

Article

A 5-Adic Ultrametric Framework for Alignment-Free Phylogenetic Analysis of Hantavirus RNA Sequences

Anselmo Torresblanca-Badillo 

Departamento de Matemáticas, Física y Ciencia de Datos, Universidad del Norte, Km 5 Vía Puerto Colombia, Barranquilla 081007, Colombia; atorresblanca@uninorte.edu.co

Abstract

We develop a non-Archimedean framework for the representation and analysis of genomic sequences based on the arithmetic and geometric structure of the ring of 5-adic integers. The proposed approach associates RNA sequences with points in a compact ultrametric space through an injective symbolic-to-arithmetic embedding that transforms genomic information into a hierarchical geometric object. We prove that the embedding is a global isometry between a natural symbolic prefix metric and the induced 5-adic metric, and we show that its image forms a compact Cantor-type subset of $\mathbb{Z}_5$. Building upon this representation, we formulate a continuous-time evolutionary model governed by a Vladimirov pseudo-differential operator. The resulting non-Archimedean diffusion equation provides a mathematically rigorous mechanism for describing evolutionary transitions across hierarchical genomic scales and admits an explicit fundamental solution obtained through 5-adic Fourier analysis. We further introduce a finite-resolution projection onto quotient rings of $\mathbb{Z}_5$ and develop an alignment-free phylogenetic inference framework based directly on the 5-adic valuation. The induced distance function is ultrametric and naturally encodes hierarchical relationships through shared symbolic prefixes. The proposed construction establishes a bridge between p -adic analysis, ultrametric geometry, pseudo-differential operators, and computational phylogenetics. As an illustration, we discuss its application to Hantavirus genomic sequences, demonstrating how hierarchical evolutionary organization can be represented within a unified non-Archimedean mathematical framework.

Keywords: p -adic analysis; genomic embedding; non-Archimedean diffusion; alignment-free phylogenetics; Hantavirus genomes; Cantor-type sets

MSC: 11S80; 35S05; 46S10; 92D15; 92D20



Received: 6 June 2026

Revised: 3 July 2026

Accepted: 7 July 2026

Published: 10 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The analysis of genomic evolution has traditionally relied on probabilistic models defined on discrete sequence spaces, multiple sequence alignments, and phylogenetic reconstruction techniques based on pairwise similarity measures. While these methodologies have achieved remarkable success in molecular evolution and comparative genomics, they often require heuristic alignment procedures and may not fully exploit the hierarchical structure naturally present in evolutionary processes. The search for mathematically rigorous frameworks capable of representing genomic organization and phylogenetic divergence has therefore become an active area of interdisciplinary research connecting mathematics, theoretical biology, and computational genomics; see, e.g., [1–6] and the references therein.

One mathematical paradigm that has attracted increasing attention in the study of hierarchical systems is the theory of ultrametric spaces and p -adic analysis. Since the pioneering works of [7–10], non-Archimedean analysis has evolved into a mature mathematical discipline with applications ranging from mathematical physics and stochastic processes to complex systems, information theory, and biological modeling; see, e.g., [11–17]. Unlike Euclidean geometry, where proximity is determined by absolute magnitude, ultrametric geometry encodes similarity through hierarchical nesting structures. Such a viewpoint is particularly appealing in evolutionary biology, where genomes derived from common ancestors naturally organize themselves into nested families reflecting successive levels of divergence.

The mathematical foundations of p -adic analysis have led to the development of a rich theory of pseudo-differential operators, harmonic analysis, stochastic dynamics, and non-local diffusion equations on ultrametric spaces. In recent years, considerable attention has been devoted to the study of generalized non-Archimedean pseudo-differential operators and their associated evolution equations. Important contributions in this direction include the investigation of generalized classes of pseudo-differential operators and their applications [18–23], the analysis of pseudo-differential operators acting on Sobolev spaces associated with negative definite functions [24], and the study of operator classes related to Bessel-type potentials [9,25]. These developments have substantially expanded the analytical framework available for modeling nonlocal phenomena on ultrametric spaces and have provided powerful tools for the investigation of hierarchical dynamical systems.

The relevance of non-Archimedean methods extends beyond pure mathematics. Ultrametric structures arise naturally in spin glasses, protein dynamics, evolutionary landscapes, cognitive models, and complex networks. In biological contexts, ultrametric spaces provide a natural language for describing evolutionary relationships because distances are determined by common ancestry rather than by geometric displacement. Consequently, p -adic models offer a mathematically consistent framework for representing phylogenetic hierarchies and evolutionary branching processes; see, for example, [26–29].

Motivated by these developments, the present work proposes a non-Archimedean framework for the analysis of Hantavirus genomic sequences. Hantaviruses constitute an important group of negative-sense single-stranded RNA viruses whose evolutionary dynamics continue to attract considerable attention due to their epidemiological relevance and remarkable genetic diversity [30–32]. Rather than adopting conventional alignment-based methodologies, we construct an ultrametric representation of genomic sequences by embedding RNA strings into the ring of 5-adic integers $\mathbb{Z}_5$. This embedding transforms symbolic genomic information into a geometric object endowed with a natural hierarchical topology.

The central idea of the paper is that sequence similarity can be represented through the 5-adic valuation. Within this framework, genomic sequences sharing long common prefixes correspond to points that are close in the ultrametric sense, while early symbolic divergences generate larger 5-adic distances. This observation leads naturally to the construction of a Cantor-type genomic space whose geometry is entirely determined by symbolic conservation patterns.

Building upon this embedding, we introduce a continuous-time evolutionary model based on the Vladimirov pseudo-differential operator. The resulting non-Archimedean diffusion equation provides a mathematically rigorous mechanism for describing transitions between genomic configurations across hierarchical scales. In contrast with classical diffusion models driven by local Euclidean interactions, the proposed dynamics are intrinsically nonlocal and naturally compatible with the ultrametric organization induced by the genomic embedding.

The contributions of the present work may be summarized as follows:

- (i) We construct an injective 5-adic embedding of RNA sequences and prove that the induced metric is globally isometric to a symbolic prefix metric.
- (ii) We establish that the image of the genomic embedding forms a compact Cantor-type subset of $\mathbb{Z}_5$, thereby providing a rigorous ultrametric geometric representation of genomic information.
- (iii) We formulate non-Archimedean evolutionary dynamics governed by a Vladimirov pseudo-differential diffusion equation and derive its associated transition kernel.
- (iv) We develop an alignment-free phylogenetic inference methodology based directly on the 5-adic valuation and finite quotient projections of $\mathbb{Z}_5$.
- (v) We illustrate how the proposed framework can be applied to Hantavirus genomic sequences, thereby establishing a bridge between p -adic analysis, ultrametric geometry, and computational phylogenetics.

In recent years, non-Archimedean methods have been increasingly employed in theoretical biology to investigate the genetic code, DNA sequence representations, hierarchical organization of biological information, and evolutionary models within ultrametric spaces [11,12,33]. These studies have demonstrated that p -adic analysis provides a natural mathematical language for describing hierarchical structures in molecular biology. The present work builds upon this line of research but pursues a different objective. Rather than focusing on the algebraic organization of the genetic code or on qualitative models of genomic hierarchy, we develop a unified mathematical framework for alignment-free phylogenetic analysis based on an injective embedding of complete RNA sequences into $\mathbb{Z}_5$, an intrinsic ultrametric distance induced by the 5-adic valuation, finite-resolution projections suitable for empirical genomic data, and a nonlocal evolutionary model governed by the Vladimirov operator. In this way, the proposed framework extends previous p -adic approaches by providing a mathematically rigorous connection between non-Archimedean analysis and phylogenetic inference.

The remainder of the paper is organized as follows: In Section 2, we review the mathematical foundations required throughout the article, including 5-adic numbers, ultrametric spaces, Haar measure, Fourier analysis, and basic tools from non-Archimedean harmonic analysis. Section 3 introduces the genomic embedding theory, establishes its injectivity and isometric properties, and proves that the image of the embedding is a Cantor-type subset of $\mathbb{Z}_5$. In Section 4, we formulate the non-Archimedean diffusion model based on the Vladimirov operator and derive its fundamental solution. Section 5 develops the alignment-free phylogenetic inference framework by projecting genomic embeddings onto finite quotient rings and defining ultrametric distance matrices suitable for hierarchical clustering. Section 6 illustrates the practical application of the proposed framework to representative Hantavirus genomic configurations, highlighting the implementation of the 5-adic embedding and ultrametric distance for phylogenetic inference. The section further provides a conceptual comparison with representative classical and alignment-free phylogenetic methodologies and discusses the robustness, biological interpretation, and current limitations of the proposed non-Archimedean framework. Finally, Section 7 illustrates the practical implementation of the proposed framework through representative Hantavirus genomic sequences, highlighting the biological interpretation of the 5-adic embedding, ultrametric distance, and phylogenetic reconstruction procedures developed throughout the paper.

2. Mathematical Preliminaries

The purpose of this section is to establish the mathematical foundations required for the construction of the proposed ultrametric framework for genomic analysis. We briefly review the theory of p -adic numbers, ultrametric spaces, Haar measure, and the Vladimirov

pseudo-differential operator, emphasizing the properties that will later be used to model Hantavirus genomic evolution.

Let p be a prime number. For any non-zero rational number $x \in \mathbb{Q}$, there exists a unique representation of the form

$$x = p^\gamma \frac{b}{a},$$

where $a, b, \gamma \in \mathbb{Z}$, with $\gcd(a, p) = \gcd(b, p) = 1$, where $\gcd$ denotes the greatest common divisor so that neither a nor b is divisible by p . The integer $\gamma := \text{ord}_p(x)$, with $\text{ord}_p(0) := \infty$, is called the p -adic order (or p -adic valuation) of x ; equivalently, it is the exponent of the highest power of p dividing x .

The p -adic absolute value of $x \neq 0$ is defined by $|x|_p = p^{-\gamma}$, where γ is determined by the above decomposition of x , while $|0|_p = 0$. This absolute value endows $\mathbb{Q}$ with the structure of a non-Archimedean normed field. The completion of $\mathbb{Q}$ with respect to the induced metric is the field of p -adic numbers, denoted by $\mathbb{Q}_p$.

