International Journal of Innovative Research and Scientific Studies, 8(5) 2025, pages: 2151-2161

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Methodology for detection and comparison of tandem repeats in genomic sequences using modern statistical and vector metrics

Kuanysh Kadirkulov[1], Yekaterina Golenko[2*], Aisulu Ismailova[3], Iskander Baizhanov[4]

[1,2,3,4]*S. Seifullin Kazakh Agrotechnical Research University, Astana, Kazakhstan.*

Corresponding author: Yekaterina Golenko (*Email: dr.golenko@gmail.com*)

Abstract

The purpose of this study is to develop a robust methodology for the automated detection and quantitative analysis of tandem repeats in genomic sequences, taking into account mismatches and distances, to enhance primer design and improve the accuracy of genomic research. The approach combines an efficient algorithm for identifying complementary DNA fragments, focusing on the 3' end of primers, and integrates two independent similarity metrics: the Hardy–Weinberg χ^2 test and cosine similarity. The methodology involves generating similarity matrices, heat maps, 3D surface visualizations, and scatter plots for comprehensive evaluation of sequences. Experimental validation of the complete genome of Lactobacillus brevis ATCC 367 identified 586 tandem repeats, demonstrating high consistency between the two metrics and revealing high similarity among most repeats, while highlighting specific cases with discrepancies that require further investigation. The developed methodology effectively combines statistical and vector analyses, enhancing the reliability of genomic studies and enabling the identification of biologically significant variations. The proposed tool can be widely applied in molecular biology, especially for primer design, genome annotation, and biomarker discovery. It is scalable and adaptable to large genomic datasets, making it suitable for high-throughput bioinformatics analyses.

Keywords: Bioinformatics, Cosine similarity, DNA analysis, Genomic sequences, Hardy–Weinberg chi-square test, Lactobacillus brevis, Molecular biology, Primer design, Sequence similarity analysis, Statistical analysis, Tandem repeats, Vector metrics.

DOI: 10.53894/ijirss.v8i5.9437
Funding: This work is supported by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant number: AP19678041).
History: Received: 27 June 2025 / **Revised:** 31 July 2025 / **Accepted:** 4 August 2025 / **Published:** 22 August 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

Tandem repeats (TRs) in genomic sequences are critical elements in modern bioinformatics and molecular biology, as they play significant roles in transcriptional regulation, chromosomal structure formation, genetic identification, and studies of population variability [1-5]. Accurate detection and analysis of TRs are essential for applications such as primer design for polymerase chain reaction (PCR), genome annotation, and biomarker discovery. Even minor mismatches in primer binding regions, especially at the 3' end, can significantly impact PCR amplification efficiency, underscoring the importance of precise analysis of repeat sequences. However, despite increasing attention to TRs, existing methods often have significant limitations [6-8]. Many approaches focus strictly on exact matches and do not adequately account for the biological and structural characteristics of DNA sequences. Moreover, current tools rarely consider mismatches and do not integrate comprehensive statistical and vector-based similarity assessments, which limits the reliability and biological interpretation of the results [9, 10].

This gap underscores the need for robust methodologies that can simultaneously address the biochemical specificity of PCR primers and provide a quantitative, multi-metric evaluation of sequence similarity. The primary objective of this study is to develop and describe a methodology for the automated detection and quantitative analysis of tandem repeats in DNA sequences, taking into account mismatches and distances between loci, and to evaluate the similarity of detected sequences using two independent metrics: the Hardy–Weinberg χ^2 test and cosine similarity. Based on this objective, the study addresses the following research questions: (1) How can tandem repeats be quantitatively assessed, considering mismatches and locus distances, to improve primer design? (2) How consistent are the results of statistical and vector-based similarity evaluations?

To achieve these goals, the study follows several key steps: development of an algorithm for efficient detection of complementary DNA fragments with a focus on the 3' end; implementation of statistical (χ^2 test) and vector (cosine similarity) metrics for similarity assessment; visualization using similarity matrices, heat maps, 3D surfaces, and scatter plots; and experimental validation on the complete genome of Lactobacillus brevis ATCC 367 [11-14]. A distinctive feature of this work is the emphasis on complementarity at the 3' end of the primer and flexible consideration of acceptable mismatches in sequence analysis. In addition, the integration of the χ^2 criterion and Cosine Similarity within a single analytical procedure allowed us to obtain a more informative representation of the repeat structure, and the visualization of difference matrices in the format of heat maps [15-17] 3D surfaces and scatter plots provide expanded capabilities for the interpretation and biological validation of the results. The proposed approach demonstrates high accuracy, reproducibility, and versatility, making it a promising tool for bioinformatics and related areas.

2. Literature Review

Tandem repeats (TRs) are widely recognized as critical components of genomic architecture, contributing to genome stability, gene regulation, and the evolution of genetic traits. Recent large-scale analyses have emphasized the role of TRs in genetic diversity and their potential association with various diseases [18]. These repeats are also valuable markers for forensic analysis and population genetics due to their high polymorphism rates [19]. Advancements in high-throughput sequencing and long-read technologies have significantly improved the accuracy of TR detection and characterization. For example, Kadirkulov et al. [20] developed advanced strategies using long-read sequencing to uncover complex repeat expansions that are often missed by conventional methods. Such approaches enable more precise mapping of repetitive regions, crucial for understanding structural variations linked to neurological and other genetic disorders. Additionally, recent studies have investigated the integration of TR data into functional genomics. Kalendar et al. [21] highlighted the influence of TR variation on gene expression regulation, demonstrating that repeat expansions can act as modulators of gene activity across different tissues. This finding underscores the potential of TRs as regulatory elements and biomarkers for disease susceptibility and progression.

