Accelerating Multiple Sequence Alignments Using Parallel Computing

Subjects: Computer Science, Information Systems

Contributor: Qanita Bani Baker, Ruba Al-Hussien, Mahmoud Al-Ayyoub

Multiple sequence alignment (MSA) stands as a critical tool for understanding the evolutionary and functional relationships among biological sequences. Obtaining an exact solution for MSA, termed *exact-MSA*, is a significant challenge due to the combinatorial nature of the problem. Using the dynamic programming technique to solve MSA is recognized as a highly computationally complex algorithm. To cope with the computational demands of MSA, parallel computing offers the potential for significant speedup in MSA.

Keywords: multiple sequence alignment ; dynamic programming ; parallel computing

1. Introduction

Sequence alignment (SA) refers to the process of arranging and comparing biological sequences, such as DNA, RNA, and proteins, with the ability to reveal meaningful information about the similarities and differences among them. SA is one of the fundamental steps in most genomic analyses [1]. The classification of sequence alignment techniques encompasses two fundamental distinctions: global versus local alignment and pairwise versus multiple alignment. Global alignment algorithms, such as the Needleman–Wunsch algorithm, align sequences from the beginning to the end [2]. By contrast, the local sequence alignment is used to compare specific regions and to find contiguous regions of high similarity, such as that performed by the the Smith–Waterman algorithm [3]. The main two types of sequence alignment are also performed in two ways: the pairwise sequence alignment (PSA) [4] and the multiple sequence alignment (MSA) [5]. PSA involves the comparison of two sequences to identify regions of similarity and dissimilarity, while MSA extends the comparison to more than two sequences, identifying conserved regions and variations across a set of sequences.

The dynamic programming technique [6] provides an efficient computational approach for optimizing alignment scores by breaking down complex problems into smaller overlapping subproblems [7]. Common dynamic programming algorithms for pairwise sequence alignment include the Needleman–Wunsch algorithm, employed for global alignment, and the Smith–Waterman algorithm, utilized for local alignment. Dynamic programming extends its utility to multiple sequence alignment algorithms, such as the progressive and iterative methods [8][9]. Aligning N sequences using dynamic programming is an NP-Hard problem [10] that stems from the complexity of considering all possible combinations and alignments among the N sequences. To address complexity challenges in MSA, heuristic methods [11] and approximation algorithms [12] are employed in practice for the MSA of a large number of sequences. In addition to these algorithms, applying parallel computing techniques offers a promising avenue to mitigate the computational demands associated with MSA [13].

2. Pairwise Sequence Alignment (PSA)

Pairwise sequence alignment (PSA) is considered an important tool for aligning biological sequences such as DNA and protein sequences [14]. Haque et al. [15] presented a comprehensive overview of both local and global pairwise sequence alignment algorithms. They also included an identification of the techniques utilized in these algorithms and discussed their respective advantages and limitations. In [16], Edgar et al. distinguished between the main three methods used to align sequences: sequence–sequence methods (like BLAST), profile–sequence methods (like PSI-BLAST), and profile–profile methods (like CLUSTALW). The survey in [17] reviewed the wide range of aligning algorithms and tools developed to assess the quality of the aligned sequences. In [18], bacterial DNA sequences were aligned using pairwise alignment and dynamic programming. **Table 1** shows an overview of the most well-known approaches utilized for PSA.

Table 1. Pairwise sequence alignment techniques.

#	Technique	Approach	Reference
1	Needleman–Wunsch	Dynamic Programming	[2]
2	Smith–Waterman	Dynamic Programming	[3]
3	Gotoh’s Algorithm	Dynamic Programming	[19]
4	FASTA Algorithm	Heuristic	[20]
5	BLAST Algorithm	Heuristic	[21]
6	EMBOSS Software	Toolkit	[22]
7	Parasail	Toolkit/Library	[23]
7	Minimap2	Toolkit/Program	[24]
9	ASCA-PSO	Heuristic	[25]
8	WFA-GPU	Toolkit	[26]

