Mathematical Biosciences and Engineering

MBE, 23(9): 2634–2654.
DOI: 10.3934/mbe.2026095
Received: 05 December 2025
Revised: 21 January 2026
Accepted: 26 January 2026
Published: 03 September 2026

http://www.aimspress.com/journal/MBE

Research article

Mathematical evaluation of the formation and evolution of hierarchical structures of repetitive DNA sequences

Ziyan Hao[1], Xiaochang Xu[2] and Xiyin Wang[1,2,3,*]

[1] College of Mathematics and Science, North China University of Science and Technology, Tangshan, China
[2] School of Life Science, North China University of Science and Technology, Tangshan, China
[3] School of Public Health and Protective Medicine, North China University of Science and Technology, Tangshan, China

* **Correspondence:** Email: wangxiyin@vip.sina.com; Tel: +13603381029.

Abstract: With the advancement of genome sequencing technologies, telomere-to-telomere (T2T) genome assemblies provide valuable opportunities to understand the formation and evolution of centromeric regions, which are enriched in highly repetitive DNA sequences. In this study, we applied a well-established alignment-free approach using k-mer frequency vectors and Euclidean distance to quantify the similarities among hierarchical repetitive units. By introducing point, insertion, and deletion mutations at varying rates, we systematically assessed their effects on the repeat unit similarity at different hierarchical levels. All analyses were based on the averages of 30 independent simulation runs, with standard deviations used to quantify the stochastic variability. Point mutations generally led to monotonic increases in the Euclidean distance, reflecting progressively reduced sequence similarities, whereas insertion and deletion mutations produced more variable patterns, highlighting the complex effects of length-altering mutations on the hierarchical structure evolution. To further evaluate the biological relevance of our simulations, representative centromeric regions from rice, *Arabidopsis*, and human genomes were analyzed, showing hierarchical patterns consistent with the simulated sequences. Overall, this study provides a reproducible and interpretable computational framework for quantitatively assessing how different mutation types influence on the formation and evolution of multi-layered repetitive sequences, thereby explicitly linking simulation results to biologically observed patterns while acknowledging the established nature of the methods.

Keywords: DNA repetitive sequences; mutation; frequency vector; Euclidean distance

1. Introduction

Deoxyribonucleic acid (DNA) is the fundamental carrier of genetic information, composed of four nucleotides (C, G, A, T), and encodes essential information related to gene expression and species evolution [1]. Repetitive sequences are pervasive within genomes and play crucial roles in maintaining genomic stability, thus ensuring accurate transmission of genetic information and regulating gene expression. In particular, DNA repetitive sequences often exhibit multi-layered similarity structures, ranging from large-scale repetitive units to smaller repetitive fragments, thus forming hierarchical features that integrate sequence similarity at both the local and global scales [2]. Therefore, systematically quantifying the similarity between repetitive units at different hierarchical levels is essential to understand the mechanisms that govern multi-layered repetitive structures.

Advances in sequencing technologies have dramatically improved the completeness and quality of genome assemblies [3]. Notably, the release of the complete human genome assembly by the T2T consortium between 2020 and 2022 unlocked highly repetitive regions, such as centromeres, available for study [4–10]. Additionally, complete T2T assemblies of plants [11], such as *Arabidopsis* [12–14] and rice [15,16], provide reliable sequence data to investigate the genetics of the model's species. These datasets enable detailed computational and mathematical analyses of repetitive regions. For example, centromeric regions in humans are composed of exceptionally long and highly similar alpha-satellite repeat arrays, which form hierarchical higher-order repeat (HOR) structures [17]. In the human X chromosome, alpha-satellite monomers organize into tandem Higher Ordered Repeat (HOR) units, with individual monomers similarity ranging from 50% to 90%, while the HOR-level similarity can exceed 95%. Such observations highlight the hierarchical nature of repetitive sequences and emphasize the need for reproducible quantitative approaches to analyze their evolution and variation.

Computational methods to evaluate the similarity in repetitive sequences continue to evolve and can be broadly classified into three categories. The first category consists of alignment-based methods [18] (such as Basic Local Alignment Search Tool (BLAST)), which compare repetitive fragments using local or global sequence alignments. These methods can encounter ambiguity and a high computational cost when applied to extremely long or highly repetitive regions. The second category includes alignment-free approaches based on k-mer frequencies [19,20], which quantify the sequence similarity using statistical properties rather than an explicit alignment, thus enabling an efficient analysis of structurally complex repeats. The third category comprises multi-scale analytical techniques [21,22], including autocorrelation analyses, Fourier transform, and wavelet analyses, which detect periodicity, hierarchical organization, and nested repeat patterns across different scales. While these approaches have significantly contributed to repetitive sequence research, they are typically applied in a static manner and are less frequently used to systematically examine how different mutation types and rates influence similarity across hierarchical levels.

To explore these questions, the present study constructs multi-layered repetitive simulated sequences and applies a well-established alignment-free approach based on k-mer frequency vectors. Using Euclidean distance as a similarity metric, we assess how point, insertion, and deletion mutations at varying rates influence the sequence similarity across hierarchical levels. All analyses are averaged over multiple independent simulation runs to account for the stochastic variability and to ensure statistical robustness. Comparisons at the structural level, based on dot plot analyses of representative centromeric regions from rice, *Arabidopsis*, and human genomes, further demonstrate the biological plausibility of the simulated sequences. Overall, this study offers a reproducible and interpretable

quantitative perspective to assess how different mutation types influence the hierarchical repeat similarity and provide a theoretical basis to discuss the formation and evolution of centromeric structures.

2. Materials and methods

2.1. Data collection

2.1.1. Collection of real sequences

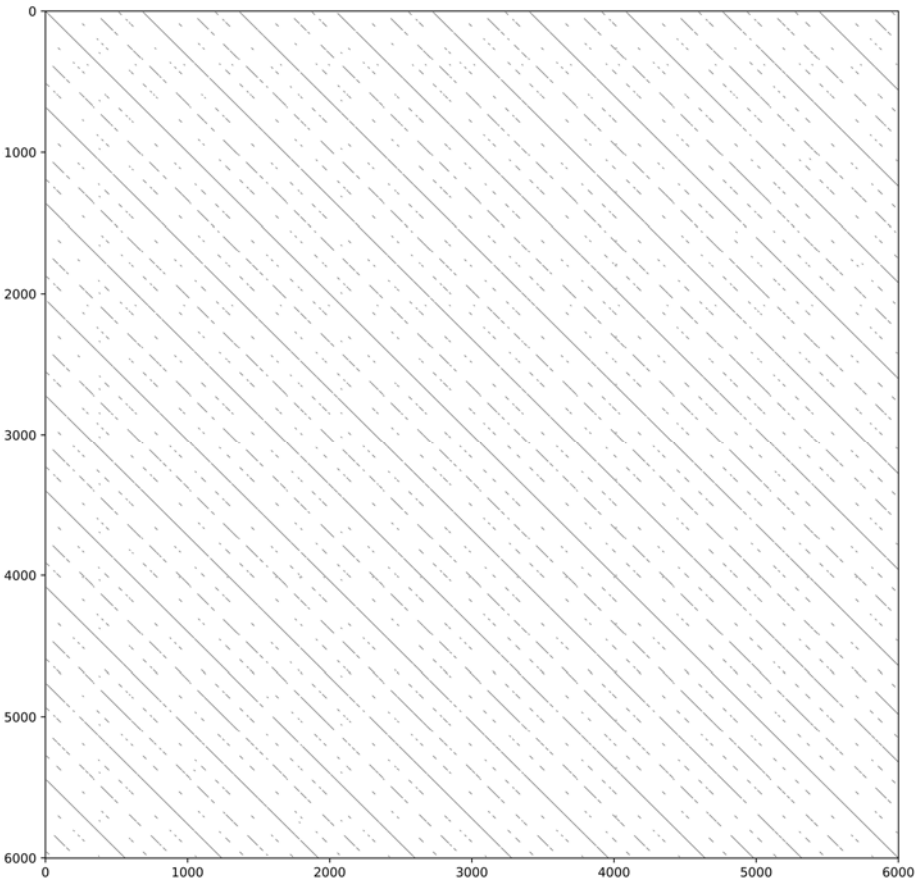

Figure 1. DNA similarity analysis results of the human sequence.