Unlike the real numbers, where distance is determined by magnitude, the p -adic metric measures divisibility by powers of the prime p . Consequently, two numbers are considered close whenever their difference is divisible by a high power of p . The induced metric is given by $d_p(x, y) = |x - y|_p$.

Since

$$|x + y|_p \leq \max\{|x|_p, |y|_p\}, \quad x, y \in \mathbb{Q}_p,$$

we have that

$$d_p(x, y) \leq \max\{d_p(x, z), d_p(z, y)\},$$

for all $x, y, z \in \mathbb{Q}_p$. A metric satisfying this condition is called an *ultrametric*.

In the present work, we specifically employ the case $p = 5$, since this choice provides a natural injective encoding of the four nucleotides together with an additional symbol used for structural gaps or segmentation markers within a single non-Archimedean digit system. This ensures a faithful representation of genomic sequences within the 5-adic expansion framework.

Every element $x \in \mathbb{Q}_5$ admits a unique series expansion of the form

$$x = \sum_{i=k}^{\infty} x_i 5^i, \quad k = \text{ord}_5(x) \in \mathbb{Z}, \quad x_i \in \{0, 1, 2, 3, 4\}, \quad x_k \neq 0.$$

The subset

$$\mathbb{Z}_5 = \{x \in \mathbb{Q}_5 : |x|_5 \leq 1\},$$

is called the ring of 5-adic integers. Elements of $\mathbb{Z}_5$ possess expansions involving only non-negative powers of 5

$$x = \sum_{i=0}^{\infty} x_i 5^i.$$

The space $\mathbb{Z}_5$ is compact, complete, and totally disconnected. Its topology is generated by balls of the form

$$B_r(a) = \{x \in \mathbb{Q}_5 : |x - a|_5 \leq r\},$$

where $r = 5^{-m}$ for some integer $m \geq 0$.

An important consequence of ultrametricity is that every point of a ball acts as its center. Moreover, any two balls are either disjoint or one is contained in the other. As a result, ultrametric spaces exhibit an intrinsic hierarchical organization that differs fundamentally from the geometry of Euclidean spaces. Their structure is characterized by nested families of balls, while every triangle is either equilateral or isosceles with a uniquely shorter base. These features naturally induce tree-like configurations and multilevel classifications.

Such properties make $\mathbb{Z}_5$ particularly well suited for modeling hierarchical biological systems, including phylogenetic trees and mutation networks. In evolutionary settings, genomes descending from common ancestors tend to form nested groups corresponding to successive levels of divergence. The ultrametric framework captures this organization directly through its topology, without the need for additional clustering procedures.

In the 5-adic representation adopted in this work, two genomic sequences sharing a long common initial segment are associated with points whose ultrametric distance is exponentially small. Consequently, increasing sequence similarity corresponds to increasing proximity in $\mathbb{Z}_5$, allowing evolutionary conservation patterns to be encoded naturally within the underlying ultrametric topology.

Integration on p -adic spaces requires the introduction of a translation-invariant measure. Consider the additive locally compact group $(\mathbb{Q}_5, +)$. By Haar's theorem, there exists a unique translation-invariant measure dx , known as the *Haar measure*, which can be normalized by the condition

$$\int_{\mathbb{Z}_5} dx = 1.$$

If $A \subset \mathbb{Q}_5$ is a measurable set and $a \in \mathbb{Q}_5$, then the Haar measure satisfies the invariance property $\mu(A + a) = \mu(A)$.

This measure constitutes a fundamental tool in p -adic harmonic analysis, providing the basis for the definition of integral transforms, pseudo-differential operators, and non-Archimedean diffusion models.

From the perspective of genomic evolution, the Haar measure supplies a natural probabilistic framework on the ultrametric sequence space. In particular, it allows one to describe the distribution of genomic configurations and to formulate stochastic evolutionary dynamics within the hierarchical geometry of $\mathbb{Z}_5$.

For every $x \in \mathbb{Q}_5$, we define its fractional part, denoted by $\{x\}_5$, as

$$\{x\}_5 = \begin{cases} 0, & \text{if } x = 0 \text{ or } \text{ord}_5(x) \geq 0, \\ 5^{\text{ord}_5(x)} \sum_{j=0}^{-\text{ord}_5(x)-1} x_j 5^j, & \text{if } \text{ord}_5(x) < 0. \end{cases}$$

A complex-valued function φ on $\mathbb{Q}_5$ is said to be *locally constant* if, for every $x \in \mathbb{Q}_5$, there exists an integer $l = l(x)$ such that $\varphi(x + x') = \varphi(x)$ for any $x' \in B_l$.

We denote by $\mathcal{E}(\mathbb{Q}_5)$ the vector space of all locally constant complex-valued functions on $\mathbb{Q}_5$.

A function $\varphi : \mathbb{Q}_5 \rightarrow \mathbb{C}$ is called a *Bruhat–Schwartz function* (or *test function*) if it is locally constant and has compact support. The space of all such functions is denoted by $\mathcal{D}(\mathbb{Q}_5)$.

If $\varphi \in \mathcal{D}(\mathbb{Q}_5)$, then there exists an integer l such that

$$\varphi(x + x') = \varphi(x), \quad x' \in B_l, \quad x \in \mathbb{Q}_5.$$

The largest integer with this property is called the *parameter of constancy* of φ and is denoted by $l(\varphi)$.

The additive character on $\mathbb{Q}_5$ is defined by

$$\chi_5(y) = \exp(2\pi i \{y\}_5), \quad y \in \mathbb{Q}_5.$$

It is a continuous group homomorphism from $(\mathbb{Q}_5, +)$ into the unit circle $S = \{z \in \mathbb{C} : |z| = 1\}$ satisfying

$$\chi_5(x+y) = \chi_5(x)\chi_5(y), \quad x, y \in \mathbb{Q}_5.$$

Moreover, the group of additive characters of $\mathbb{Q}_5$ is isomorphic to $(\mathbb{Q}_5, +)$ through the correspondence

$$\xi \longmapsto \chi_5(\xi x).$$

For $f \in L^1(\mathbb{Q}_5)$, the Fourier transform is defined by

$$(\mathcal{F}f)(\xi) = \widehat{f}(\xi) = \int_{\mathbb{Q}_5} \chi_5(\xi x) f(x) dx, \quad \xi \in \mathbb{Q}_5.$$

The inverse Fourier transform is given by

$$(\mathcal{F}^{-1}f)(x) = \int_{\mathbb{Q}_5} \chi_5(-x\xi) f(\xi) d\xi, \quad x \in \mathbb{Q}_5.$$

The Fourier transform defines a linear isomorphism of $\mathcal{D}(\mathbb{Q}_5)$ onto itself and satisfies

$$(\mathcal{F}(\mathcal{F}f))(\xi) = f(-\xi),$$

for every $f \in \mathcal{D}(\mathbb{Q}_5)$.

For a more detailed introduction to p -adic analysis, the reader may refer to [7,9,10].

3. Genomic Embedding Theory

In this section, we develop the mathematical representation of Hantavirus genomic sequences within the non-Archimedean space $\mathbb{Z}_5$. The proposed approach associates RNA sequences with 5-adic integers, thereby inducing an ultrametric geometry that naturally captures hierarchical evolutionary relationships. We further establish the main theoretical properties of this embedding and discuss its relevance for alignment-free phylogenetic analysis and classification.

Let

$$\mathcal{A} := \{A, C, G, U\}$$

denote the RNA nucleotide alphabet. We introduce the encoding map

$$\phi : \mathcal{A} \rightarrow \{1, 2, 3, 4\},$$

defined by

$$\phi(A) = 1, \quad \phi(C) = 2, \quad \phi(G) = 3, \quad \phi(U) = 4.$$

The digit 0 is deliberately omitted from the nucleotide encoding and retained for auxiliary purposes such as representing structural gaps, sequence padding, or delimitation markers. This choice ensures a one-to-one correspondence between the four nucleotides and the four non-zero digits of the base-5 system.

Given a finite RNA sequence

$$S = s_0 s_1 s_2 \cdots s_{n-1},$$

with $s_i \in \mathcal{A}$, we define its associated 5-adic representation by

$$\Phi(S) = \sum_{i=0}^{n-1} \phi(s_i) 5^i.$$

For infinite symbolic sequences, this construction extends naturally to

$$\Phi(S) = \sum_{i=0}^{\infty} \phi(s_i) 5^i,$$

which converges in $\mathbb{Z}_5$ as a consequence of the completeness of the 5-adic metric.

This construction defines a symbolic-to-arithmetic embedding that transforms genomic information into a non-Archimedean geometric object.

Let $S, T \in \mathcal{A}^{\mathbb{N}}$ be two RNA sequences and define

$$x = \Phi(S), \quad y = \Phi(T).$$

Suppose the sequences coincide in their first k nucleotides:

$$s_i = t_i, \quad 0 \leq i \leq k-1,$$

while

$$s_k \neq t_k.$$

Then

$$x - y = 5^k u,$$

where $u \in \mathbb{Z}_5$ is not divisible by 5. Consequently,

$$|x - y|_5 = 5^{-k}.$$

Therefore, the distance between two sequences decreases exponentially with the length of their common prefix.

This property creates a direct link between evolutionary conservation, symbolic similarity, and ultrametric proximity. In this framework, sequences that share longer conserved regions are mapped to points that are geometrically closer within $\mathbb{Z}_5$.

We now establish a fundamental structural property of the proposed encoding.

Lemma 1. *The mapping $\Phi : \mathcal{A}^{\mathbb{N}} \rightarrow \mathbb{Z}_5$ defined by*

$$\Phi(S) = \sum_{i=0}^{\infty} \phi(s_i) 5^i,$$

is injective.

Proof. Assume

$$\Phi(S) = \Phi(T),$$

for two sequences

$$S = s_0 s_1 s_2 \cdots, \quad T = t_0 t_1 t_2 \cdots.$$

Then

$$\sum_{i=0}^{\infty} (\phi(s_i) - \phi(t_i)) 5^i = 0. \quad (1)$$

Suppose $S \neq T$. Let k be the smallest index such that $s_k \neq t_k$.