Several studies have aimed to accelerate the performance and the accuracy of the tools used in sequence alignment by using several parallelization techniques [27]. For example, Fakirah et al. [28] utilized a diagonal traversing approach to enhance the Needleman–Wunsch algorithm by utilizing the iterations used to fill the scoring matrix. Balhaf et al. [29] enhanced the Levenshtein edit distance algorithm’s performance by using the diagonal traversing approach, and the performance was enhanced using both CPU and GPU. Jararweh et al. [30] accelerated the Levenshtein and Damerau algorithms by using parallel implementation on a GPU. Jararweh et al. showed that using unified memory resulted in the best performance. Shehab et al. [31] enhanced the performance of multiple pairwise alignments in protein sequences by utilizing a hybrid CPU-GPU implementation. In [26], Puig et al. utilized a GPU (graphics processing unit) to compute exact gap-affine alignments based on the wavefront alignment (WFA) algorithm. They showed that the proposed tool is up to 29× faster than other GPU implementations.

3. Multiple Sequence Alignment

Numerous studies have employed various techniques to address the challenge of multiple sequence alignments (MSAs) [9]. One widely adopted technique is progressive alignment [32]. The progressive alignment method initially starts with pairwise alignments and progressively builds an MSA alignment through a series of pairwise alignments, producing accurate results for moderately sized sequence sets [33]. In addition to the progressive methods, iterative approaches have also played a crucial role in improving the accuracy of MSA [34]. Iterative methods generally refine alignments applying successive cycles of alignment improvement. Iterative refinement involves realigning sequences based on the initial solution and gradually converging toward a more accurate alignment. Iterative techniques often outperform progressive methods in terms of alignment accuracy, especially in cases where sequences are more distant [35]. Lupyan et al. proposed a hybrid algorithm that combined the progressive and iterative algorithms for MSA. The hybrid approach provided a significant advancement compared to earlier methods involving a notable decrease in computational cost.

In addition to progressive and iterative methods, several studies focus on the utilization of metaheuristics techniques for performing MSA [11]. Ali et al. [36] reviewed the landscape of metaheuristics in bioinformatics highlighting various metaheuristic approaches, including tabu search [37], simulated annealing [38], and particle swarm optimization [39], showcasing their applications in computational biology problems and MSA. Hatzou et al. [9] provided valuable insights centered on the heuristic-based progress of MSA methods. Similarly, Chowdhury et al. [40] offered an overview of MSA methods with a focus on the multi-objective approach. In contrast, Vega et al. [41] provided a comparative analysis of different formulations of multi-objective metaheuristics for MSA. In **Table 2**, the researchers present some well-known tools for MSA and show the general techniques used for each.

Table 2. Multiple sequence alignment tools with techniques.

#	Technique	Approach	Heuristics	Ref.
1	Recursive MAGUS	Divide-and-Conquer Alignment	Guide Tree	[42]
2	ClipKIT	Trimming Strategies	IQ-TREE Hill-Climbing	[43]
3	Kalign	Progressive Alignment	Guide Tree	[44]

#	Technique	Approach	Heuristics	Ref.
4	**ProbCons**	**Probabilistic Consistency**	**Probabilistic Modeling**	[45]
5	**MUSCLE**	**Progressive Alignment**	**Guide Tree**	[46]
6	**MAFFT**	**Progressive Alignment**	**Guide Tree**	[47]
7	**T-Coffee**	**Various**	**Various**	[48]
8	**DIALIGN**	**Local Multiple Alignment**	**Pairwise Alignments**	[49]
9	**CLUSTAL W**	**Progressive Alignment**	**Guide Tree**	[50]

Limited studies have been directed towards seeking exact solutions for multiple sequence alignment due to the time complexity associated with obtaining the optimal results. Mojbak et al. [51] proposed an *exact-MSA* approach using forward dynamic programming. Also, in the comprehensive exploration of exact solutions, Hosseininasab et al. [52] proposed a framework employing a dynamic programming approach to construct a multivalued decision diagram, representing all PSAs. The synchronization of PSAs with the proposed decision diagram effectively incorporates modeling the MSA problem within polynomial space complexity. Moreover, Domínguez [53] delves into statistical and biological concepts employed in the MSAProbs-MPI tool to complete the alignments where high-performance computing techniques are employed for alignment acceleration. Additionally, Ju et al. [54] introduced an end-to-end deep neural network and called it CopulaNe, designed to directly estimate residue co-evolution from MSA, representing a cutting-edge approach in the finding of exact solutions for MSA.