To place our simulation framework in a realistic genomic context, we extracted 6000-bp DNA segments from the T2T genome assemblies of rice, *Arabidopsis thaliana*, and human obtained from the NCBI database (https://www.ncbi.nlm.nih.gov). Dot plot analyses were performed as a qualitative and visualization-based approach to identify multi-level similarity patterns and to provide illustrative references for plausible hierarchical repeat architectures observed in real genomes. Figure 1 shows the dot plot of the human sequence, in which approximately nine repeat units of ~640 bp can be observed. Each of these units contains smaller-scale repetitive substructures: a 640-bp unit consists of two ~320-bp repeat segments, and each 320-bp segment is further composed of two ~160-bp repeat fragments. This nested organization (160 bp→320 bp→640 bp) illustrates a typical hierarchical repeat architecture commonly observed in centromeric regions. These real genomic examples are solely used

to provide the biological motivation and structural context for the design of simulated hierarchical sequences, rather than to serve as the input for quantitative k-mer based distance analyses. The dot plots for rice and *Arabidopsis thaliana* are provided in the Supplementary.

Guided by prior reports that typical repeat repetitive units span 100–400 bp [23], we selected a 100-bp fragment as a representative lower-bound repetitive unit from a human genomic sequence as the seed to construct zero-, first-, and second-order hierarchical repeats in the simulated dataset. This choice represents a compromise between biological relevance and computational tractability, thus enabling the controlled evaluation of cross-level similarity while remaining consistent with observed genomic repeat scales. The seed sequence is as follows:

"ACATCTTGTGGCCTTCGTTGGAAACGGGATTTCTTCATATTCTGCTAGACAGAAGAA TTCTCAGTAACTTCCTTGTGTTGTGTGTATTCAACTCACAGAG".

2.1.2. Generation of simulated sequences

To emulate key features of the hierarchical repeat organization observed in real genomes, we generated a simulated dataset in two stages: (i) construct monomers or repetitive units at different orders by concatenating mutated copies of lower-ordered repetitive units; and (ii) constructed sequences were generated by repeating these repetitive units and sequences from different hierarchical orders were subsequently concatenated to form a final multi-level repetitive sequence. This simulation design enables a controlled and systematic exploration of how local repeat units contribute to higher-level sequence organization under predefined mutation regimes and provides a standardized input for the downstream quantitative similarity analysis.

The rules and definitions for monomers of different orders are as follows:

Zero-order repetitive unit: preset as the selected seed repetitive unit.

First-order repetitive unit: formed by concatenating multiple consecutive zero-order repetitive units that have undergone a mutation (with a mutation rate of r_1), and used as a new repeat unit.

Second-order repetitive unit: formed by concatenating multiple consecutive first-order repetitive units that have undergone mutation (with a mutation rate of r_2), and used as a new repeat unit.

Three mutation types are considered—point, mutation, insertion, and deletion—with preset mutation-rate levels.

The rules and definitions for sequences of different orders are as follows:

Zero-order sequence: obtained by repeating the zero-order repetitive unit multiple times, and is denoted as S_1.

First-order sequence: obtained by repeating the first-order repetitive unit multiple times, and is denoted as S_2.

Second-order sequence: obtained by repeating the second-order repetitive unit multiple times, and is denoted as S_3.

Concatenating sequences of different orders yields the final multi-level repetitive sequence (F), where its mathematical expression is as follows:

$$F = S_1 \oplus S_2 \oplus S_3, \tag{2.1}$$

where $\oplus$ denotes sequence concatenation.

2.1.3. Schematic illustration of the hierarchical construction of repetitive sequences

A DNA fragment (A) is selected as the zero-order repeat unit. In the simulation, mutated copies of the zero-order repeat unit are generated at the mutation rate r_1 and duplicated twice to form a first-order repeat unit (B). Then, the first-order repeat unit is treated as a new repeat unit, and through the same simulation procedure, mutated copies are generated at the mutation rate r_2 and duplicated twice to form a second-order repeat unit (C). The final multi-layered repetitive sequence is constructed by concatenating repeat units from different hierarchical levels, as schematically illustrated in Figure 2.

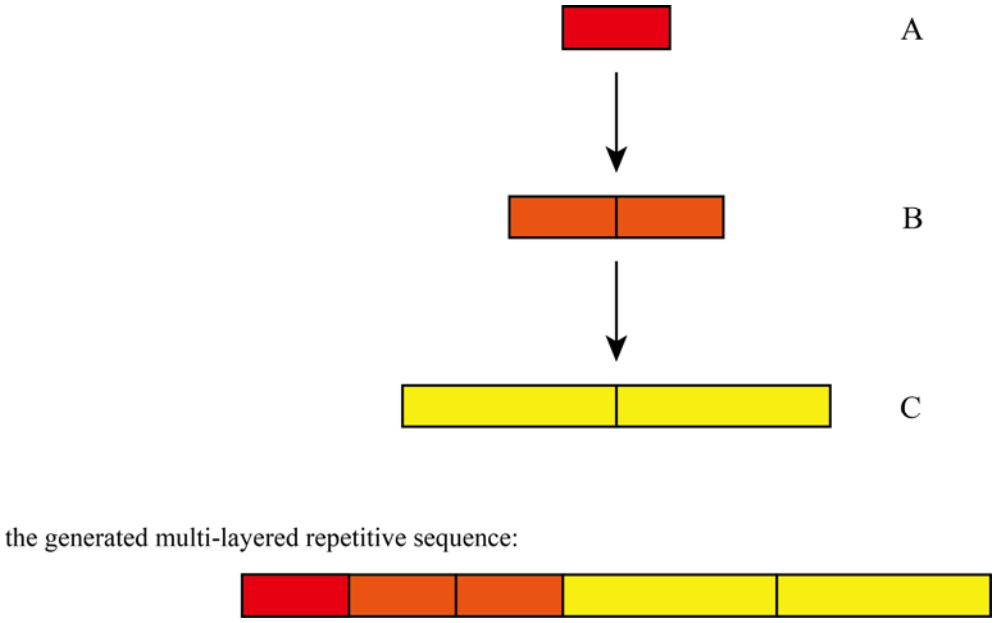

Figure 2. Schematic diagram of multi-layered repetitive sequence.

2.2. Identification of multi-level structures in sequences

Dot plots were generated using the CentriVision platform to visualize similarity patterns in both real and simulated sequences [24]. For real genomic sequences, dot plots were solely used as a qualitative visualization tool to illustrate the presence of multi-level repetitive patterns and to provide biological context for the simulation design. No quantitative segmentation or distance analysis was performed on real sequences. For simulated sequences, dot plot visualization was used to confirm the intended hierarchical organization of the repeat units. A quantitative similarity analysis was subsequently and exclusively performed on simulated sequences using alignment free k-mer frequency vectors.

2.3. Rules for mutation operations

Genomic mutations refer to changes in DNA sequences and are commonly classified into three primary types: insertions, deletions, and point mutations. Insertion refers to the addition of new nucleotide sequences, deletion involves the removal of existing nucleotides, and a point mutation represents a substitution of one nucleotide by another [25,26]. These mutation types are well known to contribute to genomic variation and evolutionary processes, and are widely studied in diverse

biological and medical contexts [27–33]. In this study, mutation processes are modeled using simplified stochastic rules designed to capture the core effects of different mutation types on the sequence composition and similarity, while maintaining computational tractability and experimental control. Although real genomic mutations may involve more complex mechanisms (such as multi-base insertion and deletion or slipped-strand mispairing), the mutation rules defined below are intentionally simplified to enable the systematic and reproducible evaluation of mutation-driven changes in hierarchical repeat structures under controlled conditions.

