LCSKPOA: Enabling banded semi-global partial order alignments via efficient and accurate backbone generation through extended LCSK++

Minindu Weerakoon[1*], Christopher T Saunders[2], Haynes Heaton[1*]

[1] Auburn University, Auburn, AL 36849

[2] Pacific Biosciences, Menlo Park, CA 94025

* Correspondence: haynesheaton@auburn.edu, wmw0016@auburn.edu;

Abstract

Background

Most multiple sequence alignment and string-graph alignment algorithms focus on global alignment, but many applications exist for semi-global and local string-graph alignment. Long reads require enormous amounts of memory and runtime to fill out large dynamic programming tables. Effective algorithms for finding the backbone and thus defining a band of an alignment such as the longest common subsequence with kmer matches (LCSk++) exist but do not work with graphs. This study introduces an adaptation of the Longest Common Subsequence with kmer matches (LCSk++) algorithm tailored for graph structures, particularly focusing on Partial Order Alignment (POA) graphs. POA graphs, which are directed acyclic graphs, represent multiple sequence alignments and effectively capture the relationships between sequences. Current state of the art methods like ABPOA and SPOA, while improving POA, primarily focus on global alignment and thus are limited in local and semi-global banding scenarios. Our approach addresses these limitations by extending the LCSk++ algorithm to accommodate the complexities of graph-based alignment.

Results

Our extended LCSk++ algorithm integrates dynamic programming and graph traversal techniques to detect conserved regions within POA graphs, termed the LCSk++ backbone. This backbone enables precise banding of the POA matrix for local and semi-global alignment, significantly enhancing the construction of consensus sequences. Compared to unbanded semi-global POA, our method demonstrates substantial memory savings (up to 98%) and significant run-time reductions (up to 37-fold), particularly for long sequences. The method maintains high alignment scores and proves effective across various string lengths and datasets, including synthetic and PacBio

HiFi reads. Parallel processing further enhances runtime efficiency, achieving up to 150x speed improvements on conventional PCs.

Conclusion

The extended LCSk++ algorithm for graph structures offers a substantial advancement in sequence alignment technology. It effectively reduces memory consumption and optimizes run times without compromising alignment quality, thus providing a robust solution for local and semi-global alignment in POA graphs. This method enhances the utility of POA in critical applications such as multiple sequence alignment for phylogeny construction and graph-based reference alignment .

Keywords: LCSKPOA, graph sequence alignment, multiple sequence alignment, partial order alignment, LCSK++, PacBio HiFi

Background

Graph-based aligners, such as GraphAligner[1], have become increasingly important in genomics and bioinformatics due to their ability to provide a more flexible and comprehensive representation of genetic variation compared to linear sequence-based reference mapping techniques like minimap2[2], HISAT2[3], and BWA[4]. These graph-based reference alignment techniques decrease reference bias and enhance the accuracy of variant detection and genome assembly. Graph-based reference alignment, despite its advantages, is much slower than linear sequence-based methods. Since the creation of graph-based references depends on multiple sequence alignment, speeding up and improving multiple sequence alignment will help reduce the performance gap. While multiple sequence alignment is a well-established problem with many publications addressing and improving it, challenges still exist.

This study introduces a novel adaptation of the Longest Common Subsequence with k-mer matches (LCSk++) algorithm specifically designed for graph structures, with a focus on enabling banded local and semi-global Partial Order Alignment (POA) graphs [5]. POA graphs are directed acyclic graphs that represent multiple sequence alignments, capturing the relationships between sequences and accommodating variations such as insertions, deletions, and substitutions more flexibly and comprehensively than traditional alignment methods.

Existing methods like ABPOA [6] and SPOA [7] improve POA by using greedy banding and computing the full matrix with use of SIMD optimizations, respectively. However, these methods are limited and cannot effectively apply to local and semi-global alignment, which are crucial for many alignment use cases. Our approach extends the LCSk++ algorithm [8, 9] to address these complexities, enabling efficient banded semi-global POA.