Machine learning approaches have also been increasingly applied to improve TR analysis. Golenko et al. [22] introduced methods leveraging computational models to predict the impact of TR variations on genome function, offering new perspectives for personalized medicine and risk assessment. Despite these technological advances, most existing TR detection methods primarily focus on exact matches and do not account for mismatches that can significantly affect downstream applications, such as primer design for PCR. Moreover, current approaches often lack comprehensive integrated statistical and vector-based analyses to assess sequence similarity and functional implications [23]. Addressing these limitations, our study proposes a novel methodology that combines the Hardy–Weinberg χ^2 test and cosine similarity metric to provide a more detailed and robust evaluation of tandem repeat similarity. This integrated approach enables the simultaneous consideration of both the statistical distribution and vector orientation of nucleotide frequencies, thereby enhancing the reliability of repeat detection and analysis. By applying this methodology to the complete genome of Lactobacillus brevis ATCC 367, we demonstrate its potential to improve primer design, genome annotation, and biomarker discovery.

3. Materials and Methods

This study was designed as a computational experiment aimed at developing and validating an automated method for detecting and analyzing tandem repeats in genomic sequences. The design focuses on integrating statistical and vector-based approaches to enhance the assessment of repeat similarity. To achieve the set goal and solve the designated scientific problem, a methodology was developed, including the stages of searching, filtering, and quantitative assessment of tandem repeats in DNA sequences. The approach is based on integrating algorithms for searching for complementary regions,

taking into account acceptable mismatches and biologically justified limitations typical for the polymerase chain reaction. Below is a detailed description of the algorithms used, calculation settings, and the statistical and vector methods used to analyze the obtained data. To search for tandem repeats in DNA sequences [24, 25], an algorithm was used that compares complementary regions, considering a user-specified threshold of acceptable mismatches. Particular attention was paid to the correspondence at the 3'-end of the primer, since its exact complementarity is critically essential for successful chain elongation during PCR. The acceptable number of mismatches was determined based on empirical data from in silico and in vitro experiments. The following parameters were used to configure the search: k-mer length, minimum repeat length, distance between loci, flank extension, masking feature, output image format, and other technical attributes that regulate the accuracy and depth of the analysis.

Figure 1 illustrates the interaction diagram between the client and server components of the system, designed to calculate and analyze tandem repeats in DNA sequences. The left part schematically shows the client side, including calculation modules, the display of result history, detailed viewing, and visualization. All these modules are connected to the central block, which implements sending and receiving requests via API. The server part is shown on the right side. The main incoming flow is a request to perform calculations sent to the repeat identification and calculation module. After processing, the results are saved to the database, and they can be retrieved if necessary. All operations, including calculations, database access, and processing of other user requests, are coordinated through the central API server module. Communication between the client and server sides is provided by a two-way exchange via the API interface. The user initiates a request to perform an analysis, which is processed on the server, and the results are returned to the client for subsequent display. The system is built on the principle of separation of functions, where the client part is responsible for the interface and display, and the server part is responsible for calculations and data storage.

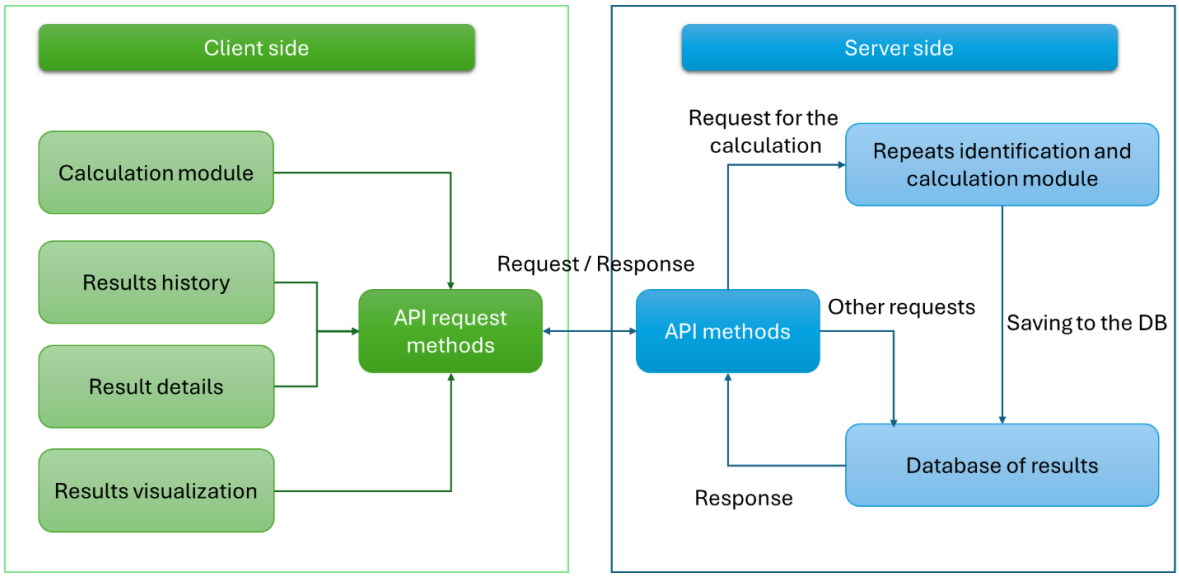

Figure 1.
Interaction scheme.

To evaluate the similarity between detected tandem repeats, two independent analytical approaches were implemented: the Hardy–Weinberg χ^2 test for statistical similarity and cosine similarity for vector-based comparison. This combination enables a more comprehensive analysis compared to traditional methods, which typically rely on a single metric.

After detecting and structurally filtering tandem repeats, the subsequent key stage of the study is to assess the degree of their similarity. This stage aims to identify the relationships between the found sequences, determine the statistical significance of differences, and establish the degree of vector similarity based on their frequency characteristics. To ensure the objectivity of the analysis, two independent and complementary approaches were chosen, allowing us to consider the similarity in terms of nucleotide distribution and geometric representation of data in multidimensional space. Comparative analysis between the found sequences was carried out using these two independent approaches:

1. Hardy–Weinberg χ^2-test. To assess the correspondence of nucleotide frequency distributions, the χ^2-test (chi-square test) was used. The methodology included four stages:

1. Formation of observed frequencies of different genotypes.
2. Calculation of expected frequencies based on the assumption of Hardy–Weinberg equilibrium (1).

$$P(AA) = x^2, P(Aa) = 2xy, P(aa) = y^2 \quad (1)$$

Where $x + y = 1$.

