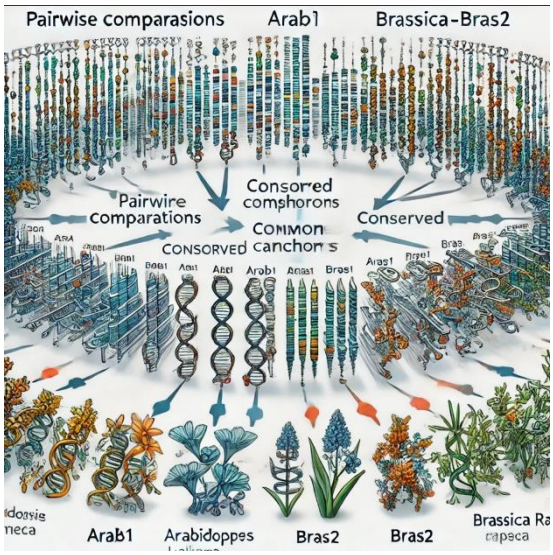
Hellenic Open University

Bioinformatics and Neuroinformatics

Postgraduate Dissertation

Extending AnchorWave for Multiple Sequence Alignment: A Novel Approach for Comparative Genomics

Andriana Sielli



Supervisor: Charidimos Kondylakis

Patras, Greece, January 2025

Extending AnchorWave for Multiple Sequence Alignment: A Novel Approach for Comparative Genomics

Andriana Sielli

Supervisor:
Kondylakis Charidimos

Patras, Greece, January 2025

"This thesis is dedicated to Leto, whose unwavering support and guidance have been a source of strength. I am truly grateful."

Abstract

Our method extends AnchorWave to support multiple sequence alignment (MSA), enabling the identification of conserved regions across multiple sequences while maintaining computational efficiency. By leveraging an anchor-based approach, it minimizes the computational overhead typically associated with MSA, making it suitable for large-scale genomic studies. The implementation is optimized to run on low-capacity resources, ensuring accessibility for researchers working with limited computational power. This extension provides a practical and scalable solution for comparative genomics, particularly in analyzing plant genomes, which often contain complex structural variations and repetitive regions.

Keywords: Multiple Sequence Alignment, AnchorWave, Computational Efficiency, Memory Optimization, Comparative Genomics, Plant Genomes, Structural Variations, Bioinformatics

Table of Contents

List of figures

List of Tables

List of Abbreviations and Acronyms

1.Introduction

Advances in sequencing technologies have revolutionized genomics, enabling large-scale, high-throughput sequencing for various applications, including phylogenetics, comparative genomics, and protein structure prediction (Chao, Tang, & Xu, 2022). At the heart of these applications is sequence alignment, a fundamental bioinformatics task that identifies regions of similarity between sequences. Sequence alignment methods are categorized into two primary types: pairwise alignment and multiple sequence alignment (MSA). Pairwise alignment, which compares two sequences, is further divided into local and global alignment. Local alignment identifies regions of high similarity within sequences, while global alignment aligns sequences across their entire length. MSA, in contrast, simultaneously aligns three or more sequences, making it indispensable for studying evolutionary relationships, predicting protein function, and identifying conserved regulatory elements (Edgar & Batzoglou, 2006).

The increasing volume and complexity of genomic data have prompted continuous improvements in MSA algorithms, enhancing both speed and accuracy. Traditional alignment methods rely on optimizing sequence similarity, often utilizing substitution matrices to score aligned residue pairs and gap penalties to account for insertions and deletions (Edgar & Batzoglou, 2006). However, aligning sequences from complex genomes, such as those of plants, presents unique challenges. Plant genomes are characterized by extraordinary diversity, large size variations, and distinct evolutionary features, including frequent polyploidization, high transposable element activity, and dispensable genomic regions (Shi, Tian, Lai, & Huang, 2023). These characteristics make MSA computationally demanding, necessitating more advanced algorithms capable of effectively handling large-scale sequence comparisons.

AnchorWave, a state-of-the-art pairwise sequence alignment tool, has demonstrated exceptional performance in aligning genomes with high sequence diversity, structural polymorphisms, and whole-genome duplications (Song et al., 2022). By employing a two-piece affine gap cost strategy and a longest-path dynamic programming algorithm, AnchorWave achieves high alignment accuracy, particularly for structurally complex genomes. However, its limitation to pairwise alignment restricts its applicability in comparative genomics studies involving multiple species, where MSA is essential to infer evolutionary relationships more effectively. Existing MSA tools, such as Clustal Omega, MAFFT, and MUSCLE, while widely used, often struggle with large and highly variable genomes due to scalability constraints or sensitivity to structural variations. Adapting AnchorWave for MSA could offer a novel approach that preserves its accuracy in handling genomic rearrangements while enabling the simultaneous alignment of multiple sequences.