1) Insertion. Insert the random sequence (T) into (S_0) at the k-th position to obtain the sequence (S_I); the mathematical expression is defined as follows:

$$S_I = S_0[1:k] + T + S_0[k+1:L_0], \tag{2.2}$$

where the length of sequence (S_0) is (L_0). For example, given an original sequence ATTGC, inserting a random nucleotide A at the third position results in the mutated sequence ATATGC.

2) Deletion. Starting at the k-th position of (S_0), delete (m) nucleotides to obtain the sequence (S_D); the mathematical expression is defined as follows:

$$S_D = S_0[1:k-1] + S_0[k+m:|S_0|], \tag{2.3}$$

where ($|S_0|$) denotes the length of (S_0). For example, given the original sequence ATTGC, deleting the nucleotide T starting from the third position results in the mutated sequence ATGC.

3) Point mutation. Apply point mutations to (S_0) and denote the mutated sequence as (S_M). For the nucleotide at the (i)-th position in (S_0), the probability (p) represents the selection of a nucleotide at random from (E) to perform the mutation, and the probability ($1-p$) represents the change that a mutation will occur; the mathematical expression is defined as follows:

$$S_M[i] = \begin{cases} Random(\{A,C,G,T\} \setminus \{S_0[i]\}), & probability: p \\ S_0[i], & probability: 1\text{-}p \end{cases}, \tag{2.4}$$

where $i = 1, 2, ..., \{A,G,C,T\} \setminus \{s_i\}$ denotes the three nucleotides other than the current base (s_i), and ($Random(\cdot)$) denotes a uniformly random selection of one nucleotide from a set of four. For example, given the original sequence ATTGC, a point mutation at the third position replaces T with A, which results in the mutated sequence ATAGC.

2.4. Construction of base combination frequency vectors and distance calculation

To quantify the sequence similarity in an alignment free manner, this study employs established base-combination (k-mer) frequency vectors to represent DNA sequences as multidimensional numerical vectors. For each sequence, the frequencies of different base combinations are calculated and normalized, thus enabling sequences of different lengths and hierarchical orders to be compared on a common compositional basis. Then, the Euclidean distance is used to quantify the compositional divergence between the frequency vectors of different sequences [34]. In this framework, smaller Euclidean distances indicate higher similarities in the base-combination composition, whereas larger

distances reflect increased divergences induced by mutation processes. Although information-theoretic measures such as the Kullback–Leibler or Jensen–Shannon divergences are widely used to compare the probability distributions, Euclidean distance is adopted here as a pragmatic and interpretable choice in the context of controlled simulations, owing to its computational simplicity and intuitive geometric interpretation in compositional space. In particular, when analyzing hierarchical repeats under insertion and deletion mutations, zero-frequency components in higher-order k-mers are common, and the Euclidean distance provides a straightforward way to handle such sparsity without requiring additional regularization, rather than implying superiority over divergence-based measures. However, it should be noted that the Euclidean distance captures differences in sequence composition rather than positional information, representing a known limitation shared by alignment-free approaches.

2.4.1. Design of different base combinations

To characterize the sequence composition across multiple granularities, this study adopts three widely used base-combination schemes: mononucleotide, dinucleotide, and trinucleotide combinations. These schemes describe compositional features ranging from single-base usage to short contiguous base patterns, thus enabling a progressive assessment of the sequence similarity at increasing levels of local context. All three schemes are well established in the alignment-free sequence analysis and have been extensively applied in comparative genomic studies, providing commonly used and interpretable representations for multi-scale compositional characterization.

1) Mononucleotide: to consider the mononucleotide composition feature in a studied DNA sequence, we define a set of four nucleotides as follows:

$$P = \{p_1, p_2, p_3, p_4\} = \{A, C, G, T\}, \tag{2.5}$$

where $p_i (i = 1, 2, 3, 4)$ represents each nucleotide, A stands for adenine, G stands for guanine, T stands for thymine, and C stands for cytosine.

2) Dinucleotide: to consider the dinucleotide composition feature in a studied DNA sequence, with a total of 16 dinucleotide combinations, we define a set of dinucleotides as follows:

$$B = \{b_1 ... b_{16}\} = \{AA, AT, AC, AG, TT, TA, TG, TC, GG, GC, GA, GT, CC, CG, CT, CA\}, \tag{2.6}$$

where $b_i (i = 1, 2, ..., 16)$ represents a specific string of two neighboring nucleotides.

3) Trinucleotide: to consider neighboring three nucleotides in the studied sequence a total of 64 trinucleotide form a set as follows:

$$C = \{c_1 ... c_{64}\} = \{AAA, AAT, AAC, AAG, ..., TTG\}, \tag{2.7}$$

where $c_i (i = 1, 2, ..., 64)$ represents a trinucleotide combination.

2.4.2. Counts of different base combinations and frequency vector design

For each DNA sequence, occurrences of all possible base combinations were counted according to standard k-mer counting rules using a sliding window with a step size of one nucleotide. To enable

direct and meaningful comparisons between sequences of different lengths, the raw k-mer counts were normalized by the total number of k-mer in each sequence, yielding compositional frequency vectors that sum to one. Thus, each DNA sequence was represented as a normalized k-mer frequency vector suitable for the subsequent distance-based similarity analysis.

1) Counts (n_{p_i}) and frequency vector (V_P) for mononucleotides

Counting rule: for a given DNA sequence X of length L, the mononucleotide n_{p_i} represents the total count of the i-th nucleotide in the set a P. The total number of nucleotides is L.

Frequency vector: the frequency vector is a multidimensional vector composed of the frequencies of each nucleotide, and is defined as follows:

$$V_P = (\frac{n_{p_1}}{L}, \frac{n_{p_2}}{L}, \frac{n_{p_3}}{L}, \frac{n_{p_4}}{L}), \tag{2.8}$$

where $\frac{n_{p_i}}{L}$ represents the frequency of the i-th nucleotide, and $\sum_{i=1}^{4} \frac{n_{p_i}}{L} = 1$.

2) Counts (n_{b_i}) and frequency vector (V_B) for dinucleotides

Counting rule: for a given DNA sequence X of length L, the dinucleotide combination n_{b_i} represents the total count of the i-th dinucleotide combination in B. Since dinucleotides consist of two consecutive bases, the total number of dinucleotides is $L-1$.

Frequency vector: the frequency vector is a multidimensional vector composed of the frequencies of each dinucleotide combination, and is defined as follows:

$$V_B = (\frac{n_{b_1}}{L-1}, \frac{n_{b_2}}{L-1}, ..., \frac{n_{b_{16}}}{L-1}), \tag{2.9}$$

where $\frac{n_{b_i}}{L-1}$ represents the frequency of the specific dinucleotide combination, and $\sum_{i=1}^{16} \frac{n_{b_i}}{L-1} = 1$ holds.

3) Counts (n_{c_i}) and frequency vector (V_C) for trinucleotides

Counting Rule: for a given DNA sequence X of length L, the trinucleotide combination n_{c_i} represents the total count of the i-th trinucleotide combination in C. Since trinucleotides consist of three consecutive bases, the total number of trinucleotides is $L-2$.

Frequency Vector: the frequency vector is a multidimensional vector composed of the frequencies of each trinucleotide combination, and is defined as follows:

$$V_C = (\frac{n_{c_1}}{L-2}, \frac{n_{c_2}}{L-2}, ..., \frac{n_{c_{64}}}{L-2}), \tag{2.10}$$

where $\frac{n_{c_i}}{L-2}$ represents the frequency of a specific trinucleotide combination, and $\sum_{i=1}^{64} \frac{n_{c_i}}{L-2} = 1$ holds.

By representing DNA sequences as normalized base-combination frequency vectors, compositional differences among sequences of varying lengths and hierarchical orders can be consistently and quantitatively compared within a unified vector space, providing a common basis for the subsequent distance-based similarity analysis.

