ViralGeneClock

A Tool for Estimating the Relative Mutation Rate of Different Genes across Viral Strains using Phylogenetic Analyses.

- Merit Kayastha

Background

- *ViralGeneClock* is a web application, where the users can submit the FASTA sequence of viral strains, and get the outputs emailed to their address.
- *ViralGeneClock* utilizes the Neighbor-joining (NJ) algorithm for phylogenetic analyses.
- NJ is a bottom-up clustering method for estimating genetic distances and branch lengths, and creating phylogenetic trees.

Rationale

- Identify regions of the viral genome that are highly conserved across different strains. Could be potentially useful for drug/vaccine design.
- Understand the phylogenetic relationship between different strains of the virus.
- Identify regions with rapid mutation, indicating that the region is under strong selective pressure.

Tools and Materials

- Prokka

- A Linux-based rapid prokaryotic genome annotation tool, for viral annotation.

- MUSCLE

- Multiple sequence alignment tool which uses progressive alignment algorithm.

- HTML and CSS: for designing front end and different tabs of the webpage.
- Ajax in JavaScript: to retrieve live updates from the command line.

- Biopython: to manage and organize FASTA files.
- Flask: wrap the CLI with a user friendly web framework.
- Flask-mail: to automatically email users with results once the analysis is complete.
- Matplotlib: for generating a phylogenic tree image file.

Methods

- Full Sequence Analysis
- Genome Annotation & Multiple Sequence Alignment
- Neighbor-Joining Algorithm to calculate genetic distance and branch length to reference strain.

- Gene Analysis
- Gene Grouping & Multiple Sequence Alignment for each Gene
- NJ algorithm to calculate genetic distance to reference strain for a gene.
- Estimation of Relative Mutation Rate.

Tool Workup (Video)

- https://ramapo.yuja.com/V/Video?v=10349598&node=44630422&a=161497724

Results

Sample 1: 10 SARS-CoV-2 viral strains

locus_tag	ftype	length_bp	gene	EC_number	COG	product
NAFDKFMI_00001	CDS	13218	1a			Replicase polyprotein 1a
NAFDKFMI_00002	CDS	7788	rep			Replicase polyprotein 1ab
NAFDKFMI_00003	CDS	3822	S			Spike glycoprotein
NAFDKFMI_00004	CDS	828	3a			Protein 3a
NAFDKFMI_00005	CDS	669	M			Membrane protein
NAFDKFMI_00006	CDS	186				hypothetical protein
NAFDKFMI_00007	CDS	366	7a			Protein 7a
NAFDKFMI_00008	CDS	366				hypothetical protein
NAFDKFMI_00009	CDS	1260	N			Nucleoprotein

Fig 1: Prokka annotation of SARS-CoV-2 virus.

