Journal of Bio-Digital and Forensic Informatics (JBDFI)
Vol. 1, Issue. 1, 2026, pp. 13–22
DOI: XX.XX.XXX
E-ISSN: xxxx-xxxx

Research

AI-Based Sequence Similarity Analysis as Digital Genetic Evidence: A Pilot Study on Growth-Related Genes

Rosyid R. Al-Hakim[1,2,*], Galih Samodra[3], Yanuar Z. Arief[4], Reina Melani[5], Lukman Hakim[6], Fiqih Nurkholis[7]

[1]Dept. of Information System, Universitas Harapan Bangsa, Indonesia
[2]The Institution of Engineers Indonesia (PII), Indonesia
[3]Dept. of Pharmacy, Universitas Harapan Bangsa, Indonesia
[4]Dept. of Electrical & Electronic Engineering, Universiti Malaysia Sarawak, Malaysia
[5]Dept. of Pharmacy, Universitas Harapan Bangsa, Indonesia
[6]Dept. of Pharmacy, Universitas Harapan Bangsa, Indonesia
[7]Dept. of Pharmacy, Universitas Harapan Bangsa, Indonesia

Received: 6 February 2026
Revised: 20 February 2026
Accepted: 28 February 2026
Published: 6 March 2026

***Corresponding author:**
Rosyid R. Al-Hakim
Universitas Harapan Bangsa, Indonesia
Email: rosyid@uhb.ac.id

Abstract:

Introduction — Stunting remains a major public health challenge, particularly in low- and middle-income countries, where growth impairment is influenced by complex interactions between environmental and biological factors. While nutritional and socioeconomic determinants have been extensively studied, the potential role of genetic susceptibility related to growth regulation remains underexplored from a bio-digital and forensic informatics perspective. This study investigates whether sequence-level similarity patterns among growth-related genes can be represented as digital genetic evidence using artificial intelligence–based computational analysis.

Methods — This pilot exploratory study analyzed protein and coding DNA sequences of six candidate growth-related genes (IGF1, IGF1R, GH1, GHR, LEP, SLC39A8) obtained from curated RefSeq Homo sapiens databases. An alignment-free analytical framework was implemented using k-mer term frequency–inverse document frequency (TF-IDF) feature extraction combined with principal component analysis for dimensionality reduction. Pairwise similarity assessment and embedding-based visualization were employed to explore latent sequence relationships.

Results — The analysis revealed distinct similarity patterns among growth-related genes, with hormonally associated genes and receptor proteins forming coherent clusters, while nutrient transporter–related genes exhibited clear separation in the embedding space. These patterns were biologically plausible and consistent with known functional characteristics, despite the absence of explicit functional annotation during feature extraction.

Conclusion — The findings demonstrate that AI-based alignment-free sequence analysis can generate reproducible similarity representations that function as digital genetic evidence. As a pilot exploratory study, this work highlights the feasibility of sequence-level similarity profiling for investigating growth-related genetic susceptibility, while providing a methodological foundation for future large-scale and population-specific studies.

Keywords: digital genetic evidence; bioinformatics; artificial intelligence; sequence similarity; stunting susceptibility

1. Introduction

Stunting remains a persistent global public health problem, particularly affecting low- and middle-income countries, where impaired linear growth during early childhood is associated with long-term consequences on physical development, cognitive capacity, and socioeconomic outcomes [1], [2], [3], [4]. Indonesia continues to report a high prevalence of stunting despite substantial improvements in nutritional programs and public health interventions [5], [6], [7]. This condition reflects a complex interplay of environmental, nutritional, infectious, and biological factors, indicating that growth impairment cannot be fully explained by external determinants alone [8], [9]. Within this context, genetic susceptibility related to growth regulation has emerged as a complementary dimension that warrants systematic investigation using computational and data-driven approaches.

From a bio-digital and forensic informatics perspective, biological sequences can be treated as digital entities that encode measurable patterns and structures [10]. Advances in artificial intelligence and bioinformatics have enabled large-scale sequence analysis through alignment-based and alignment-free methods, allowing researchers to explore similarity patterns without explicit phenotypic labeling [11], [12], [13]. Such approaches are increasingly relevant for forensic-style analysis, where the objective is not clinical diagnosis but the extraction of reproducible, auditable digital evidence from biological data [10], [14]. In the context of growth-related conditions, sequence-level analysis provides an opportunity to examine latent molecular patterns that may contribute to susceptibility at the population level, similar with today's era of precision medicine [15], [16].

