Analysis of Nipah Virus Evolution Using ATG Genomic Walks

Guennadi A. Kouzaev[1]

Abstract. In this paper, 32 complete genomic sequences of human-origin Nipah viruses registered in GenBank® from 1999 to 2025 are analyzed by tracking distributions of ATG triplets. The trajectories of these triplets are constructed, and their divergence parameter is calculated, revealing an increased value typical of highly mutated viruses that can form separate families with distinct properties. This conclusion aligns with the large deviation observed in the fractal dimension parameters of the distribution of inter-ATG-triplet distances in genomic sequences. The simulation results are compared with earlier data for SARS-CoV-2, MERS-CoV, dengue, and Ebola viruses.

Keywords. Virus genomics, genomic walks, ATG walks

1. Introduction

The Nipah virus, or *Henipavirus Hipahense*, was first identified in Malaysia in 1998 and subsequently spread to several Southeast Asian countries, including Bangladesh, India, Indonesia, the Philippines, and Singapore, among humans, pigs, and dogs. The last time an infection was registered in India was at the beginning of 2026 [1].

This zoonotic virus originated in bats and spread among humans and animals through contaminated fruits. The disease is also transmitted between people through biological fluids and respiratory droplets in crowded hospital settings. The mortality rate ranges from 40% to 80% [2]-[6], reaching 91% during the 2018-2019 outbreak in India [2],[6]. The virus mainly affects the brain (encephalitis) and the lungs. Currently, no effective treatment or vaccines are available.

Due to random mutations, the virus varies; for example, the genome of the Malaysian strain (NiV-M) evolves at a rate of 4.64×10^{-4} substitutions per site per year [4]. Additionally, the potential formation of new clusters was shown by studying the genomes of bat and human viruses during the Kerala (India) 2018-2019 outbreak [6]. Therefore, monitoring genomic variation is essential because of this virus's high mutation rate and the need to develop adaptive treatments and vaccines for this severe infection.

Genomic studies start with instrumented sequencing of isolated viral RNAs, followed by numerical analysis of the resulting sequences to identify genes and build phylogenetic trees of mutated viral samples, comparing them symbol by symbol and calculating neighborhood distances. It provides criteria for dividing sequences into distinct clades that may have different pathogenic effects. From a mathematical standpoint, this process involves problems that are exponentially complex, and their solutions can demand increasing computational resources. Besides the described approach, genomic sequences can also be compared using non-alignment techniques that do not require symbol-to-symbol comparisons. Some of these involve creating images of viral sequences and analyzing them qualitatively and quantitatively [7].

---

[1] Faculty of Information Technologies and Electroengineering. Norwegian University of Science and Technology (NTNU), Trondheim, Norway. E-mail: guennadi.kouzaev@ntnu.no

The goal of this contribution is to study the NiV's genomic sequences using a recently developed non-alignment approach [7]-[9], allowing for tracing and imaging the ATG-triplets positions in genomic sequences in the form of analytically simple trajectories, with the following calculation of divergence of these trajectories and fractal parameters of their distributions along analyzed genomic traces.

Although the method tracks any repeating patterns, including individual nucleotides, the ATG observations focus on severe mutations and on the variation in the number of ATG triplets and codons in the studied sequences.

The paper is organized as follows. Section 2 provides a brief review of existing sequencing instrumentation and software tools. A list of studied human-origin complete genomic sequences is included in the Appendix, related to this and the following parts of the paper. Simulation results are shown in Section 3 and discussed in Section 4.

2. Materials and Methods

Because the virus constantly mutates, monitoring its genomes and biological traits is crucial. Genomic sequences are obtained through various instruments and software tools, with accuracy improving over time [10],[11]. Much of the virus data collected worldwide is stored in GenBank® [12]. Unfortunately, only a limited number of human-origin Nipah virus sequences are available, and about one-third of these were analyzed in this study. The list of these genomes and their calculated parameters is provided in Table 1A of the Appendix.