Our approach extends the LCSk++ algorithm, originally developed for linear string sequences, to handle the intricacies of graph data structures. By integrating dynamic programming and graph traversal techniques, we enable the detection of conserved regions within POA graphs. This conserved region, which we call the LCSk++ backbone, allows us to band the POA matrix to accurately locally and semi-globally align sequences. This method significantly enhances the construction of consensus sequences with high accuracy and efficiency. Additionally, it allows for other critical applications of POA, such as multiple sequence alignment (MSA) for phylogeny construction and graph-based reference alignment.

Significantly, our method achieves a 37-fold increase in speed and a 98% reduction in memory usage for sufficiently long sequences. Additionally, we can achieve up to 150x speed improvements on conventional PCs through parallel processing without compromising alignment quality, a feat that is not possible with unbanded POA. This advancement allows for fast and efficient local and semi-global alignment on the banded POA graph, enhancing its utility in bioinformatics.

Overall, the extended LCSk++ algorithm for graph structures represents a significant step forward in sequence alignment technology, offering improved performance and accuracy for the analysis of complex biological datasets.

Implementation

Basic dynamic programming

Finding the optimal graph to sequence LCSK++ path can be done using basic dynamic programming recursive relation similar to that of [8] with few modifications. In our implementation, we adopt the definitions as described by [10] with some additional definitions to accommodate the graph structure as presented below.

Definitions:

dp(i, j) = entry in dynamic programming tables row i and column j, where i is the ith topologically ordered node in the graph G and j is the jth entry in query Y.

$p^{i,m}$ = predecessor node which is m edges back of the ith topologically ordered node in graph G.

$$dp(i, j) = \max \begin{cases} 0 \\ dp(i, j - 1) \\ dp(p^{i,1}, j) \\ dp(p^{i,q}, j - q) + q \text{ for all } q \geq k \text{ s.t. } G_{piq..pi1} = Y_{j-q..j-1} \end{cases}$$

The time complexity of this algorithm is O(nm * min(n, m)) where n is query length and m is number of nodes in graph G. Performing basic dynamic programming and obtaining the LCSk++ path for banded POA is not a viable solution, as the time complexity of basic dynamic programming for LCSk++ alone is higher than that of full POA. An example of the algorithm is present in the supplement.

Efficient algorithm

An efficient algorithm presented by [8] achieves a significant speed-up from the basic dynamic programming algorithm and is a crucial component in banded sequence-to-sequence dynamic programming alignment[11]. We propose a query-to-graph version that could achieve similar improvements in partial order alignment. The definition of a match pair is described as follows to accommodate a graph structure, modified from the original definition [10].

Definitions:

$s^{i,m}$ = successor node which is m edges in front of the ith topologically ordered node in graph G.

$$kMatch(i, j) = \begin{cases} 1 \text{ if } G_{si,f} = Y_{j+f}, \text{ for every } 0 <= f <= k - 1 \\ 0 \text{ otherwise} \end{cases}$$

If kMatch(i, j) = 1, we call (i, j) a match pair. (i, j) is called the start of the match pair and ($s^{i,k-1}$, j + k - 1) is called the end of the match pair, this in contrast to the end match pair of the original definition as there could be multiple ends if ($s^{i,k}$, j + k) is used. We adjust the algorithm to incorporate this change. (Fig. 1) shows an example of the efficient algorithm.