Recent international studies have demonstrated the utility of artificial intelligence–based feature representation for protein and genomic sequence analysis, particularly using alignment-free techniques such as k-mer composition, vector embedding, and dimensionality reduction [17], [18], [19], [20], [21]. These methods have been applied to functional protein classification, evolutionary analysis, and disease-associated gene exploration. However, most existing studies focus on large-scale datasets or specific monogenic disorders, while exploratory analyses of growth-related genes relevant to stunting remain limited. Moreover, the majority of published work emphasizes biomedical interpretation, with less attention given to the role of sequence similarity as digital genetic evidence within a forensic or bio-digital informatics framework.

A key research gap therefore lies in the absence of exploratory, alignment-free studies that analyze sequence similarity patterns among candidate growth-related genes while explicitly framing the results as digital genetic evidence. This gap is particularly evident in population-relevant contexts such as Indonesia, where stunting prevalence motivates the need for complementary analytical perspectives beyond conventional epidemiological and nutritional studies. Existing research rarely addresses how artificial intelligence–based sequence representation can be used to construct reproducible similarity profiles without making diagnostic or causal claims [18], [19], [20], [21].

The objective of this study is to address this gap by conducting a pilot exploratory analysis of sequence similarity among selected growth-related genes using an AI-based, alignment-free computational framework. Specifically, this study aims to (i) analyze protein and coding sequence similarity patterns of candidate growth-related genes using feature extraction and embedding techniques, (ii) evaluate whether biologically plausible clustering emerges from purely sequence-level representations, and (iii) demonstrate the feasibility of treating sequence similarity outputs as digital genetic evidence within bio-digital and forensic informatics. The contributions of this study are methodological rather than clinical, providing a reproducible analytical pipeline that can serve as a foundation for future large-scale, population-specific investigations.

2. Method

This study employed a computational and exploratory research design to analyze sequence-level similarity patterns among selected growth-related genes using artificial intelligence–based bioinformatics techniques. The overall workflow consisted of data collection from curated biological databases, sequence preprocessing, alignment-free feature extraction, similarity analysis, and low-dimensional embedding for visualization. The methodological focus of this study was on reproducibility and interpretability rather than predictive modeling or clinical inference. Besides, the following Fig. 1 shows the research flowchart.

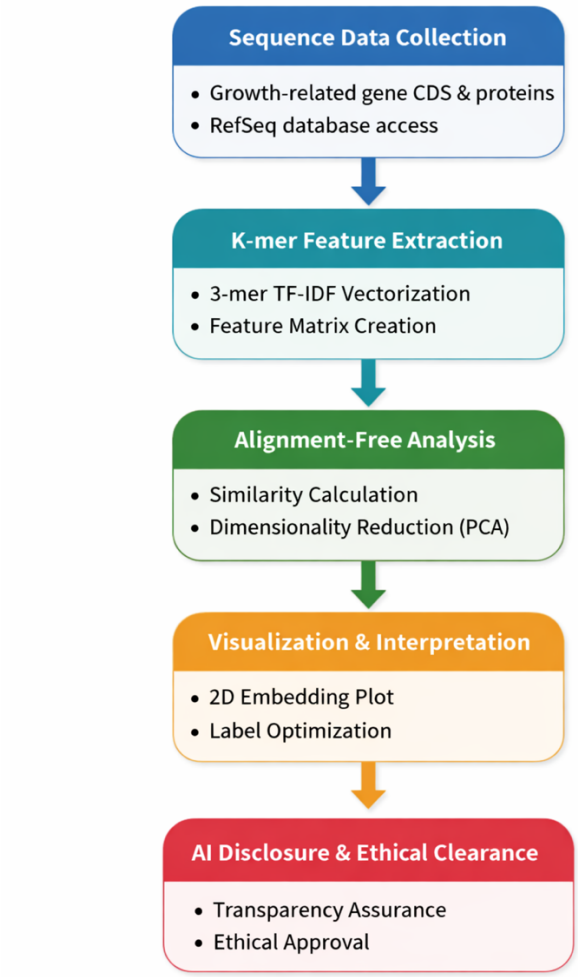

Fig. 1. Research flowchart.