So far, all Nipah sequences are classified into several main groups: NiV-M (Malaysia), NiV-BD1, and -BD-2 (Bangladesh). Another group mentioned in the literature is NiV-I (India), identified after the 2018-2019 outbreak in Kerala, India [6]. This classification was based on analyses of phylogenetic trees created using different mathematical methods and software tools [2]-[6],[13], along with observations of disease symptoms, death rates, and other factors.

Beyond purely numerical phylogenetic tree structures derived from time-consuming alignment methods, viral genomes can also be examined through constructing geometric representations or genomic walks [7]. Typically, the positions of amino acids along a sequence form a set of points that can be connected by lines to create a curve known as a DNA walk.

In some cases, depending on the visualization techniques used, these sets of points are easy to analyze visually or with simple statistical and signal-processing methods (as reviewed in Ref. [7]), without needing algorithms that are exponentially complex.

In Ref. [8], a new algorithm of this kind was proposed and examined, enabling the identification of consecutive amino acid sequences or their repeated combinations in DNAs. These points are connected by imaginary stretches, arranged along a simple curve that winds around a straight line that is easy to analyze [14]. Depending on the desired level of detail, the separate trajectories for each nucleic acid or any repeated amino acid patterns can be plotted in a single figure. One of these visualization techniques involves tracing ATG triplets, which start DNA codons or are parts of them.

Using these ATG walks, the evolution of SARS-CoV-2, MERS-CoV-1, dengue, and Ebola viruses was studied in 2021 [8]. Compared to the last two viruses, it was predicted that SARS-CoV-2 has a relatively stable ATG structure, meaning it is less affected by significant variations in codon length and their counts in viral sequences. Despite ongoing mutations, these virus properties remain valid today [7]-[9].

The techniques discussed above are applied here to analyze the Nipah virus sequences available from GenBank® [12].

3. Results

The ATG walks were derived from 32 complete viral genome samples of NiV-M, NiV-BD, and NiV-I. The simulation results are shown in Figures 1a,b, and in Table 1A (Appendix). It is observed that most traces (NiV-M and NiV-BD) are very similar, although the cumulative differences between them increase toward the end of the sequences.

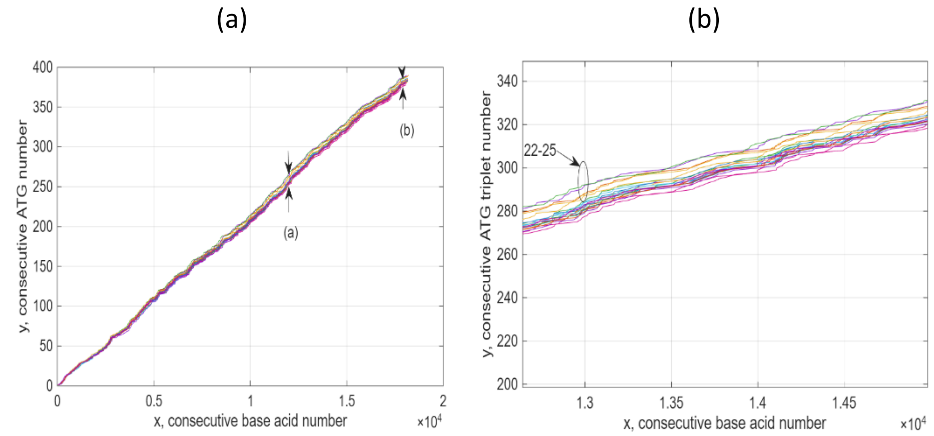

Fig. 1. ATG walks for 32 NiV sequences (a) and its enlarged plot (b) with a mark on the trajectories (rows 22-25, Table A1, Appendix) of Indian Kerala's outbreak samples.

Apart from the sequences mentioned, four DNA samples collected in Kerala in 2018 were modeled, and the curves labeled 22-25 are shown in the enlarged Figure 1b. It was observed that these curves are positioned relatively farther from the other 28s, especially beyond the centers of the sequences. This indicates that, in this region, the inter-ATG intervals are shorter than those in the NiV-M and NiV-BD samples.

