Matrix-Based Modeling of DNA/RNA Virus Sequences for Mutation Differences and Similarity Analysis Using Euclidean Distance and Cosine Similarity

Varistha Devi - 13524135[1,2]
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
[1]13524135@itb.ac.id, [2]varistha.devi@gmail.com

***Abstract*—A virus is an infectious microbe that consists of short sequences of nucleic acid, either DNA or RNA, wrapped in a protein shell. These genetic molecules are composed of nitrogen bases such as adenine, guanine, cytosine, and thymine (or uracil in RNA), which form the fundamental units of genetic information. This paper proposes a mathematical and computational approach to model short viral genetic sequences as numerical vectors using one-hot encoding and matrix representations. Euclidean distance and cosine similarity are then applied to quantify mutation differences and structural similarity between sequences. A Python-based program is developed to implement this methodology and facilitate experimentation on both synthetic and real sequence data. The results demonstrate that the proposed model can effectively reflect increasing mutation differences through higher Euclidean distances and lower cosine similarity values. This study is intended as a conceptual and educational exploration of how linear algebra can be applied to symbolic biological data, rather than as a biological or evolutionary analysis of viruses.**

***Keywords*—Cosine similarity, Euclidean distance, Genetic sequence modeling, One-hot encoding.**

I. Introduction

A virus is an infectious microbe that consists of short sequences of nucleic acid wrapped in a protein shell. These nucleic acids may be either DNA or RNA, both of which are composed of nitrogen bases that serve as the basic units of genetic information. DNA consists of Adenine (A), Guanine (G), Thymine (T), and Cytosine (C), while RNA replaces Thymine with Uracil (U) and retains the remaining components. From a computational and mathematical standpoint, these genetic sequences can be viewed as symbolic data rather than biological mechanisms.

As genetic sequences undergo mutations, differences arise between virus strains. Measuring the similarity or difference between such strains is an important task in various analytical contexts, such as comparative genomic studies and epidemiological surveillance. While these analyses are commonly approached from a biological perspective, the underlying sequence data can also be modeled and examined using mathematical tools, particularly those from linear algebra.

In this paper, we explore a matrix-based representation of DNA/RNA virus sequences to analyze variation differences and structural similarity between virus strains. By encoding sequences into vectors in Euclidean space, we apply Euclidean distance to quantify mutation differences and cosine similarity to measure similarity between sequence representations.

To support the analysis and computation, a Python-based program is developed, and constraints on sequence length are introduced to ensure computational feasibility and conceptual clarity. Specifically, this study considers short sequence segments of approximately 30–50 bases, represented as strings of characters and transformed into matrices prior to analysis. This restriction reflects the educational scope of the study and does not aim to model complete viral genomes.

Therefore, this paper aims to model DNA/RNA sequences using matrix representations, apply Euclidean distance and cosine similarity to quantify mutation differences and similarity, and provide a comparative discussion of the interpretation of these metrics within a mathematical framework

II. Theoretical Framework

A. Brief Background on Viruses and Genetic Sequences

In definition, a virus is an infectious microbe consisting of a segment of nucleic acid, either DNA or RNA, encased in a protein shell that protects and delivers its genetic material into host cells. Viruses cannot replicate on their own and depend entirely on the cellular machinery of a host organism to reproduce and make copies of themselves, distinguishing them from free-living organisms such as bacteria and other single-cell

microbes.

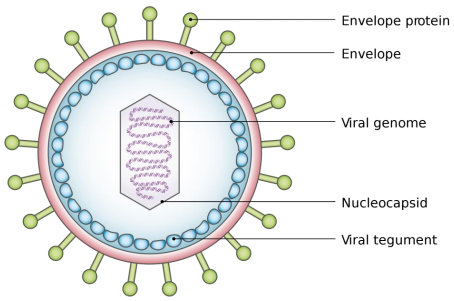

Figure 1. Basic Virus Structure
Source:
https://open.oregonstate.education/generalmicrobiology/chapter/introduction-to-viruses/

The complete infectious form of a virus outside a host cell is referred to as a virion, which includes both the genetic material and the surrounding structural proteins. This genetic material encodes the information unique to each virus and is essential for initiating infection once the virion enters a susceptible cell. In most viruses, the nucleic acid may be either DNA or RNA, but not both at the same time, and can vary in structure (e.g., single-stranded or double-stranded) depending on the virus type.

