Extended Abstract

André Salgado[1,2][0000−0002−2940−0767], Francisco Fernandes[1][0000−0003−2546−2669], and Ana Teresa Freitas[1,2][0000−0002−2997−5990]

[1] Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento (INESC-ID), R. Alves Redol 9, 1000-029 Lisboa, Portugal
[2] Instituto Superior Técnico (IST/UL), Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal

Abstract. In the realm of Bioinformatics, the comparison of DNA sequences is essential for tasks such as phylogenetic identification, comparative genomics, and genome reconstruction. Methods for estimating sequence similarity have been successfully applied in this field. The application of these methods to circular genomic structures, common in nature, poses additional computational hurdles. In the advancing field of metagenomics, innovative circular DNA alignment algorithms are vital for accurately understanding circular genome complexities. Aligning circular DNA, more intricate than linear sequences, demands heightened algorithms due to circularity, escalating computation requirements and runtime. This paper proposes CSA-MEM, an efficient text indexing algorithm to identify the most informative region to rotate and cut circular genomes, thus improving alignment accuracy. The algorithm uses a circular variation of the FM-Index and identifies the longest chain of non-repeated maximal subsequences common to a set of circular genomes, enabling the most adequate rotation and linearisation for multiple alignment. The effectiveness of the approach was validated in five sets of mitochondrial, viral and bacterial DNA. The results show that CSA-MEM significantly improves the efficiency of multiple sequence alignment, consistently achieving top scores compared to other state-of-the-art methods. This tool enables more realistic phylogenetic comparisons between species, facilitates large metagenomic data processing, and opens up new possibilities in comparative genomics.

Keywords: Circular DNA · Multiple Alignment · Text Indexing.

1 Introduction

Recent advances in metagenomics have propelled the field to new frontiers, allowing researchers to explore microbial communities with unprecedented depth and breadth [12]. However, as datasets become larger and more intricate, the challenge of addressing circular DNA alignment, efficient data processing, accurate taxonomic classification, and integration of multiomics data is of paramount importance [10]. Overcoming these challenges will not only enhance our understanding of microbial ecosystems, but also pave the way for innovative applications in fields such as biotechnology, environmental science, and personalised medicine [6].

With the expansion in scope and scale of the metagenomics field comes the urgent need to tackle the challenges posed by the analysis of large metagenomic datasets. One of the main challenges is the efficient alignment of circular DNA within these complex datasets. Circular genomes are common in many microorganisms, and accurately aligning them is crucial for understanding their structure and function. Circular DNA molecules, known as plasmids, play an important role in conferring adaptive advantages, such as antibiotic resistance and virulence, to bacteria [5]. Circular mitochondrial DNA plays an essential role in the survival and energy production of eukaryotic cells [17] and has long been used for phylogenetic analyses [16]. Smaller structures known as extrachromosomal circular DNA are considered a hallmark of genomic flexibility in eukaryotes [20]. Furthermore, the circular nature of DNA in some viruses has a major impact on their replication and infection strategies [19]. Understanding these complex mechanisms is crucial for the development of antiviral therapies and vaccines [22].

Existing DNA alignment algorithms, often designed for linear genomes, such as ClustalW [18], may struggle to handle the unique characteristics of circular DNA, leading to misinterpretations and inaccuracies in the analysis [21]. Furthermore, as metagenomic datasets increase in size and complexity, issues related to data management and analysis, processing speed, and computational resources become increasingly pressing. Efficient algorithms are needed to address challenges related to assembly quality, binning, and functional annotation, which are vital for extracting meaningful biological information from the sheer volume of metagenomic data [15].

The special importance of circular DNA presents a unique challenge in comparing its sequences. Because circular sequences can start from any point, it adds complexity. This distinctiveness feature makes traditional linear-centric multiple alignment algorithms inadequate because they lack the adaptability to effectively cope with the inherent circular structure. The outcome is a potential loss of critical genetic information during the alignment process [8].