Then

$$\phi(s_k) - \phi(t_k) \neq 0,$$

and therefore

$$\sum_{i=0}^{\infty} (\phi(s_i) - \phi(t_i))5^i = 5^k u,$$

where u is not divisible by 5. Hence,

$$\left| \sum_{i=0}^{\infty} (\phi(s_i) - \phi(t_i))5^i \right|_5 = |5^k u|_5 = 5^{-k} \neq 0,$$

which is in contradiction with (1).

Therefore, $S = T$ and the mapping is injective. $\square$

The previous result allows us to interpret genomic sequences as points in a genuine ultrametric space, while the following lemma shows that genomic clusters emerge naturally from the metric structure itself, without the need to resort to external heuristic clustering criteria:

Lemma 2. *Let $S, T, R \in \mathcal{A}^{\mathbb{N}}$. Then the induced metric*

$$d_5(S, T) := |\Phi(S) - \Phi(T)|_5$$

satisfies the ultrametric inequality

$$d_5(S, T) \leq \max\{d_5(S, R), d_5(R, T)\}.$$

Proof. Since $\Phi(S), \Phi(T), \Phi(R) \in \mathbb{Z}_5$, the result follows directly from the ultrametric property of the 5-adic norm

$$|x - y|_5 \leq \max\{|x - z|_5, |z - y|_5\},$$

where

$$x = \Phi(S), \quad y = \Phi(T), \quad z = \Phi(R).$$

Thus,

$$d_5(S, T) \leq \max\{d_5(S, R), d_5(R, T)\}.$$

$\square$

The ultrametric structure established in Lemma 2 reveals that the image of the genomic embedding inherits a strong hierarchical organization induced by the 5-adic norm. In particular, the metric topology on $\mathcal{X} := \Phi(\mathcal{A}^{\mathbb{N}})$ is governed entirely by shared symbolic prefixes, leading to a self-similar geometric structure. This observation naturally suggests that $\mathcal{X}$ should exhibit fractal-like properties within $\mathbb{Z}_5$, which we formalize in the following theorem, where we show that the embedded genomic space is, in fact, topologically equivalent to a classical Cantor-type set:

Theorem 1. *Let $\mathcal{X}$ be the image of the genomic embedding. Then $\mathcal{X}$ is a compact, totally disconnected, and perfect subset of $\mathbb{Z}_5$. Consequently, $\mathcal{X}$ is topologically homeomorphic to the classical Cantor ternary set.*

Proof. The space $\mathcal{A}^{\mathbb{N}}$ equipped with the product topology (where $\mathcal{A}$ has the discrete topology) is compact by Tychonoff's theorem; see [34]. Since Φ is an injective and continuous mapping from a compact space into the metric space $\mathbb{Z}_5$, it is a topological embedding. Thus, $\mathcal{X}$ is compact. Furthermore, since $\phi(\mathcal{A}) = \{1, 2, 3, 4\}$, $\mathcal{X}$ consists precisely of those 5-adic integers whose canonical expansion contains no zero digits, which is structurally identical to the construction of a Cantor-like set in $\mathbb{Z}_5$. Total disconnectedness and the

absence of isolated points follow directly from the properties of the subshift and the 5-adic metric. $\square$

The Cantor-type structure of the embedded space $\mathcal{X}$ reflects a purely symbolic notion of proximity between genomic sequences, determined exclusively by the position of the first mismatch. This observation motivates the introduction of the following intrinsic symbolic metric on $\mathcal{A}^{\mathbb{N}}$.

$$d_{\text{symb}}(S, T) = 5^{-\min\{i \in \mathbb{N} : s_i \neq t_i\}}.$$

This metric captures the combinatorial structure of sequence divergence and is naturally compatible with the 5-adic valuation induced by the embedding Φ . In fact, this compatibility is strong enough to yield a precise metric equivalence between the symbolic and the arithmetic representations.

Corollary 1. *The genomic embedding $\Phi : (\mathcal{A}^{\mathbb{N}}, d_{\text{symb}}) \rightarrow (\mathbb{Z}_5, d_5)$ is a global isometry. That is, for every pair of genomic sequences $S, T \in \mathcal{A}^{\mathbb{N}}$,*

$$d_5(S, T) = d_{\text{symb}}(S, T).$$

Proof. Let $S, T \in \mathcal{A}^{\mathbb{N}}$, and let $k = \min\{i \in \mathbb{N} : s_i \neq t_i\}$. By construction of the embedding Φ , we have

$$\Phi(S) - \Phi(T) = 5^k u,$$

where $u \in \mathbb{Z}_5$ is not divisible by 5. Therefore,

$$d_5(S, T) = |\Phi(S) - \Phi(T)|_5 = |5^k u|_5 = 5^{-k}.$$

By definition of the symbolic metric,

$$d_{\text{symb}}(S, T) = 5^{-k}.$$

Hence,

$$d_5(S, T) = d_{\text{symb}}(S, T),$$

which establishes that Φ is a global isometry. $\square$

The geometric interpretation of the embedding is illustrated in Figure 1. Rather than serving only as a schematic representation, the figure provides an intuitive guide to the mathematical construction developed in this section. It illustrates how a finite RNA sequence is transformed into a point of the 5-adic space through the encoding map ϕ , how symbolic prefixes determine the hierarchical organization of the embedded sequences, and how the induced ultrametric geometry naturally reflects their relative similarity. In particular, sequences sharing longer common prefixes occupy nearby regions of $\mathbb{Z}_5$, whereas early symbolic divergences produce larger 5-adic distances. Consequently, the diagram complements the formal definitions by providing a visual interpretation of the correspondence between symbolic sequence structure and non-Archimedean proximity.

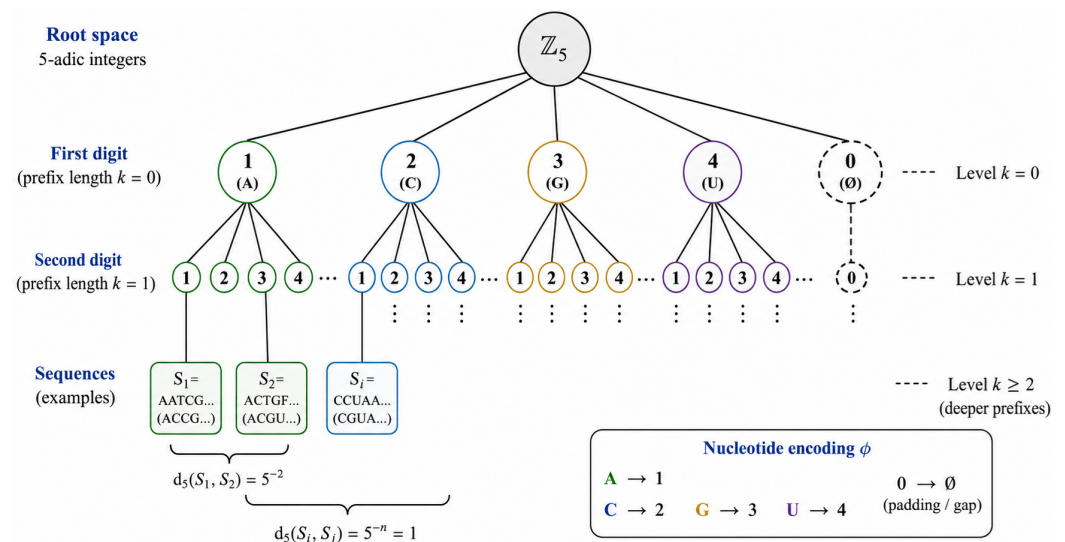


Figure 1. Hierarchical organization of genomic sequences induced by the embedding $\Phi : \mathcal{A}^{\mathbb{N}} \rightarrow \mathbb{Z}_5$. RNA sequences are represented as infinite paths in a rooted 5-adic tree, where each level corresponds to an additional nucleotide position. Shared symbolic prefixes determine the lowest common ancestor of two sequences and therefore their ultrametric distance. The image of the embedding forms a Cantor-type subset $\mathcal{X} \subset \mathbb{Z}_5$, and the distance between sequences is entirely governed by the length of their common prefix, providing a geometric realization of the global isometry established in Corollary 1. The ellipses indicate that each subset contains additional elements following the same pattern. They are included only to simplify the graphical representation and do not denote omitted or incomplete information.

Figure 1 suggests that the hierarchy of genomic clusters emerges directly from the arithmetic structure of the embedding. This observation motivates the study of the intrinsic topological properties of the image of Φ , beginning with the injectivity of the encoding map.

The exclusion of 0 from the nucleotide encoding map $\phi : \mathcal{A} \rightarrow \{1, 2, 3, 4\}$ is a deliberate structural choice rather than a biological assumption imposed on $\mathbb{Z}_5$. In this framework, the digit 0 is reserved exclusively as a neutral placeholder for sequence completion and structural alignment operations (e.g., padding or gap representation in finite-to-infinite extensions).

This convention does not alter the intrinsic 5-adic geometry of nucleotide information, but it plays a distinct role at the level of sequence representation. In particular:

- (i) *Prefix invariance under padding:* Appending trailing zeros to a finite sequence does not affect the leading non-zero digits of its 5-adic expansion. Consequently, the induced ultrametric distance between two sequences remains determined solely by their first point of divergence, independently of length normalization.
- (ii) *Separation of substitution and structural operations:* Within the symbolic-to-arithmetic embedding, nucleotide substitutions correspond to changes in non-zero digits of the 5-adic expansion, whereas gap insertions (modeled via 0) act as neutral elements that do not interfere with the valuation structure of existing prefixes. This separation allows substitution-driven divergence to be captured by the valuation $v_5(\Phi(S) - \Phi(T))$, while structural adjustments are encoded without modifying established prefix geometry.

Therefore, the role of 0 is not to introduce biological semantics into $\mathbb{Z}_5$, but to provide a consistent algebraic mechanism for extending finite genomic representations into a complete 5-adic space without disrupting the ultrametric encoding of symbolic information.

