Evolution and Population Diversity Study of Influenza D Virus (IDV) using BEAST Algorithm

Dangat Ashitosh Vilas

Project Guide : Dr. Mohan Kale

Department of Statistics
Savitribai Phule Pune University
Pune - 411007

ST 402 Project
Fourth Semester 2023
End Term Examination
Friday, 22^{th} May 2023

Table of Contents

1. Motivation
2. Introduction
3. Objectives
4. Data Description
5. Nucleotide Substitution Models and Molecular Clocks
6. Softwares Used
7. Flow Chart
8. Statistical Methods
9. Results
10. Conclusion
11. Bibliography

Motivation

Cattle industry is extremely important as it :

- Sustains self-sufficiency of rural economy by providing auxiliary income to farmers
- Initiates economic growth
- Provides employment
- Contributes to the GDP

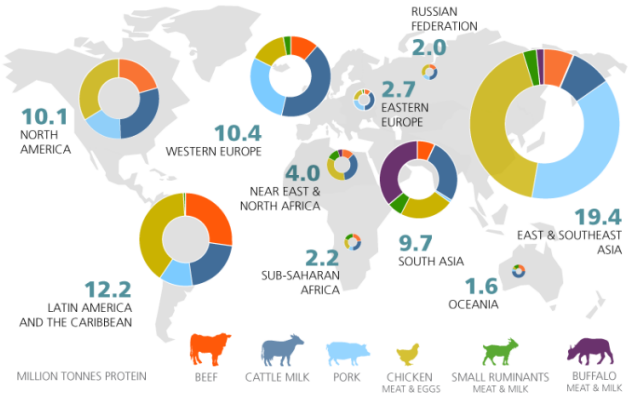


Figure 1: www.fao.org/gleam/dashboard-old/en/

Introduction

- Influenza D Virus (IDV) was initially isolated from diseased pigs but bovine(cattle) is now main host.
- The prevalence of IDV in cattle is higher, especially in calves.
- IDV detection rates are higher in clinically sick cattle than in apparently healthy cattle.

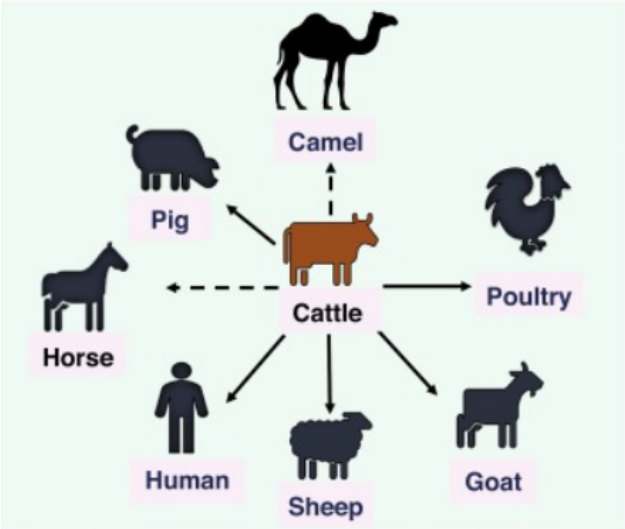


Figure 2: www.ncbi.nlm.nih.gov/pmc/articles/PMC7755673/

Objectives

- To estimate the rates of substitutions (per site,per year).
- To study the time of evolution from the most recent common ancestor.
- To conduct Molecular Phylogeny Analysis.
- To perform Biodiversity study.

Data Description

- IDV has a segmented genome structure.
- It consists of 7 gene segments namely PB2, PB1, P3, HEF, NP, P42 and NS. NS splits into 2 segments NS1 and NS2.
- The letters are A, C, G, and T representing the four nucleotide bases of a DNA strand – adenine, cytosine, guanine, and thymine.
- Data consists of characters, country of isolation, and date of isolation of each entry.

Cont...