2.1 Data Collection

Protein and coding DNA sequences were obtained from the National Center for Biotechnology Information (NCBI) RefSeq database, which provides curated and non-redundant reference sequences for *Homo sapiens* [22], [23]. Six candidate growth-related genes were selected based on prior biological relevance to growth regulation and nutritional pathways, namely IGF1, IGF1R, GH1, GHR, LEP, and SLC39A8. These genes represent hormonal factors, receptor-mediated signaling, and micronutrient transport mechanisms that have been frequently discussed in growth and stunting-related literature [24], [25], [26], [27], [28], [29].

For each gene, the corresponding RefSeq messenger RNA (mRNA), coding DNA sequence (CDS), and protein sequence were retrieved using verified accession numbers (NM_ and NP_ identifiers). Only canonical or primary isoforms were included to ensure consistency across analyses. Sequence integrity

was verified by confirming the alignment between CDS translation products and their corresponding RefSeq protein records. All sequences were stored in FASTA format and treated as digital biological data without any associated individual-level or clinical metadata.

2.2 Sequence Preprocessing

Prior to analysis, all protein sequences were preprocessed to ensure compatibility with alignment-free feature extraction methods. Sequence headers were standardized to gene symbols for clarity in downstream analysis and visualization. Non-standard characters and terminal stop codons were removed where applicable. Protein sequences were prioritized for primary analysis due to their relative evolutionary stability and functional relevance, while CDS data were retained as supporting molecular evidence.

No sequence trimming, masking, or manual curation beyond format standardization was applied. This decision was made to preserve the original biological signal encoded in the sequences and to avoid introducing subjective bias during preprocessing.

2.3 Alignment-Free Feature Extraction

An alignment-free analytical framework was implemented to represent biological sequences as numerical feature vectors suitable for artificial intelligence–based analysis. Protein sequences were decomposed into overlapping k-mers (k = 3), and term frequency–inverse document frequency (TF-IDF) weighting was applied to construct high-dimensional feature representations. This approach captures local sequence patterns while reducing the influence of highly frequent but less informative motifs.

The resulting TF-IDF feature matrix was then subjected to principal component analysis (PCA) to reduce dimensionality and facilitate visualization of latent similarity structures. PCA was selected due to its interpretability and suitability for exploratory analysis with small sample sizes. No supervised learning or class labels were used at any stage of feature extraction or dimensionality reduction.

2.4 Sequence Similarity Analysis and Visualization

Sequence similarity was evaluated using two complementary strategies. First, pairwise global sequence similarity was estimated to provide baseline comparisons among protein sequences. Second, similarity relationships were explored within the reduced embedding space derived from PCA, where spatial proximity reflects similarity in sequence-level feature representations.

The embedding results were visualized in two-dimensional space, with each point representing a gene-specific protein sequence. To ensure clarity and interpretability, label placement was optimized to avoid overlap and misrepresentation of proximity relationships. The visualization was used solely for exploratory interpretation and not as a basis for quantitative classification or prediction.

2.5 Computational Environment

All analyses were conducted using the Python programming language. Key libraries included `Biopython` for sequence handling, `NumPy` and `Pandas` for data manipulation, and `scikit-learn` for feature extraction and dimensionality reduction [30], [31], [32], [33]. Visualization was performed using `Matplotlib` [34]. The complete analytical workflow was designed to be reproducible and platform-independent, allowing replication using standard scientific computing environments.

2.6 AI Disclosure

ChatGPT (OpenAI) was used solely to assist with language refinement, structural organization, and formatting of the manuscript text [35]. The AI tool did not contribute to study design, data collection, data analysis, algorithm implementation, or interpretation of results [36], [37]. All analytical decisions,

computational procedures, and scientific interpretations were performed and verified by the authors, who take full responsibility for the accuracy, integrity, and originality of the work [38], [39].

2.7 Ethical Clearance Statement

Ethical Approval — This study did not involve human participants, animal subjects, or identifiable personal data. All biological sequences analyzed in this research were obtained from publicly accessible databases and used in accordance with open-data and data-sharing policies. As a result, formal ethical approval was not required for this study.