In summary, the algebraic framework constructed herein establishes that Hantavirus genomic evolution admits a representation as a structured trajectory on a non-Archimedean tree induced by the 5-adic valuation. The isometric nature of Φ ensures that local combinato-

rial constraints in the genomic primary structure are faithfully reflected as global ultrametric constraints in $\mathbb{Z}_5$. This provides a rigorous mathematical foundation for alignment-free phylogenetic inference within the proposed ultrametric framework.

4. Evolutionary Dynamics and 5-Adic Diffusion

Having established the algebraic structure of the genomic embedding Φ and the topological properties of its image $\mathcal{X}$ within the ring of 5-adic integers $\mathbb{Z}_5$, we now transition from a static geometric representation to a dynamical framework of genomic evolution. Classical molecular evolution models typically conceptualize phylogenetic divergence through discrete Markovian processes defined on graphs or heuristic tree topologies. While effective in many settings, such approaches often lack a continuous-space formulation capable of integrating local mutational constraints with global hierarchical organization in a unified analytic framework; see, e.g., [35–40] and the references therein.

In this section, we employ the translation-invariant Haar measure and the non-Archimedean analytic machinery developed in Section 2 to model Hantavirus genomic evolution through a continuous-time non-Archimedean diffusion process formulated on the ambient space $\mathbb{Z}_5$ and interpreted through the genomic embedding $\mathcal{X} \subset \mathbb{Z}_5$. By framing genomic divergence as a non-Archimedean diffusion process, we show that the Vladimirov pseudo-differential operator naturally encodes long-range transitions between distinct hierarchical genomic regions without requiring an explicitly specified discrete tree topology.

Although genomic sequences are represented by points of the Cantor-type subset

$$\mathcal{X} = \Phi(\mathcal{A}^{\mathbb{N}}) \subset \mathbb{Z}_5,$$

the evolutionary dynamics considered here are formulated on the ambient ultrametric space $\mathbb{Z}_5$. The subset $\mathcal{X}$ provides the intrinsic genomic geometry through the isometric embedding established in Section 3, while $\mathbb{Z}_5$ supplies the analytic structure required for the definition of Haar integration, Fourier analysis, and Vladimirov pseudo-differential operators. Consequently, the diffusion model introduced below should be viewed as an effective evolutionary dynamics on the ambient non-Archimedean space whose geometric interpretation is induced by the genomic embedding Φ .

Let $u(x, t)$ denote the probability density of observing a given Hantavirus genomic configuration at a point $x \in \mathbb{Z}_5$ at evolutionary time $t \geq 0$, with respect to the normalized Haar measure dx . We describe the temporal evolution of this distribution via the 5-adic pseudo-differential evolution equation

$$\frac{\partial u(x, t)}{\partial t} + b D^\alpha u(x, t) = 0, \quad x \in \mathbb{Z}_5, \quad t > 0, \quad (2)$$

subject to the initial condition $u(x, 0) = u_0(x)$, where $u_0 \in \mathcal{D}(\mathbb{Q}_5)$ is a Bruhat–Schwartz test function satisfying $\int_{\mathbb{Z}_5} u_0(x) dx = 1$. Here, $b > 0$ is an intrinsic mutation rate parameter, and D^α denotes the Vladimirov pseudo-differential operator of order $\alpha > 0$.

The operator D^α is defined on the space of test functions $\mathcal{D}(\mathbb{Q}_5)$ via the integral representation

$$D^\alpha u(x, t) = \frac{5^\alpha - 1}{1 - 5^{-\alpha-1}} \int_{\mathbb{Q}_5} \frac{u(x, t) - u(y, t)}{|x - y|_5^{\alpha+1}} dy,$$

where dy denotes the Haar measure on $\mathbb{Q}_5$. Since the genomic domain is restricted to the compact space $\mathbb{Z}_5$, and $\mathcal{X} \subset \mathbb{Z}_5$ contains no zero digits by construction, the integral naturally localizes to nested ultrametric ball structures associated with sequence divergence.

The Vladimirov operator admits a natural interpretation in this context that contrasts with classical Euclidean diffusion models. In standard settings, diffusion is governed by the Laplacian Δ , which encodes local interactions through infinitesimal spatial displacements. In contrast, D^α is intrinsically non-local: its value at a point x depends on global differences $u(x, t) - u(y, t)$ across the entire space $\mathbb{Z}_5$, weighted by the kernel $|x - y|_5^{-(\alpha+1)}$; see, e.g., [10,12,33] and the references therein.

Recalling the isometric property established in Corollary 1, the distance satisfies $|x - y|_5 = 5^{-k}$, where k is the length of the longest common prefix between the corresponding genomic sequences. Consequently, the operator kernel assigns transition weights that decay according to the depth of the shared ancestral structure in the ultrametric hierarchy.

The parameter $\alpha > 0$ controls the effective separation between hierarchical levels:

- (i) *Small α ($\alpha \rightarrow 0$):* the kernel becomes less sensitive to ultrametric separation, allowing transitions between distant genomic regions with relatively higher probability, consistent with regimes of increased recombination or rapid diversification.
- (ii) *Large α ($\alpha \gg 1$):* the kernel strongly suppresses long-range transitions, effectively confining the dynamics to tight ultrametric neighborhoods corresponding to highly conserved sub-clades.

Thus, the evolution of $u(x, t)$ is naturally interpreted as a diffusion process on the hierarchical structure induced by $\mathbb{Z}_5$, rather than on a classical Euclidean space.

To analyze these dynamics, we derive the fundamental solution $\Gamma_\alpha(x, t)$, representing the transition density from an initial genomic configuration to a state x at time t . Using the 5-adic Fourier transform $\mathcal{F}$ introduced in Section 2, the Vladimirov operator diagonalizes as

$$\mathcal{F}(D^\alpha f)(\xi) = |\xi|_5^\alpha \hat{f}(\xi), \quad \xi \in \mathbb{Q}_5.$$

Applying the Fourier transform to (2) yields

$$\frac{\partial \hat{u}(\xi, t)}{\partial t} + b|\xi|_5^\alpha \hat{u}(\xi, t) = 0,$$

whose solution is

$$\hat{u}(\xi, t) = \hat{u}_0(\xi) \exp(-bt|\xi|_5^\alpha).$$

By the convolution theorem on locally compact groups, the solution in the genomic domain is

$$u(x, t) = (u_0 * \Gamma_\alpha(\cdot, t))(x),$$

where

$$\Gamma_\alpha(x, t) = \int_{\mathbb{Q}_5} \chi_5(-x\xi) \exp(-bt|\xi|_5^\alpha) d\xi.$$

The explicit structure of Γ_α is well established in non-Archimedean analysis. It exhibits regime-dependent behavior governed by the interplay between t and $|x|_5$:

$$\Gamma_\alpha(x, t) = \frac{5^\alpha - 1}{5^{\alpha+1} - 1} \frac{bt}{((bt)^{2/\alpha} + |x|_5^2)^{(\alpha+1)/2}} + \Omega(|x|_5) \Theta_\alpha(t),$$

where $\Omega(|x|_5)$ is the indicator of $\mathbb{Z}_5$, and $\Theta_\alpha(t)$ ensures normalization:

$$\int_{\mathbb{Z}_5} \Gamma_\alpha(x, t) dx = 1, \quad t \geq 0.$$

From a modeling perspective, $\Gamma_\alpha(x - y, t)$ defines an effective transition kernel for alignment-free genomic comparison. It describes the probability of transition between genomic configurations over evolutionary time t while respecting the ultrametric hierarchy induced by the shared prefix structure; see, e.g., [6,8,10,41,42].

As time increases, the probability mass propagates through larger ultrametric balls, reflecting progressive exploration of higher levels of the hierarchical space. This behavior is consistent with the underlying ultrametric organization induced by the embedding Φ and provides a continuous framework that may support alignment-free phylogenetic inference in the ultrametric setting.

5. Alignment-Free Phylogenetic Inference Framework

The non-Archimedean diffusion framework developed in Section 4 operates on the idealized space of infinite symbolic sequences $\mathcal{A}^\mathbb{N}$, mapping them homeomorphically onto the genomic Cantor set $\mathcal{X} \subset \mathbb{Z}_5$. However, empirical genomic data—such as the Small (S), Medium (M), and Large (L) RNA segments of Hantaviruses—consist of finite symbolic sequences of varying lengths. To render the 5-adic diffusion theory operational for real-world biological datasets, we must establish a rigorous bridge between the continuous infinite-dimensional topology of $\mathbb{Z}_5$ and finite-length symbolic representations used in computational genomics.

In this section, we define the algebraic projection of the genomic embedding onto finite quotient rings of $\mathbb{Z}_5$. We then construct an alignment-free distance metric derived directly from the 5-adic valuation and formulate the algorithmic workflow required to reconstruct phylogenetic trees from empirical Hantavirus sequences.

Let $S = s_0 s_1 \cdots s_{n-1}$ be a finite RNA sequence of length n , with $s_i \in \mathcal{A}$. In accordance with Section 3, its 5-adic embedding is given by $\Phi(S) = \sum_{i=0}^{n-1} \phi(s_i) 5^i$. To formalize this within $\mathbb{Z}_5$, we extend S to an infinite sequence $\tilde{S} \in \mathcal{A}^\mathbb{N}$ by setting $s_i = 0$ for all $i \geq n$ so that $\Phi(\tilde{S}) = \Phi(S)$.

In practice, sequence comparison requires fixing a resolution parameter $M \in \mathbb{N}$, corresponding to an effective truncation scale. Algebraically, this induces a projection onto the quotient ring

$$\mathbb{Z}_5 / 5^M \mathbb{Z}_5 \cong \mathbb{Z} / 5^M \mathbb{Z}.$$

We define the canonical projection $\pi_M : \mathbb{Z}_5 \rightarrow \mathbb{Z}_5 / 5^M \mathbb{Z}_5$ by

$$\pi_M \left(\sum_{i=0}^{\infty} x_i 5^i \right) = \sum_{i=0}^{M-1} x_i 5^i \pmod{5^M}.$$

The kernel $\ker(\pi_M) = 5^M \mathbb{Z}_5$ is the ultrametric ball of radius 5^{-M} . Under this projection, sequences sharing a common prefix of length M belong to the same equivalence class. This induces a natural discretization of the diffusion kernel into a finite state space indexed by genomic prefixes.