- We collected Secondary data from NCBI's GenBank.
- Data type: Texual.
- The gene sequences were further extracted, curated and aligned at the Bioinformatics Centre, Savitribai Phule Pune University and then used for analysis.
- It is collaborative work with the Department of Veterinary Disease ,Pennsylvania State University and Bioinformatics Centre, SPPU, Pune.
- >KF425669-USA-2013-02-26
  ATGTTTTTGCTTCTAGCAACAATTACAGCATAACTGCTT

SR NO.	GENE NAME	NO. OF SEQUENCES
1	PB2	175
2	PB1	169
3	P3	189
4	HEF	188
5	NP	183
6	P42	179
7	NS	147
7.1	NS1	147
7.2	NS2	147

Table 1: Dataset Size

Nucleotide Substitution Models and Molecular Clocks

- In biology substitution models also called models of DNA sequence evolution are Markov models that describe changes over evolutionary time.

Model	Exchangeability parameters	Base frequency parameters
Juke Contor (JC)	a=b=c=d=e=f	$\pi_A=\pi_C=\pi_G=\pi_T=0.25$
Kimura 2-parameter (K2P)	a=c=d=f,b=e	$\pi_A=\pi_C=\pi_G=\pi_T=0.25$
Hasegawa-Kishino-Yano (HKY)	a=c=d=f,b=e	all π_i's are free
Tamura-Nei (TN93)	a=c=d=f,b,e	all π_i's are free
General Time Reversible (GTR)	all exchangeability parameters free	all π_i's are free

$a=r_{AC}$, $b=r_{AG}$, $c=r_{AT}$, $d=r_{CG}$, $e=r_{CT}$, $f=r_{GT}$

Table 2: Substitution Model Table

The molecular clock presents a means of estimating evolutionary rates using genetic data. Following are the clocks to deal with rate variation:

- Strict Clock: Constant rate among branches.

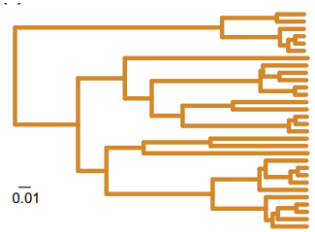


Figure 3: Strict Clock

- Uncorrelated Relax Clock: Distinct rate along each branch drawn from a chosen probability distribution.

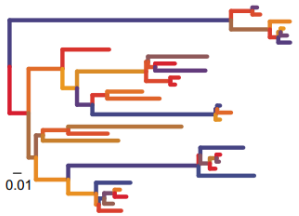


Figure 4: Uncorrelated Relax Clock

Cont...

- Fixed Local Clock: Branches having the same rate are clustered together.

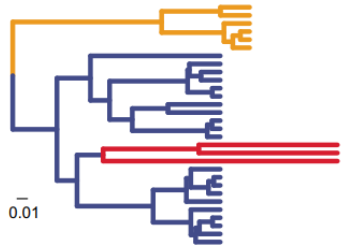


Figure 5: Fixed Local Clock

- Random Local Clock: Branches having the same rates are distributed throughout the tree rather than being locally clustered.

