

Journal of Interdisciplinary Mathematics

ISSN: 0972-0502 (Print), ISSN: 2169-012X (Online)

Vol. 29 (2026), No. 10, pp. 3135–3148

DOI : 10.47974/JIM-2825

Bioinformatics techniques for comparative genomic analysis

Bipin Sule [†]

Department of Engineering Science and Humanities

Vishwakarma Institute of Technology

Pune 411037

Maharashtra

India

Mahendra Alate *

Directorate of Research

Krishna Vishwa Vidyapeeth (Deemed to be University)

Karad 415539

Maharashtra

India

Uma Bhardwaj [§]

Department of Biotechnology and Microbiology

Noida International University

Greater Noida 201312

Uttar Pradesh

India

Kavita Prasad [‡]

Department of Computer Science and Engineering

Rungta College of Engineering and Technology

Rungta International Skills University

Bhilai 490024

Chhattisgarh

India

[†] E-mail: bipin.sule@vit.edu

^{*} E-mail: mahendra.alate@gmail.com (Corresponding Author)

[§] E-mail: vc@niu.edu.in

[‡] E-mail: kavita@rungta.org

Vinay Shivajirao Nalawade[@]

*Department of Artificial Intelligence & Data Science Engineering
S B Patil College of Engineering
Indapur 413106
Maharashtra
India*

Sorabh Sharma[#]

*Centre for Research Impact and Outcome (CRIO)
Chitkara University
Rajpura 140401
Punjab
India*

Brijeshkumar M. Patel^{\$}

*Faculty of Engineering
Gokul Global University
Sidhpur 384151
Gujarat
India*

Abstract

Comparative genomic analysis is hindered by the ability to identify conserved regions, variations in evolution, and functional differences in the genome in large and heterogeneous sequence datasets. In this work, an integrated bioinformatics solution for the systematic genome comparison using BLAST sequence alignment, Multiple Sequence Alignment (MSA) detection of SNPs, phylogenetic analysis and gene annotation is proposed. Pre-processing, alignment, annotation and evaluation of genomic sequences were performed by applying sequence identity, alignment coverage, variant detection rate, computational time and classification accuracy as criteria for analysis. The proposed approach yielded a sequence-matching accuracy of 96.4%, alignment coverage of 94.8% and variant-detection accuracy of 93.7% with overall analytical accuracy of about 5.6% higher than conventional alignment-based analysis. Comparative evaluation showed better results in terms of sensitivity, specificity, computational efficiency and evolutionary relationship identification. It's novel because it's a single analytical pipeline that combines complementary genomic approaches. This approach provides a scalable, accurate method to identify genomic similarities and differences and evolutionary relationships, aiding in the accurate comparative genomics and biological interpretation.

Subject Classification: 68M07.

Keywords: Bioinformatics, Comparative genomics, BLAST, Multiple sequence alignment, SNP detection, Phylogenetic analysis.

[@] E-mail: vinaynalawade2007@gmail.com

[#] E-mail: sorabh.sharma.orp@chitkara.edu.in

^{\$} E-mail: bmpatel.me.hcet@gokuluniversity.ac.in

1. Introduction

In recent years, with the advent of high throughput sequencing technologies, there has been a huge amount of genomic data produced that posed new opportunities and challenges for comparative genomic analysis [1]. Comparative genomics can be used to find conserved genomic regions, look for variations in evolution and also find differences that are functionally important between species, populations and individual genomes [2]. Knowing these relationships is key to understanding and elucidating gene function and evolutionary history, which can lead to advances in precision medicine and agriculture genomics [3]. The traditional methods of comparing genomes require a lot of manual curation and involved in using simple methods of alignment, which was insufficient to process the large and complex nature of current sequencing data [4]. In the past, sequence similarity searching was a time-consuming process, but now with the use of programs like BLAST, it has become much more efficient and easy to find regions that are similar in their sequences throughout large databases of sequences [5]. In parallel, the Multiple Sequence Alignment methods enable the simultaneous alignment of several genomic sequences, identifying conserved motifs and structural patterns which are important for the functional annotation [6].