2.4.3. Calculation of Euclidean distance between sequences

Based on the constructed base combination (k-mer) frequency vectors constructed for each sequence, the Euclidean distance was used to quantify the compositional similarity between pairs of sequences. In this framework, smaller Euclidean distances indicate higher similarities in the base combination composition [35], whereas larger distances reflect greater compositional divergences induced by mutation processes.

Given two sequences X and Y, their mononucleotide, dinucleotide, and trinucleotide frequency vectors $V(X)$ and $V(Y)$ are calculated, respectively. Then, the Euclidean distance between sequences X and Y is defined as follows:

$$d(X,Y)=\sqrt{\sum_{i=1}^{n}(v_i(X)-v_i(Y))^2}\,, \tag{2.11}$$

where ($n=4$, $n=16$, or $n=64$) represent the total counts of mononucleotide, dinucleotide, and trinucleotide combinations, respectively.

While information-theoretic metrics such as the Kullback–Leibler or Jensen–Shannon divergences are also commonly used to compare the probability distributions, the Euclidean distance provides a straightforward geometric interpretation and can be directly applied to compositional frequency vectors that include zero-frequency components, without implying methodological superiority. As with other alignment-free approaches, the Euclidean distance captures differences in the sequence composition rather than positional information, which represents a known limitation of the present analysis.

2.5. Comparative experiment on the similarity of different-ordered repetitive units

To systematically examine similar variation patterns between repetitive units at different hierarchical orders within multi-level repetitive sequences, the following simulation and analysis procedure was applied as follows:

1) First, the seed repetitive unit in the sequence is selected as the zero-order repetitive unit, and the zero-order repetitive unit is repeated 10 times to form a zero-order sequence, which is denoted as S_1. Next, the first-order repetitive unit is constructed by concatenating 5 consecutive zero-order repetitive units, each of which has undergone a mutation (mutation rate r_1), and this forms the new repeat unit. Then, the first-order sequence is generated by repeating the first-order repetitive unit 5 times, and is denoted as S_2. Finally, the second-order repetitive unit is formed by concatenating 3 consecutive first-order repetitive units, each of which has undergone a mutation (mutation rate r_2), and this forms the new repeat unit. Then, the second-order sequence is generated by repeating the

second-order repetitive unit 2 times, and is denoted as S_3.

2) The zero-order, first-order, and second-order repetitive units obtained from the above construction were selected as the objects for the similarity analysis. The sequence similarity between the repetitive units of different hierarchical orders was quantified by calculating Euclidean distances between their base-combination (k-mer) frequency vectors. Three mutation types—point mutations, insertions and deletions—were considered. For each mutation type, five mutation-rate levels (1%, 5%, 10%, 15% and 20%) were examined to evaluate how similarity patterns between the lower- and higher-order repetitive units vary under different mutation conditions.

3. Results

3.1. Dot plots of rice sequences

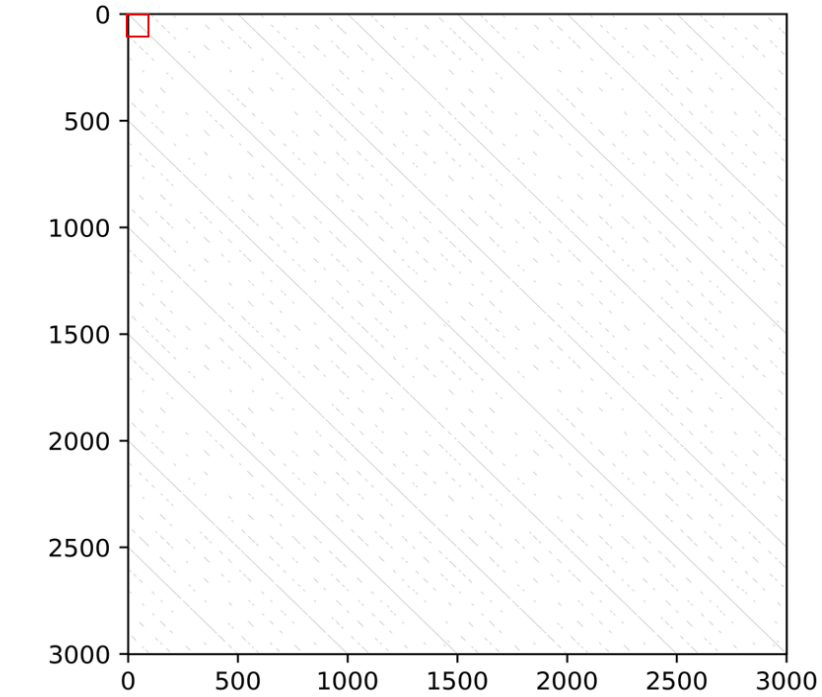

Figure 3. The top panel shows the DNA similarity analysis results of the experimental sequence, and the lower panel displays the 100 bp sequence **highlighted** by the red box.

Figure 3 presents dot plot results for a representative rice genomic sequence. The top panel shows the DNA similarity pattern of the analyzed sequence, while the bottom panel displays the 100 bp segment highlighted by the red box. To facilitate the interpretation of the dot plot, the hierarchical organization of repetitive structures observed in this sequence is summarized as follows. The dot plot reveals approximately six repeat units with lengths of 500 bp along the main diagonal. Each 500 bp unit is further composed of five sub-repeats of approximately 100 bp, thus forming a nested organization

spanning the 100 and 500 bp scales. Repeated diagonal patterns are observed at both scales, thus indicating a high sequence similarity within individual repeat units as well as between corresponding subunits across different repeats. These observations demonstrate the presence of well-defined multi-scale repetitive structures in the analyzed rice genomic sequence. Additionally, similar hierarchical repeat organizations are observed in additional rice sequences, as shown in the Supplementary.

3.2. Euclidean distance results under mononucleotide combinations

Table 1. Euclidean distances of mononucleotide composition vectors between hierarchical repeat units under different mutation types and mutation rates. Values are reported as mean ± standard deviation (n = 30 independent simulation runs).

	Mutation rate	$r_1 = 1\%, r_2 = 1\%$			$r_1 = 10\%, r_2 = 10\%$			$r_1 = 20\%, r_2 = 20\%$		
Point mutation	Inter-order comparison	Z–F	Z–S	F–S	Z–F	Z–S	F–S	Z–F	Z–S	F–S
	Euclidean distance ($\times 10^{-4}$)	0.37± 0.03	0.38± 0.04	0.38± 0.03	1.17± 0.06	2.68± 0.09	1.92± 0.08	3.61± 0.12	5.99± 0.18	3.35± 0.15
Nucleotide insertion	Inter-order comparison	Z–F	Z–S	F–S	Z–F	Z–S	F–S	Z–F	Z–S	F–S
	Euclidean distance ($\times 10^{-4}$)	0.47± 0.07	0.77± 0.10	0.33± 0.06	0.82± 0.12	1.34± 0.18	1.20± 0.17	1.77± 0.21	2.90± 0.34	1.53± 0.26
Nucleotide deletion	Inter-order comparison	Z–F	Z–S	F–S	Z–F	Z–S	F–S	Z–F	Z–S	F–S
	Euclidean distance ($\times 10^{-4}$)	0.37± 0.08	0.22± 0.07	0.38± 0.09	0.62± 0.14	0.80± 0.19	0.23± 0.11	0.78± 0.17	0.53± 0.21	0.21± 0.12

Note: Z: zero-order repeat unit, F: first-order repeat unit, and S: second-order repeat unit.