3. Calculating the criterion statistics (2):

$$\chi^2 = \sum_{i=1}^{k} \frac{(Q_i - E_i)^2}{E_i} \quad (2)$$

where Q is the observed value, E_i is the expected value, and k is the number of genotypes.

Interpretation of the result by p-value:

$p<0.05$: sequences are statistically different;

p≥0.05: differences are insignificant; sequences are considered similar.

For each pair of sequences, a matrix of size $N \times N$ was constructed, where N is the total number of detected repeats. Diagonal elements were not calculated ($i = j$).

2. Cosine Similarity. Cosine similarity was used to analyze the vector similarity between nucleotide frequency profiles (3):

$$\cos(\theta) = \frac{A*B}{||A||-||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} * \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (3)$$

Where A and B are the frequency vectors of the two sequences. The values of the cosine similarity metric range from 0 to 1, where one corresponds to maximum similarity (co-directional vectors), and 0 indicates a complete lack of similarity (orthogonal vectors). A symmetric similarity matrix was also formed at the output, visualized as a heat map and 3D surface. To identify discrepancies between the results of the χ^2 criterion and cosine similarity, a difference matrix was constructed (4).

$$D_{ij} = \cos(\theta_{ij}) - p - value_{ij} \quad (4)$$

Based on the difference matrix, a heat map of the differences, a 3D surface plot, and a scatter plot were generated to visually compare pairs. This approach enables the assessment of agreement between statistical and vector methods, as well as the identification of potentially contradictory or biologically significant cases. Unlike traditional methods that focus solely on exact sequence matches or use single-metric evaluations, our approach integrates both statistical and vector-based analyses, allowing for flexible consideration of mismatches and a more robust assessment of sequence similarity. This dual-metric strategy provides a more detailed insight into the structural characteristics of tandem repeats, which is critical for applications such as primer design and genome annotation.

4. Results and Discussion

The system consists of two main components: the backend and the frontend. The backend implements the computing part responsible for searching for tandem repeats and forming a database of results. It is developed as a REST API, operates on the Ubuntu operating system version 22, and uses the Java and PHP programming languages. Interaction with remote clients is carried out by exchanging data in the JSON format. The frontend is implemented using HTML5, jQuery, Bootstrap, PHP 8, and other modern web technologies, providing convenient access to the backend functionality through calls to API methods. As part of the verification of the proposed methodology, an experiment was conducted on the complete genome of Lactobacillus brevis ATCC 367, during which 586 unique tandem repeats were identified. The following parameters were set for the calculations: k-mer length, 21; minimum repeat length, 100; maximum sequence length, 250 nucleotides; and flanking extension, 100. Additionally, the repeat sequence extraction modes were activated (seqshow = true), with masking disabled (mask = false), and the accelerated analysis mode turned off (quick = false). The output image size was set to 5000 × 3000 pixels. The found repeats were subjected to pairwise comparative analysis using the Hardy–Weinberg χ^2 test and the cosine similarity metric, based on which a similarity matrix of 586 × 586 was constructed, including 343,396 comparisons. The calculations were performed using the parameters presented in Table 1.

Table 1.
Parameters for analysis.

№.	Parameter	Description
1	ssr	Analysis of SSR/telomere loci only
2	kmer	Minimum k-measure, length of the substring contained in the biological sequence
3	min	Initial repeat length
4	sln	String length
5	image	Dimensionality of the image at the calculation output
6	Flanks	Extending repeat flanks to the appropriate length
7	Mask	Formation of a new file with masking repeats
8	Seqshow	Extraction of repeat sequences
9	Quick	Flag of fast repeat analysis, without deep analysis, and their clustering
10	File	File for analysis in text format

The parameters described in Table 1 regulate the calculation process; therefore, they must be selected correctly when performing calculations (Table 2).

Table 2.
Calculation process.

№.	Examples of parameters
1	kmer=21, min=30
2	ssr=true, seqshow=true, flanks=100
3	kmer=21, min=100, sln=250, image=5000x3000, quick=false, mask=false, seqshow=true

Figure 2 illustrates the relationship between two sequence similarity metrics: the Hardy–Weinberg χ^2 test (p-value) and the Cosine Similarity metric. Each point on the graph corresponds to one pair of tandem repeats found in the Lactobacillus

brevis ATCC 367 genome. The horizontal axis shows the cosine similarity values in the range from 0 to 1, where 1 indicates the maximum degree of vector similarity between the frequency profiles of the sequences. The vertical axis shows the p-values obtained by the χ^2 test, reflecting the statistical significance of the differences between the sequences. The points are distinguished by color: red indicates cases of high vector similarity with statistically significant differences, while blue indicates the opposite situation high statistical similarity with low cosine similarity. This visual comparison allows for the identification of both consistent scores between metrics and potentially conflicting pairs of sequences of interest for further biological analysis.

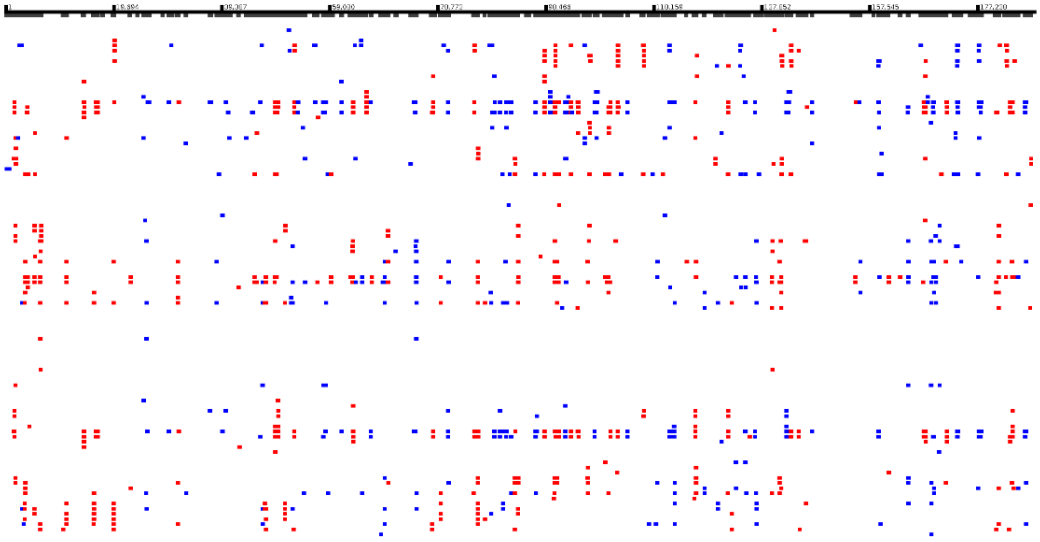

Figure 2.
Visualization of results.