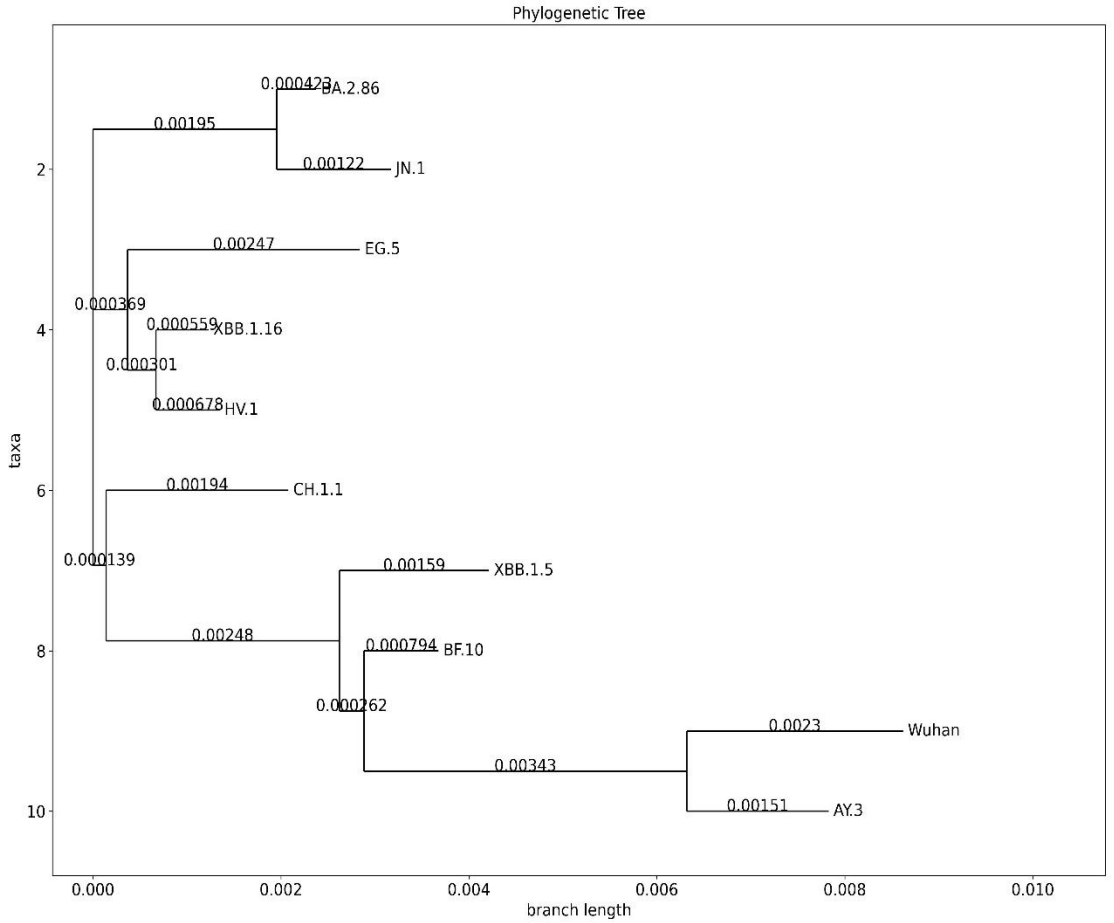

Fig 2: Phylogenic tree of SARS-CoV-2 viral strains.

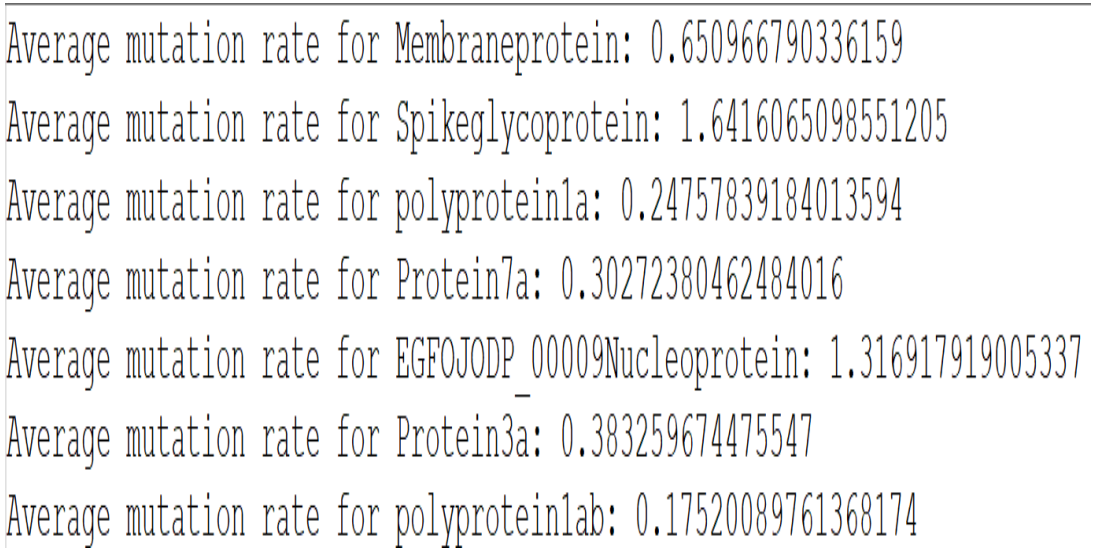

Fig 3: Relative mutation rate for annotated genes in SARS-CoV-2 from *ViralGeneClock*.

From scientific literature, S and N genes have been found with the highest mutational range.

- Mutation in S protein can alter its binding affinity with the ACE2 receptor, which increases the virus's ability to evade the immune responses.
- the N protein is responsible for packaging the viral RNA and influences the detection of the virus by the immune system.