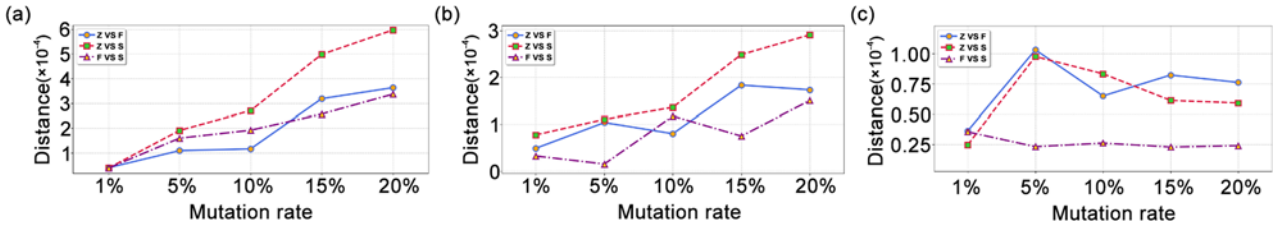

Figure 4. Changes in Euclidean distances of mononucleotide combinations between hierarchical repeat units under different mutation types and **mutation** rates. Values are reported as the mean over 30 independent simulation runs. (a) Point mutation; (b) Base insertion; (c) Base deletion. Z: zero-order repeat unit, F: first-order repeat unit, and S: second-order repeat unit.

We found that the similarity patterns between hierarchical repeat units under mononucleotide combinations strongly depend on the mutation type. Specifically, point mutations produce a consistent average increase in compositional divergence with the mutation rate, whereas insertion and deletion mutations give rise to more variable and non-monotonic behaviors across hierarchical levels.

Based on the averaged results over 30 independent simulation runs, Figure 4 and Table 1 summarize the mean Euclidean distances (± standard deviation) of mononucleotide compositions between repetitive units at different hierarchical levels under varying mutation types and mutation rates. The inclusion of multiple independent realizations enables a statistically more robust characterization of similarity trends while explicitly accounting for the stochastic nature of the mutation processes. Under point mutation conditions (Figure 4a), the mean Euclidean distances between zero-order (Z), first-order (F), and second-order (S) repeat units show a clear overall increasing trend with an increasing mutation rate. Specifically, as the mutation rate increases from 1% to 20%, the mean Z–F distance increases from 0.37×10^{-4} to 3.61×10^{-4}, the Z–S distance increases from 0.38×10^{-4} to 5.99×10^{-4}, and the F–S distance increases from 0.38×10^{-4} to 3.35×10^{-4}. The associated standard deviations remain small relative to the corresponding mean values across all mutation rates, indicating that this overall increasing trend is consistently observed across independent simulation runs. Importantly, this trend reflects an average tendency rather than a strictly deterministic relationship for individual realizations. In contrast, under nucleotide insertion mutations (Figure 4b), the mean Euclidean distances generally increase with the mutation rate but do not follow a strictly monotonic pattern. Although the Z–F and Z–S distances exhibit an overall upward tendency from low to high mutation rates, noticeable local fluctuations are observed, particularly for the F–S comparisons at intermediate mutation rates. The corresponding standard deviations are larger than those observed under point mutations, indicating an increased variability across independent realizations.

A similar but more pronounced behavior is observed for nucleotide deletion mutations (Figure 4c). The mean Euclidean distances display non-monotonic changes with increasing mutation rates, especially for the Z–S and F–S comparisons. In several cases, a substantial overlap of error bars is observed across adjacent mutation rates, thus suggesting that differences in mean distances are not always clearly separable given the inherent stochastic variability. Overall, these patterns indicate that, compared with point mutations, insertion and deletion mutations introduce a greater variability in the mononucleotide composition similarity. Such behavior is consistent with the additional complexity introduced by length-altering mutations, which may differentially affect repeat unit boundaries and the hierarchical organization, a feature commonly associated with the dynamic evolution of repetitive regions.

3.3. Euclidean distance results under dinucleotide combinations

We found that the effects of different mutation types on zero-order (Z), first-order (F), and second-order (S) repeat units under dinucleotide combinations are broadly consistent with those observed for mononucleotide combinations, while the absolute Euclidean distance values are generally higher. This increase is consistent with the fact that dinucleotide combinations capture additional local sequence context compared with the mononucleotide composition.

Based on the averaged results over 30 independent simulation runs, under point mutation conditions (Figure 5a), the mean Euclidean distances between repeat units at different hierarchical levels show a clear overall increasing trend with an increasing mutation rate. For all inter-order comparisons (Z–F, Z–S and F–S), the mean distances tend to increase as the mutation rate rises, with Z–S consistently exhibiting the largest values, followed by Z–F and F–S. The associated standard deviations remain relatively small across mutation rates, thus indicating that these trends are consistently observed across independent realizations rather than driven by stochastic variability. In contrast, under nucleotide insertion mutations (Figure 5b), the mean Euclidean distances generally increase with the mutation rate

but do not display strict monotonicity. While the Z–F and Z–S comparisons show an overall upward tendency, the F–S distances exhibit weaker increases and occasional plateaus. The corresponding standard deviations are noticeably larger than those observed under point mutations, reflecting an increased variability among independent simulation runs. A similar but more pronounced pattern is observed for nucleotide deletion mutations (Figure 5c). Although the mean Euclidean distances remain relatively low compared with those under point mutations, particularly for the F–S comparison, non-monotonic changes with the mutation rate are evident. In several cases, overlapping error bars are observed across adjacent mutation rates, suggesting that differences in mean distances are not always clearly separable given the inherent stochastic variability.

Table 2. Euclidean distances of dinucleotide composition vectors between hierarchical repeat units under different mutation types and mutation rates. Values are reported as mean ± standard deviation (n = 30 independent simulation runs).

	Mutation rate	$r_1=1\%, r_2=1\%$			$r_1=10\%, r_2=10\%$			$r_1=20\%, r_2=20\%$		
Point mutation	Inter-order comparison	Z–F	Z–S	F–S	Z–F	Z–S	F–S	Z–F	Z–S	F–S
	Euclidean distance ($\times10^{-4}$)	1.02± 0.07	1.28± 0.09	0.69± 0.06	3.48± 0.18	4.72± 0.21	2.83± 0.16	5.44± 0.26	6.85± 0.32	3.41± 0.20
Nucleotide insertion	Inter-order comparison	Z–F	Z–S	F–S	Z–F	Z–S	F–S	Z–F	Z–S	F–S
	Euclidean distance ($\times10^{-4}$)	0.81± 0.12	1.09± 0.15	0.36± 0.09	1.27± 0.19	1.78± 0.24	1.08± 0.21	1.98± 0.28	2.91± 0.36	1.16± 0.27
Nucleotide deletion	Inter-order comparison	Z–F	Z–S	F–S	Z–F	Z–S	F–S	Z–F	Z–S	F–S
	Euclidean distance ($\times10^{-4}$)	0.95± 0.14	1.08± 0.18	0.47± 0.11	1.01± 0.20	1.22± 0.23	0.38± 0.13	0.79± 0.22	0.89± 0.26	0.43± 0.15

Note: Z: zero-order repeat unit, F: first-order repeat unit, and S: second-order repeat unit.

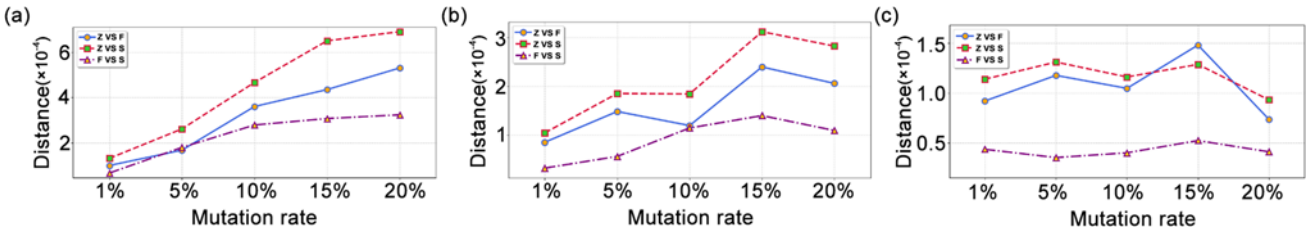

Figure 5. Changes in Euclidean distances of dinucleotide combinations between hierarchical repeat units under different mutation types and mutation rates. **Values** are reported as the mean over 30 independent simulation runs. (a) Point mutation; (b) Base insertion; (c) Base deletion. Z: zero-order repeat unit, F: first-order repeat unit, and S: second-order repeat unit.