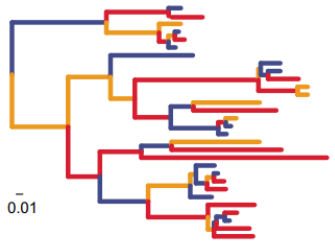


Figure 6: Random Local Clock

Softwares Used

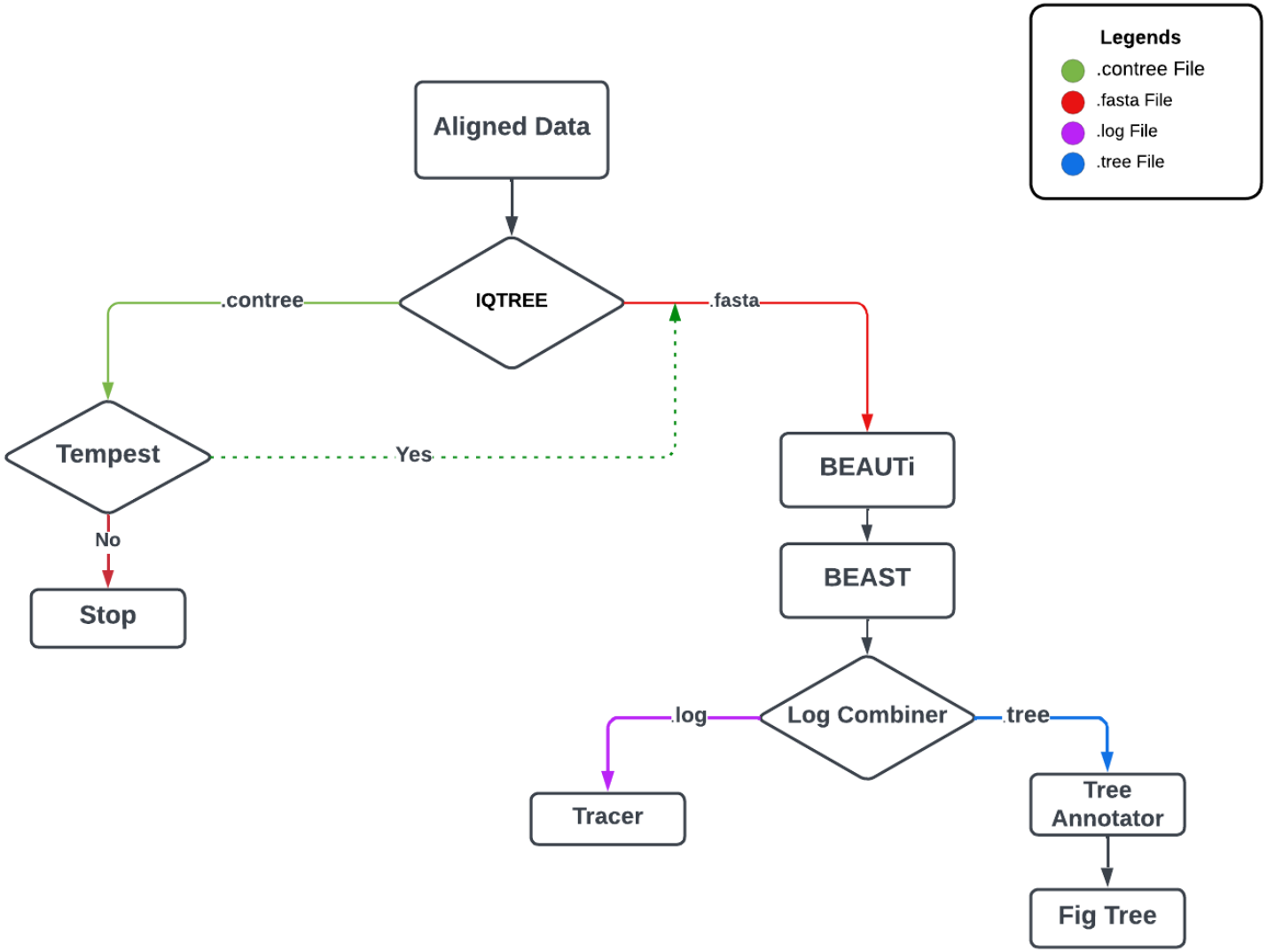
- IQTREE - Phylogenetic inference by maximum likelihood.
- Tempest v1.5.3 - This is a graphical program for looking for temporal signals in trees generated from updated sequences.
- BEAUti v1.10.4 - Bayesian Evolutionary Analysis Utility. This program is used to import data, design the analysis, and generate the BEAST control file.
- BEAST v1.10.4 - Bayesian Evolutionary Analysis Sampling Trees. This is the main program that takes a control file generated by BEAUti and performs the analysis.

Cont...

- LOGCOMBINER - Logcombiner will combine log files from different runs and reduce the sampling frequency. It allows you to combine log and tree files from multiple independent runs of BEAST.
- Tracer v1.7.2 - Tracer is a graphical program for exploring the output of BEAST, diagnosing problems, and summarizing the results.
- TreeAnnotator v1.10.4 - Tree Annotator is a post-analysis program that will produce a summary tree from the output of BEAST.
- FigTree v1.4.4 - Fig Tree is a graphical program for viewing trees, displaying summary information from Tree Annotator.

Flow Chart

The pipeline followed throughout the analysis for each gene.