In this thesis, I propose to extend the AnchorWave algorithm to support MSA by modifying its anchor-based approach to identify and align conserved regions across multiple sequences. This adaptation aims to enhance MSA efficiency and scalability, particularly for large and complex genomic datasets, while preserving evolutionary relationships and minimizing computational

overhead. By addressing the limitations of existing MSA tools and leveraging AnchorWave's strengths, this work seeks to advance comparative genomics by providing a robust and biologically meaningful alignment method for complex genomes.

2. Background and Related Work

2.1 Sequence Alignment and Its Significance

Sequence Alignment in Bioinformatics
Sequence alignment is a fundamental technique in bioinformatics, used to detect regions of similarity across DNA, RNA, or protein sequences. These alignments are pivotal for understanding evolutionary relationships, gene functions, and structural conservation, with key applications in phylogenetics, comparative genomics and protein structure prediction (Reddy & Fields, 2022). Alignment methods are typically categorized into two types: pairwise sequence alignment (PSA) and multiple sequence alignment (MSA)

Pairwise Sequence Alignment (PSA):
PSA involves comparing two sequences to determine the optimal alignment between them. This can be performed using either global or local alignment methods. Global alignment, introduced by Needleman and Wunsch (1970), aims to align the sequences across their entire length, while local alignment, developed by Smith and Waterman (1981), focuses on aligning the most similar regions within the sequences. PSA algorithms, such as Needleman-Wunsch and Smith-Waterman, rely on dynamic programming, which guarantees optimal alignments but is computationally demanding, especially for large datasets. To address these computational challenges, advancements such as the Wavefront Alignment Algorithm (WFA) and its bidirectional variant, BiWFA (Marco-Sola et al., 2021), have significantly reduced the time and memory complexity of global pairwise alignment. Other recent developments, such as minimap2 (Li, 2018), have been tailored to efficiently handle large-scale genomic data, particularly from long-read sequencing technologies.

Multiple Sequence Alignment (MSA):
MSA extends Pairwise Sequence Alignment (PSA) to align three or more sequences simultaneously, making it crucial for identifying conserved motifs, functional domains, and evolutionary patterns across different species (Edgar & Batzoglou, 2006). However, the computational complexity increases significantly as the number of sequences grows, often requiring heuristic or probabilistic methods to achieve feasible solutions (Thompson et al., 2011). Algorithms such as MAFFT-DASH (Rozewicki et al., 2019) and UPP (Nguyen et al., 2015) have enhanced the scalability and accuracy of MSA, enabling better handling of large and highly divergent datasets, thus improving alignment efficiency and accuracy in the analysis of evolutionary relationships.

Existing MSA Tools

Various MSA algorithms have been developed to balance computational efficiency with biological accuracy. Key tools include:

- **Clustal Omega** (Sievers et al., 2011): A widely used progressive alignment method that constructs a guide tree to align sequences. It performs well for moderate-size datasets but struggles with large, highly divergent sequences.
- **MAFFT** (Katoh & Standley, 2013): Uses fast Fourier transform (FFT) for rapid alignment, making it effective for large datasets, although it may be less accurate for sequences with significant structural rearrangements. The MAFFT-DASH extension (Rozewicki et al., 2019) improves its performance in large-scale alignments.
- **MUSCLE** (Edgar, 2004): A tool that employs iterative refinement to improve alignment accuracy, but can be computationally intensive for large datasets.
- **T-Coffee** (Notredame et al., 2000): Uses a consistency-based scoring system to enhance alignment accuracy, though it requires substantial computational resources.
- **PRANK** (Löytynoja & Goldman, 2005): A phylogeny-aware tool useful for handling insertions and deletions, particularly in evolutionary studies.
- **UPP** (Nguyen et al., 2015): A machine learning-based ensemble alignment method that improves accuracy for datasets with high sequence divergence.

Despite the effectiveness of these tools, challenges persist when aligning large genomes with complex structural variations, such as those seen in plant genomes. As such, developing computationally efficient and structurally robust MSA tools remains a critical challenge in comparative genomics (Shi et al., 2023). Integrating machine learning and parallel computing into these tools has shown promising results in addressing some of these issues (Zhang et al., 2022).

2.2 Anchor-Based Alignment and the Role of AnchorWave

To address the limitations of traditional alignment methods, anchor-based alignment techniques have been developed. These methods first identify high-confidence anchor regions—segments that are highly conserved across sequences—and use them as fixed reference points to guide the alignment process. Anchor-based approaches are particularly effective for aligning genomes with high sequence diversity, structural polymorphisms, and repetitive elements, offering a robust solution for complex genomic comparisons (Song et al., 2022).

AnchorWave is a state-of-the-art anchor-based alignment tool specifically designed for pairwise sequence alignment. It excels in aligning genomes with high sequence diversity, large structural polymorphisms, and whole-genome duplications. By employing a two-piece affine gap cost model and a longest-path dynamic programming approach, AnchorWave achieves high accuracy

in aligning sequences while accommodating complex genomic rearrangements. This makes it a powerful tool for comparative genomics and evolutionary studies (Song et al., 2022).