Beyond identifying the optimal rotation for each sequence in a multiple circular DNA alignment, it is also necessary to address the challenge of handling large volumes of data. In this context, it is necessary to develop new algorithms that provide better solutions in terms of space and time efficiency.

Efforts to reconcile the circular-to-linear disparity have given rise to remarkable methods. Cyclope [14] is a software designed to enhance the alignment of multiple circular sequences. However, the cubic runtime of the pairwise alignment step becomes a limiting factor in practical scenarios. CSA [8] is an algorithm based on a circular version of a generalised suffix tree. The algorithm identifies the largest chain of non-repeated longest subsequences common to a set of circular DNA sequences to determine their optimal rotations. Although very efficient, it is limited to 32 input sequences and relies on an outdated suffix tree data structure.

Other types of methods include BEAR [2], which extends existing algorithms for circular and fixed-length approximate string matching [3]. It calculates edit

distances and rotations between all pairs of sequences and then uses agglomerative hierarchical clustering to determine the most suitable rotations. A similar and more recent approach, MARS [1], is an heuristic method that computes all pairwise cyclic edit distances using a distance measurement algorithm based on q-grams [10]. It then performs classic progressive alignment of sequence profile pairs using a guide tree to refine the rotations. However, this progressive nature and the dependency on dynamic programming algorithms may render these last methods slower and less efficient when dealing with longer sequences or larger datasets.

1.1 Contribution

To effectively tackle the challenges associated with circular multiple sequence alignment on large datasets, we introduce CSA-MEM, in which we propose: (1) the adaptation of advanced data structures such as the FM-Index [9], namely a circular modification based on the implementation used in slaMEM [7] to achieve a computationally efficient exact solution for circular sequence matching, and (2) an effective identification of the longest chain of non-repeated maximal exact matches (MEMs) common to a set of circular DNA sequences, in an approach similar to the CSA tool [8]. This way, the CSA-MEM algorithm allows for a seamless rotation and linearisation process for multiple circular alignment purposes.

2 Methods

2.1 Basic Notions

We generally consider that a text is appended with a terminator sentinel symbol '\$' which is lexicographically smaller than all other characters in its alphabet Σ. For a string T of length n, the Suffix Array (SA) [13] of T is an array of integers $SA[1, n]$ where each element $SA[i]$ corresponds to the starting position in T of the i-th lexicographically smallest suffix of the string. $SA[i]$ points to the position in string T where the suffix with lexicographic rank i begins.

The Longest Common Prefix array LCP is an integer array of length n which stores information about the length of the longest common prefixes (lcp) between consecutive suffix pairs in SA [11]. It is defined as $LCP[i] = lcp(T[SA[i - 1], n], T[SA[i], n])$ for $i \neq 1$ and 0 otherwise.

The Burrows-Wheeler Transform (BWT) [4] is a data structure that consists of a reversible transformation that rearranges the original characters of T into a new string more suitable to text processing and data compression methods. In the conceptual matrix of all the lexicographically sorted rotations of a string, the BWT matches its last column L. This corresponds to the character immediately preceding each suffix starting at position $SA[i]$ in the string T and is formally defined as $L[i] = T[SA[i] - 1]$ when $SA[i] \neq 1$ and $L[i] =$ \$ otherwise.

The FM-Index [9] is an indexing data structure built on top of the BWT which can be used to search and process large volumes of text efficiently. In

addition to the BWT and SA arrays, the FM-index also uses a summation array which stores the *rank* of each character in the alphabet, meaning the number of occurrences of character c in the BWT up to position i, and represented by the function $rank_c(L, i)$. The inverse operation, $select_c(L, j)$, returns the position i in the BWT corresponding to the j-th occurrence of character c. To *locate* and *count* the occurrences of a specific pattern P, the FM-index employs a backward search strategy and maintains two pointers, identifying the start and end index positions of runs of consecutive suffixes starting with the current matched string, which are updated by iteratively applying the *LF-mapping* procedure [9].

