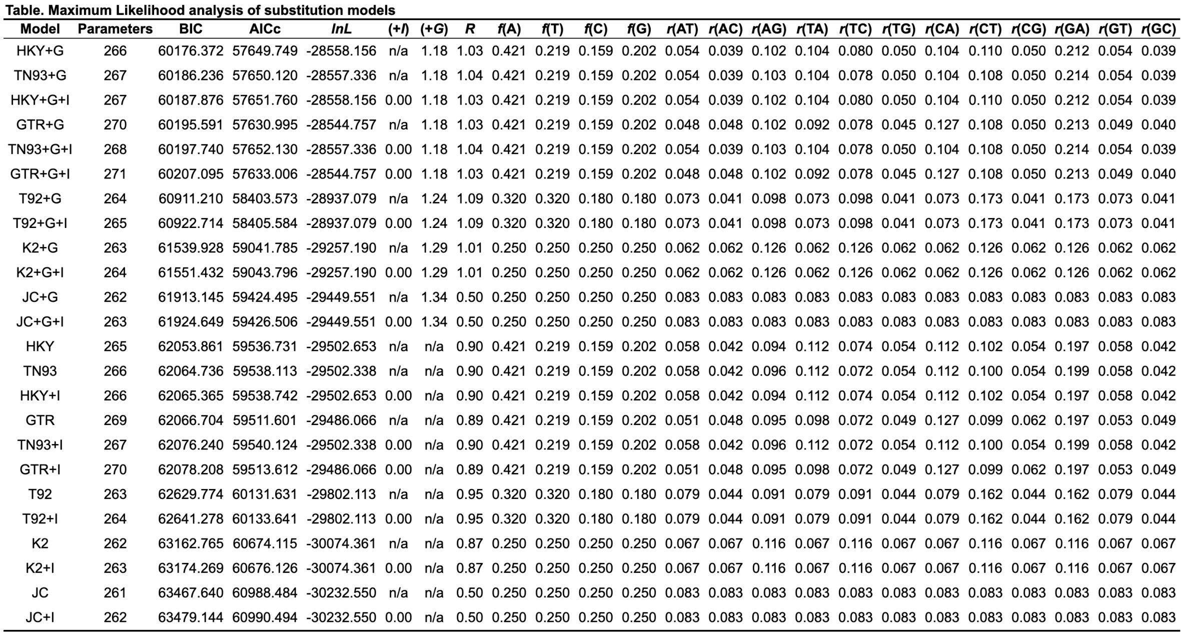

Table. Maximum Likelihood analysis of substitution models

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G)	*r*(AT)	*r*(AC)	*r*(AG)	*r*(TA)	*r*(TC)	*r*(TG)	*r*(CA)	*r*(CT)	*r*(CG)	*r*(GA)	*r*(GT)	*r*(GC)
HKY+G	266	60176.372	57649.749	-28558.156	n/a	1.18	1.03	0.421	0.219	0.159	0.202	0.054	0.039	0.102	0.104	0.080	0.050	0.104	0.110	0.050	0.212	0.054	0.039
TN93+G	267	60186.236	57650.120	-28557.336	n/a	1.18	1.04	0.421	0.219	0.159	0.202	0.054	0.039	0.103	0.104	0.078	0.050	0.104	0.108	0.050	0.214	0.054	0.039
HKY+G+I	267	60187.876	57651.760	-28558.156	0.00	1.18	1.03	0.421	0.219	0.159	0.202	0.054	0.039	0.102	0.104	0.080	0.050	0.104	0.110	0.050	0.212	0.054	0.039
GTR+G	270	60195.591	57630.995	-28544.757	n/a	1.18	1.03	0.421	0.219	0.159	0.202	0.048	0.048	0.102	0.092	0.078	0.045	0.127	0.108	0.050	0.213	0.049	0.040
TN93+G+I	268	60197.740	57652.130	-28557.336	0.00	1.18	1.04	0.421	0.219	0.159	0.202	0.054	0.039	0.103	0.104	0.078	0.050	0.104	0.108	0.050	0.214	0.054	0.039
GTR+G+I	271	60207.095	57633.006	-28544.757	0.00	1.18	1.03	0.421	0.219	0.159	0.202	0.048	0.048	0.102	0.092	0.078	0.045	0.127	0.108	0.050	0.213	0.049	0.040
T92+G	264	60911.210	58403.573	-28937.079	n/a	1.24	1.09	0.320	0.320	0.180	0.180	0.073	0.041	0.098	0.073	0.098	0.041	0.073	0.173	0.041	0.173	0.073	0.041
T92+G+I	265	60922.714	58405.584	-28937.079	0.00	1.24	1.09	0.320	0.320	0.180	0.180	0.073	0.041	0.098	0.073	0.098	0.041	0.073	0.173	0.041	0.173	0.073	0.041
K2+G	263	61539.928	59041.785	-29257.190	n/a	1.29	1.01	0.250	0.250	0.250	0.250	0.062	0.062	0.126	0.062	0.126	0.062	0.062	0.126	0.062	0.126	0.062	0.062
K2+G+I	264	61551.432	59043.796	-29257.190	0.00	1.29	1.01	0.250	0.250	0.250	0.250	0.062	0.062	0.126	0.062	0.126	0.062	0.062	0.126	0.062	0.126	0.062	0.062
JC+G	262	61913.145	59424.495	-29449.551	n/a	1.34	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+G+I	263	61924.649	59426.506	-29449.551	0.00	1.34	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
HKY	265	62053.861	59536.731	-29502.653	n/a	n/a	0.90	0.421	0.219	0.159	0.202	0.058	0.042	0.094	0.112	0.074	0.054	0.112	0.102	0.054	0.197	0.058	0.042
TN93	266	62064.736	59538.113	-29502.338	n/a	n/a	0.90	0.421	0.219	0.159	0.202	0.058	0.042	0.096	0.112	0.072	0.054	0.112	0.100	0.054	0.199	0.058	0.042
HKY+I	266	62065.365	59538.742	-29502.653	0.00	n/a	0.90	0.421	0.219	0.159	0.202	0.058	0.042	0.094	0.112	0.074	0.054	0.112	0.102	0.054	0.197	0.058	0.042
GTR	269	62066.704	59511.601	-29486.066	n/a	n/a	0.89	0.421	0.219	0.159	0.202	0.051	0.048	0.095	0.098	0.072	0.049	0.127	0.099	0.062	0.197	0.053	0.049