All sequences can be quite similar in length, as microbiologists previously mentioned [2], but RNAs carry spatial information, and the differences between sequences lie in how the nucleotides are distributed in them.

In addition to visual analysis, the method enables quantitative estimates. For this purpose, the relative width of a set of 32 trajectories can be used.

For comparison, the normalized cumulative difference of ATG curves is calculated at two points at $x = 12620$ and 17930. It gives, correspondingly, $\delta y_{1,32} = 2\left(y_{32} - y_1\right)/\left(y_1 + y_{32}\right) \approx 6.2\%,\ 3.4\%$ after more than 27 years of observed evolution (1999-2025, Table 1A, Appendix).

To compare, the divergence ATG data for SARS-CoV-2, MERS-CoV, dengue, Ebola [9], and Nipah viruses are provided in Table 1 (row 4).

Table 1. ATG walk observation results for SARS-CoV-2, MERS-CoV, dengue, Ebola [9], and Nipah viruses.

1	Virus name	SARS-CoV-2	MERS-CoV	dengue	Ebola	Nipah
2	Observation years	2020-08.2025	2012-2020	1974-2019	1976-2019	1999-2025
3	Number of studied genomes	108	20	27	15	32
4	$\delta y_{1,32}$,%	1.5	2	14	9	3.4 - 6.2
5	D_F	0.0105	0.0216	0.1944	0.1011	0.0529

Analyzing these data, it should be noted that each virus has its own rate of evolution, determined by its nature and the speed of its mutation and spread among host organisms. The record in this regard belongs to the highly pathogenic SARS-CoV-2 virus, which has now developed hundreds of lineages and clades, and has killed more than seven million people worldwide [1]. Despite this, studies conducted in Refs. [7]-[9] for 108 most widespread sequence lineages did not show significant divergence in the SARS-CoV-2 ATG curves, similar to what was observed with MERS-CoV, and these viruses were recognized temporarily as relatively stable species. Although random mutations may reverse this opinion in the future.

Other viruses listed in Table 1 have longer evolutionary histories and have been separated into different clades and families, even within closely related geographic regions. Some of these clusters can be observed building ATG curves.

Besides estimates of divergence for ATG walks, a comprehensive statistical parameter—the fractal regularization dimension of inter-ATG distances—was computed for all 32 genomes. These distances are measured in nucleotide counts, and their distributions along the genomes were analyzed using the software tool FracLab 2.2 [15], as in Refs. [7]-[9]. Figure 2 shows the calculated fractal dimensions for all 32 genomic sequences.

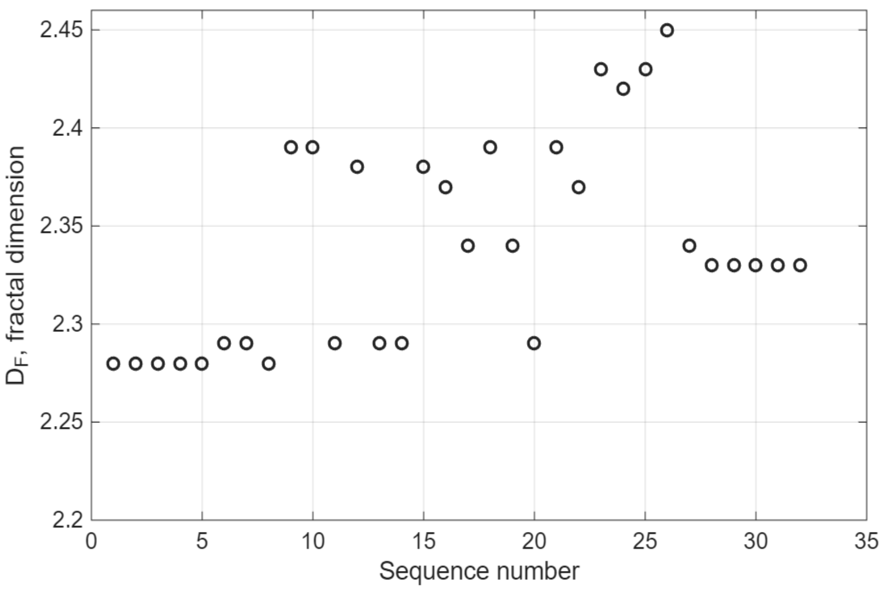

Figure 2. Fractal dimension D_F of inter-ATG distances calculated for 32 Nipah virus genomes.