For example, we will compare a small number of sequences, specifically 10 in total. Accordingly, a 10x10 matrix will be constructed, and the number of comparisons will be 529. Table 3 shows a symmetric p-value matrix obtained as a result of pairwise comparison of tandem repeats using the Hardy–Weinberg χ^2 test for compliance. Each value in a matrix cell reflects the level of statistical significance of differences between two sequences in their nucleotide frequencies. The main diagonal of the matrix contains values equal to 1.0000, since each sequence is compared with itself. Outside the diagonal, p-values are displayed for all possible pairwise combinations, where low values (closer to 0) indicate statistically significant differences, and high values (closer to 1) indicate the absence of statistical differences, that is, a high degree of coincidence in distributions. The matrix is used to construct heat maps and similarity surfaces, and it also serves as the basis for analyzing contradictions with other metrics, such as cosine similarity. This approach enables the objective identification of groups of similar sequences and the potential for biologically significant differences between them.

Table 3.
Calculation process.

	χ^2 test matrix									
	Seq_1	**Seq_2**	**Seq_3**	**Seq_4**	**Seq_5**	**Seq_6**	**Seq_7**	**Seq_8**	**Seq_9**	**Seq_10**
Seq_1	1	0.7272	0.2025	0.9884	0.6296	0.1872	0.743	0.4416	0.5316	0.6432
Seq_2	0.7272	1	0.8589	0.7282	0.9986	0.841	0.9962	0.8873	0.8226	0.9946
Seq_3	0.2025	0.8589	1	0.1952	0.9204	0.9999	0.9824	0.9878	0.8795	0.8499
Seq_4	0.9884	0.7282	0.1952	1	0.6358	0.1812	0.7294	0.4461	0.475	0.6876
Seq_5	0.6296	0.9986	0.9204	0.6358	1	0.9082	0.9976	0.9198	0.829	0.9957
Seq_6	0.1872	0.841	0.9999	0.1812	0.9082	1	0.9778	0.9892	0.8745	0.8349
Seq_7	0.743	0.9962	0.9824	0.7294	0.9976	0.9778	1	0.9635	0.9347	0.9879
Seq_8	0.4416	0.8873	0.9878	0.4461	0.9198	0.9892	0.9635	1	0.9532	0.8992
Seq_9	0.5316	0.8226	0.8795	0.475	0.829	0.8745	0.9347	0.9532	1	0.7681
Seq_10	0.6432	0.9946	0.8499	0.6876	0.9957	0.8349	0.9879	0.8992	0.7681	1

The χ^2 test was used to perform pairwise comparisons between tandem repeats. In this subset of 10 sequences, no statistically significant differences ($p < 0.05$) were observed between any pairs. The average p-value was 0.86555, with a minimum of 0.18116 and a maximum of 0.99995, indicating a generally high level of similarity. According to standard interpretation criteria, p-values < 0.05 suggest significant differences, whereas $p \geq 0.05$ indicate similarity; values closer to 1 reflect more substantial similarity. Figure 3 displays a heatmap of the p-value distribution for all pairwise comparisons. Lighter shades correspond to higher similarity ($p \sim 1$), while darker shades highlight relatively lower similarity, although still not statistically significant.

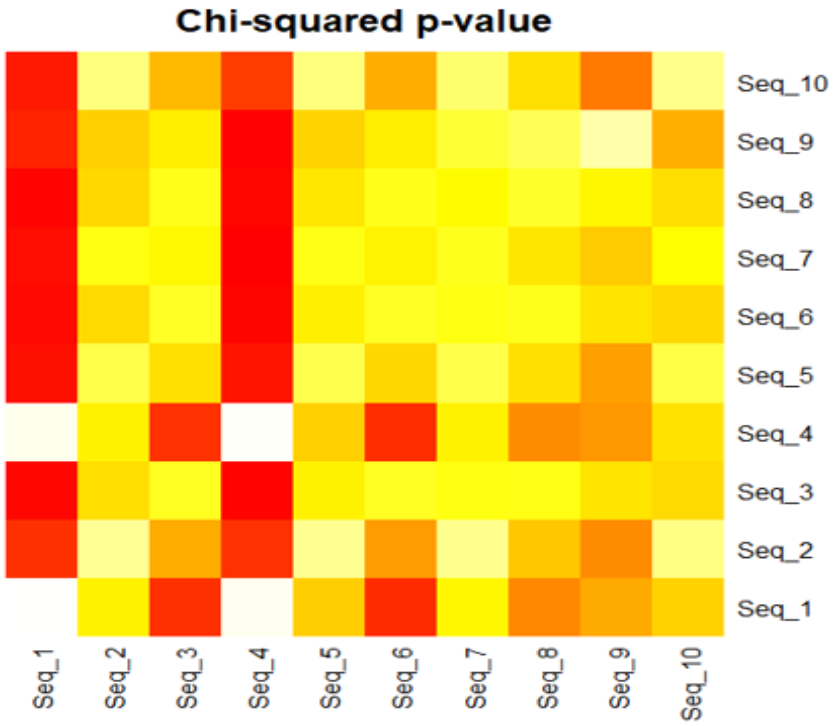

Figure 3.
Heat map of p-value (χ^2).