TN93+I	267	62076.240	59540.124	-29502.338	0.00	n/a	0.90	0.421	0.219	0.159	0.202	0.058	0.042	0.096	0.112	0.072	0.054	0.112	0.100	0.054	0.199	0.058	0.042
GTR+I	270	62078.208	59513.612	-29486.066	0.00	n/a	0.89	0.421	0.219	0.159	0.202	0.051	0.048	0.095	0.098	0.072	0.049	0.127	0.099	0.062	0.197	0.053	0.049
T92	263	62629.774	60131.631	-29802.113	n/a	n/a	0.95	0.320	0.320	0.180	0.180	0.079	0.044	0.091	0.079	0.091	0.044	0.079	0.162	0.044	0.162	0.079	0.044
T92+I	264	62641.278	60133.641	-29802.113	0.00	n/a	0.95	0.320	0.320	0.180	0.180	0.079	0.044	0.091	0.079	0.091	0.044	0.079	0.162	0.044	0.162	0.079	0.044
K2	262	63162.765	60674.115	-30074.361	n/a	n/a	0.87	0.250	0.250	0.250	0.250	0.067	0.067	0.116	0.067	0.116	0.067	0.067	0.116	0.067	0.116	0.067	0.067
K2+I	263	63174.269	60676.126	-30074.361	0.00	n/a	0.87	0.250	0.250	0.250	0.250	0.067	0.067	0.116	0.067	0.116	0.067	0.067	0.116	0.067	0.116	0.067	0.067
JC	261	63467.640	60988.484	-30232.550	n/a	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
JC+I	262	63479.144	60990.494	-30232.550	0.00	n/a	0.50	0.250	0.250	0.250	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083

Table 1 Maximum Likelihood analysis of substitution models showing AICc, BIC, and parameter estimates for tested evolutionary models.

This stands for General Time Reversible with Gamma distribution, which is a complex model that accounts for different rates of mutation. The GTR+G model had a BIC value of 60195.591 and an AICc value of 57630.995, which were the lowest values among all the models tested. This model estimates several important parameters from the sequences. The base frequencies showed that adenine (A) makes up 42.1% of the nucleotides, thymine (T) is 21.9%, guanine (G) is 20.2%, and cytosine (C) is only 15.9%. This means HIV-1 envelope genes have way more A's than C's, which is interesting because it shows a strong compositional bias. The model also calculated different substitution rates for each type of mutation. For example, AG changes happen at a rate of 0.102, which is more than twice as fast as CG changes at 0.045. This makes sense because transitions (like A to G) are chemically easier than transversions (like C to G). The gamma parameter was 1.03, which means that different parts of the envelope gene evolve at different speeds. Some regions are highly conserved because they're essential for the virus to function, while other regions called variable loops change really fast to help the virus escape the immune system.