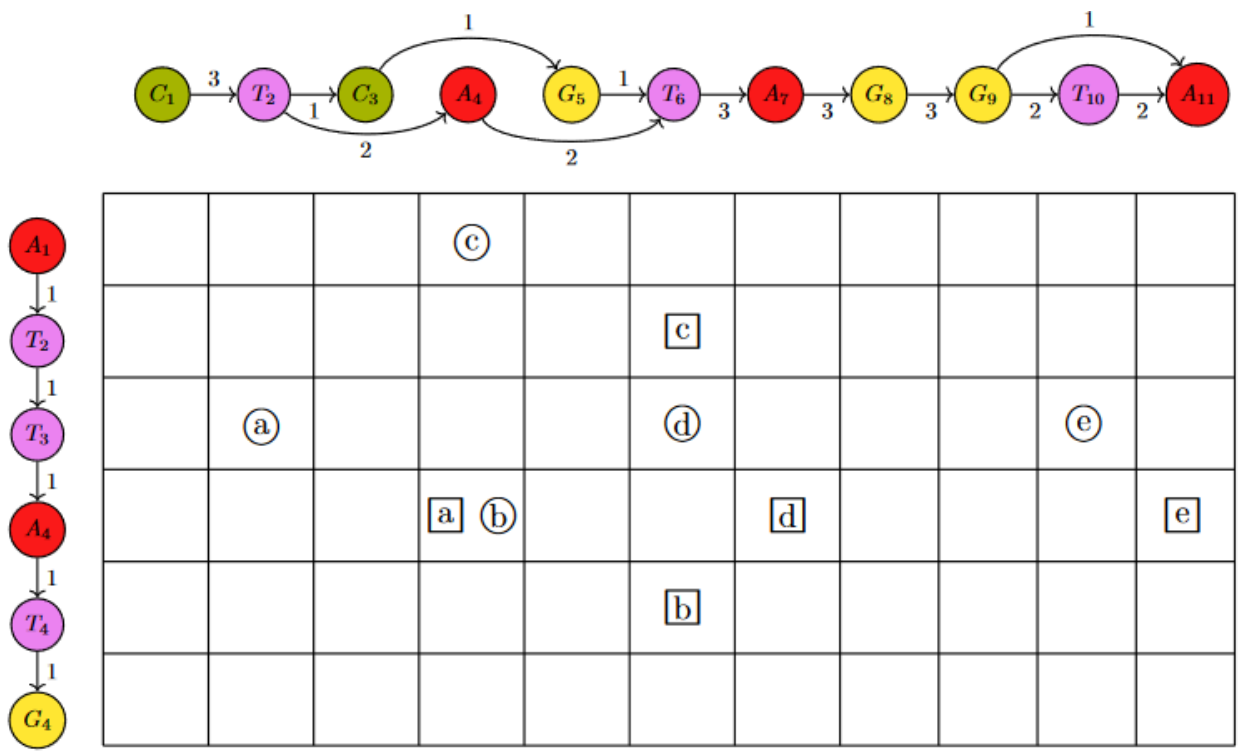

Fig. 1 Matching kmers in query to graph, k = 2; 5 kmer match pairs are present denoted a to e. Starts are presented by the circles and ends by squares. “b continues a”, “c precedes e” applies.

The formula for the efficient algorithm is as follows,

$$dp(P) = \begin{cases} k \\ \max_G dp(G) + 1 \text{ over all G preceding P and path available} \\ dp(G) + 1 \text{ if P continues G and share same path} \end{cases}$$

kmer match pair P can start its k-common subspace, or extend a k-common subspace if there is a path available from start of P to end of G or extend a k-common subsequence by enlarging k if the continuing nodes are on the same path.