Unlike classical alignment-free approaches based on k -mer statistics, the 5-adic embedding preserves positional hierarchy via exponential weighting; see, e.g., [5,6,43] and the references therein.

Given a set of sequences $\mathcal{S} = \{S_1, \dots, S_N\}$, we define the pairwise distance matrix D using the 5-adic valuation

$$D_{jl} = d_5(\pi_M(\Phi(S_j)), \pi_M(\Phi(S_l))). \quad (3)$$

If S_j and S_l first differ at position $k < M$, then $D_{jl} = 5^{-k}$. Otherwise, $D_{jl} = 0$ at resolution M , indicating indistinguishability at the chosen scale.

Lemma 3. *The computation of the distance matrix D over N sequences of length M has complexity $\mathcal{O}(N^2M)$ and can be reduced to $\mathcal{O}(NM)$ using trie-based hierarchical indexing.*

Proof. For each pair, the first mismatch is detected in at most M steps, yielding $\mathcal{O}(M)$ per comparison. A trie representation of $\mathcal{A}^N$ encodes sequences as root-to-leaf paths, allowing lowest common ancestor queries to determine distances in linear preprocessing time. $\square$

This approach avoids substitution scoring matrices and relies solely on the intrinsic ultrametric structure of the embedding.

The induced distance matrix D satisfies the strong triangle inequality and is therefore ultrametric in the mathematical sense. This property is consistent with hierarchical evolutionary organization under the proposed framework; see, e.g., [37,44].

To summarize the computational implementation of the proposed framework, Figure 2 presents the complete alignment-free phylogenetic inference workflow. The figure is intended as a visual guide to the methodology, illustrating how each mathematical component introduced in the previous sections contributes to the overall phylogenetic reconstruction process. Beginning with empirical RNA sequences, the workflow successively performs the 5-adic embedding, finite-resolution projection, ultrametric distance computation, and hierarchical clustering, thereby showing how abstract non-Archimedean concepts are translated into a practical alignment-free inference procedure. In this way, the diagram complements the formal mathematical development by providing an intuitive overview of the logical connections among the different stages of the proposed framework.

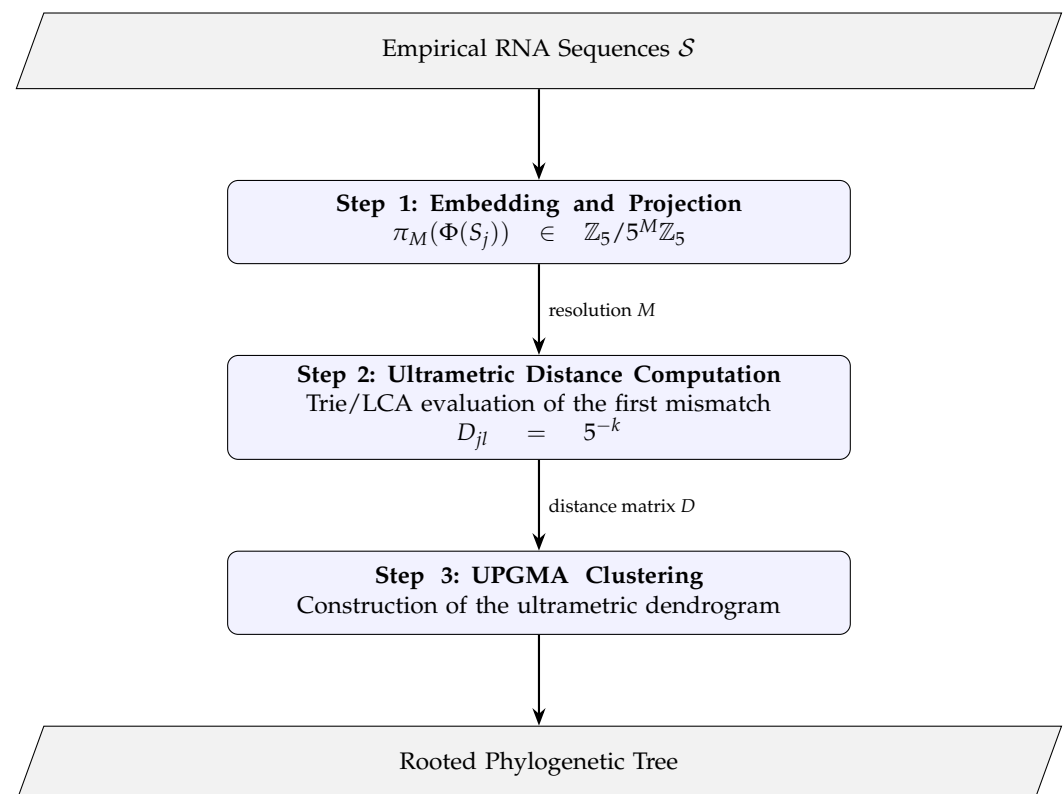


Figure 2. Workflow of the proposed alignment-free phylogenetic inference framework. Finite RNA sequences are embedded into the quotient space $\mathbb{Z}_5/5^M\mathbb{Z}_5$, compared through ultrametric distances derived from the 5-adic valuation, and subsequently clustered using UPGMA to obtain a rooted phylogenetic tree.

We reconstruct phylogenetic relationships via the following pipeline:

Step 1: Embedding and truncation. Compute $\Phi(S_j)$ for each sequence and fix resolution M according to the conserved genomic core length.

Step 2: Distance computation. Construct D using (3). The resulting metric does not rely on substitution scoring heuristics such as PAM or BLOSUM, but directly reflects ultrametric divergence.

Step 3: Hierarchical clustering. Apply UPGMA to (S, d_5) . Due to the ultrametric structure of the space, the resulting dendrogram corresponds to a canonical hierarchical clustering consistent with the nesting structure of balls in $\mathbb{Z}_5$.

The resulting phylogenetic tree provides a mathematically grounded representation of viral diversification. In this framework, clades are determined by arithmetic invariants of the 5-adic valuation, offering a principled alternative to purely heuristic distance-based phylogenetic reconstruction.

6. Illustrative Application to Hantavirus Genomes

To illustrate the practical implementation of the proposed framework, we consider representative genomic configurations arising in the study of Hantaviruses. Hantaviruses are negative-sense, single-stranded RNA viruses belonging to the family *Hantaviridae*, whose genome consists of three segments, commonly denoted by Small (S), Medium (M), and Large (L), encoding the nucleocapsid protein, glycoproteins, and RNA-dependent RNA polymerase, respectively; see, for example, [30–32].

The objective of this section is not to perform a comprehensive phylogenetic analysis of a particular dataset, but rather to demonstrate how the mathematical machinery developed in the previous sections can be applied to biologically relevant genomic sequences. The computational workflow underlying this procedure is summarized in Figure 2, and the present section illustrates how each stage of that framework may be interpreted in a biologically meaningful setting.

Consider a collection of finite RNA sequences obtained from distinct Hantavirus strains

$$\mathcal{S} = \{S_1, S_2, \dots, S_N\},$$

representing genomic segments from distinct Hantavirus strains. Using the encoding map

$$\phi : \mathcal{A} \rightarrow \{1, 2, 3, 4\},$$

each sequence is transformed into a 5-adic integer through the embedding

$$\Phi(S_j) = \sum_{i=0}^{n_j-1} \phi(s_i)5^i,$$

where n_j denotes the length of the corresponding sequence.

To ensure a uniform representation, a finite resolution parameter M is selected. This parameter determines the effective observational scale and induces the projection

$$\pi_M : \mathbb{Z}_5 \longrightarrow \mathbb{Z}_5 / 5^M \mathbb{Z}_5.$$

The projected representations

$$\pi_M(\Phi(S_j))$$

retain the first M digits of the 5-adic expansion and therefore preserve the hierarchical prefix information encoded by the ultrametric structure.

From a practical perspective, the finite resolution parameter M determines the genomic scale at which symbolic similarities are observed. Small values of M emphasize broad ancestral relationships by considering only short conserved prefixes, whereas larger values progressively incorporate finer nucleotide-level differences. Thus, M acts as a multiscale observational parameter rather than an empirical fitting constant. In practical applications, its selection should be guided by the desired level of phylogenetic resolution and by the sequence lengths available in the dataset. Increasing M refines the hierarchical partition while preserving the coarser ultrametric relationships established at lower resolutions.

At this resolution, genomic similarity is measured through the 5-adic distance introduced in Section 5. For every pair of sequences, the distance is defined by

$$D_{jl} = d_5(\pi_M(\Phi(S_j)), \pi_M(\Phi(S_l))).$$

If the first mismatch between S_j and S_l occurs at position $k < M$, then

$$D_{jl} = 5^{-k}.$$

Consequently, sequences sharing longer common prefixes are assigned smaller ultrametric distances and therefore occupy nearby regions of the genomic space.

The resulting distance matrix possesses an intrinsic hierarchical organization inherited from the strong triangle inequality. In particular,

$$d_5(A, B) \leq \max\{d_5(A, C), d_5(C, B)\}.$$

for any three genomic configurations A , B and C . This property implies that clusters are naturally determined by shared ancestral prefixes rather than by optimization procedures imposed externally to the data.

Applying the UPGMA algorithm to the ultrametric distance matrix yields a rooted dendrogram whose branching structure reflects the nested family of balls in the underlying space ($\mathbb{Z}_5$). Unlike conventional alignment-based approaches, the proposed methodology does not require substitution scoring matrices, gap penalties, or sequence alignment procedures. Instead, hierarchical relationships emerge directly from arithmetic properties of the 5-adic valuation.