The phylogenetic tree I created shows all 133 sequences arranged in a circular pattern based on their evolutionary relationships.

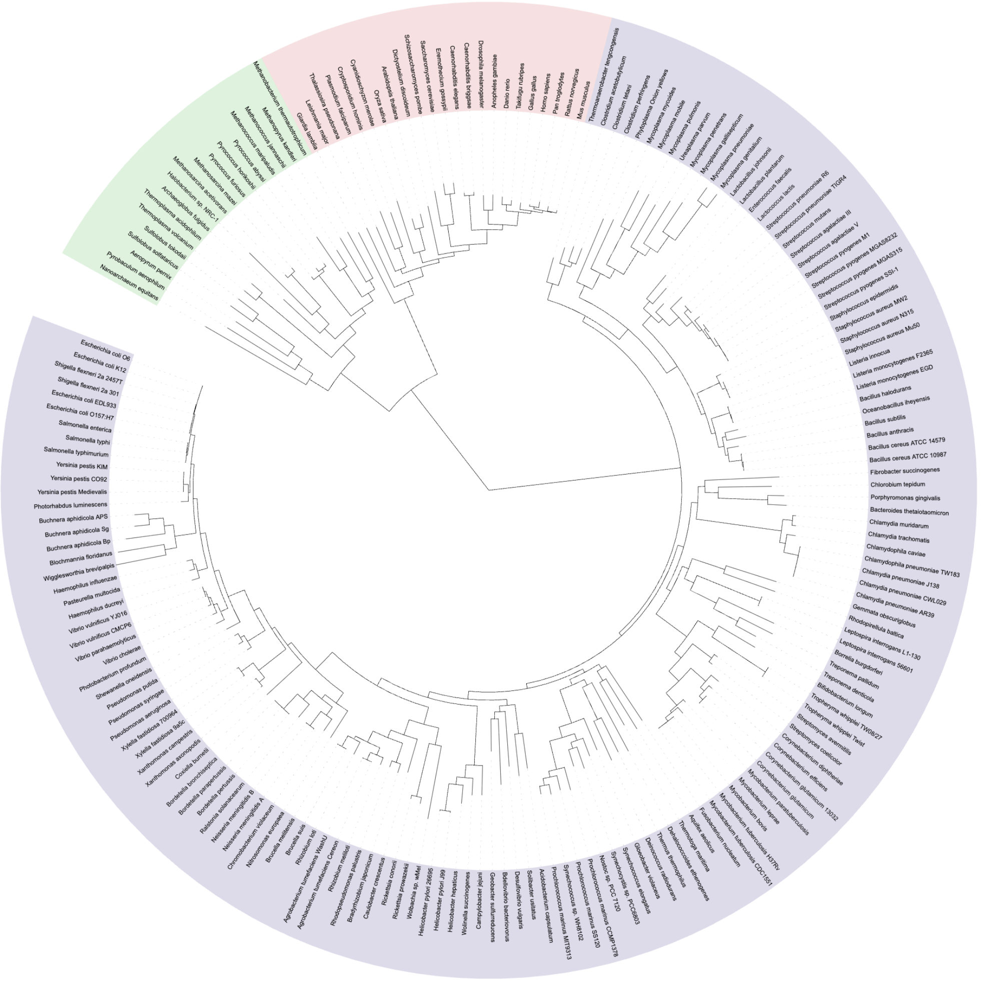

Figure 1. Maximum Likelihood phylogenetic tree of 133 HIV-1 envelope gene sequences constructed using the GTR+G model. Scale bar represents 1.0 substitutions per site. Sequences are colored by clusters.

The tree has a scale bar of 1.0 substitutions per site, and you can see that some branches are much longer than others, indicating different amounts of evolutionary change. My query sequence

AY156744.1 clusters together with other HIV-1 isolates from the USA, which confirms it's part of the North American HIV epidemic. The tree shows multiple distinct clusters of sequences, which probably represent different transmission networks or groups of related infections. Overall, the tree reveals that there's a lot of genetic diversity among these HIV-1 envelope sequences, even though they're all from similar geographic areas.

Discussion

The GTR+G model is much more sophisticated than the simple models used during sequence alignment. During alignment, Clustal Omega uses basic assumptions to make the process faster, but the GTR+G model captures the real biology of HIV evolution much better. It allows for unequal base frequencies instead of assuming each nucleotide is equally common, it lets different types of mutations happen at different rates, and it accounts for the fact that some parts of the gene evolve faster than others. These features make the model more realistic for describing how HIV actually evolves. The reason different sites evolve at different rates is because of selective pressure. Parts of the envelope that are critical for binding to cells or fusing with cell membranes can't change much without breaking the virus, so they evolve slowly. But the variable loops are under constant attack from the immune system, so the virus benefits from mutations in those regions that help it escape antibodies.