Because viral genomes are relatively short compared to those of cellular organisms, they often consist of only a few thousand to tens of thousands of nucleotide bases. These bases — adenine (A), cytosine (C), guanine (G), thymine (T) for DNA, and uracil (U) replacing thymine in RNA — form ordered sequences that encode proteins and regulatory elements critical for virus replication and assembly. Viewed from a computational perspective, these linear sequences of nucleotide symbols provide a convenient framework for mathematical representation and analysis, separate from their biological function.

B. Representation of Genetic Sequences as Strings

Genetic sequences such as DNA or RNA can be viewed as ordered sequences of symbols drawn from a finite set of nucleotide bases. This set consists of {A, G, C, T} for DNA, while RNA replaces the T with U. From a computational perspective, this means that a genetic sequence can naturally be represented as a string over a finite alphabet, similar to how texts are represented as strings of characters.

This abstraction allows genetic data to be separated from its biological context and instead treated as structured symbolic data. Such representations are common in bio-informatics and computational biology, where sequence data is often handled in the form of strings before being transformed into numerical representations for further analysis. By viewing them this way, we obtain a format that is compatible with formal mathematical and computational tools.

C. Vector Spaces and Euclidean Space

In mathematics, a vector space is a set of objects, called vectors, that can be added together and multiplied by scalars while satisfying certain axioms. One of the most familiar examples is the Euclidean space $\mathbb{R}^n$, whose elements are n-dimensional vectors with real-valued components.

Once genetic sequences are transformed into numerical form, they can be interpreted as vectors in a high-dimensional Euclidean space. This makes it possible to analyze them using geometric and algebraic concepts such as distance, magnitude, and angle. In this framework, each sequence corresponds to a point in space, and comparisons between sequences can be interpreted as geometric relationships between these points.

D. Matrix and Vector Representation of Sequences

To move from symbolic strings to numerical objects, genetic sequences are encoded using a method known as one-hot encoding. In this scheme, each nucleotide base is represented by a binary vector of length four. For example, the bases A, C, G, and T are encoded as (1,0,0,0), (0,1,0,0), (0,0,1,0), and (0,0,0,1), respectively.

A sequence of length n can therefore be represented as a binary matrix of size 4 × n, where each column corresponds to the encoded form of a nucleotide at a specific position. This matrix can then be flattened into a single vector in $\mathbb{R}^{4n}$, allowing the sequence to be treated as a vector suitable for linear algebraic operations.

E. Vector Norm and Euclidean Distance

The norm of a vector provides a measure of its magnitude. In Euclidean space, the Euclidean norm of a vector $x = (x_1, x_2, \ldots x_n)$ is defined as:

$$\| x \| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$$

Using this definition, the Euclidean distance between two vectors, a and b, can be evaluated as:

$$d(a, b) = \| a - b \|$$

When applied to encoded genetic sequences, this distance reflects the overall difference between two sequences. Differences in nucleotide positions result in differences in vector components, and the Euclidean distance captures the cumulative effect of these differences. In this study, Euclidean distance is used as a measure of mutation difference between sequences.

F. Dot Product and Cosine Similarity

The dot product of two vectors $a, b \in \mathbb{R}^n$ is defined as:

$$a \cdot b = \sum_{i=1}^{n} a_i b_i$$

From this, the cosine similarity of those two vectors

can be evaluated as:

$$cos(a, b) = \frac{a \cdot b}{\| a \| \| b \|}$$

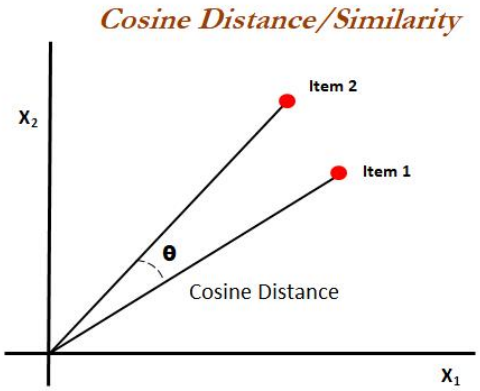

Figure 2. Geometric interpretation of cosine similarity as the angle between two vectors
Source: https://pub.aimind.so/understanding-cosine-similarity-and-cosine-distance-in-depth-cc91eac3ef2

Cosine similarity measures the cosine of the angle between two vectors and therefore captures how similar their directions are, independent of their magnitude. In the context of encoded genetic sequences, cosine similarity emphasizes structural similarity between sequences rather than the absolute number of differences. This makes it a useful complement to Euclidean distance in analyzing relationships between sequences.