Efficient algorithm implementation

Algorithm 1 Sequence to graph LCSK++ computation

Require:
Query X
Partial order graph
Sequences $Seq_{i\epsilon 0..Num}$

1: Seq_paths $\leftarrow$ find the $Seq_{i\epsilon 0..Num}$ paths in graph
2: MaxColDp $\leftarrow$ 2D array filled with n zeros * Num, where n is the number of nodes in graph
3: MatchPairs $\leftarrow$ find all k-match pairs between X and $Seq_{i\epsilon 0..Num}$, save the corresponding paths $paths[]$ and graph nodes j
4: events $\leftarrow$ all starts, ends, and paths of Match Pairs sorted in row-major order, if some start S = (i_S, j_S, $paths[]_S$) and some end E = (i_E, j_E, $paths[]_E$) share the same indices, E should come first
5: **for** all event ϵ events **do**
6: **if** event is a start P = (i_p, j_p, $paths[]_p$) **then**
7: **for** for all path ϵ $paths[]_p$ do **do**
8: dp(P) $\leftarrow$ k + $\max_{x\epsilon 0...jP}$ MaxColDp[path][x]
9: **end for**
10: **else if** event is an end P = (i_p + k, j_p + k, $paths[]_p$) **then**
11: **if** $\exists$G s.t P continues G and share same path **then**
12: dp(P) $\leftarrow$ maxdp(P), dp(G) + 1
13: **end if**
14: **for** for all path ϵ $paths[]_p$ do **do**
15: MaxColDp[path][j_p + k] $\leftarrow$ maxMaxColDp[path][j_p + k], dp(P)
16: **end for**
17: **end if**
18: **end for**
19: **return** $max_p dp(p)$

Definitions:

V vertices in graph

E edges in graph

Num number of sequences

n query length

m max sequence length

r number of kmer matches between query and sequences

The complexity of various operations can be analyzed as follows. Finding sequence paths in the graph has a time complexity of O(Num * (V + E)). Finding match pairs involves hashing the query in O(n) and matching in O(Num * m), resulting in O(n + Num * m). Sorting events has a complexity of O(r * log(r)). Lines 7 to 15 are handled by Num Fenwick trees, which operate in O(Num * log n). Processing all events takes O(r * Num * log n). Consequently, the total time complexity is O(Num * (V + E) + n + Num * m+ r log r + r * Num * log n), and the total space complexity is O(r). By integrating sequence finding in the graph using DFS while recording nodes during the POA, the time complexity can be reduced to O(n + Num * m + r log r + r * Num * log n), significantly

enhancing performance when aligning a large number of sequences in the POA graph.

Using LCSK++ backbone in POA

After obtaining the LCSk++ path, which is the path with the most k-mers, we need an optimal method to use this path to create the bands for the POA dynamic programming matrix. Initially, we band the entire section until the first k-mer is encountered. Similarly, we band from the last k-mer to the end, as illustrated by the yellow-colored region in (Fig. 2). This ensures the region necessary for semi-global and local alignment is present. Second, for the k-mer matches, we create a specified band around them, indicated by the orange region in (Fig. 2). Third, in the intersections between k-mer matches, we band the entire area, as shown in the purple region in (Fig. 2). The band size is chosen based on the sequence length; for most PacBio datasets, a band size of 200 has proven effective.