2) Sample 2: 4 Influenza A viral strains.

locus_tag	ftype	length_bp	gene	EC_number	COG	product
OIPHMFNG_00001	CDS	2280	PB2			Polymerase basic protein 2
OIPHMFNG_00002	CDS	2274	PB1	2.7.7.48		RNA-directed RNA polymerase catalytic subunit
OIPHMFNG_00003	CDS	2118	PA	3.1.-.-		Polymerase acidic protein
OIPHMFNG_00004	CDS	1701	HA			Hemagglutinin
OIPHMFNG_00005	CDS	1497	NP			Nucleoprotein
OIPHMFNG_00006	CDS	1368	NA	3.2.1.18		Neuraminidase
OIPHMFNG_00007	CDS	759	M			Matrix protein 1
OIPHMFNG_00008	CDS	660	NS			Non-structural protein 1

Fig 4: Prokka annotation of Influenza A virus.

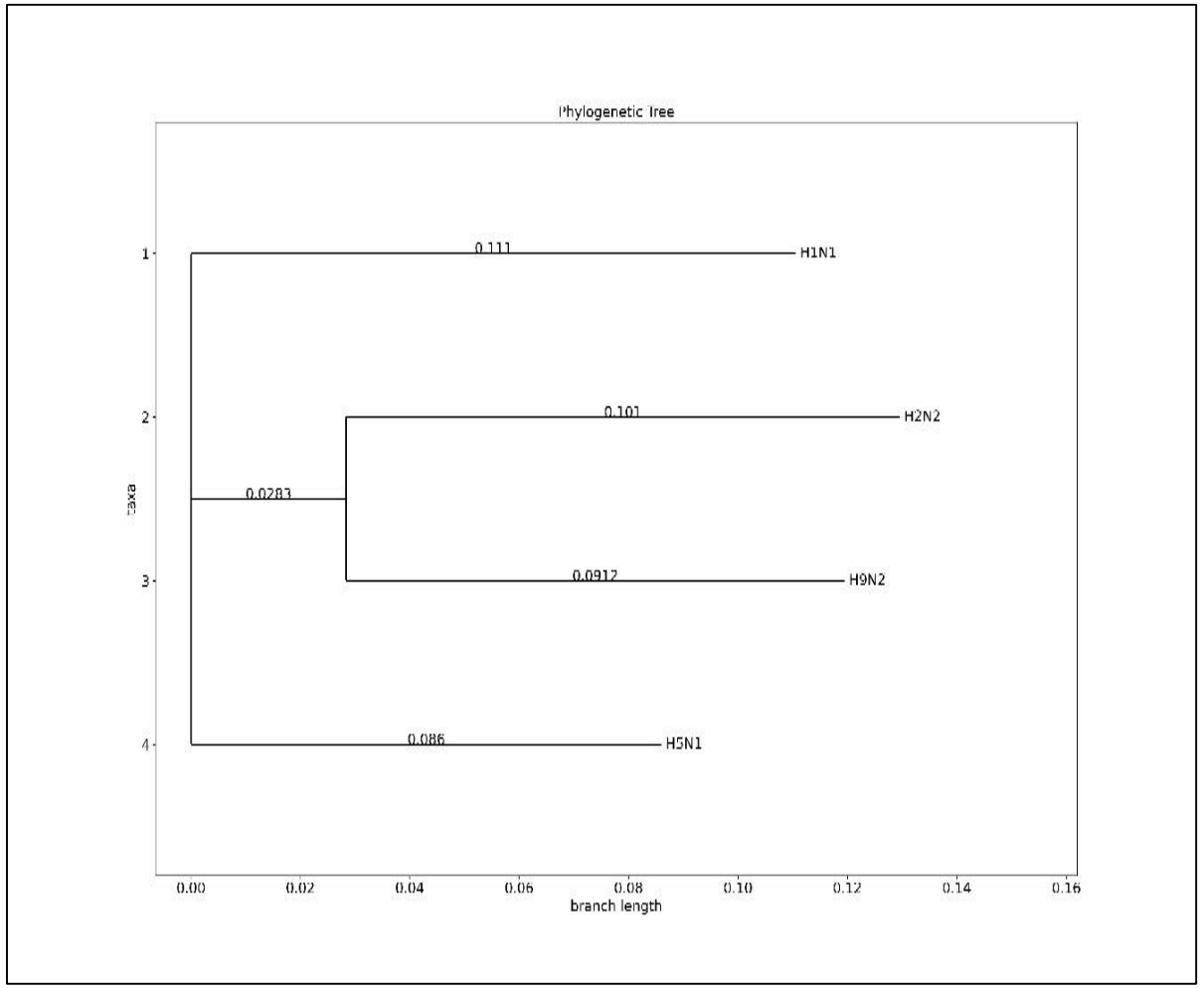

Fig 5: Phylogenic tree of Influenza A viral strains.

```
Average mutation rate for NMCHHEOD_00005Nucleoprotein: 0.7519412408169522
Average mutation rate for protein1: 1.2006538657037895
Average mutation rate for acidicprotein: 0.6421659772966403
Average mutation rate for catalyticsubunit: 0.5975132953040165
Average mutation rate for protein1: 0.43340307740013245
Average mutation rate for GBOAADJJ_00006Neuraminidase: 1.7229239933563172
Average mutation rate for NMCHHEOD_00004Hemagglutinin: 1.7463257465660276
Average mutation rate for protein2: 0.7141895603686917
```

Fig 6: Relative mutation rate for annotated genes in Influenza A from *ViralGeneClock*.

From scientific literature, Hemagglutinin and Neuraminidase genes have been found to evolve more rapidly.

- Surface protein genes HA and NA evolve rapidly than internal protein genes.
- HA, in particular, helps in attaching to and entering the cells in the respiratory tract, and regular mutations in the HA gene help the virus evade the immune response.

Conclusion

- With a lack of literature on quantitative mutation rates across different genes in viral strains, verifying the accuracy of *ViralGeneClock*'s quantitative relative mutation rates is difficult. However, the tool’s findings align with the qualitative information in the literature, as observed in the examples of SARS-CoV-2, HIV-1, and Influenza 1.
- *ViralGeneClock* is effective in identifying the evolutionary relationship between different viral strains and can be utilized for pinpointing viral regions with varying levels of conservation.

References:

- Cardona-Ospina, J. A., Rojas-Gallardo, D. M., Garzón-Castaño, S. C., Jiménez-Posada, E. V., & Rodríguez-Morales, A. J. (2021). Phylodynamic analysis in the understanding of the current COVID-19 pandemic and its utility in vaccine and antiviral design and assessment. *Human Vaccines & Immunotherapeutics*, *17*(8), 2437–2444. https://doi.org/10.1080/21645515.2021.1880254
- Hegde, S., Tang, Z., Zhao, J., & Wang, J. (2021). Inhibition of SARS-CoV-2 by targeting conserved viral RNA structures and sequences. *Frontiers in Chemistry*, *9*. https://doi.org/10.3389/fchem.2021.802766
- Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*(4), 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454
- Kuhner, M. K., & Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, *11*(3), 459–468. https://doi.org/10.1093/oxfordjournals.molbev.a040126
- Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics (Oxford, England), 30(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153
- Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1). https://doi.org/10.1186/1471-2105-11-119
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340
- Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163
- *Flask*. Readthedocs.org. Retrieved April 6, 2024, from https://readthedocs.org/projects/flask/
- *Flask-mail — Flask-Mail 0.9.1 documentation*. Pythonhosted.org. Retrieved May 1, 2024, from https://pythonhosted.org/Flask-Mail/
- *AJAX introduction*. W3schools.com. Retrieved May 1, 2024, from https://www.w3schools.com/js/js_ajax_intro.asp
- Hirabara, S. M., Serdan, T. D. A., Gorjao, R., Masi, L. N., Pithon-Curi, T. C., Covas, D. T., Curi, R., & Durigon, E. L. (2022). SARS-COV-2 variants: Differences and potential of immune evasion. *Frontiers in Cellular and Infection Microbiology*, *11*. https://doi.org/10.3389/fcimb.2021.781429
- Cuevas, J. M., Geller, R., Garijo, R., López-Aldeguer, J., & Sanjuán, R. (2015). Extremely high mutation rate of HIV-1 in vivo. *PLoS Biology*, *13*(9), e1002251. https://doi.org/10.1371/journal.pbio.1002251
- Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M., & Kawaoka, Y. (1992). Evolution and ecology of influenza A viruses. *Microbiological Reviews*, *56*(1), 152–179. https://doi.org/10.1128/mr.56.1.152-179.1992