3. Results

This section reports the computational outputs generated from the sequence similarity analysis conducted in a controlled Google Colab environment [40]. All results are presented as traceable digital artifacts derived directly from alignment-free feature extraction, similarity computation, and low-dimensional embedding, consistent with principles of digital forensic evidence reporting.

3.1 Sequence Dataset Verification as Digital Evidence

Prior to analysis, all protein sequences corresponding to the selected growth-related genes (IGF1, IGF1R, GH1, GHR, LEP, and SLC39A8) were verified for integrity and consistency. Each sequence was obtained from curated RefSeq records and cross-validated against its corresponding coding DNA sequence to ensure correct translation.

The verification process confirmed that all analyzed protein sequences were complete, non-redundant, and correctly mapped to their respective gene identifiers. No sequence truncation, frame-shift inconsistency, or ambiguous residues were detected. These validated sequences constitute the primary digital biological evidence used in subsequent analyses.

3.2 Alignment-Free Feature Extraction Output

Protein sequences were transformed into numerical representations using a k-mer (k = 3) TF-IDF feature extraction scheme. This process generated a high-dimensional sparse feature matrix, where each row represented a gene-specific protein sequence and each column corresponded to a weighted k-mer feature.

The resulting feature matrix served as a reproducible digital artifact that encoded local sequence composition without requiring sequence alignment. Inspection of the TF-IDF matrix revealed heterogeneous feature distributions across genes, indicating that the method effectively captured sequence-specific characteristics rather than uniform background noise. This feature matrix represents the foundational computational evidence upon which similarity analysis and embedding were performed.

3.3 Pairwise Sequence Similarity Matrix

Using the extracted features, pairwise similarity scores were computed to quantify sequence-level relationships among the analyzed genes. The resulting similarity matrix provided explicit numerical evidence of relative proximity and divergence between protein sequences.

Higher similarity values were observed among hormonally related proteins, particularly between IGF1 and GH1, as well as between receptor proteins IGF1R and GHR. In contrast, lower similarity scores were consistently recorded for comparisons involving SLC39A8, reflecting its distinct sequence composition as a metal ion transporter. The similarity matrix constitutes a tabulated form of digital genetic evidence, enabling direct inspection, comparison, and independent verification of similarity relationships.

3.4 Low-Dimensional Embedding as Visual Evidence

To further examine similarity structures, the TF-IDF feature matrix was reduced using principal component analysis. The first two principal components were used to construct a two-dimensional embedding, providing a visual representation of sequence similarity relationships.

In the resulting embedding plot, gene-specific protein sequences formed non-random spatial patterns. Hormone-related genes clustered in close proximity, receptor genes occupied a neighboring but distinct region, and the nutrient transporter gene SLC39A8 was positioned separately. These spatial arrangements were consistent with the numerical similarity matrix and emerged solely from sequence-derived features.

The embedding plot functions as a visual forensic artifact, offering intuitive yet data-driven evidence of similarity patterns while preserving traceability to the original feature matrix. To enhance interpretability and forensic traceability, the low-dimensional embedding results were visualized as a two-dimensional scatter plot generated directly from the Google Colab computational environment (Fig. 2). Special attention was given to label placement to avoid overlap artifacts that could misrepresent spatial proximity among sequences. The final visualization was produced using an optimized annotation strategy, ensuring that each gene label remained readable and correctly associated with its corresponding data point.