Workflow of AnchorWave

The workflow of AnchorWave relies on three key input files: the reference genome in FASTA format, the corresponding gene annotation in GFF3 format, and the query genome in FASTA format. The process begins with the extraction of full-length coding sequences (CDS) from the reference genome, using the gene annotation data to precisely define these regions. These CDS positions are then mapped onto the query genome using a splice-aware alignment tool, such as minimap2, which ensures accurate alignment across introns and splice junctions. This approach allows AnchorWave to handle complex genomic features with high precision.
AnchorWave excels in aligning highly diverse genomes, achieving up to three times more positional matches or indels compared to other methods. It also demonstrates superior performance in transcription factor-binding site recall, with rates ranging from 1.05 to 74.85 times higher than competing tools, while maintaining significantly lower false-positive alignment rates (Song et al., 2022). These capabilities make AnchorWave a powerful tool for analyzing genomes with dispersed repeats, active transposable elements, and whole-genome duplications.

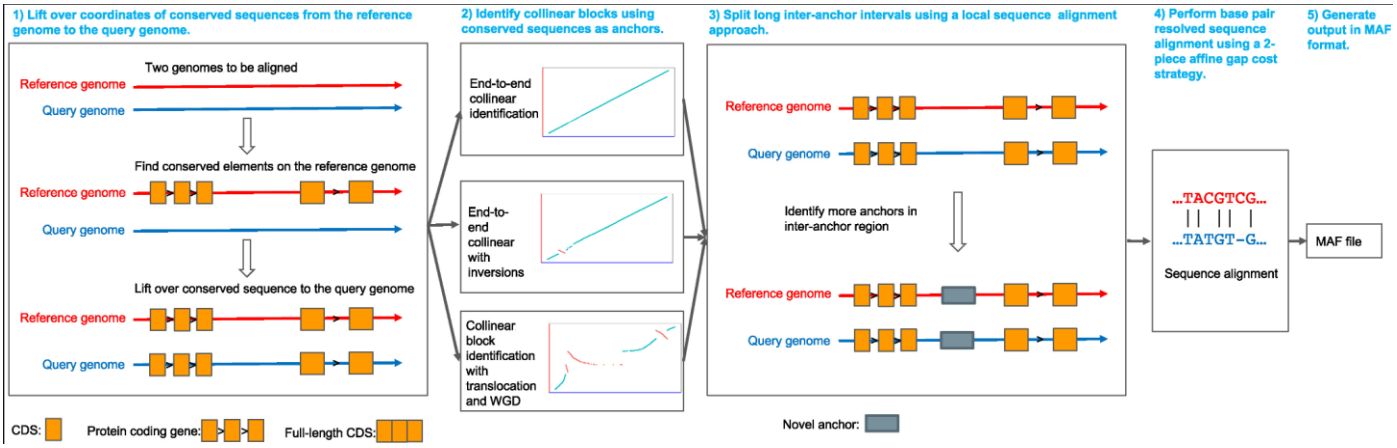


Figure 1: AnchorWave process

However, AnchorWave is limited to pairwise alignment, making it unsuitable for analyses requiring simultaneous alignment of multiple genomes. This limitation poses significant challenges for comparative genomics studies involving multiple species, where identifying conserved regions across all genomes is essential. This restriction hinders the applicability of

AnchorWave in large-scale multi-species comparative studies, highlighting the need for extending its capabilities to support MSA.

2.3 Motivation for Extending AnchorWave to MSA

As already mentioned, despite the success of AnchorWave in PSA, its lack of MSA support limits its applicability in comparative genomics studies requiring the simultaneous alignment of multiple species. Existing MSA tools either struggle with highly diverse genomes or require extensive computational resources. By extending AnchorWave to support MSA, this work aims to:

- Leverage AnchorWave's strength in handling large structural variations to enable MSA of highly diverse and structurally complex genomes, addressing limitations of existing tools.
- Develop an anchor-based MSA approach that preserves conserved regions while reducing computational complexity.
- Provide a scalable solution for aligning large and structurally complex genomes, particularly in plants.

By integrating anchor-based strategies with MSA methodologies, this research seeks to advance comparative genomics by enabling more accurate and efficient multi-genome alignment.

3. Methods

Overview

The goal of this study was to extend AnchorWave for multiple sequence alignment (MSA) by leveraging pairwise comparisons to identify conserved regions across multiple species. The approach was designed to efficiently handle large genomic datasets while minimizing memory usage and computational overhead, with an additional focus on increasing alignment accuracy. The methodology consisted of two main phases: identifying conserved anchors through pairwise alignments and refining these anchors to construct a robust multi-species alignment. Each phase was carefully designed to address the challenges of aligning large genomes, ensuring both accuracy and scalability. Eventually, a multiple sequence alignment was performed using MAFFT, which was applied to both the anchor and inter-anchor regions.By incorporating both anchor and inter-anchor regions into the alignment, the study aimed to enhance the overall process, offering a potentially more reliable framework for studying evolutionary relationships and genomic conservation.