In addition to the previously mentioned approaches, several parallelization strategies have been employed to tackle the challenges associated with MSA [55]. Some of these strategies focus on the parallelization of dynamic programming algorithms, such as in [56]. Other strategies aim to parallelize the progressive alignment [57]. Several studies focus on the parallelization of heuristic algorithms, such as [58]. Recently, many studies utilized GPU acceleration for MSA [59]. The optimization of parallel MSA is characterized by continuous innovation in algorithmic design and adaptation to emerging hardware architectures [55].

References

1. Diab, S.; Nassereldine, A.; Alser, M.; Gómez Luna, J.; Mutlu, O.; El Hajj, I. A framework for high-throughput sequence alignment using real processing-in-memory systems. Bioinformatics 2023, 39, btad155.
2. Needleman, S.B.; Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 1970, 48, 443–453.
3. Smith, T.F.; Waterman, M.S. Identification of common molecular subsequences. J. Mol. Biol. 1981, 147, 195–197.
4. Agrawal, A.; Huang, X. Pairwise statistical significance of local sequence alignment using sequence-specific and position-specific substitution matrices. IEEE/ACM Trans. Comput. Biol. Bioinform. 2009, 8, 194–205.
5. Edgar, R.C.; Batzoglou, S. Multiple sequence alignment. Curr. Opin. Struct. Biol. 2006, 16, 368–373.
6. Bellman, R.E.; Dreyfus, S.E. Applied Dynamic Programming; Princeton University Press: Princeton, NJ, USA, 2015; Volume 2050.
7. Chao, J.; Tang, F.; Xu, L. Developments in algorithms for sequence alignment: A review. Biomolecules 2022, 12, 546.
8. Saeed, F.; Khokhar, A. An Overview of Multiple Sequence Alignment Systems. arXiv 2009, arXiv:0901.2747.
9. Chatzou, M.; Magis, C.; Chang, J.M.; Kemena, C.; Bussotti, G.; Erb, I.; Notredame, C. Multiple sequence alignment modeling: Methods and applications. Briefings Bioinform. 2016, 17, 1009–1023.
10. Zemali, E.a.; Boukra, A. A new hybrid bio-inspired approach to resolve the multiple sequence alignment problem. In Proceedings of the 2016 International Conference on Control, Decision and Information Technologies (CoDIT), Saint Julian’s, Malta, 6–8 April 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 108–113.
11. Amorim, A.R.; Zafalon, G.F.D.; de Godoi Contessoto, A.; Valêncio, C.R.; Sato, L.M. Metaheuristics for multiple sequence alignment: A systematic review. Comput. Biol. Chem. 2021, 94, 107563.
12. Bafna, V.; Lawler, E.L.; Pevzner, P.A. Approximation algorithms for multiple sequence alignment. Theor. Comput. Sci. 1997, 182, 233–244.

13. Nowicki, M.; Bzhalava, D.; BaŁa, P. Massively parallel implementation of sequence alignment with basic local alignment search tool using parallel computing in java library. J. Comput. Biol. 2018, 25, 871–881.

14. Chiaromonte, F.; Yap, V.B.; Miller, W. Scoring pairwise genomic sequence alignments. In Biocomputing 2002; World Scientific: Singapore, 2001; pp. 115–126.

15. Haque, W.; Aravind, A.; Reddy, B. Pairwise sequence alignment algorithms: A survey. In Proceedings of the 2009 Conference on Information Science, Technology and Applications, Sliema, Malta, 11–16 October 2009; pp. 96–103.

16. Edgar, R.C.; Sjölander, K. A comparison of scoring functions for protein sequence profile alignment. Bioinformatics 2004, 20, 1301–1308.

17. Li, H.; Homer, N. A survey of sequence alignment algorithms for next-generation sequencing. Briefings Bioinform. 2010, 11, 473–483.