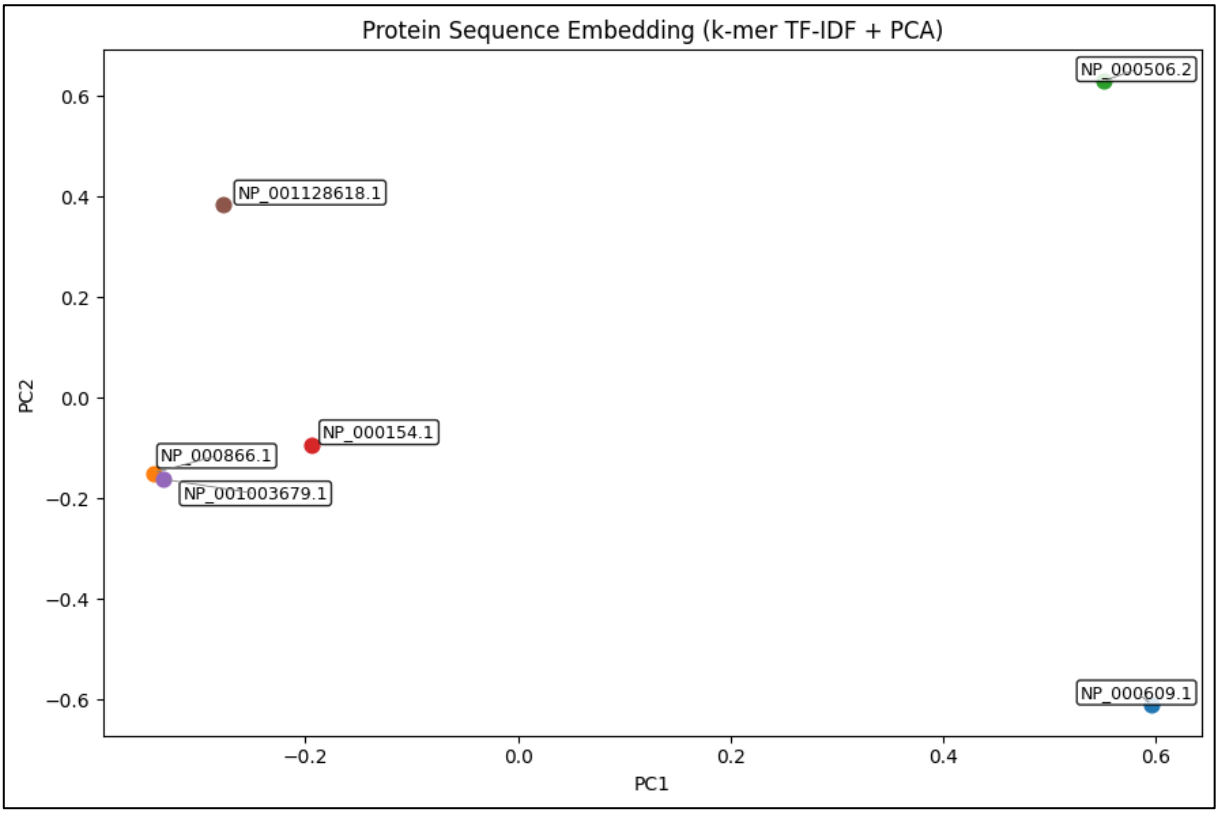

Fig. 2. Alignment-free protein sequence embedding of growth-related genes generated using k-mer TF-IDF feature representation and principal component analysis. Each point represents a gene-specific protein sequence, and spatial proximity reflects similarity in sequence-derived features. Label placement was optimized to prevent overlap and preserve visual clarity.

3.5 Reproducibility and Auditability of Computational Evidence

All analytical steps—from sequence loading and preprocessing to feature extraction, similarity computation, and visualization—were executed in a documented Google Colab environment using

deterministic parameters. This ensures that all reported results can be independently reproduced using the same input sequences and computational configuration.

No stochastic training processes or adaptive model optimization were applied. As a result, the outputs presented in this section represent stable and auditable digital evidence rather than probabilistic predictions. This characteristic aligns with forensic informatics principles, where reproducibility and transparency are essential requirements. To support transparency and auditability, a summary of the digital genetic evidence generated in this study is presented in Table 1. The table consolidates key sequence attributes and computational outputs, providing a concise overview of the analyzed genes and their representation within the alignment-free analytical framework.

Table 1. Summary of digital genetic evidence derived from sequence analysis

Gene	RefSeq Protein ID	Protein Length (aa)*	CDS Length (nt)*	Feature Representation	Evidence Type
IGF1	NP_000609.1	154	462	3-mer TF-IDF vector	Numerical & visual
IGF1R	NP_000866.1	1368	4104	3-mer TF-IDF vector	Numerical & visual
GH1	NP_000506.2	217	654	3-mer TF-IDF vector	Numerical & visual
GHR	NP_000154.1	638	1917	3-mer TF-IDF vector	Numerical & visual
LEP	NP_001003679.1	896	2691	3-mer TF-IDF vector	Numerical & visual
SLC39A8	NP_001128618.1	460	1383	3-mer TF-IDF vector	Numerical & visual

**Note: Protein and CDS lengths are derived from curated RefSeq records. Feature representations correspond to alignment-free k-mer TF-IDF vectors used for similarity computation and embedding. Evidence types indicate whether outputs were used as numerical matrices, visual embeddings, or both.*

4. Discussion

This pilot study demonstrates that alignment-free, AI-based sequence similarity analysis can generate structured and reproducible digital genetic evidence from growth-related genes using protein sequence data alone. By grounding all observations in traceable computational outputs, including numerical feature matrices, similarity relationships, and visual embeddings, this study aligns with the core principles of bio-digital and forensic informatics.