Maximal Exact Matches (MEMs) are substrings which simultaneously belong to both a reference text T and a query text R, and that cannot be extended in either direction without producing a mismatch, i.e. $R[i, j] = Q[i', j']$ and $R[i-1] \neq Q[i'-1] \wedge R[j+1] \neq Q[j'+1]$. This type of subtrings is often used in genomic comparison as they provide common blocks between the sequences that can be used as anchors to detect similar regions in the alignments. Such MEMs can be retrieved by matching query Q over the FM-Index of text R using an efficient algorithm, such as slaMEM[7].

2.2 The approach

CSA-MEM is organised into a meticulous three-step process with the primary aim of identifying and extracting the longest common chain of nucleotides from a comprehensive sequence dataset. It should be noted that the CSA algorithm [8] has previously established the state-of-the-art accuracy in finding the starting positions of this chain to improve the alignment of multiple circular sequences. However, CSA-MEM advances the field by removing CSA's limitations regarding both sequence number and size by using more efficient indexing data structures and searching for maximal exact matches.

The first step of CSA-MEM involves the building of specialised indexing data structures. This is achieved by designing an FM-Index variant modified to excel in detecting circular DNA patterns. After establishing this foundational step, a thorough search over the index is conducted to identify all Maximal Exact Matches (MEMs) that are common across all sequences. These MEMs serve as universally applicable puzzle pieces for each sequence. Subsequently, these puzzle pieces are systematically organised to obtain the longest common chain of nucleotides, which will define a new cutting region which optimises the alignment of the sequences.

2.3 Circular FM-Index

Classified: This research was presented at the **International Symposium on Bioinformatics Research and Applications (ISBRA)** held in Poland and subsequently published in the corresponding proceedings of the **Springer Lecture Notes in Computer Science book series** (LNBI, volume 14248).

2.4 Most Significant Common Subsequence Chain of MEMs

Classified: This research was presented at the **International Symposium on Bioinformatics Research and Applications (ISBRA)** held in Poland and subsequently published in the corresponding proceedings of the **Springer Lecture Notes in Computer Science book series** (LNBI, volume 14248).

3 Results and Discussion

Classified: This research was presented at the **International Symposium on Bioinformatics Research and Applications (ISBRA)** held in Poland and subsequently published in the corresponding proceedings of the **Springer Lecture Notes in Computer Science book series** (LNBI, volume 14248).

4 Conclusion

Understanding the genetic relationships and evolutionary history encoded in circular DNA molecules has a profound impact on human health, environmental studies, and understanding the complexity and diversity of life. The CSA-MEM tool demonstrates that the use of efficient indexing data structures and string matching algorithms for circular sequences yields superior benchmark scores with minimal computational demands. This novel approach not only enhances rotation strategies, but also optimises space and time complexities, enabling a more thorough analysis of circular DNA sequences. Future work will include building a database with various datasets for circular genomes that can be used to characterise the boundaries of the algorithms tested in comparative genomics studies. The software source code, scripts, and datasets used in this work are available for download at: https://github.com/andre99salgado/CSA-MEM .

Acknowledgements

The authors acknowledge the support of Fundação para a Ciência e a Tecnologia, projects PRELUNA (Grant PTDC/CCIINF/4703/2021) and UIDB/50021/2020.

References

1. Ayad, L.A., Pissis, S.P.: Mars: improving multiple circular sequence alignment using refined sequences. BMC genomics **18**(1), 1–10 (2017)
2. Barton, C., Iliopoulos, C.S., Kundu, R., Pissis, S.P., Retha, A., Vayani, F.: Accurate and efficient methods to improve multiple circular sequence alignment. In: Bampis, E. (ed.) Experimental Algorithms. pp. 247–258. Springer (2015)
3. Barton, C., Iliopoulos, C.S., Pissis, S.P.: Fast algorithms for approximate circular string matching. Algorithms for Molecular Biology **9**, 1–10 (2014)
4. Burrows, M.: A block-sorting lossless data compression algorithm. SRS Research Report **124** (1994)