18. Abbasi, M.; Paquete, L.; Liefooghe, A.; Pinheiro, M.; Matias, P. Improvements on bicriteria pairwise sequence alignment: Algorithms and applications. Bioinformatics 2013, 29, 996–1003.

19. Gotoh, O. An improved algorithm for matching biological sequences. J. Mol. Biol. 1982, 162, 705–708.

20. Pearson, W.R. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzym. 1990, 183, 63–98.

21. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. J. Mol. Biol. 1990, 215, 403–410.

22. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European molecular biology open software suite. Trends Genet. 2000, 16, 276–277.

23. Daily, J. Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments. BMC Bioinform. 2016, 17, 81.

24. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics 2018, 34, 3094–3100.

25. Issa, M.; Hassanien, A.E.; Oliva, D.; Helmi, A.; Ziedan, I.; Alzohairy, A. ASCA-PSO: Adaptive sine cosine optimization algorithm integrated with particle swarm for pairwise local sequence alignment. Expert Syst. Appl. 2018, 99, 56–70.

26. Aguado-Puig, Q.; Marco-Sola, S.; Moure, J.C.; Matzoros, C.; Castells-Rufas, D.; Espinosa, A.; Moreto, M. WFA-GPU: Gap-affine pairwise alignment using GPUs. bioRxiv 2022.

27. Kaur, K.; Chakraborty, S.; Gupta, M.K. Accelerating Smith-Waterman Algorithm for Faster Sequence Alignment using Graphical Processing Unit. In Proceedings of the Journal of Physics: Conference Series; IOP Publishing: Bristol, UK, 2022; Volume 2161, p. 012028.

28. Fakirah, M.; Shehab, M.A.; Jararweh, Y.; Al-Ayyoub, M. Accelerating needleman-wunsch global alignment algorithm with gpus. In Proceedings of the 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), Marrakech, Morocco, 17–20 November 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1–5.

29. Balhaf, K.; Shehab, M.A.; Wala'a, T.; Al-Ayyoub, M.; Al-Saleh, M.; Jararweh, Y. Using gpus to speed-up levenshtein edit distance computation. In Proceedings of the 2016 7th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 5–7 April 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 80–84.

30. Jararweh, Y.; Al-Ayyoub, M.; Fakirah, M.; Alawneh, L.; Gupta, B.B. Improving the performance of the needleman-wunsch algorithm using parallelization and vectorization techniques. Multimed. Tools Appl. 2019, 78, 3961–3977.

31. Shehab, M.A.; Ghadawi, A.A.; Alawneh, L.; Al-Ayyoub, M.; Jararweh, Y. A hybrid CPU-GPU implementation to accelerate multiple pairwise protein sequence alignment. In Proceedings of the 2017 8th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 4–6 April 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 12–17.

32. Sievers, F.; Higgins, D.G. Clustal Omega, accurate alignment of very large numbers of sequences. Mult. Seq. Alignment Methods 2014, 1079, 105–116.

33. Boyce, K.; Sievers, F.; Higgins, D.G. Instability in progressive multiple sequence alignment algorithms. Algorithms Mol. Biol. 2015, 10, 1–10.

34. Wallace, I.M.; Higgins, D.G. Evaluation of iterative alignment algorithms for multiple alignment. Bioinformatics 2005, 21, 1408–1414.

35. Notredame, C. Recent evolutions of multiple sequence alignment algorithms. PLoS Comput. Biol. 2007, 3, e123.

36. Ali, A.F.; Hassanien, A.E. A survey of metaheuristics methods for bioinformatics applications. In Applications of Intelligent Optimization in Biology and Medicine: Current Trends and Open Problems; Springer: Berlin/Heidelberg, Germany, 2015; pp. 23–46.