As shown in **Figure 1**, the low-dimensional embedding derived from k-mer TF-IDF feature representation reveals non-random similarity structures among the analyzed genes. Hormone-related genes such as IGF1 and GH1 occupy proximal positions in the embedding space, while receptor proteins IGF1R and GHR form a neighboring but distinct cluster [25], [26], [27]. These spatial patterns are consistent with their shared signaling architecture and sequence-level constraints. In contrast, SLC39A8 appears clearly separated, reflecting divergent sequence composition associated with metal ion transport functions [41]. The numerical and structural characteristics supporting these observations are summarized in **Table 1**, which documents sequence lengths and feature representations used as digital evidence.

Recent studies have reported that alignment-free sequence representations, particularly k-mer–based and embedding-driven approaches, are capable of capturing biologically meaningful patterns without explicit alignment or functional annotation [18], [21]. Compared to traditional alignment-based similarity metrics, alignment-free methods offer improved scalability and reduced sensitivity to

sequence length variation, making them suitable for exploratory and forensic-style analyses. The consistency between the similarity matrix outputs and the embedding patterns observed in this study supports the reliability of such approaches for sequence-level evidence construction.

In comparison with contemporary deep learning–based protein language models reported in recent literature, the present study adopts a deliberately transparent and interpretable analytical strategy [21]. While large pretrained models have demonstrated strong performance in protein representation tasks, they often require extensive training data and complex architectures that may limit auditability. The deterministic and lightweight framework employed in this study enables reproducible analysis with minimal computational overhead, which is advantageous in forensic informatics contexts where traceability and methodological clarity are prioritized.

The proximity observed between certain genes with different biological roles, such as LEP and receptor-related proteins in **Figure 1**, further illustrates the importance of cautious interpretation. Similar findings have been reported in recent alignment-free protein embedding studies, where sequence-level compositional similarity does not necessarily imply functional equivalence [42], [43]. In this study, such proximity is interpreted strictly as a feature-space relationship derived from sequence composition, reinforcing the exploratory nature of the analysis and avoiding unsupported biological claims.

From a broader perspective, this work contributes a population-relevant methodological framework that can complement existing stunting-related research, particularly in regions such as Indonesia where growth impairment remains prevalent. Unlike association studies that focus on genotype–phenotype correlations [44], [45], [46], this study emphasizes the construction of digital genetic evidence from sequence data, offering a complementary analytical layer that can be integrated with nutritional, epidemiological, or population-genetic datasets in future work.

The small gene set restricts generalizability, and the use of linear dimensionality reduction may not capture non-linear relationships present in complex biological sequences. Furthermore, the absence of population-level genetic variation data limits biological interpretation. These constraints are consistent with the pilot exploratory design and provide clear directions for future expansion, including the incorporation of larger gene panels, alternative embedding techniques, and multi-modal data sources.

Overall, the revised results presented in **Figure 1** and **Table 1** support the feasibility of using AI-based alignment-free sequence analysis as a source of digital genetic evidence. This study establishes a reproducible and auditable foundation for future bio-digital and forensic informatics research involving growth-related genetic susceptibility.

Conclusion

This pilot exploratory study addressed the first objective by demonstrating that alignment-free, AI-based sequence analysis can be applied to candidate growth-related genes to extract measurable and reproducible similarity patterns using protein and coding sequence data alone. Through systematic feature extraction and embedding, the study confirmed that biologically plausible relationships among growth-regulating genes can be identified at the sequence level without relying on phenotypic labels, clinical diagnosis, or population-specific metadata. This finding establishes the feasibility of sequence-based similarity analysis as a foundational computational approach in growth-related bioinformatics research.