Pairwise Alignment and Anchor Identification

The first step in the methodology involved performing pairwise comparisons between a reference species and each of the other species in the dataset. This was achieved using the ProAli command, which generated pairwise alignments and identified conserved anchors between each pair of species. Anchors are regions of high similarity and structural conservation that serve as reliable markers for alignment. By breaking the problem into pairwise comparisons, this approach avoids the memory-intensive task of aligning all sequences simultaneously, making it scalable for large genomes.

Pairwise alignments are computationally less demanding than full MSA because they only compare two sequences at a time. This reduces the memory footprint and allows for parallel processing, making it feasible to handle large datasets. Additionally, anchors represent regions of high confidence in the alignment, as they are based on conserved sequences and structural features. By focusing on anchors, the approach prioritizes biologically meaningful regions, reducing the risk of misalignment in less conserved areas. This approach increases the biological relevance of the alignment by focusing on areas that are functionally important, such as functional domains or regulatory elements.

Filtering and Refining Anchors

Once the pairwise anchor files were generated, the dataset was refined to identify anchors conserved across all species. This involved extracting anchors from the reference species that were shared in all pairwise comparisons, specifically those with common starting and ending points relative to the reference genome. The filtering step ensured that the final dataset included only regions conserved across all species, thus reducing noise and improving

alignment accuracy. Shared anchors are likely to represent biologically important regions, making them ideal for guiding the alignment.
By focusing on these highly conserved regions, the approach minimizes computational resources, which is critical when working with large genomes that require substantial memory and processing power. This filtering process ensures that the alignment captures the most significant and conserved parts of the genome, enhancing the overall alignment's biological relevance.

Constructing the Multi-Species Alignment
The multi-species alignment was constructed using the refined anchors as a scaffold. This method ensured that the conserved anchors, which represent high-confidence regions, were treated as the foundational framework for the alignment. **Inter-anchor regions, the genomic areas between the anchors, were aligned around these conserved regions, helping to ensure the overall accuracy and integrity of the alignment.**
By prioritizing biologically relevant areas, the approach targets regions that are likely to be evolutionarily significant and important for comparative genomics. The reference species provided a consistent framework, ensuring coherence and meaningful alignment across distantly related species. This approach helps to accurately reflect the evolutionary relationships between species by focusing on regions that are conserved across the genomes.

Integration with MAFFT
Once the anchors and inter-anchor regions were identified and refined, a final multiple sequence alignment was performed using the MAFFT (Multiple Alignment using Fast Fourier Transform) algorithm. MAFFT was applied separately to both the anchor and inter-anchor regions, but this was done for a small subset of four sequences rather than the entire genome. This approach ensured that the anchor regions, assumed to be highly conserved, were aligned first, followed by alignment of the inter-anchor regions.
By utilizing the AnchorWave MSA tool, we generated alignments for both anchors and inter-anchor regions, allowing us to focus on smaller parts of the sequences instead of attempting to align the whole genome. This approach is especially useful in situations with limited computational resources, where full-genome alignment would not be feasible. An average alignment score was obtained from MAFFT for the four sequences. This score can provide a comparative measure of the alignment quality, offering insight into the accuracy of the alignments for the given data. This flexibility enables researchers to work with conserved regions or inter-anchor regions individually, enhancing scalability and alignment accuracy even with limited computational capacity.

Advantages of the Approach
This methodology offers several key advantages. First, by relying on pairwise comparisons, it avoids the memory limitations typically associated with traditional MSA tools, making it more

suitable for large-scale genomic analyses. Pairwise alignments are inherently less memory-intensive and can be parallelized, allowing the approach to handle datasets that would be infeasible for traditional MSA tools.
Second, the filtering step ensures that the final alignment is based on highly conserved regions, potentially improving both accuracy and biological relevance. By focusing on regions of high conservation, the approach prioritizes functionally important parts of the genome, which are often the focus of downstream analyses. This may enhance the alignment's overall biological significance.
Finally, these advantages make the approach a powerful tool for large-scale genomic alignment, particularly when computational resources are limited or traditional MSA tools face challenges. By incorporating both anchor and inter-anchor regions into the alignment process, this study offers a more scalable and efficient framework for studying evolutionary relationships and genomic conservation.

4.Results

Data Selection

The dataset for this study was carefully selected from Ensembl to evaluate the performance of the extended AnchorWave method for multiple sequence alignment (MSA). We focused on four plant species: *Arabidopsis thaliana* (Arab1), *Arabidopsis halleri* (Arab2), *Brassica juncea* (Bras1), and *Brassica rapa* (Bras2). These species were chosen due to their evolutionary relationships, with *Arabidopsis thaliana* serving as the reference genome. *Arabidopsis halleri* is closely related to *Arabidopsis thaliana*, while *Brassica juncea* and *Brassica rapa* are more distantly related, providing a range of genetic diversity to test the method's robustness.