Overall, these results indicate that point mutations tend to produce relatively consistent and cumulative changes in dinucleotide composition differences between hierarchical repeat units, whereas

insertion and deletion mutations introduce greater variability. This increased dispersion under insertion and deletion mutations are consistent with the additional complexity introduced by length-altering events, which may differentially affect the local sequence context and hierarchical organization, thus leading to more heterogeneous dinucleotide frequency distributions across the hierarchical levels.

3.4. Euclidean distance results under trinucleotide combinations

Table 3. Euclidean distances of trinucleotide composition vectors between hierarchical repeat units under different mutation types and mutation rates. **Values** are reported as mean ± standard deviation (n = 30 independent simulation runs).

	Mutation rate	$r_1 = 1\%, r_2 = 1\%$			$r_1 = 10\%, r_2 = 10\%$			$r_1 = 20\%, r_2 = 20\%$		
Point mutation	Inter-order comparison	Z–F	Z–S	F–S	Z–F	Z–S	F–S	Z–F	Z–S	F–S
	Euclidean distance ($\times 10^{-4}$)	1.42± 0.11	1.71± 0.13	0.81± 0.09	4.02± 0.22	6.14± 0.27	3.41± 0.19	7.58± 0.34	9.92± 0.41	4.28± 0.26
Nucleotide insertion	Inter-order comparison	Z–F	Z–S	F–S	Z–F	Z–S	F–S	Z–F	Z–S	F–S
	Euclidean distance ($\times 10^{-4}$)	1.15± 0.18	1.46± 0.21	0.39± 0.12	1.89± 0.27	2.58± 0.32	1.17± 0.24	2.98± 0.36	3.84± 0.43	1.31± 0.29
Nucleotide deletion	Inter-order comparison	Z–F	Z–S	F–S	Z–F	Z–S	F–S	Z–F	Z–S	F–S
	Euclidean distance ($\times 10^{-4}$)	1.26± 0.20	1.48± 0.24	0.51± 0.15	1.44± 0.29	1.69± 0.33	0.46± 0.18	1.09± 0.31	1.32± 0.37	0.54± 0.21

Note: Z: zero-order repeat unit, F: first-order repeat unit, and S: second-order repeat unit.

We found that the effects of different mutation types on zero-order (Z), first-order (F), and second-order (S) repeat units under trinucleotide combinations are broadly consistent with those observed for mono- and dinucleotide combinations. Because trinucleotide combinations capture richer local sequence arrangement information, the absolute Euclidean distance values are generally higher, which is consistent with their enhanced responsiveness to local sequence changes.

Based on the averaged results over 30 independent simulation runs, under point mutation conditions (Figure 6a), the mean Euclidean distances between repeat units at different hierarchical levels show a clear overall increasing trend with an increasing mutation rate. For all inter-order comparisons (Z–F, Z–S and F–S), the mean distances tend to increase as the mutation rates rise, with Z–S consistently exhibiting the largest values, followed by Z–F, while F–S remains comparatively smaller. The associated standard deviations are relatively small across the mutation rates, indicating that these trends are consistently observed across independent realizations rather than driven by stochastic fluctuations. In contrast, under nucleotide insertion mutations (Figure 6b), the mean Euclidean distances generally increase with the mutation rate but do not exhibit strict monotonicity. While the Z–F and Z–S comparisons show an overall upward tendency, the F–S distances display weaker growth and occasional plateaus. Moreover, the corresponding standard deviations are noticeably larger than those observed under point mutations, thus reflecting an increased variability among independent simulation runs. A similar but more pronounced pattern is observed under

nucleotide deletion mutations (Figure 6c). Although the mean Euclidean distances remain relatively low compared with those under point mutations, particularly for the F–S comparison, non-monotonic changes across mutation rates are evident. In several cases, increases in the mutation rate do not correspond to higher mean distances, and overlapping error bars are observed across adjacent mutation rates, thus indicating substantial stochastic variability.

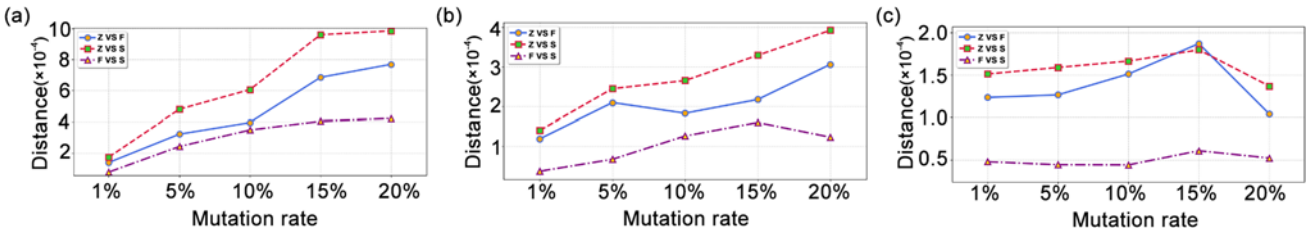

Figure 6. Changes in Euclidean distances of trinucleotide combinations between hierarchical repeat units under different mutation types and mutation rates. Values are reported as the mean over 30 independent simulation runs. (a) Point mutation; (b) Base insertion; (c) Base deletion. Z: zero-order repeat unit, F: first-order repeat unit, and S: second-order repeat unit.

Taken together, the integrated analysis of mono-, di-, and trinucleotide combinations indicates that, within the tested parameter ranges, point mutations are associated with relatively consistent and cumulative increases in the compositional differences between hierarchical repeat units when assessed using averaged Euclidean distances. In contrast, insertion and deletion mutations introduce greater variability and non-monotonic behavior, consistent with the additional complexity introduced by length-altering events, which may differentially affect the nucleotide arrangement patterns across hierarchical levels.

4. Discussion

4.1. Multi-level structural features of real sequences

Analyzing the similarity patterns of centromeric repeat sequences provides important insights into their structural organization and offers a biological context for the design and interpretation of simulated sequences. Through a dot plot analysis of representative genomic regions from rice, *Arabidopsis thaliana*, and the human genome, this study demonstrates that multi-layered repetitive architectures are widely observed across diverse eukaryotic genomes. These architectures are characterized not only by local short-range repeat units but also by long-range tandem repeats and nested repeat patterns spanning multiple length scales. In the human genome, centromeric regions are predominantly composed of α-satellite monomers that are further organized into higher-order repeat units that exhibit clear hierarchical nesting. This structural organization is consistent with observations reported in recent telomere-to-telomere T2T genome assemblies, thus providing a biologically grounded reference for the hierarchical repeat models adopted in our simulations. Similar multi-scale similarity patterns observed in rice and *Arabidopsis* further suggest that a hierarchical repeat organization is a common structural feature of repeat-rich genomic regions. The widespread presence

of such multi-layered repeat architectures suggests that a hierarchical organization represents a stable outcome of long-term genomic processes, shaped by the interplay of replication, mutation, sequence homogenization, and selective constraints. Rather than arising from isolated mutational events, these structures are qualitatively consistent with cumulative evolutionary processes such as concerted evolution and unequal recombination that are known to operate in repetitive genomic regions. Importantly, the real genomic examples analyzed here are not intended to serve as direct targets for the quantitative validation of the proposed distance-based framework. Instead, they provide biologically realistic reference patterns that motivate, inform, and constrain the construction and interpretation of the simulated hierarchical repeat sequences examined in this study.