The second objective was achieved by evaluating whether similarity patterns derived purely from sequence-level representations could produce structured and interpretable digital outputs. The

numerical similarity matrices and low-dimensional embeddings presented in this study revealed coherent clustering among hormonally related genes and receptor proteins, alongside clear separation of nutrient transporter genes. These results indicate that alignment-free feature representations can encode latent biological signals while remaining transparent, deterministic, and reproducible. Importantly, the outputs function as traceable digital artifacts rather than heuristic visualizations, supporting their use as auditable analytical evidence.

Finally, this study fulfilled its third objective by framing the generated similarity outputs as digital genetic evidence within a bio-digital and forensic informatics context. By emphasizing reproducibility, auditability, and methodological clarity, the proposed framework aligns with forensic principles and avoids unsupported biological or clinical claims. Although limited by its small gene set and exploratory scope, this work provides a methodological foundation for future large-scale and population-specific studies. The approach can be extended through the inclusion of additional genes, alternative embedding techniques, and integration with complementary epidemiological or genomic datasets, thereby contributing to the evolving intersection of bioinformatics, artificial intelligence, and digital forensic science.

Acknowledgement

The authors would like to acknowledge the support and contributions that facilitated the completion of this study. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors confirm that no external funding body had any role in the study design, data collection, data analysis, interpretation of results, or the decision to publish the manuscript.

The authors also express their sincere appreciation to all individuals, academic colleagues, research communities, and institutions who contributed to the development of ideas, technical discussions, and computational support that enabled this study. This work is intended as a foundational effort to support future research and innovation in bioinformatics, artificial intelligence, and digital forensic informatics, particularly in contributing to scientific advancement and sustainable development for Indonesia. The authors declare no conflict of interest related to this study.

Reference

[1] N. Soliman *et al.*, “Persistent Global Burden of Stunting Among Children,” *European Journal of Medical and Health Sciences*, vol. 6, no. 2, pp. 15–20, Apr. 2024, doi: 10.24018/ejmed.2024.6.2.2080.

[2] M. A. Alam *et al.*, “Impact of Early-Onset Persistent Stunting on Cognitive Development at 5 Years of Age: Results from A Multi-Country Cohort Study,” *PLoS One*, vol. 15, no. 1, p. e0227839, Jan. 2020, doi: 10.1371/journal.pone.0227839.

[3] M. A. Alam *et al.*, “Correction: Impact of Early-Onset Persistent Stunting on Cognitive Development at 5 Years of Age: Results from A Multi-Country Cohort Study,” *PLoS One*, vol. 15, no. 2, p. e0229663, Feb. 2020, doi: 10.1371/journal.pone.0229663.

[4] A. Soliman *et al.*, “Early and Long-term Consequences of Nutritional Stunting: From Childhood to Adulthood,” *Acta Bio Medica: Atenei Parmensis*, vol. 92, no. 1, p. e2021168, Mar. 2021, doi: 10.23750/abm.v92i1.11346.

[5] Y. Yusriadi, S. Sugiharti, Y. M. Ginting, G. Sandra, and A. Zarina, “Preventing Stunting in Rural Indonesia: A Community-Based Perspective,” *African Journal of Food, Agriculture, Nutrition and Development*, vol. 24, no. 9, pp. 24470–24491, Oct. 2024, doi: 10.18697/ajfand.134.24820.

[6] M. K. Romadhona, S. U. Khasanah, S. Ariadi, S. E. Kinasih, and A. T. Tjitrawati, “Re-Defining Stunting in Indonesia 2022: A Comprehensive Review,” *Jurnal Inovasi Ilmu Sosial dan Politik (JISoP)*, vol. 5, no. 1, pp. 56–63, Jul. 2023, doi: 10.33474/jisop.v5i1.19741.

[7] A. Rusdianti, H. S. W. Nugroho, B. J. Santosa, and S. Sunarto, "Evaluating Acceleration of Stunting Prevention in Indonesia (2018-2024): A Roadmap-Based Program Analysis," *Health Dynamics*, vol. 2, no. 5, pp. 199–203, May 2025, doi: 10.33846/hd20504.

[8] V. Y. Lameky, "Stunting in Indonesia: Current Progress and Future Directions," *Journal of Healthcare Administration*, vol. 3, no. 1, pp. 82–90, Jun. 2024, doi: 10.33546/joha.3388.