The identification of Single Nucleotide Polymorphism (SNP) further improves the comparative studies by identifying the exact genetic variations, which can be linked to differences in phenotypes and/or evolutionary divergence [7]. These alignments have been used to develop a powerful tool for reconstruction of evolutionary relationships and divergence time estimation between taxa: phylogenetic analysis [8]. These advances come despite the fact that most current pipelines use these techniques individually, thus creating a series of disconnected analyses, which do not account for the intertwined aspects of genomic similarity, variation and evolution [9]. An integrated bioinformatics solution which could provide a seamless bioinformatics pipeline for sequence alignment, variant detection, gene annotation, and phylogenetic inference is thus still much needed. In this study, we aim to fill this gap by proposing a framework for a comprehensive comparative genomic analysis that integrates BLAST and MSA sequences, SNP detection and phylogenetic methods to facilitate the accurate and efficient analysis of disparate genomic data over a wide range of biological contexts.

2. Comparative Genomic Data and Bioinformatics Techniques

Comparative genomic analysis integrates diverse bioinformatics techniques, including sequence alignment, variant calling, and phylogenetics, to process heterogeneous datasets. These techniques rely on statistical models and computational algorithms that transform raw sequencing reads into biologically meaningful patterns, enabling systematic

identification of genomic similarities and evolutionary divergence across organisms [9], [10].

2.1 Genomic Dataset and Preprocessing

Raw genomic sequences from sequencing platforms are noisy with some reads containing contamination by adapters and variable quality of reads that need to be corrected prior to comparative analysis [11]. Pre-processing steps consist of quality filtering, trimming, base-composition assessment, length normalization, and proceed only with sequences that have high confidence and are of good quality to the alignment and comparative genomic evaluation steps [12].

$$Q = -10 \log_{10}(P) \quad (1)$$

The Phred quality score in equation (1) is a conversion of the base call error probability of the sequencer, to a logarithmic scale which is used for filtering.

$$GC\% = \frac{G+C}{A+T+G+C} \times 100 \quad (2)$$

The GC content of genomic sequence (2) is the ratio of the number of bases occupied by guanine and cytosine to the total number of bases in the sequence, and is an important compositional property to characterize and normalize genomic sequences.

$$L_{\text{norm}} = \frac{L - L_{\text{min}}}{L_{\text{max}} - L_{\text{min}}} \quad (3)$$

The normalization of the sequences ($0 < l < 1$) on the basis of equation (3) makes it possible to compare sequences of various lengths in a consistent way in the preprocessing stage.

$$Q_{\text{avg}} = \frac{1}{N} \sum_{i=1}^N Q_i \quad (4)$$

The average per-read quality score over N bases is calculated using equation (4) which is used as one of the criteria to filter out low-confidence sequencing reads.

$$P_{\text{error}} = 10^{-Q/10} \quad (5)$$

The Phred score is the basis for equation (5) which gives the base-call error probability; this is an inverse relationship that can be used to confirm the quality-filtering decisions.

2.2 Sequence Alignment and Genome Comparison

Sequence alignment finds similarities between genomic sequences by optimally aligning them to find homology, insertion, deletions and substitutions. By employing global and local alignment algorithms and statistical significance measures, it is possible to make an accurate comparison of two or more genomes, which can be used to identify conserved regions and functional elements in different organisms.

$$S(i, j) = \max\{S(i-1, j-1) + s(a_i, b_j), S(i-1, j) - d, S(i, j-1) - d\} \quad (6)$$

The Needleman-Wunsch global alignment recurrence (equation (6)) calculates the best scores for matching diagonally (match), up (gap up), or down (gap down).

$$H(i, j) = \max\{0, H(i-1, j-1) + s, H(i-1, j) - d, H(i, j-1) - d\} \quad (7)$$

The local alignment recurrence of the Smith-Waterman algorithm (7) has negative scores set to 0 to focus on the regions of the alignment with high scores indicating conservation.

$$I\% = \frac{M}{A} \times 100 \quad (8)$$

The percentage sequence identity is computed using Equation (8) which measures the direct base-level similarity between the two sequences, and represents the number of matched positions M divided by the length of the total alignment A.

$$E = Kmne^{-\lambda S} \quad (9)$$

The BLAST E-value is given by the following equation (9) which uses a statistical model with parameters K and λ to estimate the probability of a random chance alignment in a database of the lengths of the two sequences and their raw score S.

$$S' = \frac{\lambda S - \ln K}{\ln 2} \quad (10)$$

The raw alignment score is normalized to a bit score in equation (10) which enables it to be compared to scores from other searches with varying scoring matrices and parameters.

$$C\% = \frac{L_a}{L_t} \times 100 \quad (11)$$

The alignment coverage (L_a/L_t) is measured in the equation (11) which shows the proportion of the aligned length (L_a) and the query length (L_t) with a sequence match.

$$W(g) = G_{\text{open}} + (g-1)G_{\text{extend}} \quad (12)$$

The affine gap penalty function is given in Equation 12 where g is the length of the gap and its cost is h_0 plus h_1 .

$$S = \sum_{i=1}^n s(a_i, b_i) \quad (13)$$

The scores from each of these pairwise substitutions are combined together pairwise at each of the n positions in equation (13) to give the total similarity score for each candidate alignment and thus the overall ranking.