5. Carattoli, A.: Plasmids and the spread of resistance. International Journal of Medical Microbiology **303**(6), 298–304 (2013)
6. Dulanto, C.A., Dekker, J.P.: From the Pipeline to the Bedside: Advances and Challenges in Clinical Metagenomics. The Journal of Infectious Diseases **221**(Supplement 3), S331–S340 (2019)
7. Fernandes, F., Freitas, A.T.: slamem: efficient retrieval of maximal exact matches using a sampled lcp array. Bioinformatics **30**(4), 464–471 (2014)
8. Fernandes, F., Pereira, L., Freitas, A.T.: Csa: an efficient algorithm to improve circular dna multiple alignment. BMC bioinformatics **10**(1), 1–13 (2009)
9. Ferragina, P., Manzini, G.: Opportunistic data structures with applications. In: Proceedings 41st annual symposium on foundations of computer science. pp. 390–398. IEEE (2000)
10. Grossi, R., Iliopoulos, C.S., Mercas, R., et al.: Circular sequence comparison: algorithms and applications. Algorithms Molecular Biology **11**(12) (2016)
11. Gusfield, D.: An "increment-by-one" approach to suffix arrays and trees. Report. CSE-90-39, Computer Science Division, University of California, Davis (1990)
12. Laudadio, I., Fulc, V., Stronati, L., Carissimi, C.: Next-generation metagenomics: Methodological challenges and opportunities. OMICS **23**(7), 327–333 (2019)
13. Manber, U., Myers, G.: Suffix arrays: a new method for on-line string searches. siam Journal on Computing **22**(5), 935–948 (1993)
14. Mosig, A., Hofacker, I.L., Stadler, P.F.: Comparative analysis of cyclic sequences: Viroids and other small circular rnas. Lecture Notes in Informatics. Proceedings German Conference on Bioinformatics (2006)
15. Pan, S., Zhao, X.M., Coelho, L.P.: SemiBin2: self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing. Bioinformatics **39**(Supplement 1), i21–i29 (2023)
16. Pereira, L., Freitas, F., Fernandes, V., Pereira, J.B., Costa, M.D., Costa, S., Máximo, V., Macaulay, V., Rocha, R., Samuels, D.C.: The diversity present in 5140 human mitochondrial genomes. The American Journal of Human Genetics **84**(5), 628–640 (2009)
17. Pohjoismäki, J.L.O., Goffart, S.: Of circles, forks and humanity: Topological organisation and replication of mammalian mitochondrial dna. BioEssays **33**(4), 290–299 (2011)
18. Thompson, J.D., Gibson, T.J., Higgins, D.G.: Multiple sequence alignment using clustalw and clustalx. Current protocols in bioinformatics (1), 2–3 (2003)
19. Tisza, M.J., Pastrana, D.V., Welch, N.L., Stewart, B., Peretti, A., Starrett, G.J., Pang, Y.Y.S., Krishnamurthy, S.R., Pesavento, P.A., McDermott, D.H., et al.: Discovery of several thousand highly diverse circular dna viruses. Elife **9** (2020)
20. Yang, L., Jia, R., Ge, T., Ge, S., Zhuang, A., Chai, P., Fan, X.: Extrachromosomal circular DNA: biogenesis, structure, functions and diseases. Signal transduction and targeted therapy **7**(1), 342 (2022)
21. Zhang, Y., Zhang, Q., Zhou, J., Zou, Q.: A survey on the algorithm and development of multiple sequence alignment. Briefings in Bioinformatics **23**(3) (2022)
22. Zhao, L., Rosario, K., Breitbart, M., Duffy, S.: Chapter three - eukaryotic circular rep-encoding single-stranded dna (cress dna) viruses: Ubiquitous viruses with small genomes and a diverse host range. Advances in Virus Research, vol. 103, pp. 71–133. Academic Press (2019)