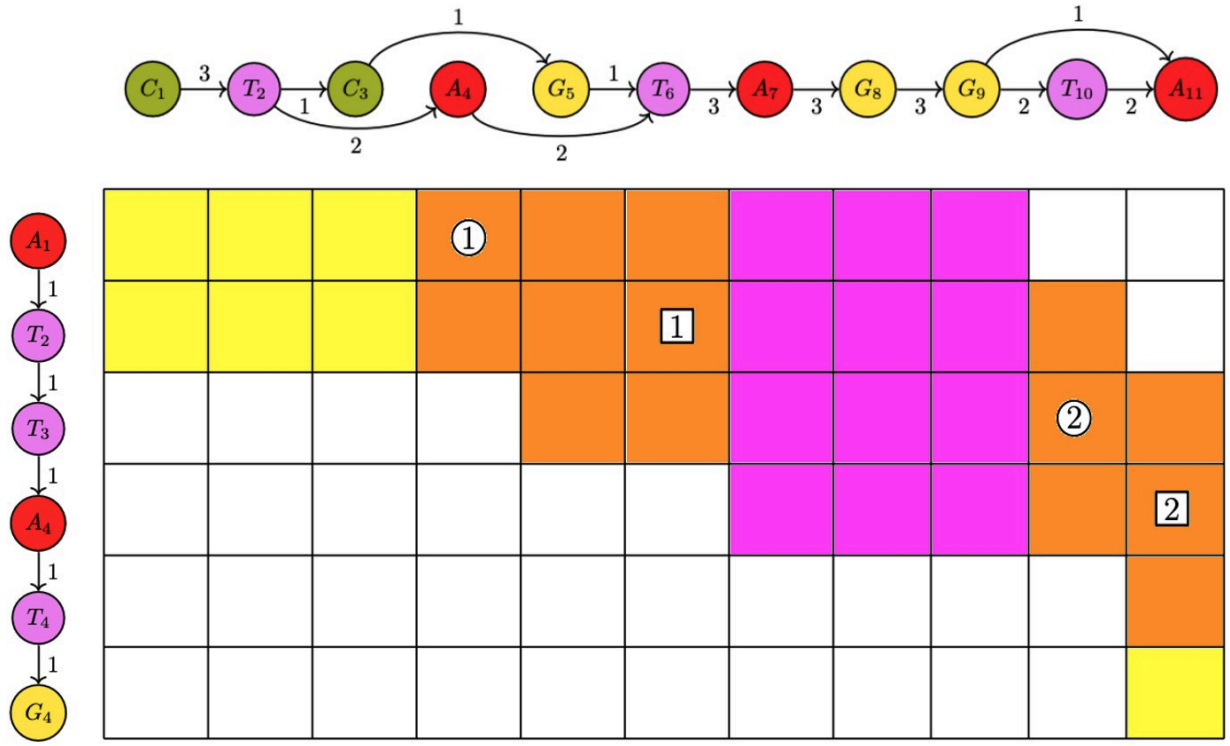

Fig.2 Banding using LCSK++ path, yellow represents the start end banding, orange represents the LCSK++ path anchor banding, and purple indicates the intermediate section banding.

To further optimize runtime and resource efficiency, parallel processing is employed. The graph and query sequences are divided into smaller sections based on anchors or nodes that support all previous sequences along the LCSk++ path. These anchors serve as partitioning points, enabling independent processing of each section through

parallel POA computations. This method leverages modern computational capabilities to achieve significant speed improvements without sacrificing alignment quality, making it particularly beneficial for large-scale genomic datasets.

Results

The synthetic data set was generated by the custom sequence generator for this project. The custom code introduces variations like insertions, deletions, and substitutions, simulating realistic genetic changes for testing our methods. The PacBio dataset used in this study was obtained from real sequenced data, the same as that utilized in topoqual [12].