Statistical Methods

- Chi-Square Tests for the test of homogeneity of character composition.
- Molecular phylogeny was done using the maximum likelihood method.
- The nucleotide substitution model selection was performed based on the Bayesian Information Criterion (BIC) from the provided AIC, BIC, and corrected AIC values.
- Bayesian analysis method such as MCMC.
- Marginal likelihood Calculations using path sampling and stepping stone sampling.

Results

The following table summarizes our analysis, including the best models for each gene, the substitution rate, and the time for a most recent ancestor.

Dataset	Gene name	Substitution Model	Clock	R^2	Correlation coefficient
1	PB2	TIM+F+G4	UCLN	0.7447	0.863
2	PB1	TIM+F+G4	STRICT	0.0019	0.045
3	P3	TIM+F+I+G4	STRICT	0.4904	0.7003
4	HEF	TVM+F+G4	UCLN	0.5927	0.7699
5	NP	TN+F+G4	STRICT	0.6533	0.8082
6	P42	GTR+F+G4	STRICT	0.064	0.2531
7	NS	HKY+F+G4	UCED	0.2622	0.512
7.1	NS1	HKY+F+G4	UCLN	0.2939	0.5421
7.2	NS2	TN93 + F + G4	UCLN	0.1826	0.4273

Table 3: Result Table

Dataset	Gene name	MEAN NSR (sub/site/year)	NSR (95% HPD Interval)	tMRCA
1	PB2	1.35E-03	(1.122E-3,1.5935E-3)	1997.47
2	PB1	1.16E-03	(9.9688E-4,1.3341E-3)	1997.04
3	P3	1.31E-03	(1.1203E-3,1.4832E-3)	1997.53
4	HEF	1.66E-03	(1.4047E-3,2.0118E-3)	1996.09
5	NP	1.47E-03	(1.2389E-3,1.6881E-3)	1998.97
6	P42	1.35E-03	(1.106E-3,1.6039E-3)	1997.3
7	NS	1.31E-03	(9.0492E-4,1.7197E-3)	1997.36
7.1	NS1	1.32E-03	(9.572E-4,1.7028E-3)	1998.62
7.2	NS2	1.63E-03	(1.0584E-3,2.2291E-3)	2003.16

Table 4: Result Table

Cont...

- Analysis of HEF resulted in the five distinct lineages.