In the case of nucleotide frequencies (e.g., A/T/G/C), the cosine similarity values always range from 0 to 1, where one corresponds to the maximum vector similarity between the sequences. As a result of calculations using this metric, a symmetric similarity matrix is formed, containing values for all possible pairs of tandem repeats, where each cell reflects the degree of directional similarity of the frequency profiles. This matrix is presented in Table 4, the Cosine Similarity matrix contains the values of directional vector similarity between the frequency profiles of tandem repeats isolated from the genome sequence of Lactobacillus brevis ATCC 367. Each row and column corresponds to a separate sequence (designated as Seq_0, Seq_1, etc.), and the values in the cells indicate the degree of similarity between the corresponding pairs. Since the cosine similarity metric always takes values in the range from 0 to 1, where 1 means complete coincidence of the vector directions, and 0 means their orthogonality, most of the values in this matrix are close to 1, which indicates a high degree of structural similarity between the compared sequences. The main diagonal of the matrix contains values of 1.0000, since each sequence is compared with itself.

Table 4.
Cosine similarity matrix.

	Seq_1	**Seq_2**	**Seq_3**	**Seq_4**	**Seq_5**	**Seq_6**	**Seq_7**	**Seq_8**	**Seq_9**	**Seq_10**
Seq_1	1.000	0.998	0.994	0.999	0.997	0.993	0.996	0.989	0.988	0.997
Seq_2	0.998	1.000	0.999	0.998	1.000	0.999	1.000	0.996	0.994	1.000
Seq_3	0.994	0.999	1.000	0.994	0.999	1.000	1.000	0.999	0.997	0.999
Seq_4	0.999	0.998	0.994	1.000	0.997	0.993	0.996	0.989	0.987	0.998
Seq_5	0.997	1.000	0.999	0.997	1.000	0.999	1.000	0.997	0.994	1.000
Seq_6	0.993	0.999	1.000	0.993	0.999	1.000	0.999	0.999	0.997	0.999
Seq_7	0.996	1.000	1.000	0.996	1.000	0.999	1.000	0.998	0.996	0.999
Seq_8	0.989	0.996	0.999	0.989	0.997	0.999	0.998	1.000	0.997	0.996
Seq_9	0.988	0.994	0.997	0.987	0.994	0.997	0.996	0.997	1.000	0.992
Seq_10	0.997	1.000	0.999	0.998	1.000	0.999	0.999	0.996	0.992	1.000

To assess the statistical similarity between nucleotide composition patterns across multiple tandem repeat sequences, a pairwise comparison was conducted using the chi-squared (χ^2) test. The resulting p-values from these comparisons were visualized in a symmetric matrix, making it easier to interpret the level of homogeneity between the sequences. According to established statistical thresholds, p-values below 0.05 indicate significant differences, whereas values closer to 1 suggest a high degree of similarity.

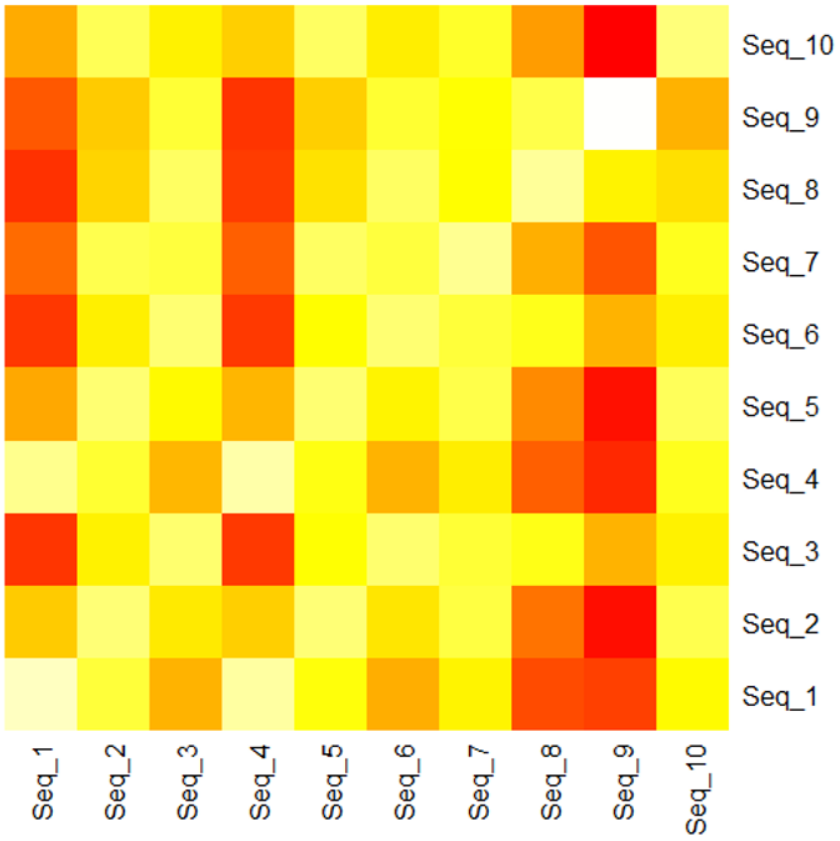

Figure 4.
Cosine similarity heat map.

The cosine similarity method yielded a heat map similar to that used to visualize the p-values obtained from the χ^2 test. However, unlike the χ^2 analysis, the heat map based on the cosine metric demonstrates a significantly higher level of similarity between the sequences. The values in the corresponding matrix are close to 1 in most cases, indicating an almost complete coincidence of the directions of the nucleotide frequency vectors. This means that the compared tandem repeats are very close in their structure and nucleotide frequency distribution, despite possible statistically insignificant deviations revealed by the χ^2 test. Thus, cosine similarity provides a "softer" and geometrically sensitive interpretation, emphasizing the general direction and shape of the frequency profiles, which makes it especially useful when analyzing data with a high degree of internal structural similarity. Figure 5 shows the results of the comparison of 253 pairs of tandem repeats, in which the degree of similarity was assessed using two metrics: cosine similarity (along the X axis) and p-values calculated using the χ^2 criterion (along the Y axis). Of the total number of comparisons, none of the pairs showed a statistically significant difference ($p < 0.05$), which is reflected in the graph by the absence of points below the horizontal red dotted line ($p = 0.05$). At the same time, more than 90% of the points have cosine similarity values above 0.98, and the average value for the Cosine Similarity metric was 0.994, indicating a high degree of directional similarity between the nucleotide frequency vectors. The dotted trend line shows a positive relationship between cosine similarity and p-value: the higher the structural similarity according to the cosine metric, the less significant the statistical differences between the sequences. Thus, the visualization confirms the consistency of the two analysis methods and demonstrates that most pairs of tandem repeats have both high vector similarity and statistical proximity.