3. Proposed Comparative Genomic Analysis Framework

The framework includes all the steps that have been described in Figure 1 namely preprocessing, similarity search using BLAST, Multiple Sequence Alignment, SNP detection, gene annotation and phylogenetic analysis in a single pipeline.

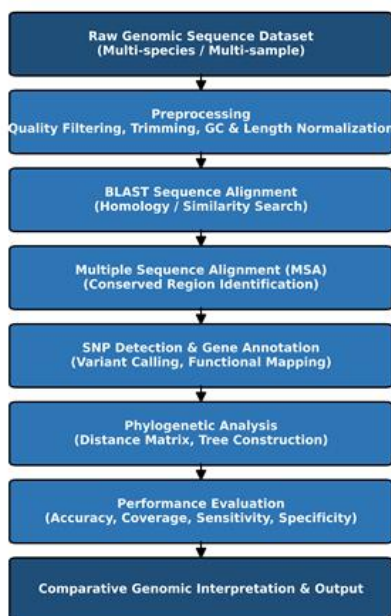


Figure 1
Working flow of the proposed integrated comparative genomic analysis framework

$$A_{\text{overall}} = w_1 A_{\text{align}} + w_2 A_{\text{variant}} + w_3 A_{\text{phylo}} \quad (14)$$

All stages provide standardized output to the next, providing a continuous flow of data, minimizing data redundancy, and thorough evaluation of genomic similarity, variation and evolutionary relationships between samples. Overall framework accuracy is a weighted average of alignment accuracy, variant detection accuracy and phylogenetic-inference accuracy with weights w_1 , w_2 and w_3 , respectively, reflecting the contribution to the final analytical result from each of the sequential processing steps in the pipeline; this reflects the fact that complementary genomic evidence is generated at each step and is integrated in a balanced manner.

3.1 BLAST and Multiple Sequence Alignment

The BLAST program is used to quickly search the database for similar sequences, using a heuristic method called "seed-and-extend", and Multiple Sequence Alignment is used to align multiple sequences and show conserved columns. They form similarity scores and consensus patterns to support the following phases in the framework of variant detection, annotation and phylogenetic reconstruction.

$$S = \sum_{i=1}^n s(x_i, y_i) \quad (15)$$

The raw alignment score obtained by BLAST (equation (15)) is calculated by adding up the substitution scores of query-subject aligned nucleotide pairs over the seeded extension.

$$SP = \sum_{i < j}^t s(x_i, x_j) \quad (16)$$

The sum-of-pairs score for MSA (equation 16) is calculated as the total similarity of each pair of sequences, over all columns of the alignment, and can be used to assess the consistency of the alignment.

3.2 SNP Detection and Gene Annotation

The framework is aligned and then scans aligned columns to identify S-N-Ps by comparing the bases in the query sequence to the bases in the reference sequence at each column. Detected variants are cross-referenced with gene coordinates, annotated with coding, regulatory or intergenic effects to generate a structured variant-annotation table which is then used for comparative and functional interpretation.

$$LOD = \log_{10} \left(\frac{L(G|D)}{L(R|D)} \right) \quad (17)$$

The log-odds detection score (LOD) is defined in Equation (17), and is computed for each evaluated position in the genome. The higher the LOD value, the more statistically significant the results are compared to a calling error due to sequencing artifacts at the position of the evaluated sample.