From a biological perspective, the framework provides a mathematically coherent mechanism for organizing viral genomes according to progressively finer levels of symbolic similarity. Closely related strains correspond to points contained in small ultrametric neighborhoods, whereas more divergent lineages occupy distinct branches of the associated hierarchical tree.

Although the present study focuses on the theoretical foundations of the method, the proposed construction establishes a direct pathway toward large-scale alignment-free phylogenetic analyses of Hantavirus genomes and other segmented RNA viruses. Future work will address comprehensive computational implementations, large-scale genomic datasets, statistical validation, and quantitative comparisons with established alignment-based and alignment-free phylogenetic reconstruction techniques.

6.1. Illustrative Example Using Published Hantavirus Genome Segments

To further illustrate the applicability of the proposed framework to real biological data, we consider representative Hantavirus genomic segments publicly available in the NCBI GenBank database [45]. Rather than constructing a comprehensive phylogenetic study, the purpose of this example is to demonstrate how the mathematical pipeline developed in Section 5 can be applied directly to authentic viral genomic sequences.

As a representative example, we consider the complete M genomic segment of the prototype Hantaan virus strain 76–118 (GenBank accession Y00386.1), retrieved from GenBank [45] and originally reported by Yoo and Kang [46]. This RNA segment contains 3616 nucleotides and encodes the envelope glycoprotein precursor that is subsequently processed into the G1 and G2 glycoproteins. Similar genomic sequences are publicly available for the S and L segments of Hantaan virus and numerous additional members of the genus *Orthohantavirus*, providing a suitable source of experimentally validated sequence data for the proposed alignment-free framework.

Given an RNA sequence

$$S = s_0 s_1 \cdots s_{n-1},$$

each nucleotide is first encoded through the mapping

$$\phi : \mathcal{A} \rightarrow \{1, 2, 3, 4\},$$

followed by the 5-adic embedding

$$\Phi(S) = \sum_{i=0}^{n-1} \phi(s_i) 5^i.$$

The resulting integer is projected onto the finite quotient ring

$$\pi_M(\Phi(S)) \in \mathbb{Z}_5 / 5^M \mathbb{Z}_5,$$

where the truncation parameter M determines the observational resolution. Repeating this procedure for every genomic segment yields a collection of embedded points in the ultrametric space.

Pairwise genomic distances are then computed according to

$$D_{ij} = d_5(\pi_M(\Phi(S_i)), \pi_M(\Phi(S_j))),$$

which depends exclusively on the first nucleotide position at which two sequences differ. Since the resulting distance is intrinsically ultrametric, the corresponding distance matrix can be supplied directly to a hierarchical clustering algorithm such as UPGMA, producing a rooted dendrogram without requiring sequence alignment, substitution models, or gap penalties.

The workflow followed in this illustrative example is summarized schematically in Figure 3. Although no claims are made regarding biological superiority over existing phylogenetic reconstruction methods, this example demonstrates that the proposed non-Archimedean framework operates directly on experimentally determined Hantavirus genome segments available from public genomic repositories such as GenBank [45].

The availability of complete S, M, and L genomic segments for numerous Hantavirus species therefore provides a natural setting for future large-scale computational studies, where the proposed ultrametric representation can be quantitatively compared with both classical multiple-sequence-alignment approaches and modern alignment-free methodologies.

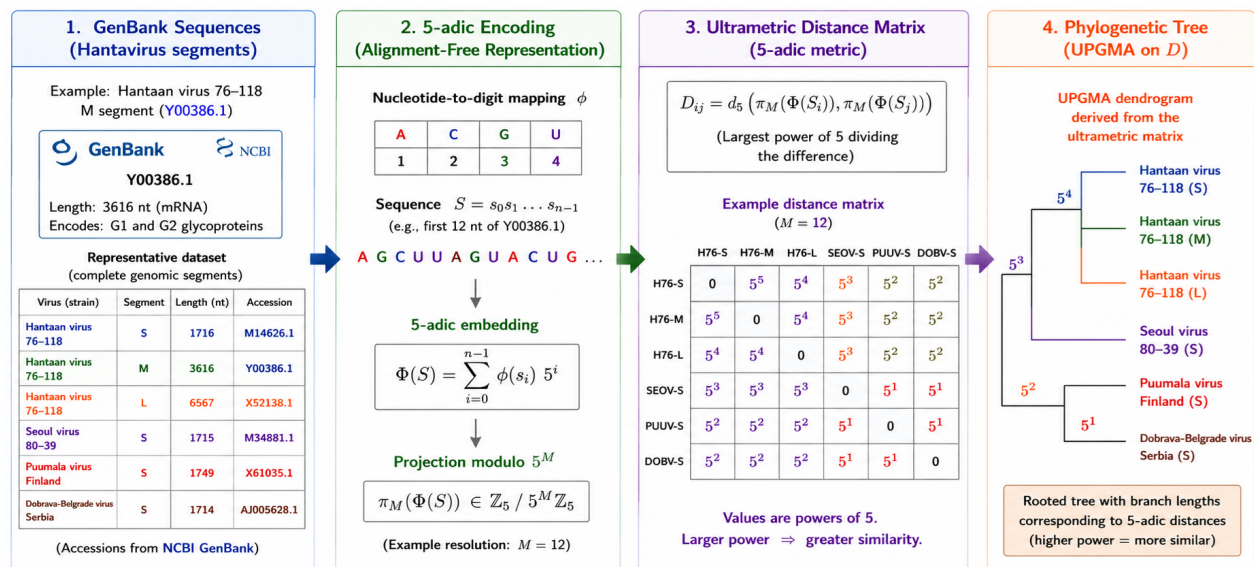


Figure 3. Workflow illustrating the application of the proposed 5-adic ultrametric framework to representative Hantavirus genome segments retrieved from GenBank [45]. The pipeline comprises sequence retrieval, 5-adic encoding, ultrametric distance computation, and hierarchical tree reconstruction. The ellipses are intentional and indicate that the nucleotide sequence continues beyond the portion shown.

6.2. Comparative Assessment with Classical and Alignment-Free Phylogenetic Methods

The mathematical framework introduced in the previous sections differs substantially from both conventional alignment-based phylogenetic methodologies and existing alignment-free approaches. Although all these methods aim to infer evolutionary relationships among genomic sequences, they rely on fundamentally different mathematical representations of sequence similarity.

Classical phylogenetic reconstruction typically begins with the computation of a multiple sequence alignment (MSA), followed by the estimation of evolutionary distances under a selected substitution model. The resulting distance matrix is then used to reconstruct a phylogenetic tree by means of clustering algorithms such as UPGMA or Neighbor-Joining, or alternatively through maximum likelihood or Bayesian inference techniques [2,37,40,44]. The accuracy of these procedures depends critically on the quality of the alignment, the selected substitution model, and the treatment of insertion and deletion events.

Alignment-free methods avoid explicit sequence alignment by representing genomes through numerical descriptors such as word-frequency vectors, k -mer distributions, information-theoretic measures, or statistical moments [5,6,43]. These approaches generally provide substantial computational advantages for large genomic datasets and have become increasingly popular in comparative genomics. However, most alignment-free techniques rely on statistical representations that do not explicitly preserve the hierarchical positional structure of the original nucleotide sequence.

The present framework follows a different mathematical philosophy. Rather than comparing nucleotide frequencies or optimizing sequence alignments, each RNA sequence is embedded into the ultrametric space $\mathbb{Z}_5$ through the injective mapping developed in Section 3. Sequence similarity is therefore determined directly by the 5-adic valuation, which measures the position of the first symbolic divergence between two genomic sequences.

Although the present approach also belongs to the broader area of p -adic mathematical biology, its objective differs substantially from previous p -adic genomic models. Earlier studies have primarily investigated p -adic representations of the genetic code, numerical encodings of DNA sequences, and hierarchical descriptions of genomic organization. In contrast, the present work develops a complete mathematical framework for alignment-free

phylogenetic inference, integrating an injective genomic embedding, an intrinsic ultrametric distance, finite-resolution projections suitable for empirical sequences, and a nonlocal evolutionary model based on the Vladimirov operator. Consequently, the contribution of the present paper is not a new p -adic encoding itself, but the development of a unified framework connecting p -adic analysis with phylogenetic reconstruction.

Consequently, the proposed distance possesses several distinctive mathematical properties. First, it is intrinsically ultrametric and therefore satisfies the strong triangle inequality without requiring any subsequent approximation or tree-fitting procedure. Second, the distance admits an explicit arithmetic interpretation through the 5-adic valuation. Third, the embedding preserves the complete positional hierarchy of nucleotide prefixes, allowing hierarchical relationships to emerge naturally from the underlying non-Archimedean geometry.

From a computational perspective, the proposed methodology avoids multiple sequence alignment, substitution scoring matrices, affine gap penalties, and iterative optimization procedures. Once the genomic embedding has been computed, pairwise distances are obtained directly from the arithmetic representation, making the overall computational workflow particularly simple while preserving an exact hierarchical interpretation.

It should be emphasized, however, that the proposed framework is not intended to replace established phylogenetic methodologies. Instead, it provides an alternative mathematical representation of genomic similarity that is especially well suited for studying hierarchical evolutionary organization within an ultrametric setting. Consequently, the proposed approach should be regarded as complementary to both classical alignment-based phylogenetics and modern alignment-free methodologies rather than as a direct substitute.

To facilitate the comparison between the proposed non-Archimedean framework and representative phylogenetic methodologies currently employed in computational genomics, Figure 4 summarizes their principal mathematical and computational characteristics. Rather than providing a performance ranking, the objective of this comparison is to emphasize the conceptual differences between the underlying sequence representations, the corresponding distance constructions, and the computational requirements of each methodology. The comparison is based on the general methodological principles described in the phylogenetics and alignment-free literature [2,5,6,37,43,44].