III. METHODOLOGY

This study adopts a computational and mathematical approach to analyze genetic sequence similarity and variation using linear algebraic tools. The overall process consists of preparing genetic sequence data, transforming it into numerical representations, computing similarity and distance metrics, and interpreting the resulting values. All computations and experiments are implemented using the Python programming language.

The genetic data used in this study consists of short DNA or RNA sequences represented as strings composed of the characters A, C, G, and T (or U for RNA). For the purposes of this project, the length of each sequence is restricted to approximately 30 to 50 bases in order to ensure computational simplicity and ease of interpretation. Two types of data are considered: synthetic sequences constructed manually to simulate mutations, and example sequences obtained from public biological databases for testing and validation. Each sequence is treated purely as an ordered string of symbols, independent of its biological function. To preserve biological plausibility, sequences are restricted to contain either thymine (T) or uracil (U), but not both, and comparisons are only performed between sequences of the same type (DNA–DNA or RNA–RNA).

To enable mathematical processing, the symbolic sequences are transformed into numerical form using one-hot encoding. Each nucleotide base is mapped to a four-dimensional binary vector, where A is represented as (1,0,0,0), C as (0,1,0,0), G as (0,0,1,0), and T or U as (0,0,0,1). A sequence of length n is therefore represented as a binary matrix of size 4 × n, with each column corresponding to one nucleotide position. This matrix is then flattened into a single vector in $\mathbb{R}^{4n}$, allowing the use of standard vector operations from linear algebra.

Once the sequences are encoded as vectors, two quantitative measures are applied to compare them. The first is Euclidean distance, which measures the overall magnitude of difference between two sequences by computing the norm of their difference vector. The second is cosine similarity, which measures the angle between two vectors and therefore reflects how structurally similar they are, independent of magnitude. These two measures are used together in order to capture both the degree of mutation and the structural alignment between sequences.

The experimental procedure consists of selecting a reference sequence and generating mutated variants by altering one or more nucleotide positions. Each mutated sequence is then compared with the original sequence using both Euclidean distance and cosine similarity. By repeating this process with different numbers and positions of mutations, it becomes possible to observe how increasing variation affects the resulting similarity and distance values. If invalid inputs are provided, such as sequences of unequal length or incompatible types, the program rejects the input and prompts the user to re-enter the data.

All encoding, transformation, and computation steps are implemented using Python, with the NumPy library used for efficient matrix and vector operations. The program is designed in a modular manner so that real biological data can later be substituted for synthetic data without modifying the underlying methodology.

Finally, it is important to note that this study is intentionally limited in scope. The model does not account for biological properties such as gene function, evolutionary constraints, or biochemical interactions. The analysis therefore reflects purely mathematical similarity between symbolic representations of genetic sequences rather than biological relatedness. The results should be interpreted within this computational and educational context.

IV. RESULTS AND DISCUSSIONS

To evaluate the proposed sequence representation and similarity measures, several test cases were constructed to reflect different biological relationships between genetic sequences. These range from identical sequences to unrelated viruses, allowing the behavior of Euclidean distance and cosine similarity to be observed across varying degrees of genetic similarity. The resulting values were then analyzed to determine whether the mathematical model aligns with intuitive expectations of

biological relatedness.

Test Case Overview

The following sequence pairs were used:

- Case A: SARS-CoV-2 reference sequence (NC_045512.2) compared with itself (control).

Table 1. Genetic Sequences and Genomic Regions for Test Case A

Seq. Number	Sequence	Region
1	ATGTTTGTTTTTCTTGTTTTATTGCCA CTAGTCTCTAGTCAGTGTGTTAA	21563-21612
2	ATGTTTGTTTTTCTTGTTTTATTGCCA CTAGTCTCTAGTCAGTGTGTTAA	21563-21612

- Case B: SARS-CoV-2 reference (NC_045512.2) compared with a Delta variant isolate (MZ437368.1).

Table 2. Genetic Sequences and Genomic Regions for Test Case B

Seq. Number	Sequence	Region
1	ATGTTTGTTTTTCTTGTTTTATTGCCA CTAGTCTCTAGTCAGTGTGTTAA	21563-21612
2	GTTTGTTTTTCTTGTTTTATTGCCACT AGTCTCTAGTCAGTGTGTTAATC	21563-21612

- Case C: Delta variant (MZ437368.1) compared with Omicron variant (OL672836.1).