37. Riaz, T.; Wang, Y.; Li, K.B. Multiple sequence alignment using tabu search. In Proceedings of the Second Conference on Asia-Pacific Bioinformatics, Dunedin, New Zealand, 18–22 January 2004; Volume 29, pp. 223–232.
38. Kim, J.; Pramanik, S.; Chung, M.J. Multiple sequence alignment using simulated annealing. Bioinformatics 1994, 10, 419–426.
39. Xu, F.; Chen, Y. A method for multiple sequence alignment based on particle swarm optimization. In Proceedings of the Emerging Intelligent Computing Technology and Applications. With Aspects of Artificial Intelligence: 5th International Conference on Intelligent Computing, ICIC 2009, Ulsan, Republic of Korea, 16–19 September 2009; Proceedings 5. Springer: Berlin/Heidelberg, Germany, 2009; pp. 965–973.
40. Chowdhury, B.; Garai, G. A review on multiple sequence alignment from the perspective of genetic algorithm. Genomics 2017, 109, 419–431.
41. Zambrano-Vega, C.; Nebro, A.J.; García-Nieto, J.; Aldana-Montes, J.F. Comparing multi-objective metaheuristics for solving a three-objective formulation of multiple sequence alignment. Prog. Artif. Intell. 2017, 6, 195–210.
42. Smirnov, V. Recursive MAGUS: Scalable and accurate multiple sequence alignment. PLoS Comput. Biol. 2021, 17, e1008950.
43. Steenwyk, J.L.; Buida III, T.J.; Li, Y.; Shen, X.X.; Rokas, A. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. PLoS Biol. 2020, 18, e3001007.
44. Lassmann, T.; Sonnhammer, E.L. Kalign—An accurate and fast multiple sequence alignment algorithm. BMC Bioinform. 2005, 6, 298.
45. Do, C.B.; Mahabhashyam, M.S.; Brudno, M.; Batzoglou, S. ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res. 2005, 15, 330–340.
46. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004, 32, 1792–1797.
47. Katoh, K.; Misawa, K.; Kuma, K.i.; Miyata, T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002, 30, 3059–3066.
48. Notredame, C.; Higgins, D.G.; Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 2000, 302, 205–217.
49. Morgenstern, B. DIALIGN 2: Improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 1999, 15, 211–218.
50. Thompson, J.D.; Higgins, D.G.; Gibson, T.J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994, 22, 4673–4680.
51. Mojbak, J.; Pedersen, C. Exact Multiple Sequence Alignment Using Forward Dynamic Programming; Bioinformatics Research Center: Singapore, 2010.
52. Hosseininasab, A.; van Hoeve, W.J. Exact multiple sequence alignment by synchronized decision diagrams. INFORMS J. Comput. 2021, 33, 721–738.
53. González-Domínguez, J. Fast and Accurate Multiple Sequence Alignment with MSAProbs-MPI. In Multiple Sequence Alignment; Springer: Berlin/Heidelberg, Germany, 2021; pp. 39–47.
54. Ju, F.; Zhu, J.; Shao, B.; Kong, L.; Liu, T.Y.; Zheng, W.M.; Bu, D. CopulaNet: Learning residue co-evolution directly from multiple sequence alignment for protein structure prediction. Nat. Commun. 2021, 12, 2535.
55. Almanza-Ruiz, S.H.; Chavoya, A.; Duran-Limon, H.A. Parallel protein multiple sequence alignment approaches: A systematic literature review. J. Supercomput. 2023, 79, 1201–1234.
56. Helal, M.; El-Gindy, H.; Mullin, L.; Gaeta, B. Parallelizing optimal multiple sequence alignment by dynamic programming. In Proceedings of the 2008 IEEE International Symposium on Parallel and Distributed Processing with Applications, Sydney, NSW, Australia, 10–12 December 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 669–674.
57. Hung, C.L.; Lin, Y.S.; Lin, C.Y.; Chung, Y.C.; Chung, Y.F. CUDA ClustalW: An efficient parallel algorithm for progressive multiple sequence alignment on Multi-GPUs. Comput. Biol. Chem. 2015, 58, 62–68.
58. Ishikawa, M.; Toya, T.; Hoshida, M.; Nitta, K.; Ogiwara, A.; Kanehisa, M. Multiple sequence alignment by parallel simulated annealing. Bioinformatics 1993, 9, 267–273.
59. Blazewicz, J.; Frohmberg, W.; Kierzynka, M.; Wojciechowski, P. G-MSA—A GPU-based, fast and accurate algorithm for multiple sequence alignment. J. Parallel Distrib. Comput. 2013, 73, 32–41.