Criterion	Classical MSA	Alignment-Free (k-mer)	Proposed 5-adic Method
Multiple sequence alignment	✓	✗	✗
Substitution model	✓	✗	✗
Gap penalties	✓	✗	✗
Positional information preserved	High	Partial	Exact (prefix-based)
Ultrametric by construction	✗	✗	✓
Mathematical space	Euclidean/statistical	✗	$\mathbb{Z}_5$
Mathematical space	Euclidean/statistical	Feature space	$\mathbb{Z}_5$
Hierarchy intrinsic	✗	✗	✓
Computational scalability	Moderate	High	High

Figure 4. Conceptual comparison of representative phylogenetic reconstruction strategies and the proposed 5-adic ultrametric framework.

The comparison presented in Figure 4 shows that the proposed framework differs from existing methodologies primarily through its underlying mathematical representation. Classical phylogenetic approaches reconstruct evolutionary relationships after estimating

distances from multiple sequence alignments, whereas alignment-free methods generally rely on statistical descriptors derived from sequence composition. By contrast, the proposed approach embeds genomic sequences directly into the ultrametric space $\mathbb{Z}_5$, where distances are determined by the arithmetic properties of the 5-adic valuation. As a consequence, hierarchical organization is encoded intrinsically in the representation itself rather than being introduced by the reconstruction algorithm. At the same time, the proposed methodology preserves the computational advantages of alignment-free approaches while providing an explicit mathematical interpretation of sequence similarity within a non-Archimedean framework [2,5,6,37,44].

From a computational perspective, the proposed framework also benefits from its alignment-free construction. The 5-adic embedding of each genomic sequence requires a single traversal of the nucleotide string, while the computation of the complete pairwise distance matrix involves $\mathcal{O}(N^2)$ comparisons, each requiring at most M symbolic evaluations, leading to an overall computational complexity of $\mathcal{O}(N^2M)$. Unlike multiple sequence alignment procedures, each pairwise distance evaluation is independent, making the proposed methodology naturally amenable to parallel implementation. Although the present work focuses on the mathematical foundations of the framework rather than computational optimization, large-scale runtime benchmarks and scalability analyses constitute an important direction for future research.

6.3. Robustness, Limitations, and Biological Interpretation

The mathematical framework developed in the previous sections provides an exact ultrametric representation of genomic sequences through the 5-adic embedding Φ . Although the construction is primarily theoretical, several observations can be made regarding its robustness, biological interpretation, and current limitations.

A fundamental property of the proposed representation is that the induced distance depends exclusively on the position of the first nucleotide at which two sequences differ. Consequently, the embedding is highly stable with respect to mutations occurring after long conserved prefixes. Small variations located near the end of a sequence produce only minor changes in the associated 5-adic representation, whereas substitutions occurring close to the beginning of the sequence induce substantially larger ultrametric distances. This behavior follows directly from the valuation structure of $\mathbb{Z}_5$ and reflects the hierarchical organization encoded by the embedding rather than any externally imposed weighting scheme.

From a biological perspective, the proposed metric naturally emphasizes conserved ancestral structures. Sequences sharing long common prefixes are interpreted as belonging to the same ultrametric neighborhood, while early symbolic divergence corresponds to separation at higher hierarchical levels. This interpretation is consistent with the general principle that increasing sequence similarity is associated with increasing evolutionary relatedness, although the present framework measures similarity through symbolic prefix preservation rather than through alignment scores or probabilistic substitution models.

An important advantage of the proposed methodology is that its mathematical formulation is independent of empirical scoring parameters. Classical alignment-based phylogenetic methods typically require the specification of substitution matrices, affine gap penalties, or explicit evolutionary models, whereas many alignment-free approaches rely on statistical descriptors such as k -mer frequencies. In contrast, the present framework derives its distance directly from the arithmetic properties of the 5-adic valuation. As a consequence, the resulting ultrametric geometry is uniquely determined once the nucleotide encoding is fixed, eliminating the need for heuristic parameter calibration.

The framework also possesses favorable structural robustness. Since the image of the embedding is a compact Cantor-type subset of $\mathbb{Z}_5$, neighboring genomic configura-

tions remain organized within nested ultrametric balls under finite-resolution projections. Therefore, increasing the truncation parameter M refines the hierarchical partition without altering the relationships already established at coarser resolutions. This multiscale consistency is an intrinsic consequence of the ultrametric topology.

Nevertheless, several limitations should also be acknowledged. The current formulation emphasizes prefix preservation and therefore does not explicitly model insertions, deletions, genomic rearrangements, or recombination events. Although the reserved digit 0 provides a mathematically consistent mechanism for representing structural gaps, the biological consequences of complex structural variation remain outside the scope of the present theoretical framework. Likewise, the continuous diffusion model introduced in Section 4 should be interpreted as an idealized description of evolutionary dynamics in the ambient ultrametric space rather than as a direct stochastic model of molecular evolution.

The diffusion exponent α governs the strength of nonlocal evolutionary transitions in the Vladimirov diffusion model. Smaller values of α promote relatively long-range transitions across the ultrametric hierarchy, whereas larger values increasingly favor local transitions between closely related genomic configurations. In the present theoretical framework, α should therefore be interpreted as a mathematical parameter controlling the scale of hierarchical evolutionary propagation rather than as a directly measurable biological quantity. The calibration of α from empirical mutation rates or phylogenetic data remains an important topic for future investigation.

Although the diffusion model is formulated in purely mathematical terms, its qualitative behavior admits a natural biological interpretation. The evolution of the probability density may be viewed as describing the redistribution of genomic configurations within the ultrametric space as mutations accumulate over evolutionary time. In this context, the Vladimirov operator does not model individual nucleotide substitutions directly; rather, it represents transitions between hierarchically organized genomic states. Consequently, the diffusion parameter α controls the effective scale of these transitions, while additional biological mechanisms such as mutation rates, natural selection, recombination, or host adaptation are not explicitly incorporated into the present formulation. Instead, these processes should be regarded as factors that could influence the effective dynamics in future biologically calibrated versions of the model.

Furthermore, the present study focuses on establishing the mathematical foundations of the proposed approach rather than providing a comprehensive computational benchmark. Consequently, no claims are made regarding empirical superiority over existing phylogenetic methodologies. Instead, the proposed framework should be viewed as a complementary mathematical representation that provides an exact ultrametric geometry for genomic sequences and offers an alternative perspective for alignment-free evolutionary analysis.

Future work will concentrate on large-scale computational implementations involving complete Hantavirus genomes and other segmented RNA viruses. Such studies will enable quantitative comparisons with established alignment-based and alignment-free phylogenetic methods, statistical assessments of reconstruction accuracy, and investigations of the relationship between the proposed 5-adic geometry and experimentally inferred evolutionary histories.

7. Conclusions

In this work, we introduced a novel non-Archimedean framework for the analysis of genomic sequences based on the arithmetic and geometric structure of the ring of 5-adic integers. The proposed approach establishes a rigorous connection between symbolic genomic information and ultrametric geometry through an injective embedding of RNA

sequences into $\mathbb{Z}_5$. This construction transforms sequence data into a hierarchical geometric object whose topology is naturally determined by shared symbolic prefixes and the associated 5-adic valuation.

From a mathematical perspective, the embedding was shown to induce an ultrametric distance that is globally isometric to a symbolic prefix metric, thereby providing an exact correspondence between combinatorial sequence similarity and non-Archimedean proximity. Furthermore, the image of the embedding was characterized as a compact Cantor-type subset of $\mathbb{Z}_5$, revealing an intrinsic fractal and hierarchical organization of genomic information within the ultrametric setting.

Building upon this geometric representation, we formulated a continuous-time evolutionary model governed by the Vladimirov pseudo-differential operator. The resulting non-Archimedean diffusion equation provides a natural mechanism for describing evolutionary transitions across hierarchical genomic scales. In contrast to classical Euclidean diffusion models, the proposed dynamics are intrinsically nonlocal and reflect the nested structure of ultrametric neighborhoods induced by the genomic embedding. The associated transition kernel supplies a mathematically consistent description of evolutionary propagation within the hierarchical space.

We further developed an alignment-free phylogenetic inference framework based directly on the 5-adic valuation. By projecting genomic embeddings onto finite quotient rings of $\mathbb{Z}_5$, we obtained a computationally tractable methodology for comparing finite RNA sequences without the need for sequence alignment, substitution scoring matrices, or gap-penalty schemes. The resulting distance matrix is ultrametric by construction, allowing phylogenetic relationships to emerge naturally from arithmetic invariants rather than from externally imposed optimization procedures.

Although the present study is primarily theoretical, it demonstrates how tools from p -adic analysis, ultrametric geometry, harmonic analysis, and pseudo-differential equations can be integrated into a unified framework for genomic modeling. The illustrative application to Hantavirus RNA sequences highlights the potential of the approach as a mathematically grounded alternative for investigating hierarchical evolutionary structures in viral genomes.

Several directions for future research naturally arise from the results presented here. These include the development of large-scale computational implementations, the analysis of complete genomic datasets, statistical validation against established phylogenetic methodologies, and the incorporation of more general classes of non-Archimedean pseudo-differential operators associated with negative definite functions and Bessel-type potentials. Another promising direction consists of extending the framework to stochastic evolutionary models, genomic networks, and broader classes of RNA and DNA viruses exhibiting hierarchical evolutionary organization.

The comparative assessment presented in Section 6 highlights that the proposed methodology should be viewed as complementary to existing phylogenetic approaches rather than as a replacement for them. In contrast to alignment-based methods, which rely on multiple sequence alignments and evolutionary substitution models, and statistical alignment-free approaches based on sequence composition, the present framework derives genomic similarity directly from the arithmetic structure of the 5-adic valuation. This provides an exact ultrametric representation in which hierarchical relationships emerge naturally from the underlying non-Archimedean geometry.

At the same time, the present work has several limitations that should be acknowledged. The current formulation is primarily intended as a mathematical framework and therefore does not explicitly address complex biological processes such as insertions, deletions, recombination, or genomic rearrangements. Moreover, although the proposed

construction establishes a rigorous theoretical basis for alignment-free phylogenetic inference, comprehensive computational validation on large genomic datasets remains an important direction for future research. Such studies will enable quantitative comparisons with established phylogenetic methodologies and provide further insight into the biological applicability of the proposed ultrametric representation.