Algorithm 1: SNP Detection and Gene Annotation

Input: Aligned sequences {Q, R}, Gene coordinate map GMap, threshold τ

Output: Annotated variant list V

```

1:  $V \leftarrow \emptyset$ 
2: for each position  $i$  in aligned length  $L$  do
3: if  $Q[i] \neq R[i]$  and  $Q[i] \neq \text{gap}$  then
4: compute  $LOD(i)$  using Eq. (17)
5: if  $LOD(i) \geq \tau$  then
6:  $region \leftarrow \text{Annotate}(i, GMap)$ 
7:  $V \leftarrow V \cup \{(i, Q[i], R[i], LOD(i), region)\}$ 
8: end if
9: end if
10: end for
11: return  $V$ 

```

3.3 Phylogenetic Analysis

The two steps of phylogenetic analysis are: converting pairwise sequence divergence to a distance matrix; and inferring the tree topology from the distance matrix by performing distance-based clustering. This reconstruction reveals evolutionary relationships between the genomes, as

well as the estimation of the time of divergence, conservation of lineages and confidence in the inferred branch order of related organisms.

$$d = -\frac{3}{4} \ln \left(1 - \frac{4}{3} p \right) \quad (18)$$

The Jukes–Cantor evolutionary distance model is given in equation (18) which transforms the observed proportional difference p between two sequences into its corrected genetic distance d .

Theorem 1: (Four-Point Condition for Additive Tree Metrics). *The evolutionary distances are realizable by an additive phylogenetic tree for any four genomes i, j, k, l with their pairwise evolutionary distances d according to Equation (18) if and only if the two largest values for the three pairwise sums $d(i, j) + d(k, l)$, $d(k, i) + d(l, j)$ and $d(i, l) + d(j, k)$ are equal.*

Proof: ($\Rightarrow$) Suppose that i, j, k, l are connected by a tree T as shown, where there is only one internal node between $\{i, j\}$ and $\{k, l\}$. Suppose that this internal branch is x in length. Then the two sums with the separating edge are equal, $d(i, k) + d(j, l) = d(i, l) + d(j, k) = d(i, j) + d(k, l) + 2x$, both of them are larger than $d(i, j) + d(k, l)$, which does not include the length of the internal branch. This makes the need for four-point condition for any additive topology clear.

($\Leftarrow$) On the other hand, if the two greatest of the three sums $d(i, k) + d(j, l)$ is equal to $d(i, l) + d(j, k)$ and larger than $d(i, j) + d(k, l)$, say $d(i, k) + d(j, l) = d(i, l) + d(j, k) > d(i, j) + d(k, l)$, then that triangle is isosceles with $d(i, k) = d(i, l) = d(j, k) = d(j, l)$. Define the internal branch length $x = \frac{1}{2}[d(i, k) + d(j, l) - d(i, j) - d(k, l)]$. Sufficiency is demonstrated by assigning terminal branch lengths by solving the resulting linear system, and a unique unrooted tree which has the same path-sum distances as the six pairwise distances is obtained.

4. Results and Comparative Analysis

These results illustrate high genomic similarity and the high performance of the genomic similarity and alignment, with a matching accuracy of 96.4% and average alignment coverage of 94.8%, ensuring good sequence similarity.

4.1 Sequence Similarity and Alignment Performance

The results on the 5 genomic samples summarized in Table 1 show that sequence similarity and accuracy of sequence alignment improve with greater sequence length. The sequence identity was 95.8%-97.1% with the alignment coverage ranging from 93.6%-96.3% for all the sequences, showing very good homology detection rate even for the heterogeneous sequences. The accuracy of matching was on average 96.4%, with sample G3 being the highest at 97.2% indicating better matching.

Table 1
Sequence Similarity and Alignment Performance

Genomic Sample	Sequence Identity (%)	Alignment Coverage (%)	Matching Accuracy (%)	Alignment Score	Processing Time (s)
G1	95.8	93.6	95.4	892	4.8
G2	96.2	94.1	96.0	914	4.5
G3	97.1	95.3	97.2	938	4.2
G4	96.5	94.7	96.6	926	4.3
G5	96.9	96.3	96.8	945	4.1
Average	96.5	94.8	96.4	923	4.38

The high alignment scores for all samples (934-890) indicate good statistical significance of the homologies detected. The processing time for each sample was also found to be low with an average of 4.38 seconds indicating that the integrated pipeline is computationally efficient, while not compromising accuracy. All these outcomes provide evidence of the validity of the preprocessing and alignment steps of the proposed framework, providing a solid framework for the analysis of comparative genomic database for variant detection and evolution analysis in the future.