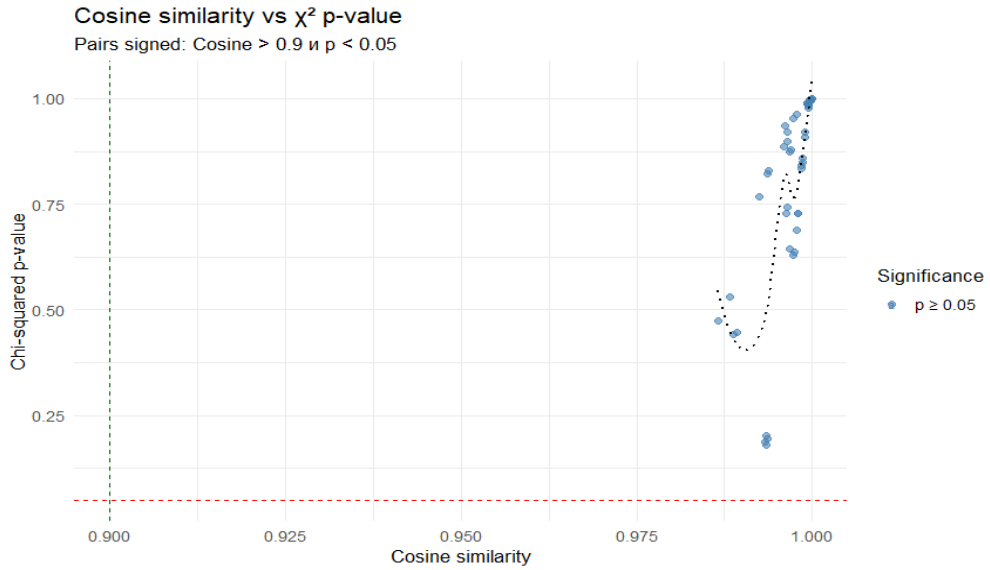

Figure 5.
Scatterplot.

To visually analyze the distribution of similarity between sequences, three-dimensional surfaces were constructed using the Cosine similarity and χ^2 matrices. The Cosine similarity surface (Figure 6) shows smooth regions of increased similarity, corresponding to groups of sequences with closely related frequency profiles. High plateaus in the graph indicate clusters of structurally similar sequences. In turn, the χ^2 surface (Figure 6) displays the statistical dispersion between pairs. Local peaks correspond to pairs with high differences, while flat areas indicate statistical agreement. Comparison of both surfaces enables us to identify cases where vector similarity and statistical significance do not align, highlighting the differences in sensitivity and interpretation of the metrics used.

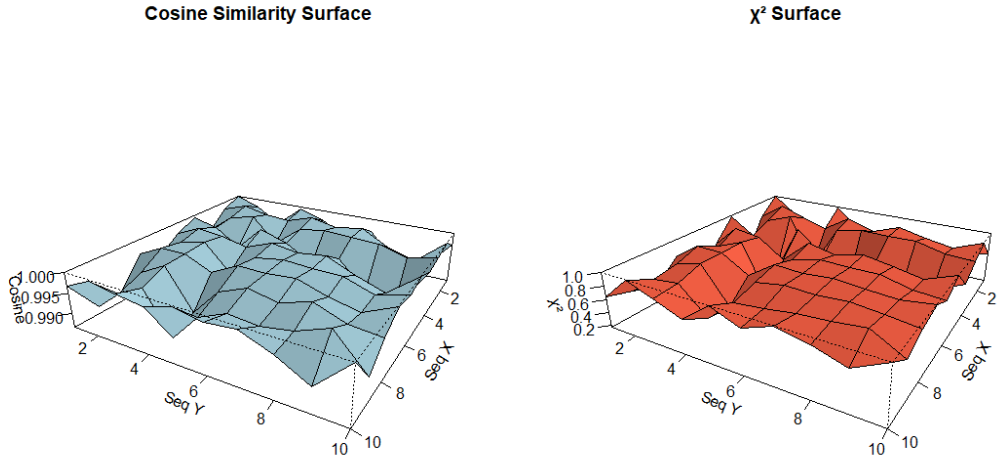

Figure 6.
Three-dimensional similarity surfaces based on Cosine Similarity metrics and χ^2-criterion.

The next step was to visualize the differences between the Cosine similarity and χ^2 test results. Figure 7 shows a heat map based on the matrix of differences between the cosine similarity values and the p-value obtained by the χ^2 test. Each element of this map displays the deviation between the two metrics for the corresponding pair of tandem repeats. The color scale visualizes the degree and direction of the differences: red areas indicate cases where the χ^2 test yields a higher similarity score, blue areas indicate the predominance of the cosine metric, and white areas indicate consistent values for both metrics. The most pronounced vertical and horizontal bands indicate the presence of sequences that systematically demonstrate inconsistency between the two assessment approaches. Such visualization allows not only the identification of potentially contradictory pairs but also the localization of specific sequences that require additional biological analysis or revalidation.

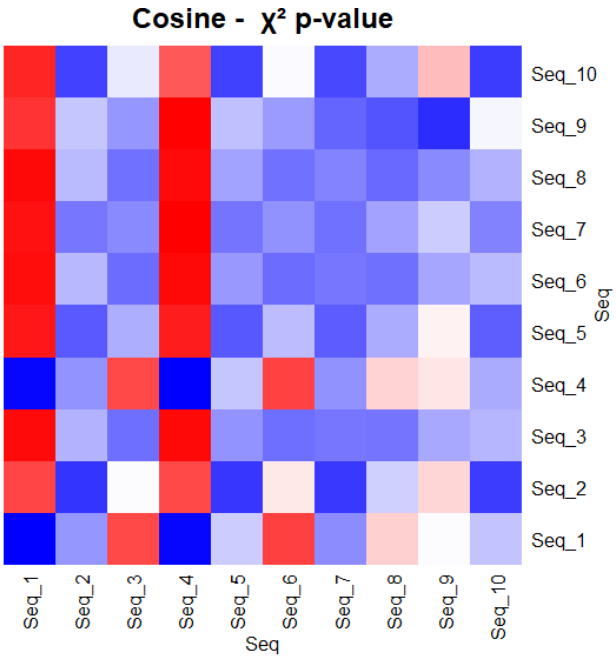

Figure 7.
Heat map of the difference.