The proposed framework should also be viewed in the context of previous p -adic approaches to genomics. Whereas earlier contributions established the usefulness of non-Archimedean methods for modeling the genetic code and hierarchical biological organization, the present work extends this line of research toward a mathematically rigorous framework for alignment-free phylogenetic inference and hierarchical evolutionary modeling.

More broadly, the results suggest that ultrametric geometry and non-Archimedean analysis provide a fertile mathematical setting for the study of biological information. By linking symbolic sequence structure, hierarchical topology, and evolutionary dynamics within a single analytic framework, the proposed approach opens new avenues for interaction between modern p -adic mathematics and computational genomics.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Durbin, R.; Eddy, S.R.; Krogh, A.; Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*; Cambridge University Press: Cambridge, UK, 1998. [\[CrossRef\]](#)
2. Felsenstein, J. *Inferring Phylogenies*; Sinauer Associates Inc.: Sunderland, MA, USA, 2004.
3. Kimura, M. *The Neutral Theory of Molecular Evolution*; Cambridge University Press: Cambridge, UK, 1983. [\[CrossRef\]](#)
4. Nei, M.; Sudhir, K. *Molecular Evolution and Phylogenetics*; Oxford Academic: New York, NY, USA, 2000. [\[CrossRef\]](#)
5. Vinga, S.; Almeida, J. Alignment-free sequence comparison—A review. *Bioinformatics* **2003**, *19*, 513–523. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Zielezinski, A.; Vinga, S.; Almeida, J.; Karlowski, W.M. Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biol.* **2017**, *18*, 186. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Albeverio, S.; Khrennikov, A.Y.; Shelkovich, V.M. *Theory of p -Adic Distributions: Linear and Nonlinear Models*; London Mathematical Society Lecture Note Series, 370; Cambridge University Press: Cambridge, UK, 2010.
8. Kochubei, A.N. *Pseudo-Differential Equations and Stochastic over Non-Archimedean Fields, Pure and Applied Mathematics*; Marcel Dekker: New York, NY, USA, 2001.
9. Taibleson, M.H. *Fourier Analysis on Local Fields*; Princeton University Press: Princeton, NJ, USA, 1975.
10. Vladimirov, V.S.; Volovich, I.V.; Zelenov, E.I. *p -Adic Analysis and Mathematical Physics*; World Scientific: Singapore, 1994.
11. Dragovich, B.; Khrennikov, A.Y.; Kozyrev, S.V.; Mišić, N.Ž. p -Adic mathematics and theoretical biology. *Biosystems* **2021**, *199*, 104288. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Dragovich, B.; Mišić, N.Ž. p -Adic hierarchical properties of the genetic code. *Biosystems* **2019**, *185*, 104017. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Khrennikov, A.Y.; Kochubei, A.N. p -Adic Analogue of the Porous Medium Equation. *J. Fourier Anal. Appl.* **2018**, *24*, 1401–1424. [\[CrossRef\]](#)
14. Khrennikov, A.; Oleschko, K.; Correa López, M.D.J. Modeling Fluids Dynamics with Master Equations in Ultrametric Spaces Representing the Treelike Structure of Capillary Networks. *Entropy* **2016**, *18*, 249. [\[CrossRef\]](#)
15. Torresblanca-Badillo, A.; Torresblanca-Badillo, V.P. Environmental contaminant dispersion models based on the Vladimirov-Taibleson p -adic pseudo-differential operator. *J. Pseudo-Differ. Oper. Appl.* **2025**, *16*, 59. [\[CrossRef\]](#)
16. Zúñiga-Galindo, W.A. Dynamic mean-field theory for continuous random networks. *J. Phys. A Math. Theor.* **2025**, *58*, 125201. [\[CrossRef\]](#)
17. Zúñiga-Galindo, W.A. Non-Archimedean White Noise, Pseudodifferential Stochastic Equations, and Massive Euclidean Fields. *J. Fourier Anal. Appl.* **2017**, *23*, 288–323.
18. Aguilar-Arteaga, V.; Gutiérrez García, I.; Torresblanca-Badillo, A. Energy landscapes and non-Archimedean pseudo-differential operators as tools for studying the spreading of infectious diseases in a situation of extreme social isolation. *Kragujev. J. Math.* **2024**, *48*, 827–844. [\[CrossRef\]](#)

19. Antoniouk, A.V.; Khrennikov, A.Y.; Kochubei, A.N. Multidimensional nonlinear pseudo-differential evolution equation with p -adic spatial variables. *J. Pseudo-Differ. Oper. Appl.* **2019**, *11*, 311–343. [\[CrossRef\]](#)
20. Athira, N.; Lineesh, M.C. Vanishing Moments of Wavelets on p -adic Fields. *Complex Anal. Oper. Theory* **2025**, *19*, 9. [\[CrossRef\]](#)
21. Serdiuk, M. Estimation of the Fundamental Solution for a New Class of Non-Archimedean Pseudodifferential Equations. *Ukr. Math. J.* **2025**, *76*, 1731–1737. [\[CrossRef\]](#)
22. Torresblanca-Badillo, A. On some generalizations of non-archimedean pseudo-differential operators and their applications. *J. Math. Anal. Appl.* **2022**, *512*, 126168. [\[CrossRef\]](#)
23. Torresblanca-Badillo, A.; Arroyo-Ortiz, E.; Barrios-Garizao, R. Pseudo-differential operators in several p -adic variables and sub-Markovian semigroups. *J. Pseudo-Differ. Oper. Appl.* **2024**, *15*, 51. [\[CrossRef\]](#)
24. Torresblanca-Badillo, A. Non-archimedean pseudo-differential operators on Sobolev spaces related to negative definite functions. *J. Pseudo-Differ. Oper. Appl.* **2021**, *12*, 7. [\[CrossRef\]](#)
25. Gutiérrez García, I.; Torresblanca-Badillo, A. Some classes of non-archimedean pseudo-differential operators related to Bessel potentials. *J. Pseudo-Differ. Oper. Appl.* **2020**, *11*, 1111–1137. [\[CrossRef\]](#)
26. Antoniouk, A.V.; Oleschko, K.; Kochubei, A.N.; Khrennikov, A.Y. A stochastic p -adic model of the capillary flow in porous random medium. *Phys. A Stat. Mech. Appl.* **2018**, *505*, 763–777. [\[CrossRef\]](#)
27. Avetisov, V.A.; Bikulov, A.K. Ultrametricity of the fluctuation dynamic mobility of protein molecules. *Proc. Steklov Inst. Math.* **2009**, *265*, 75–81. [\[CrossRef\]](#)
28. Avetisov, V.A.; Bikulov, A.K.; Osipov, V.A. p -adic description of characteristic relaxation in complex systems. *J. Phys. A* **2003**, *36*, 4239–4246. [\[CrossRef\]](#)
29. Bikulov, A.K. On solution properties of some types of p -adic kinetic equations of the form reaction diffusion. *P-Adic Numbers Ultrametr. Anal. Appl.* **2010**, *2*, 187–206. [\[CrossRef\]](#)
30. Jonsson, C.B.; Figueiredo, L.T.; Vapalahti, O. A Global Perspective on Hantavirus Ecology, Epidemiology, and Disease. *Clin. Microbiol. Rev.* **2010**, *23*, 412–441. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Schmaljohn, C.S.; Nichol, S.T. (Eds.) *Hantaviruses; Current Topics in Microbiology and Immunology (CT MICROBIOLOGY)*; Springer: Berlin/Heidelberg, Germany, 2001; Volume 256.
32. Vaheri, A.; Strandin, T.; Hepojoki, J.; Sironen, T.; Henttonen, H.; Mäkelä, S.; Mustonen, J. Uncovering the mysteries of hantavirus infections. *Nat. Rev. Microbiol.* **2013**, *11*, 539–550. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Dragovich, B.; Dragovich, A.Y. A p -adic model of DNA sequence and genetic code. *P-Adic Numbers Ultrametr. Anal. Appl.* **2009**, *1*, 34–41. [\[CrossRef\]](#)
34. Kelley, J.L. *General Topology*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 1975.
35. Billera, L.J.; Holmes, S.P.; Vogtmann, K. Geometry of the Space of Phylogenetic Trees. *Adv. Appl. Math.* **2001**, *27*, 733–767. [\[CrossRef\]](#)
36. Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **1981**, *17*, 368–376. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Gascuel, O. *Mathematics of Evolution and Phylogeny*; Oxford Academic: Oxford, UK, 2005. [\[CrossRef\]](#)
38. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **1980**, *16*, 111–120. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **1999**, *401*, 877–884. [\[CrossRef\]](#) [\[PubMed\]](#)
40. Yang, Z. *Molecular Evolution: A Statistical Approach*; Oxford Academic: Oxford, UK, 2014. [\[CrossRef\]](#)
41. Bikulov, A.K.; Zubarev, A.P. Classical first passage problems for p -adic stochastic processes. *Phys. Stat. Mech. Its Appl.* **2026**, *690*, 131450. [\[CrossRef\]](#)
42. Murtagh, F. On Ultrametricity, Data Coding, and Computation. *J. Classif.* **2004**, *21*, 167–184. [\[CrossRef\]](#)
43. Song, K.; Ren, J.; Zhai, Z.; Liu, X.; Deng, M.; Sun, F. Alignment-Free Sequence Comparison Based on Next Generation Sequencing Reads: Extended Abstract. In *Research in Computational Molecular Biology. RECOMB; Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7262. [\[CrossRef\]](#)
44. Semple, C.; Mike, S. *Phylogenetics*; Oxford Academic: Oxford, UK, 2003. [\[CrossRef\]](#)
45. Benson, D.A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D.J.; Ostell, J.; Sayers, E.W. GenBank. *Nucleic Acids Res.* **2013**, *41*, D36–D42. [\[CrossRef\]](#) [\[PubMed\]](#)
46. Yoo, D.W.; Kang, C.Y. Nucleotide sequence of the M segment of the genomic RNA of Hantaan virus 76–118. *Nucleic Acids Res.* **1987**, *15*, 6299–6300. [\[CrossRef\]](#) [\[PubMed\]](#) [\[PubMed Central\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.