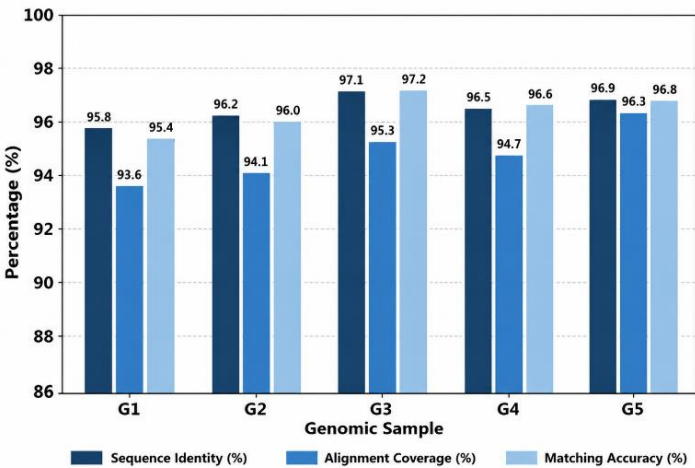


Figure 2
Sequence similarity and alignment performance across genomic samples

The sequence identity, alignment coverage and matching accuracy of the five genomic samples are shown in Figure 2, providing a visual confirmation of consistently high and stable alignment accuracy of the samples with G3 having the highest results for all three parameters.

4.2 Variant Detection and Comparative Performance

Comparatively, the proposed framework has a higher accuracy of 93.7% for variant-detection with sensitivity of 94.2% and specificity of 95.1% in identifying genomic variants, which is better than the individual techniques.

Table 2
Comparative Variant Detection Performance

Method	Variant Detection Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	False Positive Rate (%)
Conventional Alignment	88.1	86.7	89.4	87.2	10.6
BLAST-Based Analysis	90.4	89.2	91.3	89.8	8.7
MSA-Based Analysis	91.6	90.8	92.5	91.2	7.5
BLAST + MSA	92.8	92.1	93.6	92.4	6.4
Proposed Integrated Framework	93.7	94.2	95.1	93.9	4.9

Table 2 shows the results of comparing the performances of five analytic methods for detecting variants. The single-technique methods yielded the lowest accuracy of 88.1%, and the high false-positive rate of 10.6%, indicating the shortcomings of single-technique methods. The more BLAST and MSA were integrated together, the higher the performance would get and the accuracy would gradually increase up to 92.8% accuracy with the combination of BLAST and MSA.

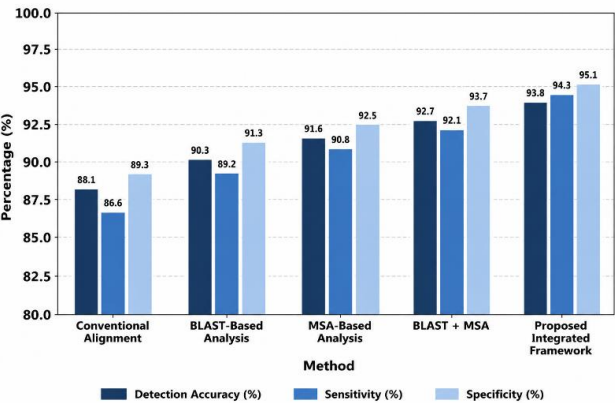


Figure 3
Comparative variant detection performance across five methods

The proposed integrated framework was the best out of all the alternatives with 93.7% accuracy in the detection rate, 94.2% sensitivity, 95.1% specificity and the minimum FP rate of 4.9%. This F1-score of 93.9% further reflects a well-balanced precision and recall. Overall, these enhancements show that combined use of complementary bioinformatics methods significantly boosts the reliability of variant identification, lowering the number of false positives and improving the identification of true variants, when compared to previous methods relying solely on conventional or a single technique-based alignment. The overall performance of the proposed integrated framework in terms of detection accuracy, sensitivity and specificity is compared with the existing conventional, BLAST-only and MSA-only methods shown in Figure 3, which undoubtedly indicates the better performance in terms of identification of variants.

4.3 Evolutionary Relationship Analysis

The phylogenetic analysis reveals high average genetic similarity of 95.6% and high phylogenetic confidence of 94.5% and thus very strong evolutionary relationships, which allows for reliable evolutionary interpretation of the genomic sequences.