Table 3. Genetic Sequences and Genomic Regions for Test Case C

Seq. Number	Sequence	Region
1	GTTTGTTTTTCTTGTTTTATTGCCACT AGTCTCTAGTCAGTGTGTTAATC	21563-21612
2	CAATTACCCCCTGCATACACTAATTC TTTCACACGTGGTGTTTATTACCC	21563-21612

- Case D: SARS-CoV-2 (NC_045512.2) compared with Human coronavirus OC43 (NC_006213.1, same virus family, different species).

Table 4. Genetic Sequences and Genomic Regions for Test Case D

Seq. Number	Sequence	Region
1	ATGTTTGTTTTTCTTGTTTTATTGCCA CTAGTCTCTAGTCAGTGTGTTAA	21563-21612
2	AACATGTTTTTGATACTTTTAATTTCC TTACCAACGGCTTTTGCTGTTAT	23640-23689

- Case E: SARS-CoV-2 (NC_045512.2) compared with Human adenovirus 2 (AC_000007.1, unrelated virus).

Table 5. Genetic Sequences and Genomic Regions for Test Case E

Seq. Number	Sequence	Region
1	ATGTTTGTTTTTCTTGTTTTATTGCCACTAG TCTCTAGTCAGTGTGTTAA	21563-21612
2	TCGCTACGGCGGCGGCGGAGTTTGGCCGT AGGTGGCGCCCTCTTCCTCCC	10000-10049

All sequences were extracted from GenBank using the same or corresponding genomic regions to ensure comparability, and all sequences were truncated to equal length prior to computation.

Results and Interpretation

The control experiment in Case A produced a Euclidean distance of 0.0 and a cosine similarity of approximately 1.0, confirming that the encoding and comparison process behaves correctly when identical sequences are compared. This serves as a baseline for interpreting the remaining results.

In Case B, comparing the SARS-CoV-2 reference with a Delta variant isolate yielded a Euclidean distance of 7.87 and a cosine similarity of 0.38. This reflects the presence of mutations between the two sequences, resulting in a noticeable but not extreme level of divergence. The moderate distance and reduced similarity indicate that the sequences remain closely related but are no longer structurally identical in the vector space.

Case C, which compares two different variants of SARS-CoV-2 (Delta and Omicron), shows a further increase in Euclidean distance to 8.37 and a decrease in cosine similarity to 0.30. This suggests that two independently evolved variants differ more substantially from each other than either does from the original reference strain. This behavior aligns with biological expectations and demonstrates that the model is sensitive to increasing levels of mutation.

In Case D, the comparison between SARS-CoV-2 and Human coronavirus OC43 yielded a Euclidean distance of 7.35 and a cosine similarity of 0.46. Interestingly, while the distance is comparable to that observed in Case B, the cosine similarity is higher. This suggests that although the two sequences differ at many positions, their overall structural composition in the encoded vector space remains relatively aligned due to shared evolutionary origins within the coronavirus family. This highlights an important distinction between the two metrics: Euclidean distance captures positional differences, whereas cosine similarity reflects structural or directional similarity.

Case E compares SARS-CoV-2 with Human adenovirus 2, which is biologically unrelated. This produced the largest Euclidean distance, 8.60, and the lowest cosine similarity, 0.26. This indicates a high degree of dissimilarity and confirms that the model effectively distinguishes unrelated sequences from those that are evolutionarily related.

General Trends

Across all test cases, a consistent trend is observed: as biological relatedness decreases, Euclidean distance

increases and cosine similarity decreases. This monotonic relationship suggests that the proposed vector representation combined with these similarity measures captures meaningful patterns of variation in genetic sequences, even within the simplified framework of this study.

It is important to emphasize that the numerical values obtained do not represent biological or evolutionary distance in a strict sense. The model does not account for functional genomics, selection pressure, or biochemical relevance of mutations. Instead, it measures similarity between symbolic representations of genetic sequences, and its purpose is illustrative rather than predictive. Nevertheless, the results demonstrate that linear algebraic tools can be effectively applied to genetic sequence data in a computational context.