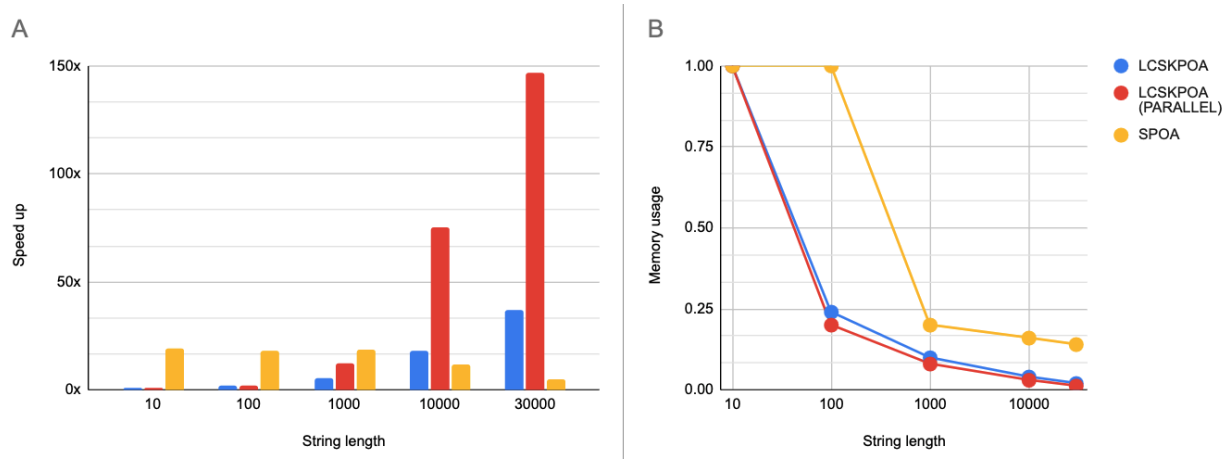

Fig.3 Resulting graphs of running LCSKPOA with Synthetic data. A.speed up of SPOA, LCSKPOA and LCSKPOA parallel compared with vanilla semi-global POA B. Memory usage of SPOA, LCSKPOA and LCSKPOA parallel compared with vanilla semi-global POA.

(Fig. 3A) shows the speed-up factors for LCSkPOA, LCSkPOA (PARALLEL), and SPOA compared to vanilla unbanded POA across different string lengths. LCSkPOA (PARALLEL) achieves dramatic speed increases, reaching up to 150x for the longest strings, while LCSkPOA shows modest improvements. SPOA achieves high speed-up for short sequences but remains low for long sequences. (Fig. 3B) illustrates memory usage for the same methods, with LCSkPOA and its parallel version displaying steep declines in memory consumption for longer strings and stabilizing at low levels. In contrast, SPOA initially uses high memory but reduces significantly for longer strings due to architectural and language optimizations. However, LCSkPOA (PARALLEL) still maintains memory usage that is tenfold lower than SPOA for a string length of 30,000. These results highlight the efficiency of

the LCSkPOA algorithm, especially with parallel processing for longer sequences, in providing substantial speed-up and maintaining low memory usage, making it a robust solution for local and semi-global alignment in Partial Order Alignment (POA) graphs.