4.2. Advantages of the base combination vector analysis method

Existing computational approaches to analyze the similarity in repetitive DNA sequences can be broadly grouped into three categories. The first category is comprised of alignment-based approaches [18] (such as BLAST), which assess the sequence similarity through explicit positional matching of nucleotides. While intuitive and effective for moderately repetitive regions, these methods often suffer from alignment ambiguity and a high computational cost when applied to long, highly repetitive, or heavily mutated genomic regions, such as centromeres. The second category includes alignment-free methods based on k-mer, frequency representations [19,20], which characterize sequences using compositional features rather than an explicit alignment and are therefore well suited for large-scale or structurally complex repetitive datasets. However, most extending k-mer based studies focus on global sequence comparisons and do not explicitly quantify similarity differences among repeat units organized at different hierarchical levels. The third category consists of multi-scale analytical techniques [21,22], such as autocorrelation analyses, Fourier transforms, and wavelet-based methods, which are effective in revealing periodicity, nesting, and long-range organization, but are generally less sensitive to fine-scale compositional changes and are not designed to distinguish the effects of specific mutation types.

Within this methodological landscape, the present study does not aim to introduce a fundamentally new sequence comparison paradigm, nor does it claim empirical superiority over alignment-based or other alignment-free methods. Instead, the primary contribution of the proposed framework lies in the systematic application of established k-mer frequency representations within an explicit hierarchical repeat context. By combining mononucleotide, dinucleotide, and trinucleotide composition vectors with the Euclidean distance, the framework enables dimensionally consistent and computationally efficient comparisons of similarity across different organizational levels, while remaining robust to the alignment ambiguity in long and repetitive sequences. This hierarchical, composition-based design allows for a controlled investigation of how different mutation types influence the similarity patterns across repeat orders. Point mutations alter the nucleotide identity without affecting the sequence length or overall structural organization. As a result, their effects on base-combination frequencies tend to gradually accumulate along the sequence, thus leading to smooth average shifts in the compositional similarity as the mutation rates increase. When assessed using averaged Euclidean distances, these effects manifest as an overall increasing tendency in the compositional divergence across the hierarchical levels. In contrast, insertion and deletion mutations modify the sequence length and local positional context, thus introducing additional complexity into the compositional structure of repeat units. Even single-base insertions and deletions can induce

downstream positional shifts, perturb repeat boundaries, or alter local k-mer distributions in a context-dependent manner. Consequently, the effects of insertion and deletion on the base-combination frequencies are inherently nonlinear and more variable, thus resulting in non-monotonic similarity patterns and increased dispersion across independent simulation runs. Importantly, these fluctuations should not be interpreted merely as analytical noise. Rather, they may reflect meaningful signatures of repeat boundary erosion, local structural instability, or "gappy" decay between hierarchical repeat units. From a biological perspective, the contrasting behaviors observed under point mutations versus insertion and deletion mutations are qualitatively consistent with known properties of centromeric repeat evolution. Substitution-driven divergence tends to gradually accumulate while largely preserving higher-order repeat organization, whereas insertions, deletions, and recombination-related processes can reshape repeat architectures and introduce structural heterogeneity across hierarchical levels. The present framework does not seek to directly model in vivo evolutionary processes, but instead provides a simplified and interpretable system to examine how different mutation regimes can influence the hierarchical similarity patterns.

In summary, the base-combination vector analysis employed here serves as a complementary tool rather than a replacement for existing alignment-based or multi-scale methods. It retains the computational efficiency and robustness of alignment-free approaches, explicitly incorporates hierarchical organization, and enables the systematic investigation of mutation-type-specific effects. As such, this framework offers a quantitative bridge between microscopic mutation processes and a macroscopic repeat architecture, and may support future extensions that integrate more complex mutation models or biologically informed validation strategies.

4.3. Proposed validation strategy using real genomic sequences

Although the present study focuses on controlled simulations to systematically characterize mutation-type-specific effects on the hierarchical repeat similarity, thereby extending the proposed framework to real genomic data represents a natural and important direction for future work. In the current study, simulations were intentionally adopted as the primary analytical setting, as they allow precise control over repeat architecture and mutation processes, which is difficult to achieve in real centromeric regions where repeat units are often overlapping, degenerated, or variable in length and copy number. Nevertheless, a feasible strategy to apply the proposed k-mer frequency and Euclidean distance analysis to real genomic sequences can be outlined. First, representative centromeric regions can be selected from high-quality telomere-to-telomere genome assemblies, such as those available for human, rice, and *Arabidopsis thaliana*. Second, hierarchical repeat units, including monomers and higher-order repeats, can be identified using established repeat annotation tools and previously reported higher-order repeat definitions. Third, for each identified repeat unit at different hierarchical levels, mononucleotide, dinucleotide, or trinucleotide frequency vectors can be constructed following the same procedures used in the simulation framework. Finally, the Euclidean distances between repeat units of different orders can be computed to examine whether real genomic data exhibit compositional similarity patterns qualitatively comparable to those observed in the simulations. Within such an analysis pipeline, dot plot visualization can provide an initial qualitative reference for the presence of hierarchical organization, while the k-mer based Euclidean distance analysis offers a complementary quantitative measure of compositional divergence across repeat levels. In this context, the simulation results presented in the current study serve as a controlled baseline, against which patterns observed in

real genomic sequences may be interpreted. The integration of simulation-based reference models with a real genomic data analysis would help clarify the extent to which the simplified hierarchical repeat structures and mutation regimes considered here capture essential features of repeat-rich genomic regions. Such an extension lies beyond the scope of the present work but represents a logical next step to further assess the biological relevance and generalizability of the proposed analytical framework.

5. Conclusions

This study shows that representative centromeric regions from rice, *Arabidopsis thaliana*, and the human genome exhibit pronounced multi-layered repetitive architectures, thus providing biologically grounded reference patterns that support the plausibility of modeling higher-order repeats through hierarchical constructions based on lower-order repetitive units. These real genomic examples serve to contextualize and motivate the simulation framework employed in this work rather than to provide direct quantitative validation targets.

Based on systematic simulations averaged over 30 independent runs, we quantitatively examined how different mutation types influence the sequence similarity across hierarchical repeat levels. Within the tested parameter ranges, point mutations were associated with relatively consistent average changes in the sequence similarity: for mononucleotide, dinucleotide, and trinucleotide composition vectors, the mean Euclidean distances showed a clear overall increasing trend with increasing mutation rates, thus reflecting a gradual reduction in compositional similarity on average. In contrast, insertion and deletion mutations gave rise to more variable similarity responses. Under these mutation types, the mean Euclidean distances did not exhibit a strict monotonic behavior but instead displayed noticeable fluctuations that are often comparable in magnitude to their associated standard deviations, thus highlighting the sensitivity of length-altering mutations to mutation positions, mutation scale, and local sequence context. Overall, the base-combination frequency vector representation combined with a Euclidean distance analysis provides a reproducible and interpretable framework to characterize mutation-induced similarity changes in hierarchical repeat sequences. Rather than replacing existing alignment-based or multi-scale analytical approaches, this framework complements them by enabling controlled and statistically grounded comparisons of mutation-type-specific effects across multiple organizational levels. By linking averaged mutation behaviors to hierarchical repeat architectures, the proposed approach offers a quantitative basis that may facilitate future investigations into the evolution, organization, and structural properties of complex repetitive genomic regions.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by programs from the National Natural Science Foundation of China (no. 32070669), Beijing-Tianjin-Hebei Medical-Engineering Integration Education, Research and Industry (ZD-YG-JBGS202512) and the Talented Scientist program from the Tangshan Municipal Bureau of Human Resources and Social Security to XW (no. 16013601).