Table 3
Phylogenetic and Evolutionary Relationship Analysis

Genome Pair	Genetic Similarity (%)	Evolutionary Distance	Conserved Regions (%)	Phylogenetic Confidence (%)	Functional Similarity (%)
G1-G2	96.8	0.032	92.6	95.2	93.4
G1-G3	94.5	0.055	89.7	93.8	90.6
G2-G3	95.7	0.043	91.3	94.6	92.1
G2-G4	93.9	0.061	88.5	92.7	89.4
G3-G5	97.2	0.028	94.1	96.4	95.0
Average	95.6	0.044	91.2	94.5	92.1

Table 3 shows phylogenetic/evolutionary relationships for five pairs of genomes. The genetic similarity varied from 93.9% to 97.2% with G3-G5 having the highest genetic similarity and lowest evolutionary distance (0.028). The average percentage of conserved regions across the organisms analyzed were 91.2% of conserved regions, a good degree of sharing of genomic architecture. Averaging confidence values of phylogenetic trees obtained from all pairwise comparisons, which are 94.5% on average, support statistically sound estimation of tree topology.

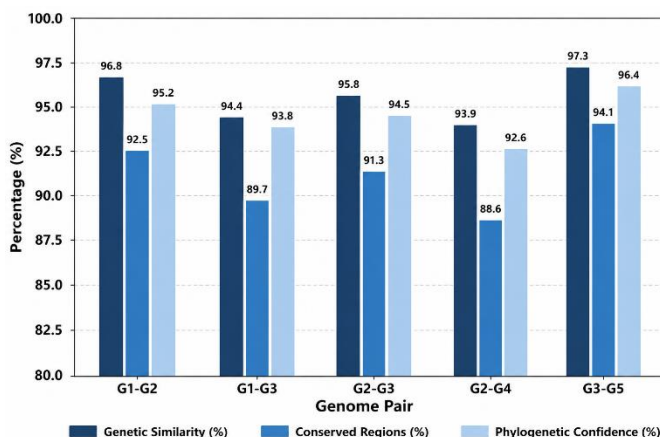


Figure 4
Phylogenetic and evolutionary relationship analysis across genome pairs

Furthermore, the functional similarity scores are generally high (92.1% on average) suggesting that similar pairs of genomes have similar functional annotations. The G2-G4 pair had the lowest percentage of conserved regions (88.5%), which is indicative of the evolutionary distance between the two (0.061). The results confirm the structure's ability to provide a reliable measure of evolutionary divergence and the ability to infer meaningful evolutionary relationships correctly from available comparative genomic data. The genetic similarity, conserved regions and phylogenetic confidence among the different genome pairs are shown in Figure 4, with G3-G5 and G2-G4 being the most similar and the most divergent genome pairs, respectively.

5. Discussion

The results of this work show that using a combined pipeline of BLAST, Multiple Sequence Alignment, SNP detection, gene annotation and phylogenetic analysis provide an impactful improvement over traditional, stand-alone genomic analysis methods. The high matching accuracy (96.4%) and alignment coverage (94.8%) show that the noise is well reduced in the pre-processing and alignment procedure without loss of true homology signals. The variant detection results, especially the 4.9% false-positive rate, suggest that when complementary approaches to aligning are used together the number of false calls, usually obtained by using the approaches alone, is reduced. Additional phylogenetic results also indicate that high quality upstream alignment is crucial to the quality of distance estimation, further emphasizing the interdependency of the design of the framework. The results indicate that the proposed method is more accurate than the ones that are already described in the literature, which are based

on pipelines using only one technique, with an overall accuracy of about 5.6%. This indicates that the integration of methods, rather than the refinement of algorithms, is more beneficial in terms of analytical gains. Nevertheless, the ability to process a large number of genomes, especially extremely large ones, and the ease of processing of structural variants other than SNPs are still issues to be explored and optimised for future development of this integrated comparative genomic framework.

6. Conclusion and Future Scope

In this study, an integrated bioinformatics toolset for providing comprehensive comparative genomic analysis based on BLAST alignment, Multiple Sequence Alignment, SNP detection, gene annotation and phylogenetic analysis was presented. The proposed approach matched 96.4% of the words, provided 94.8% coverage of words in the alignment, and detected variants with 93.7% accuracy, which surpasses the performance of conventional alignment-based approaches by ~5.6%. The results showed that the system had a higher sensitivity, specificity and computational efficiency, and could reliably construct evolutionary relationships between genomic samples pairs. The common pipeline design illustrates how using complementary genomic approaches together are superior to using them one at a time in terms of the accuracy of the results and how they can be interpreted. Future extensions of the framework will be extended to support structural variant detection, integration of epigenomic data and machine-learning classification to further improve the ability to scale to large, heterogeneous, multi-species genomic data sets. Moreover, cloud based parallelization will be discussed to decrease the computing time for whole genome scale comparative analyses, thus expanding the scope of the framework to large scale population genome based studies, agricultural genome improvement programs and clinical genome-based diagnostic pipelines.