As in Ref. [9], the standard deviation (*std*) was calculated for these points and compared with the data for the other five viruses from Refs. [8] and [9].

It is observed that the relatively stable viruses SARS-CoV-2 and MERS-CoV have considerably smaller standard deviations in their fractal dimension parameters compared to dengue, Ebola, and Nipah viruses, which have evolved toward clustering (Table 1, row 5).

4. Conclusions

This study applied a recently proposed ATG-walk method to Nipah virus genomes to identify and analyze significant mutations, as well as variations in codon length and their frequencies, across 32 human-origin genomes deposited in GenBank® between 1999 and 2025 by various contributors.

Despite a limited number of genomic samples, some properties can be generalized and applied to analyze them alongside previously studied genomes of SARS-CoV-2, MERS-CoV, dengue, and Ebola. Among them:

- Continuous mutations, along with variations in the number of ATG triplets, codons, and non-coding elements, lead to an increase in the divergence of ATG trajectories measured quantitatively.
- In some cases, specific ATG trajectories form distinct clusters associated with the virus seen visually.
- So far, a link between virus genomes capable of causing strong mutations and increased standard deviation in ATG walk fractal properties is known.

It's important to recognize that the viral landscape is highly unpredictable, and ongoing surveillance of viral genomes is strongly advised to prevent pandemics, as demonstrated with SARS-CoV-2. The ATG walk method is one of these monitoring tools used to detect severe virus mutations and can complement existing virus research methods.

Acknowledgements

The author thanks all researchers who have deposited the viral genomic sequences in the GenBank® and GISAID databases.

The author thanks the Faculty of Information Technology and Electroengineering at the Norwegian University of Science and Technology, NTNU, for a sabbatical research grant in 2026 and Mach-3dP Inc. (Burlington, Canada) for hosting during the sabbatical.

Conflict of interest

The author declares that they have no competing interests.

References

[1] Nipah virus fact sheet - World Health Organization (WHO). https://www.who.int/news-room/fact-sheets/detail/nipah-virus.

[2] D. Madhukalya, et al., Nipah virus: pathogenesis, genome, diagnosis, and treatment. *Appl Microbiol Biotechnol.* 2025, 109: 158. https://doi.org/10.1007/s00253-025-13474-6

[3] S. Whitmer S, et al., Inference of Nipah virus evolution, 1999–2015. *Virus Evol.* 2021, 7: veaa062. https://doi.org/10.1093/ve/veaa062

[4] Md. Rahman, et al., Development of a culture-independent whole-genome sequencing of Nipah virus using the MinION Oxford Nanopore platform. *Microbiol Spectrum*. 2025, 13: e02492-24. https://doi.org/10.1128/spectrum.02492-24

[5] O. Cortes-Azuero, et al., The genetic diversity of Nipah virus across spatial scales. *J Infect Dis*. 2024; 230: e1235. https://doi.org/10.1093/infdis/jiae221

[6] A. Sudeep, et al., Detection of Nipah virus in Pteropus medius in 2019 outbreak from Ernakulam district, Kerala, India. *BMC Infect Dis.* 2021, 21:162. https://doi.org/10.1186/s12879-021-05865-7

[7] A. Belinsky and G.A. Kouzaev, DNA walks in virus genomics. *JP J. Biostatistics*. 2024, 24:251-286. http://dx.doi.org/10.17654/0973514324017

[8] A. Belinsky and G.A. Kouzaev. Visual and quantitative analyses of virus genomic sequences using a metric-based algorithm. *WSEAS Trans Circ Syst.* 2022, 21:323-348. http://dx.doi.org/10.37394/23201.2022.21.35 . See as well: BioArxiv: A. Belinsky and G. Kouzaev (2021). https://www.biorxiv.org/content/10.1101/2021.06.17.448868v3 . https://doi.org/10.1101/2021.06.17.448868