While my phylogenetic tree successfully shows how these sequences are related, it's really important to understand what it can and can't tell us. The tree shows that sequences cluster together and how genetically different they are from each other, but it absolutely cannot tell us about guilt or innocence in any kind of legal case or even tell us who infected whom. Phylogenetic trees only show relationships, not direction of transmission or causation. If you wanted to use phylogenetics in a forensic case, you would need way more information. First, you'd need actual sequences from the specific people involved in the case, not just random database sequences like I used. Second, you'd need to look at within-host diversity because the person who's been infected longer will have more genetic diversity in their virus than someone they just infected. Third, you'd need temporal data with dates of infection so you could use molecular clock methods to figure out timing. Fourth, you'd need epidemiological information like medical records, contact histories, and other contextual evidence.

Several important papers have shown both the potential and the limitations of using phylogenetics for transmission cases. Metzker and colleagues published a paper in 2002 that showed how analyzing the diversity of virus within each person could help determine who infected whom in a criminal case. Romero-Severson and colleagues wrote a paper in 2016 explaining that even when sequences cluster together closely on a tree, that's not enough proof of direct transmission by itself. You need all that other supporting evidence I mentioned. These papers make it really clear that phylogenetic similarity doesn't equal proof that one person directly infected another.

When I did a BLAST search on my sequence and read about HIV-1 envelope proteins, I learned some really interesting things about why this protein is so important. The envelope glycoprotein is the main target for vaccine development, but it's incredibly difficult to make a vaccine against it. That's because the envelope is extremely variable, with some regions showing 60-80% sequence diversity even within a single infected person.

The envelope is also covered in sugar molecules called glycans that act like a shield to hide it from antibodies. Recent research has shown that the viruses that successfully start new infections have certain properties that make them better at establishing infection while also being good at avoiding the immune system. The conformational flexibility of the envelope means it can change its shape, which helps it resist antibodies while still functioning properly to infect cells.

My analysis has some limitations that are worth mentioning. The tree doesn't show bootstrap values, which would tell us how confident we can be about specific branching patterns. Ideally, you'd want to run at least 1000 bootstrap replicates to test how robust each branch is. Also, I only analyzed one gene region, but looking at multiple genes would give a more complete picture of evolutionary relationships. The sequences came from a database rather than being specifically collected for a study, which could introduce some biases. Despite these limitations, using Maximum Likelihood with proper model selection is a solid approach for building phylogenetic trees, and the GTR+G model is one of the most realistic models commonly used for viral evolution studies.

Conclusion

This phylogenetic analysis successfully identified GTR+G as the best evolutionary model for HIV-1 envelope sequences and created a tree showing how 133 sequences are related. The model parameters reveal important biological patterns like strong compositional bias toward adenine, faster rates for transitions than transversions, and variable evolutionary rates across different parts of the gene. The phylogenetic tree demonstrates the huge amount of genetic diversity in HIV-1, which is one reason why it's so hard to develop an effective vaccine. However, it's crucial to remember that phylogenetic trees show relationships but can't determine who infected whom or establish guilt or innocence without additional evidence like temporal data, within-host diversity analysis, and epidemiological information. Understanding HIV-1 evolution is important for public health efforts, vaccine development, and tracking how the virus spreads, but phylogenetic evidence must always be interpreted carefully and in context with other types of data.

References

Checkley, M. A., Luttge, B. G., & Freed, E. O. (2011). HIV-1 envelope glycoprotein biosynthesis, trafficking, and incorporation. *Journal of Molecular Biology, 410*(4), 582-608. https://doi.org/10.1016/j.jmb.2011.04.042

Metzker, M. L., Mindell, D. P., Liu, X. M., Ptak, R. G., Gibbs, R. A., & Hillis, D. M. (2002). Molecular evidence of HIV-1 transmission in a criminal case. *Proceedings of the National Academy of Sciences, 99*(22), 14292-14297. https://doi.org/10.1073/pnas.222522599

Romero-Severson, E., Skar, H., Bulla, I., Albert, J., & Leitner, T. (2016). Phylogenetically resolving epidemiologic linkage. *Proceedings of the National Academy of Sciences, 113*(10), 2690-2695. https://doi.org/10.1073/pnas.1522930113

Tamura, K., Stecher, G., & Kumar, S. (2021). MEGA11: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution, 28*(10), 2731-2739. https://doi.org/10.1093/molbev/msab120