We specifically selected plant genomes because AnchorWave is particularly well-suited for plant genome alignments. This choice allowed us to assess the ability of the extended AnchorWave approach to handle both closely and distantly related genomes, ensuring its applicability to a wide range of comparative genomics studies. The inclusion of species with varying degrees of genetic divergence also enabled us to evaluate the method's performance in identifying conserved anchors and aligning inter-anchor regions, which are critical for understanding evolutionary relationships and genomic conservation.

Identification of Common Anchors and Anchor Extraction

The first step in extending AnchorWave for multiple sequence alignment (MSA) involved identifying common anchors between pairs of species. Using the ProAli command, pairwise comparisons were performed between Arab1 and Arab2, Arab1 and Bras1, and Arab1 and Bras2. These pairwise comparisons were critical for identifying conserved regions that could serve as reliable anchors for subsequent multi-species alignment. While ProAli provided the common anchors for each pairwise comparison, the dataset was further refined by focusing on anchors with the same starting and ending point across all three comparisons. This ensured consistency and reliability in the anchor selection process.

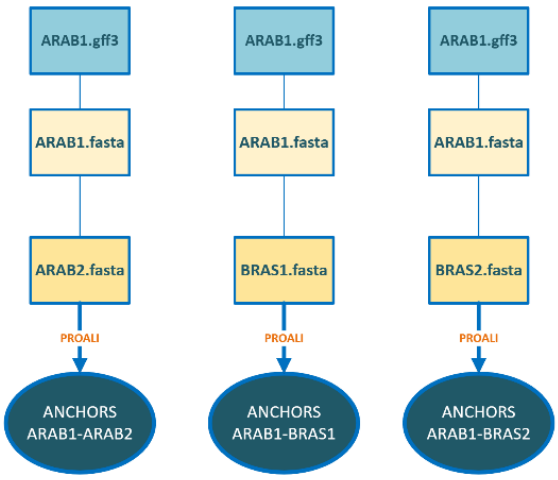


Figure 2: Extracting anchors from the pairwise alignment process of AnchorWave

Pairwise Alignment and Common Anchor Identification

Figure 1 showcases the first stage of the process, where pairwise alignments were performed between the reference genome, *Arabidopsis thaliana* (ARAB1) and the three related species. This process generated sets of anchors for each pairwise comparison (ARAB1-BRAS1, ARAB1-BRAS2, and ARAB1-BRAS3), highlighting regions of significant sequence similarity and structural conservation. The number of anchors found in each case is as follows: Arab1 and Arab2 have 35,642 anchors, while Arab1 and Bras1 have 19,580 anchors, and Arab1 and Bras2 have 18,904 anchors. This distribution is logical, as Arab1 and Arab2 are genetically much closer to each other, resulting in a higher number of common anchors. Similarly, Bras1 and Bras2 are genetically closer, which is reflected in the similar number of anchors found when compared to Arab1.

These anchors served as potential building blocks for the final MSA. However, to ensure accuracy and biological relevance, the analysis focused on regions conserved across all species. This is where the common anchor identification, depicted in *Figure2* , became essential. By comparing the anchor coordinates from the pairwise alignments, anchors that shared identical start and end positions within the ARAB1 reference were identified. These 'common anchors' represented the most highly conserved segments across all three genomes, indicating regions likely to be functionally important. As shown in the figure, only anchors meeting this stringent criterion were selected for further analysis, forming a robust foundation for constructing the multi-species alignment.

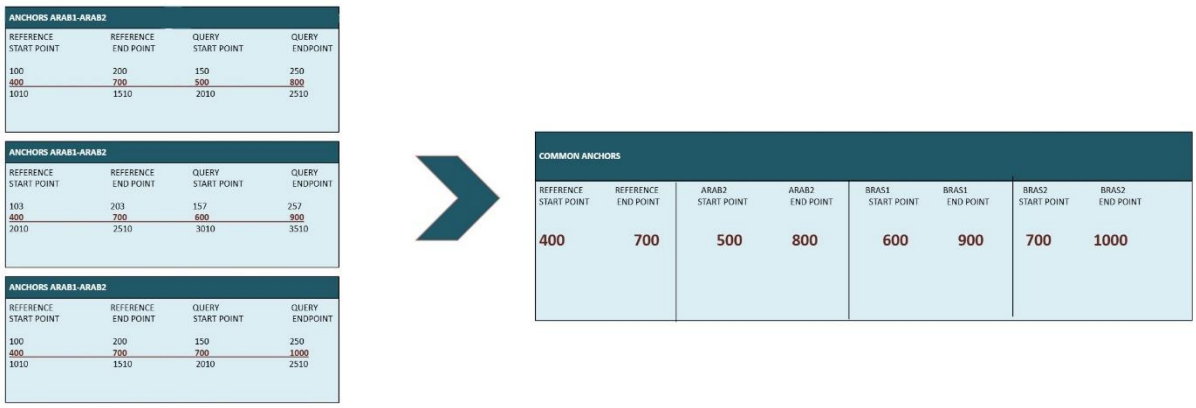


Figure 3: Selecting only the anchors that have the same start and end point mutual