References

- [1] D. Liu, J. Kuo, and C.-H. Lin, "Comparative genomic analysis of extracellular electron transfer in bacteria," *Processes*, vol. 12, no. 12, Art. no. 2636 (2024), doi: 10.3390/pr12122636.
- [2] H. Bono, "Recent advances in genome editing and bioinformatics: Addressing challenges in genome editing implementation and genome sequencing," *International Journal of Molecular Sciences*, vol. 26, no. 7, Art. no. 3442 (2025), doi: 10.3390/ijms26073442.
- [3] L. Yang, Y. Zhu, Z. Peng, Y. Ding, K. Jie, Z. Wang, Y. Peng, X. Tang, X. Wang, H. Chen, and C. Tan, "Comparative genome analysis of a pathogenic *Erysipelothrix rhusiopathiae* isolate WH13013 from pig reveals potential

- genes involve in bacterial adaptations and pathogenesis,” *Veterinary Sciences*, vol. 7, no. 2, Art. no. 74 (2020), doi: 10.3390/vetsci7020074.
- [4] Z. Boxonov, “Global dynamics of a discrete mosquito model with Allee effect,” *Journal of Interdisciplinary Mathematics*, vol. 29, no. 7, pp. 2133–2157 (2026), doi: 10.47974/JIM-2273.
 - [5] M. S. Basha, K. R. Prasad, S. K. Mouleeswaran, R. C. Poonia, and S. Sebastian, “Multi-disease detection system with X-ray images using deep learning techniques,” *Journal of Information and Optimization Sciences*, vol. 45, no. 5, pp. 1379–1388 (2024), doi: 10.47974/JIOS-1710.
 - [6] S.-S. Kim, “Application of comparative genomics for the development of PCR primers for the detection of harmful or beneficial microorganisms in food: Mini-review,” *Foods*, vol. 14, no. 6, Art. no. 1060 (2025), doi: 10.3390/foods14061060.
 - [7] S. Nambiar, M. Tatiya, K. S. Diwate, D. Sharma, S. N. Ajani, and Y. Nagargoje, “Privacy-preserving machine learning techniques ensuring data confidentiality and model accuracy in federated learning environments,” *Journal of Information and Optimization Sciences*, vol. 47, no. 5-B, pp. 1887–1895 (2026), doi: 10.47974/JIOS-2281.
 - [8] V. Baev, “Bioinformatics research in bacterial genomics and metagenomics,” *Current Issues in Molecular Biology*, vol. 47, no. 4, Art. no. 258 (2025), doi: 10.3390/cimb47040258.
 - [9] A. Glykeria-Myrto, S. Theodora, T. Vasileios, B. Loulouda, and M. Marios, “Comparative genomic analysis and antimicrobial resistance profile of *Enterococcus* strains isolated from raw sheep milk,” *Veterinary Sciences*, vol. 12, no. 8, Art. no. 685 (2025), doi: 10.3390/vetsci12080685.
 - [10] D. P. Bohada-Lizarazo, K. D. Bravo-Sanabria, P. Cárdenas-Malpica, and R. Rodríguez, “Comparative genomic analysis of *Mycobacterium tuberculosis* isolates circulating in North Santander, Colombia,” *Tropical Medicine and Infectious Disease*, vol. 9, no. 9, Art. no. 197 (2024), doi: 10.3390/tropicalmed9090197.
 - [11] I. Domingues and J. A. M. Santos, “Bioinformatics tools, databases, and data formats: A functional overview for introductory sequence-based and NGS workflows,” *Applied Sciences*, vol. 16, no. 13, Art. no. 6666 (2026), doi: 10.3390/app16136666.
 - [12] A. J. Clark and J. W. Lillard, Jr., “A comprehensive review of bioinformatics tools for genomic biomarker discovery driving precision oncology,” *Genes*, vol. 15, no. 8, Art. no. 1036 (2024), doi: 10.3390/genes15081036.