Discussion of Limitations

Several limitations should be acknowledged. The use of short sequence fragments restricts the biological representativeness of the analysis, and the one-hot encoding scheme treats all nucleotide substitutions equally, regardless of their biological impact. Additionally, insertion and deletion mutations are not handled by this model, and all sequences must be of equal length. These simplifications were necessary to keep the model computationally tractable and aligned with the educational scope of this project.

Despite these limitations, the observed results are consistent with intuitive expectations and demonstrate the feasibility of using matrix-based representations and linear algebraic similarity measures to analyze genetic variation in a simplified and controlled setting.

V. Conclusion

This paper set out to explore how genetic sequences can be represented and analyzed using tools from linear algebra. By framing DNA and RNA sequences as symbolic data and transforming them into vector representations through one-hot encoding, it became possible to apply mathematical similarity and distance measures to compare virus strains in a structured and computationally simple way.

The results show that Euclidean distance and cosine similarity behave in a manner that is consistent with intuitive expectations of genetic relatedness. Identical sequences yield zero distance and maximal similarity, while increasing levels of mutation and biological divergence correspond to larger distances and lower similarities. Comparisons between different variants, between different species within the same virus family, and between completely unrelated viruses all exhibit distinguishable patterns in the resulting metrics, suggesting that the model captures meaningful structural differences between sequences.

At the same time, this study emphasizes that these measures reflect mathematical similarity rather than biological significance. The model does not account for gene function, evolutionary pressure, or biochemical effects of mutations, and therefore its results should be interpreted within a computational and educational context. The purpose of this approach is not to replace biological analysis, but to provide an alternative lens through which genetic data can be explored and understood.

Overall, this work demonstrates that even simple linear algebraic tools can be effectively applied to biological sequence data when appropriate abstractions are used. This highlights the value of mathematical modeling as a way to structure, interpret, and experiment with complex real-world data, and shows how concepts from linear algebra can be connected to applications beyond purely theoretical settings.

VI. Appendix

Github Repository Link : https://github.com/Beeziebloop/GenSeq-Matrix-Similarity-Analyzer

VII. Acknowledgment

First and foremost, the author would like to express her deepest gratitude to Allah SWT, for His grace and guidance, without which the completion of this paper would not have been possible. Secondly, the author would like to extend her sincere thanks to Mr. Arrival Dwi Sentosa, for his guidance, patience, and dedication throughout the IF2123 Linear Algebra and Geometry course, which greatly contributed to the development of this work. The author would also like to thank her family and friends for their continuous support, understanding, and encouragement, especially during challenging moments throughout the writing process. The author is truly grateful for all those who have contributed, directly or indirectly, to the completion of this work.

References

[1] National Human Genome Research Institute, “Virus,” Genetics Glossary. Available: https://www.genome.gov/genetics-glossary/Virus. [Accessed: December 23, 2025]
[2] E. V. Koonin, V. V. Dolja, and M. Krupovic, “The logic of virus evolution,” Cell Host & Microbe, vol. 27, no. 5, pp. 749–758, May 2020. [Accessed: December 23, 2025]
[3] M. Li and J. Liu, “Representation of DNA sequences as strings and symbolic data,” MATCH Communications in Mathematical and in Computer Chemistry, vol. 60, no. 2, pp. 291–300, 2008. [Accessed: December 23, 2025]
[4] A. P. Belkacem, “Mathematical modeling of biological sequences,” African Journal of Pure and Applied Research, vol. 2, no. 1, pp. 14–21, 2021. [Accessed: December 23, 2025]
[5] J. Pei, “Sequence analysis and symbolic representation,” Information Sciences, vol. 180, no. 12, pp. 2356–2370, 2010. [Accessed:]
[6] E. W. Weisstein, “Cosine Similarity,” MathWorld — A Wolfram Web Resource. [Accessed: December 23, 2025]
[7] G. Strang, Linear Algebra and Its Applications, 4th ed. Belmont, CA: Brooks/Cole, 2006. [Accessed: December 23, 2025]
[8] R. A. Horn and C. R. Johnson, Matrix Analysis. Cambridge, U.K.: Cambridge Univ. Press, 1985. [Accessed: December 23, 2025]
[9] National Center for Biotechnology Information, “GenBank Overview,” NCBI. [Accessed: December 23, 2025]

STATEMENT

With this, I hereby declare that this paper that I have written is my own, not an adaptation or translation of someone else's work, and not plagiarized.

Bandung, 24 Desember 2025

Varistha Devi 13524135