Quality Assessment of Common Anchors

A total of 2,916 common anchors were identified, significantly strengthening the reliability of the final multi-species alignment (MSA). To assess the quality of these anchors across all species, we calculated the average alignment score, which was 0.9545. This high average score indicates strong sequence conservation, suggesting that the anchors are well-conserved among the species studied. Such conservation is essential for ensuring that the anchors are biologically relevant and can serve as reliable starting points for constructing the multi-species alignment.

ARAB1		ARAB2		BRAS1		BRAS2	
START	END	START	END	START	END	START	END
22740	24847	5288	5611	19873575	19875047	30631474	30633936
34872	37756	5612	8999	19747156	19748271	30539789	30540201
45503	46789	9000	11803	19722392	19724922	30529774	30531469
46619	48143	11804	16741	19675498	19677127	30479644	30481205
46790	47704	16742	21066	19642407	19643814	30411648	30412685
48144	51731	33951	36371	19291436	19292942	30381475	30383798
51732	54344	96196	97545	19253485	19254418	30375319	30381474
52239	54494	162375	163914	19168281	19170510	30356470	30358729
54345	54654	189685	193064	19004705	19009974	30352945	30355244
54495	57391	260710	262296	18901117	18902823	30348856	30352066

Table 1: Example of Identified Anchors Regions for Multiple Sequence Alignment.

Further Validation of Anchor Quality through MAFFT Alignment
Further validating the quality of the selected anchors, we aligned one anchor sequence from each species using MAFFT, resulting in a matrix score of 1.898. This score reflects the high degree of sequence conservation across the four species. A higher matrix score typically indicates a better alignment with fewer mismatches and gaps, which supports the idea that these regions are biologically important and evolutionarily conserved. Thus, the high quality of the alignment reinforces the reliability of these anchors for subsequent analysis in the multi-species alignment.

Phylogenetic Tree Analysis
Additionally, a phylogenetic tree was generated from the MAFFT alignment, revealing a high degree of similarity between the species, which further corroborates the robustness of the conserved regions identified.

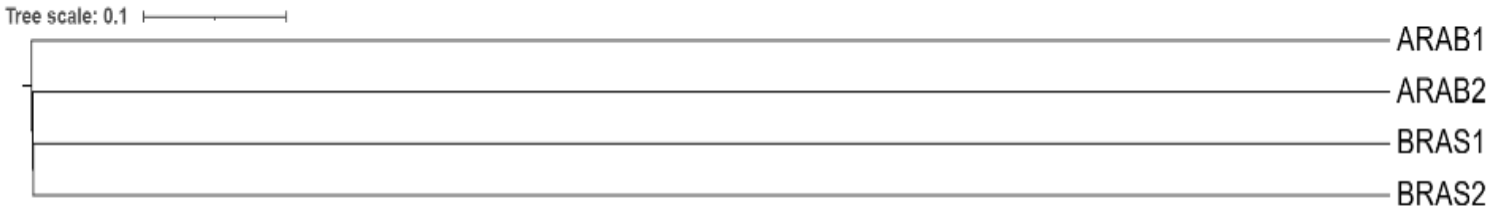


Figure 4: Phylogenetic tree based on four anchor sequences

Analysis of Inter-Anchor Regions
In addition to the conserved anchors, the analysis of inter-anchor regions was essential for refining the alignment accuracy across multiple species. These regions, located between the anchors, offer crucial contextual information about the spacing and relationships between anchors, contributing to a more precise alignment. The inter-anchor regions were identified for *Arabidopsis thaliana* (Arab1), *Arabidopsis halleri* (Arab2), *Brassica juncea* (Bras1), and *Brassica rapa* (Bras3), and the corresponding DNA sequences were extracted for each species. By incorporating both the conserved anchors and inter-anchor regions, the dataset for multi-species alignment was enriched, extending the pairwise AnchorWave method into a multi-

species context. While the alignment score for the inter-anchor regions (1.884) was slightly lower than that achieved with the anchors alone, this finding is consistent with expectations, as inter-anchor regions typically exhibit more variability in alignment. Moreover, the phylogenetic tree derived from this alignment showed a slight difference in the evolutionary relationships between the species compared to the tree based on the anchors, reflecting the expected evolutionary divergence. This consistency between the alignment scores and the tree suggests that the inclusion of inter-anchor regions strengthens the reliability of our approach, further supporting the robustness of our methodology in capturing both conserved and variable genomic features across species.