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. G. Wang, M. K. Vasquez, Dynamic alternative DNA structures in biology and disease, *Nat. Rev. Genet.*, **24** (2023), 211–234. https://doi.org/10.1038/s41576-022-00539-9
2. H. X. Tang, Genome assembly, rearrangement, and repeats, *Chem. Rev.*, **107** (2007), 3391–3406. https://doi.org/10.1021/cr0683008
3. H. Cheng, G. T. Concepcion, X. Feng, H. Zhang, H. Li, Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm, *Nat. Methods*, **18** (2021), 170–175. https://doi.org/10.1038/s41592-020-01056-5
4. N. Altemose, G. A. Logsdon, A. V. Bzikadze, P. Sidhwani, S. A. Langley, G. V. Caldas, et al., Complete genomic and epigenetic maps of human centromeres, *Science*, **376** (2022), eabl4178. https://doi.org/10.1126/science.abl4178
5. S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, et al., The complete sequence of a human genome, *Science*, **376** (2022), 44–53. https://doi.org/10.1126/science.abj6987
6. S. Aganezov, S. M. Yan, D. C. Soto, M. Kirsche, S. Zarate, P. Avdeyev, et al., A complete reference genome improves analysis of human genetic variation, *Science*, **376** (2022), eabl3533. https://doi.org/10.1126/science.abl3533
7. M. D. Church, A next-generation human genome sequence, *Science*, **376** (2022), 34–35. https://doi.org/10.1126/science.abo5367
8. A. Gershman, M. E. G. Sauria, X. Guitart, M. R. Vollger, P. W. Hook, S. J. Hoyt, et al., Epigenetic patterns in a complete human genome, *Science*, **376** (2022), eabj5089. https://doi.org/10.1126/science.abj5089
9. S. J. Hoyt, J. M. Storer, G. A. Hartley, P. G. S. Grady, A. Gershman, L. G. de Lima, From telomere to telomere: The transcriptional and epigenetic state of human repeat elements, *Science*, **376** (2022), eabk3112. https://doi.org/10.1126/science.abk3112
10. M. R. Vollger, X. Guitart, P. C. Dishuck, L. Mercuri, W. T. Harvey, A. Gershman, et al., Segmental duplications and their variation in a complete human genome, *Science*, **376** (2022), eabj6965. https://doi.org/10.1126/science.abj6965
11. S. Goodwin, D. J. Mcpherson, R. W. Mccombie, Coming of age: Ten years of next-generation sequencing technologies, *Nat. Rev. Genet.*, **17** (2016), 333–351. https://doi.org/10.1038/nrg.2016.49
12. M. Naish, M. Alonge, P. Wlodzimierz, A. J. Tock, B. W. Abramson, A. Schmücker, et al., The genetic and epigenetic landscape of the Arabidopsis centromeres, *Science*, **374** (2021), eabi7489. https://doi.org/10.1126/science.abi7489
13. X. Hou, D. Wang, Z. Cheng, Y. Wang, Y. Jiao, A near-complete assembly of an Arabidopsis thaliana genome, *Mol. Plant*, **15** (2022), 1247–1250. https://doi.org/10.1016/j.molp.2022.05.014
14. B. Wang, X. Yang, Y. Jia, Y. Xu, P. Jia, N. Dang, et al., High-quality Arabidopsis thaliana genome assembly with nanopore and HiFi long reads, *Genomics Proteomics Bioinf.*, **20** (2022), 4–13. https://doi.org/10.1016/j.gpb.2021.08.003

15. J. M. Song, W. Z. Xie, S. Wang, Y. X. Guo, D. H. Koo, D. Kudrna, et al., Two gap-free reference genomes and a global view of the centromere architecture in rice, *Mol. Plant*, **14** (2021), 1757–1767. https://doi.org/10.1016/j.molp.2021.06.018
16. Y. Zhang, J. Fu, K. Wang, X. Han, T. Yan, Y. Su, et al., The telomere-to-telomere gap-free genome of four rice parents reveals SV and PAV patterns in hybrid rice breeding, *Plant Biol. J.*, **20** (2022), 1642–1644. https://doi.org/10.1111/pbi.13880
17. Y. Jiao, Double the genome, double the fun: Genome duplications in angiosperms, *Mol. Plant*, **11** (2018), 357–358. https://doi.org/10.1016/j.molp.2018.02.009
18. F. Stephen, Basic local alignment search tool, *J. Mol. Biol.*, **215** (1990), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
19. A. Sievers, K-mer content, correlation, and position analysis of genome DNA sequences for the identification of function and evolutionary features, *Genes*, **8** (2017), 122. https://doi.org/10.3390/genes8040122
20. J. H. Yu, Segmented K-mer and its application on similarity analysis of mitochondrial genome sequences, *Genes*, **518** (2013), 419–424. https://doi.org/10.1016/j.gene.2012.12.079
21. S. D. Huang, Normalized feature vectors: A novel alignment-free sequence comparison method based on the numbers of adjacent amino acids, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **10** (2013), 457–467. https://doi.org/10.1109/TCBB.2013.10
22. M. Randić, V. Marjan, On the similarity of DNA primary sequences, *J. Chem. Inf. Comput. Sci.*, **40** (2000), 599–606. https://doi.org/10.1021/ci9901082
23. W. Gu, F. Zhang, R. J. Lupski, Mechanisms for human genomic rearrangements, *Pathogenetics*, **1** (2008), 4. https://doi.org/10.1186/1755-8417-1-4
24. M. F. Lan, X. Y. Wang, X. C. Zhang, CentriVision: An integrated platform for multi-scale centromere analysis in plants, *Plant Commun.*, **7** (2025), 101689. https://doi.org/10.1016/j.xplc.2025.101689
25. Z. Li, L. Wang, K. Zhang, Algorithmic approaches for genome rearrangement: A review, *IEEE Trans. Syst. Man Cybern. Part C*, **36** (2006), 636–648. https://doi.org/10.1109/TSMCC.2005.855522
26. E. d'Alençon, H. Sezutsu, F. Legeai, E. Permal, S. Bernard-Samain, S. Gimenez, Extensive synteny conservation of holocentric chromosomes in Lepidoptera despite high rates of local genome rearrangements, *Proc. Natl. Acad. Sci. U.S.A.*, **107** (2010), 7680–7685. https://doi.org/10.1073/pnas.0910413107
27. B. Dorshorst, A. M. Molin, C. J. Rubin, A. M. Johansson, L. Strömstedt, M. H. Pham, et al., A complex genomic rearrangement involving the endothelin 3 locus causes dermal hyperpigmentation in the chicken, *PLoS Genet.*, **7** (2011), e1002412. https://doi.org/10.1371/journal.pgen.1002412
28. F. Cheng, J. Wu, X. Wang, Genome triplication drove the diversification of Brassica plants, *Hortic. Res.*, **1** (2014), 14024. https://doi.org/10.1038/hortres.2014.24
29. J. J. Smith, N. Timoshevskaya, C. Ye, C. Holt, M. C. Keinath, H. J. Parker, et al., The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution, *Nat. Genet.*, **50** (2018), 270–277. https://doi.org/10.1038/s41588-018-0199-4
30. J. M. Chen, D. N. Cooper, C. Férec, H. Kehrer-Sawatzki, G. P. Patrinos, Genomic rearrangements in inherited disease and cancer, *Semin. Cancer Biol.*, **20** (2010), 222–233. https://doi.org/10.1016/j.semcancer.2010.05.007
31. D. D. W. Twa, A. Mottok, F. C. Chan, S. Ben, Recurrent genomic rearrangements in primary testicular lymphoma, *J. Pathol.*, **236** (2015), 136–141. https://doi.org/10.1002/path.4522

32. F. Notta, M. Chan-Seng-Yue, M. Lemire, Y. Li, G. W. Wilson, A. A. Connor, et al., A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns, *Nature*, **538** (2016), 378–382. https://doi.org/10.1038/nature19823
33. F. Bai, T. M. Wang, On graphical and numerical representation of protein sequences, *J. Biomol. Struct. Dyn.*, **23** (2006). https://doi.org/10.1080/07391102.2006.10507078
34. J. H. Yu, D. S. Huang, Novel graphical representation of genome sequence and its applications in similarity analysis, *Physica A*, **391** (2012), 6128–6136. https://doi.org/10.1016/j.physa.2012.07.020
35. R. Mendizabal, On DNA numerical representations for genomic similarity computation, *PloS One*, **12** (2017), e0173288. https://doi.org/10.1371/journal.pone.0173288

AIMS Press