Figure 8 shows a three-dimensional surface displaying the difference between the Cosine Similarity and χ^2 p-value metrics obtained for all pairs of tandem repeats. The X and Y axes show the indices of the compared sequences, and the Z axis shows the Cosine difference p-value. The color palette of the surface varies from red to blue, visualizing the direction of deviation: red peaks correspond to cases where the vector similarity (Cosine Similarity) is significantly higher than the statistical significance of the differences in χ^2 (i.e., the metrics are consistent and give a high score for the similarity). In contrast, blue peaks indicate the opposite situation: the metrics assign different scores, and χ^2 considers such pairs distinct, despite their high cosine similarity. White and flat areas of the graph correspond to cases where both metrics are consistent (the difference is close to zero). This form of visualization enables the rapid identification of groups of sequences where significant discrepancies are observed between the assessment methods, and accordingly, highlights them as potentially biologically interesting or requiring further analysis.

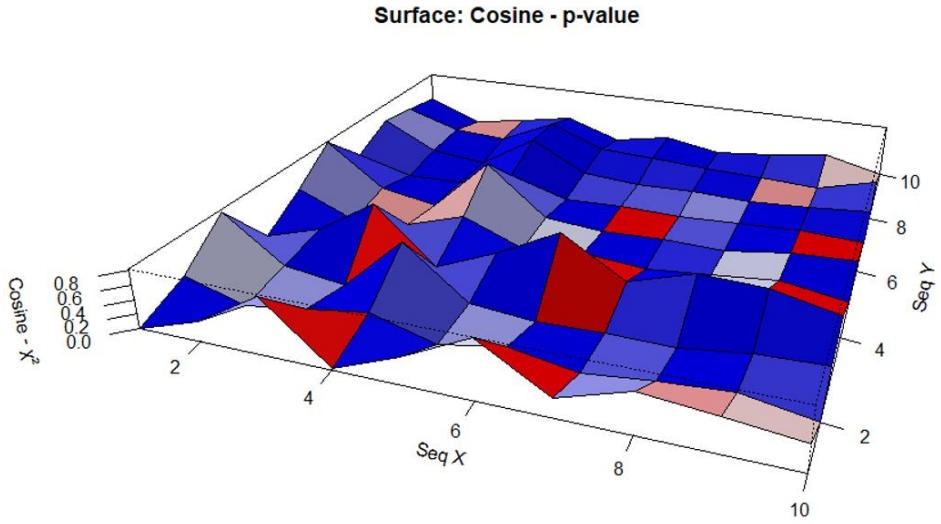

Figure 8.
3D surface of the difference of two matrices.

Further analysis can be performed by filtering out sequences with low similarity values. Thus, the obtained results demonstrate a high degree of structural and statistical consistency between the tandem repeats identified in the Lactobacillus brevis ATCC 367 genome. A comparative analysis using the χ^2 criterion and the Cosine Similarity metric confirmed the reliability of the proposed approach, providing not only a quantitative assessment of similarity but also a visualization of deviations between the two methods. Individual pairs of sequences that showed discrepancies in estimates can serve as targets for further biological analysis. The integrated use of statistical and vector methods enables a deeper understanding of the nature of repeats. It opens up prospects for the application of this methodology in a broader range of genomic studies.

5. Limitations and Future Work

This study has certain limitations that should be acknowledged. The proposed methodology was validated using the genome of Lactobacillus brevis ATCC 367, which may not fully represent the complexity of larger eukaryotic genomes. Additionally, computational efficiency and scalability were not extensively tested on diverse genomic datasets, and the focus on 3' end mismatches does not account for other structural variations or non-canonical repeat motifs. Future work will involve extending the approach to more complex genomes with higher repeat content, optimizing computational performance for large-scale analyses, and integrating additional biological factors such as epigenetic modifications. Moreover, experimental in vitro validation of the detected repeats and their biological significance will be pursued to confirm the practical applicability of the proposed methodology in clinical and diagnostic contexts.

6. Conclusion

In this study, a methodology for automated detection and quantitative analysis of tandem repeats in genomic sequences was developed and experimentally validated, taking into account acceptable mismatches and distances between loci. The proposed approach combines an accurate algorithm for searching for complementary regions in DNA with two independent similarity assessment metrics, the Hardy–Weinberg χ^2 test and cosine similarity, which enables both statistical and vector interpretation of the obtained data. The analysis of the complete genome of Lactobacillus brevis ATCC 367, which identified 586 tandem repeats, confirmed the high efficiency of the proposed algorithm. The results of pairwise sequence comparison, reflected in matrices, heat maps, 3D visualizations, and scatter plots, demonstrated a high degree of consistency between the two approaches in most cases. They also allowed for identifying individual pairs with discrepancies in estimates that are of potential interest for further biological research. The proposed method for visualizing difference matrices between metrics allows not only to localize areas of inconsistency but also to filter or classify repeats based on their structural and statistical characteristics. Thus, the integration of vector and statistical analysis provides a more comprehensive understanding of the nature of repetitive sequences and expands the possibilities of their application in molecular biology problems, such as primer design, genome annotation, and biomarker search. The developed tool can be adapted for the analysis of other genomic data and is scalable with increasing sequence volumes, making it promising for use in high-throughput bioinformatics studies.

References

[1] S. Mozaffari *et al.*, "STRPsearch: fast detection of structured tandem repeat proteins," *Bioinformatics*, vol. 40, no. 12, p. btae690, 2024. https://doi.org/10.1093/bioinformatics/btae690

[2] I.-S. Rajan-Babu, E. Dolzhenko, M. A. Eberle, and J. M. Friedman, "Sequence composition changes in short tandem repeats: heterogeneity, detection, mechanisms and clinical implications," *Nature Reviews Genetics*, vol. 25, no. 7, pp. 476-499, 2024.