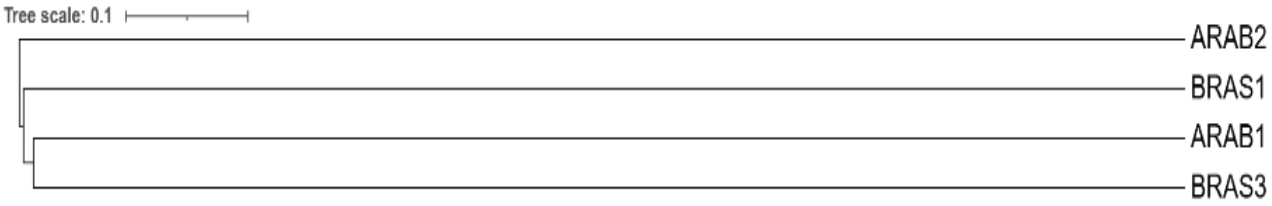


Figure 5: Phylogenetic tree based on four interanchor sequences

Computational Performance Comparison

To evaluate the efficiency of the extended AnchorWave method, we measured the time required to complete the multiple sequence alignment (MSA) and compared it to traditional MSA tools. The results, shown in Figure 6, demonstrate that our method successfully completed the MSA in 50 minutes, whereas traditional methods failed to complete the alignment within a reasonable time frame. This highlights the scalability and efficiency of our approach, making it suitable for large-scale genomic analyses

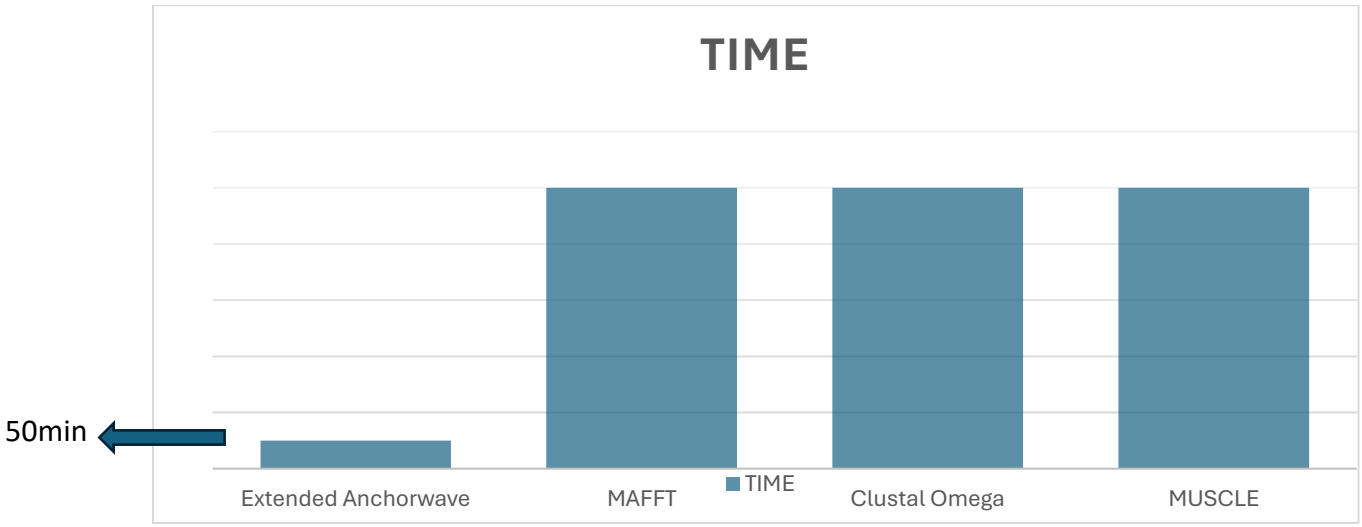


Figure 6: Time of implementation of different MSA methods

5. Conclusions

This study explores the potential of extending AnchorWave to support multiple sequence alignment (MSA), aiming to contribute to the field of comparative genomics, particularly for analyzing large and complex genomes such as those of plants. By adapting the anchor-based approach, which has shown promise in pairwise alignment, to handle multiple sequences, this work seeks to address some of the challenges associated with traditional MSA tools, such as scalability and the ability to manage structural variations. While the results are encouraging, they represent a step forward rather than a definitive solution.

A key advantage of this approach is its ability to reduce computational overhead while maintaining alignment accuracy. Traditional MSA tools often struggle with large datasets due to high memory and processing demands. By focusing on conserved anchors and breaking the problem into pairwise comparisons, the method minimizes resource requirements, making it more accessible for large genomes, even on limited computational resources. This is particularly relevant for plant genomes, which frequently contain structural variations, repetitive elements, and duplications.