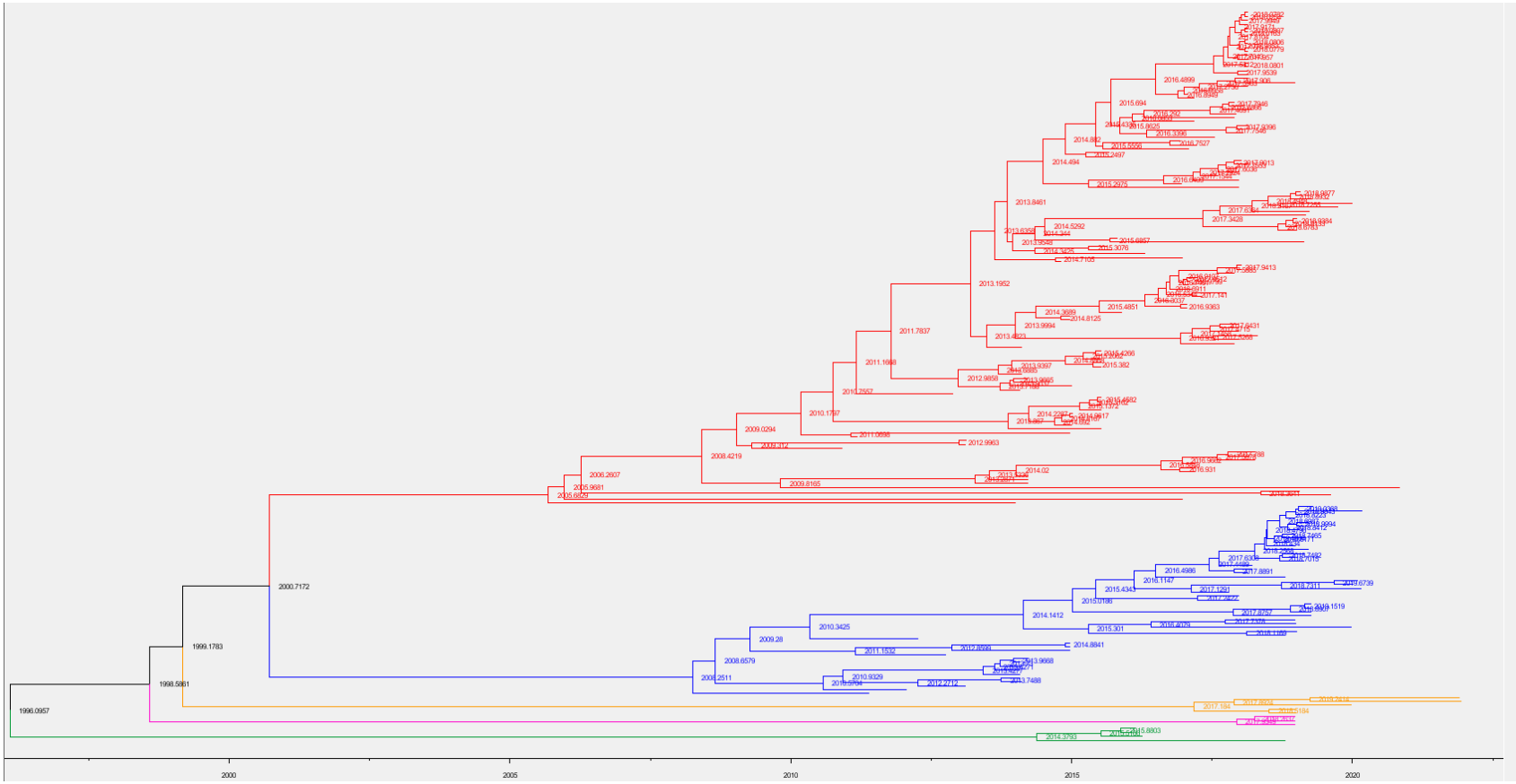


Figure 7: Best Tree with node ages

Conclusions

The Hemagglutinin-Esterase-Fusion (HEF) protein of IDV is responsible for the attachment of the virus particle to the host cells. Due to its critical role in viral entry and immune recognition, the hemagglutinin gene (HEF) is often of particular importance in influenza research.

- For HEF, the TVM+F+G4 nucleotide substitution model was identified as the best fit.
- By applying uncorrelated lognormal distribution to the clock model for HEF, we estimated the rates of molecular evolution and divergence times more accurately.
- The convergence occurred at 10^8 (MCMC chain length) for all the parameters ($\geq$200) with Effective Sample Size (ESS).

Cont...

- The mean rate for HEF is estimated to be 1.66×10^{-3} substitutions per site per year.
- The 95% HPD interval of mean substitution rates which is a credible interval is given as $(1.4047 \times 10^{-3}, 2.0118 \times 10^{-3})$. It suggests that with 95% confidence the actual rate of evolutionary change in HEF falls within this interval.
- The analysis of HEF resulted in the identification of five distinct clusters based on the Bayesian information. This suggests that the data supports the presence of five separate groups or lineages within the gene sequence.
- Since the rate of evolution is approximately 10^{-3}, we can infer that IDV is a fast evolving virus and hence developing effective control and prevention strategies is need of the hour.

Bibliography

- Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS biology. 2006 May;4(5):e88
- Rambaut, A., Drummond, A. J. (2012). Tracer v1. 7.2, obtained from the "Workshop on Molecular Evolution", Aug 2011. Tracer v1. 5.
- Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology. 2007 Dec;7(1):1-8.
- Ho SY, Duchêne S. Molecular-clock methods for estimating evolutionary rates and timescales. Molecular ecology. 2014 Dec;23(24):5947-65.
- Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus evolution. 2018 Jan;4(1):vey016.

Acknowledgment

We would like to thank **Mr. Sanket Limaye**, Bioinformatics Centre, SPPU, Pune, and **Prof. Suresh Kuchipudi**, Pennsylvania State University for their constant guidance for the successful execution of the project.

Thank You !