[9] G.A. Kouzaev. ATG walks in virus genomics. In: *Proc. 2nd Int. Conf. Infectious Diseases and Applied Microbiology and Beneficial Microbes*, Vienna, Austria, 25-26 Sept., 2025, p. 46. https://www.researchgate.net/publication/396045850_ATG_Walks_in_Virus_Genomics

[10] J. Heather and B. Chain. The sequence of sequencers: The history of sequencing DNA. *Genomics*. 2016, 107:1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

[11] *The Sequencing Buyer's Guide*. https://lp.frontlinegenomics.com/sequencing-buyers-

guide-7th-edition-2025?utm_campaign=13004439-Sequencing%20Buyer%27s%20Guide%202025&utm_source=Wordpress&utm_medium=Wordpress&utm_term=Wordpress&utm_content=Wordpress Seen on April 1, 2026.

[12] Genbank ®. www.ncbi.nkm.nih.gov.genbank

[13] B. Hall, *Phylogenetic Trees Made Easy: A How-To Manual,* 5th Edition, Sinauer Assoc., 2018.

[14] A. Belinsky A and G.A. Kouzaev. Geometrical study of virus RNA sequences. *BioRxiv preprint*: 2021.09.06.459135. https://doi.org/10.1101/2021.09.06.459135

[15] *FracLab 2.2. A Fractal Analysis Toolbox for Signal and Image Processing*. https://project.inria.fr/fraclab/ Seen on 2026.03.25.

Appendix

Table 1A. *Henipavirus hipahense* (Nipah virus) genomic sequence data

No	GenBank complete sequence number	Country	Collection year	Number of nucleotides	Number of ATG triplets	Fractal dimension D_F
1	AJ564623.1	Malaysia	1999	18246	384	2.28
2	MK673562.1	Malaysia	1999	18231	383	2.28
3	AF212302.2	Malaysia	2001	18246	384	2.28
4	AY029767.1	Malaysia	2001	18246	384	2.28
5	AY029768.1	Malaysia	2001	18246	384	2.28
6	MK673565.1	Malaysia	2004	18235	384	2.29
7	MK633567.1	Bangladesh	2004	18245	383	2.29
8	AY988601.1	Bangladesh	2005	18252	384	2.28
9	FJ513078.1	India	2007	18252	378	2.39
10	FJ513078.1	India	2007	18252	378	2.39
11	MK673568.1	Malaysia	2008	18237	386	2.29
12	JN808863.1	Bangladesh	2008	18242	372	2.38
13	MK673571.1	Bangladesh	2011	18240	382	2.29
14	MK673573.1	Bangladesh	2011	18236	382	2.29

15	MK673578.1	Bangladesh	2011	18235	381	2.38
16	MK673581.1	Bangladesh	2012	18235	381	2.37
17	MK673591.1	Bangladesh	2014	18157	380	2.34
18	MK673592.1	Bangladesh	2014	18234	380	2.39
19	MK673590.1	Bangladesh	2014	18208	383	2.34
20	MK673584.1	Bangladesh	2015	18229	385	2.29
21	MK673585.1	Bangladesh	2015	18245	381	2.39
22	MH523642.1	India	2018	18242	390	2.37
23	MH396625.1	India	2018	18210	389	2.43
24	MH523640.1	India	2018	18132	390	2.42
25	MH523641.1	India	2018	18027	388	2.43
26	PP981674.1	Bangladesh	2022	18077	381	2.45
27	PP981675.1	Bangladesh	2022	18096	381	2.34
28	PP981676.1	Bangladesh	2023	18092	378	2.33
29	PP981683.1	Bangladesh	2023	18090	382	2.33
30	PQ368168.1	Bangladesh	2023	18090	381	2.33
31	PV132707.1	Bangladesh	2024	18128	381	2.33
32	PX130166.1	Bangladesh	2025	18128	381	2.33