The results highlight the method's potential, with 2,916 common anchors identified and an average alignment score of 0.9545, indicating strong sequence conservation. The high matrix score (1.898) from MAFFT alignment further supports the quality of these conserved regions, suggesting their biological and evolutionary significance. Including inter-anchor regions, though more variable, provides additional context for the alignment, with a slightly lower score (1.884) reflecting expected variability. The phylogenetic trees derived from both anchor and inter-anchor regions align with evolutionary relationships, demonstrating the method's robustness.

In conclusion, this work represents a modest but meaningful step in extending AnchorWave for MSA, offering a scalable and efficient approach for aligning complex genomes. While challenges remain, the method shows promise in addressing some limitations of existing tools and contributing to the advancement of comparative genomics.

6. Future Directions

A key area for future research is enhancing the scalability of the AnchorWave method to accommodate large genomic datasets. As sequencing technologies advance, optimizing computational efficiency will be crucial for processing multiple genomes simultaneously. Approaches such as parallelization or cloud computing could enable high-throughput analyses, making the method applicable to large-scale comparative genomics studies.
In addition, integrating machine learning (ML) techniques could significantly improve alignment precision. ML models could be trained to predict optimal alignment parameters, such as gap penalties and anchor positions, based on genomic characteristics. These models could adaptively handle genomic variability and refine the identification of inter-anchor regions. By leveraging ML, the method could more effectively address challenges like structural variations and repetitive sequences, leading to more accurate multi-species alignments.
Focusing on scalability and ML integration will make the AnchorWave method more versatile, enabling its application to large, complex datasets and enhancing its utility in high-throughput genomic research.

References

1. Chao, J., Tang, F., & Xu, L. (2022). Developments in algorithms for sequence alignment: A review. Biomolecules, 12(4), 546. https://doi.org/10.3390/biom12040546
2. Edgar, R. C., & Batzoglou, S. (2006). Multiple sequence alignment. Current Opinion in Structural Biology, 16(3), 368–373. https://doi.org/10.1016/j.sbi.2006.04.004
3. Shi, J., Tian, Z., Lai, J., & Huang, X. (2023). Plant pan-genomics and its applications. Molecular Plant. https://doi.org/10.1016/j.molp.2022.12.009
4. Song, B., Marco-Sola, S., Moreto, M., Johnson, L., Buckler, E. S., & Stitzer, M. C. (2022). AnchorWave: Sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. Proceedings of the National Academy of Sciences, 119(1), e2113075119. https://doi.org/10.1073/pnas.2113075119
5. Reddy, B., & Fields, R. (2022). Multiple sequence alignment algorithms in bioinformatics. In Bioinformatics and Computational Biology (pp. 145–160). Springer. https://doi.org/10.1007/978-981-16-4016-2_9
6. Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 48(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4
7. Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, *147*(1), 195–197. https://doi.org/10.1016/0022-2836(81)90087-5
8. Marco-Sola, S., Moure, J. C., Moreto, M., & Espinosa, A. (2021). Fast gap-affine pairwise alignment using the wavefront algorithm. *Bioinformatics, 37*(4), 456–463. https://doi.org/10.1093/bioinformatics/btaa777
9. Li H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191
10. Thompson JD, Linard B, Lecompte O, Poch O (2011) A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. PLoS ONE 6(3): e18093. https://doi.org/10.1371/journal.pone.0018093
11. Rozewicki, J., Li, S., Amada, K. M., Standley, D. M., & Katoh, K. (2019). MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic acids research*, *47*(W1), W5–W10. https://doi.org/10.1093/nar/gkz342
12. Nguyen, N.-P. D., Mirarab, S., Kumar, K., & Warnow, T. (2015). Ultra-large alignments using phylogeny-aware profiles. *Genome Biology, 16*(1), 124. https://doi.org/10.1186/s13059-015-0688-z
13. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, *7*, 539. https://doi.org/10.1038/msb.2011.75

14. Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010
15. Edgar R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340
16. Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, *302*(1), 205–217. https://doi.org/10.1006/jmbi.2000.4042
17. Löytynoja A. (2014). Phylogeny-aware alignment with PRANK. *Methods in molecular biology (Clifton, N.J.)*, *1079*, 155–170. https://doi.org/10.1007/978-1-62703-646-7_10
18. Zhang, Y., Gao, Y., Wang, H., Wu, H., Xia, Y., & Wu, X. (2024). A secure high-order gene interaction detection algorithm based on deep neural network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 21*(4), 619–630. https://doi.org/10.1109/TCBB.2022.3214863

Author's Statement:
I hereby expressly declare that, according to the article 8 of Law 1559/1986, this dissertation is solely the product of my personal work, does not infringe any intellectual property, personality and personal data rights of third parties, does not contain works/contributions from third parties for which the permission of the authors/beneficiaries is required, is not the product of partial or total plagiarism, and that the sources used are limited to the literature references alone and meet the rules of scientific citations.