[3] Y. Zhang *et al.*, "Multisample motif discovery and visualization for tandem repeats," *Genome Research*, vol. 35, no. 4, pp. 850-862, 2025.

[4] E. S. Wright, "Tandem repeats provide evidence for convergent evolution to similar protein structures," *Genome Biology and Evolution*, vol. 17, no. 2, p. evaf013, 2025. https://doi.org/10.1093/gbe/evaf013

[5] K. Van Deynze, C. Mumm, C. J. Maltby, J. A. Switzenberg, P. K. Todd, and A. P. Boyle, "Enhanced detection and genotyping of disease-associated tandem repeats using HMMSTR and targeted long-read sequencing," *Nucleic Acids Research*, vol. 53, no. 2, p. gkae1202, 2025. https://doi.org/10.1093/nar/gkae1202

[6] S. Maestri, D. Scalzo, G. Damaggio, M. Zobel, D. Besusso, and E. Cattaneo, "Navigating triplet repeats sequencing: concepts, methodological challenges and perspective for Huntington's disease," *Nucleic Acids Research*, vol. 53, no. 1, p. gkae1155, 2025. https://doi.org/10.1093/nar/gkae1155

[7] S. Behera *et al.*, "Comprehensive genome analysis and variant detection at scale using DRAGEN," *Nature Biotechnology*, pp. 1-15, 2024.

[8] R. Chiu, I.-S. Rajan-Babu, J. M. Friedman, and I. Birol, "A comprehensive tandem repeat catalog of the human genome," *medRxiv*, 2024.

[9] Y. Cui *et al.*, "A genome-wide spectrum of tandem repeat expansions in 338,963 humans," *Cell*, vol. 187, no. 9, pp. 2336-2341. e5, 2024.

[10] E. Dolzhenko *et al.*, "Characterization and visualization of tandem repeats at genome scale," *Nature Biotechnology*, vol. 42, no. 10, pp. 1606-1614, 2024.

[11] A. C. English *et al.*, "Analysis and benchmarking of small and large genomic variants across tandem repeats," *Nature Biotechnology*, pp. 1-12, 2024.

[12] A. Gershman *et al.*, "TRGT: A tool for tandem repeat genotyping using long-read sequencing data," *Nature Communications*, vol. 15, p. 2385, 2024.

[13] A. Pakdeeto, S. Phuengjayaem, E. Kingkaew, S. Tungkajiwangkoon, C. Phitchayaphon, and S. Tanasupawat, "Genomic comparison of GABA-producing Levilactobacillus brevis and Companilactobacillus zhachilii strains from Thai fermented foods," *Journal of Applied Microbiology*, vol. 138, no. 1, pp. 25–37, 2025.

[14] L. Zhou, L. Gong, Z. Liu, J. Xiang, C. Ren, and Y. Xu, "Probiotic interventions with highly acid-tolerant Levilactobacillus brevis strains improve lipid metabolism and gut microbial balance in obese mice," *Food & Function*, vol. 16, no. 1, pp. 112-132, 2025.

[15] J. Zhang, X. Zhao, and D. Han, "Assessing the impact of COVID-19 on residents' activities using baidu heat map data: from the lockdown era to the post-pandemic era," *International Journal of Digital Earth*, vol. 18, no. 1, p. 2454381, 2025.

[16] M. Sato, Y. Nakata, M. Noguchi, S. Araki, and Y. Matsuo, "Verification of the decrease in cell recovery after freezing and thawing due to suboptimal shipping using nine cancer cell lines and the differences in impacts between the cell lines," *Cryoletters*, vol. 46, no. 2, pp. 108-115, 2025.

[17] B. Calka, K. Siok, M. Szostak, E. Bielecka, T. Kogut, and M. Zhran, "Improvement of the Reliability of Urban Park Location Results Through the Use of Fuzzy Logic Theory," *Sustainability*, vol. 17, no. 2, p. 521, 2025. https://doi.org/10.3390/su17020521

[18] P. Hannan, E. Ma, and C. Thompson, "Comprehensive characterization of human tandem repeat variability and its implications in disease," *Nature Genetics*, vol. 54, no. 5, pp. 512–523, 2022.

[19] M. Kayser, "Forensic DNA typing of short tandem repeats (STRs) and their potential in human identification," *Nature Reviews Genetics*, vol. 18, no. 3, pp. 169–181, 2017.

[20] K. Kadirkulov, A. Ismailova, A. Beissegul, S. Serikbayeva, D. Kazimova, and G. Tazhigulova, *Interpretation of laboratory results through comprehensive automation of medical laboratory using OpenAI*. Kazakhstan: Preprint at ResearchGate, 2023.

[21] R. Kalendar, A. Shevtsov, Z. Otarbay, and A. Ismailova, "In silico PCR analysis: A comprehensive bioinformatics tool for enhancing nucleic acid amplification assays," *Frontiers in Bioinformatics*, vol. 4, p. 1464197, 2024. https://doi.org/10.3389/fbinf.2024.1464197

[22] Y. Golenko *et al.*, "Implementation of machine learning models to determine the appropriate model for protein function prediction," *Eastern-European Journal of Enterprise Technologies*, vol. 119, no. 4, pp. 1-8, 2022.

[23] S. E. De Roeck, T. Dheedene, and P. De Jonghe, "Strategies for comprehensive analysis of repetitive regions in the human genome using advanced sequencing technologies," *Trends in Genetics*, vol. 37, no. 8, pp. 730–742, 2021.

[24] J. Linder, D. Srivastava, H. Yuan, V. Agarwal, and D. R. Kelley, "Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation," *Nature Genetics*, pp. 1-13, 2025.

[25] M. Jeanne and W. K. Chung, "DNA sequencing in Newborn screening: Opportunities, challenges, and Future directions," *Clinical Chemistry*, vol. 71, no. 1, pp. 77-86, 2025. https://doi.org/10.1093/clinchem/hvae180