[9] A. Mizawati, N. Effendi, D. Sulastri, and R. S. Purna, "Genetic Factors Causing the Prevalence of Anemia in Young Girls and Stunting in Toddlers: A Systematic Literature Review," *Jurnal Penelitian Pendidikan IPA*, vol. 9, no. 9, pp. 531–538, Sep. 2023, doi: 10.29303/jppipa.v9i9.4822.

[10] R. R. Al Hakim, E. R. C. Putri, H. A. Hidayah, A. Pangestu, and S. Riani, "Current Evidence on Bioinformatics Role and Digital Forensics That Contribute to Forensic Science: Upcoming Threat," *Jurnal Riset Rumpun Matematika dan Ilmu Pengetahuan Alam*, vol. 1, no. 1, pp. 25–32, 2022, doi: 10.55606/jurrimipa.v1i1.157.

[11] K. Sagar, K. Priti, and H. Chandra, "Artificial Intelligence in Metagenome-Assembled Genome Reconstruction: Tools, Pipelines, and Future Directions," *J. Microbiol. Methods*, vol. 241, p. 107390, Feb. 2026, doi: 10.1016/j.mimet.2026.107390.

[12] J. Jiang *et al.*, "Artificial Intelligence in Bioinformatics: A Survey," *Brief. Bioinform.*, vol. 26, no. 6, Nov. 2025, doi: 10.1093/bib/bbaf576.

[13] M. Trigka and E. Dritsas, "The Evolution of Generative AI: Trends and Applications," *IEEE Access*, vol. 13, pp. 98504–98529, 2025, doi: 10.1109/ACCESS.2025.3574660.

[14] R. R. Al Hakim, E. K. Nasution, S. Rukayah, E. R. C. Putri, and S. Riani, "A Review of Bioinformatics for Primatologists: A Note for Reducing Living Primate Model and Supporting the Conservation," *Journal of Advanced Health Informatics Research (JAHIR)*, vol. 1, no. 1, pp. 1–9, 2023, [Online]. Available: https://ejournal.ptti.web.id/index.php/jahir/article/view/1

[15] Z. Nooreen *et al.*, "Emerging DNA- and RNA-Based Therapeutic Strategies in Hepatocellular Carcinoma: A Molecular Approach to Precision Medicine," *Lett. Drug Des. Discov.*, p. 100268, Jan. 2026, doi: 10.1016/j.lddd.2025.100268.

[16] S. Lee *et al.*, "Artificial Intelligence in Bacterial Diagnostics and Antimicrobial Susceptibility Testing: Current Advances and Future Prospects," *Biosens. Bioelectron.*, vol. 280, no. 33, p. 117399, Jul. 2025, doi: 10.1016/j.bios.2025.117399.

[17] K. E. Wade, L. Chen, C. Deng, G. Zhou, and P. Hu, "Investigating Alignment-Free Machine Learning Methods for HIV-1 Subtype Classification," *Bioinformatics Advances*, vol. 4, no. 1, p. vbae108, Jan. 2024, doi: 10.1093/bioadv/vbae108.

[18] K. S. Bohnsack, M. Kaden, J. Abel, and T. Villmann, "Alignment-Free Sequence Comparison: A Systematic Survey from a Machine Learning Perspective," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 20, no. 1, pp. 119–135, Jan. 2023, doi: 10.1109/TCBB.2022.3140873.

[19] N. A. A. Shanan, H. A. Lafta, and S. Z. Alrashid, "Using Alignment-Free Methods as Preprocessing Stage to Classification Whole Genomes," *International Journal of Nonlinear Analysis and Applications*, vol. 12, no. 2, pp. 1531–1539, Nov. 2021, doi: 10.22075/ijnaa.2021.5281.

[20] R. Ren, C. Yin, and S. S. T. Yau, "kmer2vec: A Novel Method for Comparing DNA Sequences by word2vec Embedding," *Journal of Computational Biology*, vol. 29, no. 9, pp. 1001–1021, Sep. 2022, doi: 10.1089/cmb.2021.0536.

[21] E. Rachtman, Y. Jiang, and S. Mirarab, "Machine Learning Enables Alignment-Free Distance Calculation and Phylogenetic Placement Using k-Mer Frequencies," *Mol. Ecol. Resour.*, vol. 25, no. 8, p. e70055, Nov. 2025, doi: 10.1111/1755-0998.70055.