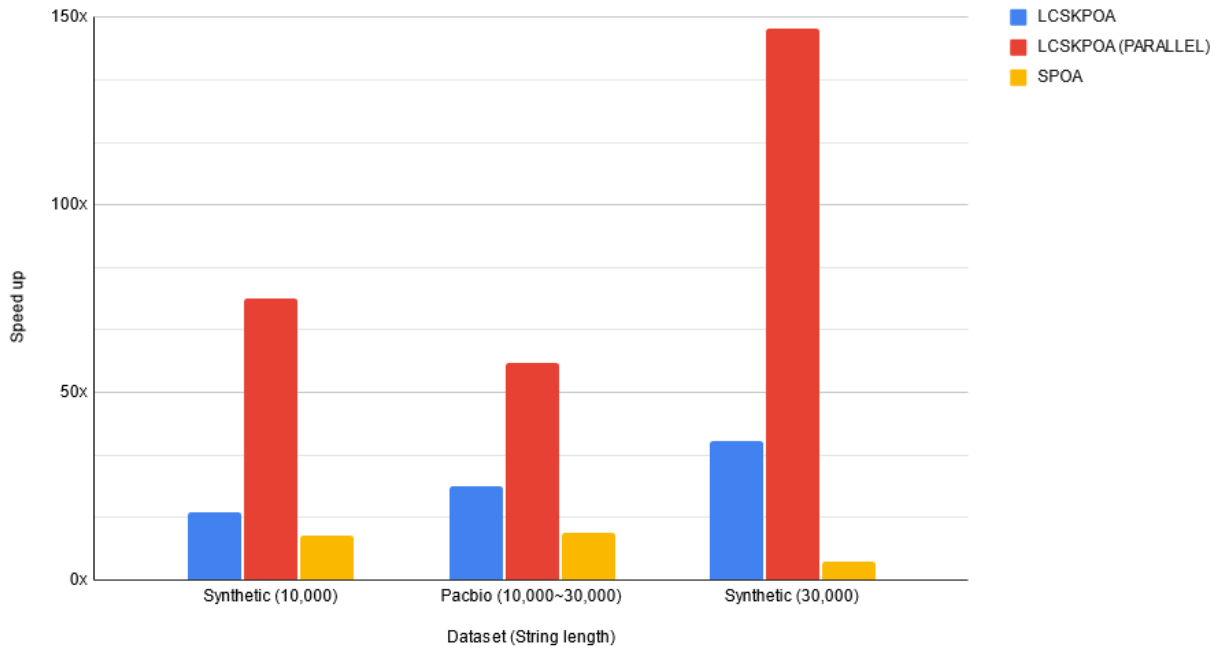

Fig.4 Comparison of Speed up of LCSKPOA and LCSKPOA parallel when run with synthetic dataset and PacBio HiFi data set.

In (Fig. 4), the single-threaded LCSKPOA exhibits a linear speedup from synthetic datasets of length 10,000 to PacBio HiFi datasets of length 10,000 to 30,000 and to synthetic datasets of length 30,000. However, the multi-threaded LCSKPOA does not demonstrate linear scaling for the PacBio HiFi datasets which are around 10,000 ~ 30,000 in length. This discrepancy is due to the presence of edges that span the majority of the graph in PacBio HiFi datasets as a result of the multiple repeating sections in the human genome, making it more challenging to find suitable anchor points for partitioning. As a result, parallel processing is less effective for PacBio HiFi datasets compared to synthetic datasets. SPOA performance becomes weaker as the string lengths increase.

Table. 1 Comparison of run time, memory usage, and accuracy with vanilla semi-global poa with LCSKPOA with different K values (Averages of 20 runs, Synthetic data and PacBio HiFi data)

	Unbanded POA			**LCSKPOA (Parallel)**			
String length	**Score**	**Run time**	**Memory**	**k**	**Score**	**Run time**	**Memory**
10	16	4µs	8KB	2	16 (100%)	21µs (100%+)	4KB (50.8%)
				4	16 (100%)	14µs (100%+)	4KB (50.8%)
100	142	458µs	462KB	4	142 (100%)	499.85µs (100%+)	86KB (18.7%)
				8	128 (98.6%)	326µs (71.2%)	107KB (23.2%)
1,000	1,387	46ms	40MB	4	1,300 (100%)	4.3ms (9.3%)	8.7MB (21.7%)
				8	1,362 (99.1%)	3.7ms (8.1%)	2.5MB (6.25%)
10,000	13,944	4.6s	5.3GB	8	13,945 (100%)	807.02ms (17.5%)	182MB (3.4%)
				12	13,820 (99.1%)	64.5ms (1.4%)	230MB (4.3%)
30,000	41,715	41.3s	55GB	8	41,715 (100%)	1,720ms (4.1%)	827MB (1.55%)
				12	41,351 (99.1%)	301ms (0.7%)	694MB (1.25%)
10,000~30,000 (PacBio HiFi)	29,165	17s	18.8GB	8	29,458 (100%)	3,960ms (23.2%)	677MB (3.76%)
				12	29,458 (100%)	400ms (2.3%)	299MB (1.58%)

The table compares vanilla unbanded semi-global POA and LCSkPOA (Parallel) across various string lengths, emphasizing the impact of the parameter k on scores, run times, and memory usage. For short string lengths, such as 10 and 100, vanilla POA achieves high scores with low run times and minimal memory usage. In comparison, LCSkPOA (Parallel) also achieves high scores, but with slightly higher run times and significantly lower memory usage. As the string length increases to 1,000, vanilla POA's run time and memory usage grow substantially, whereas LCSkPOA (Parallel) maintains high scores and demonstrates significant efficiency

improvements. Notably, as k increases from 4 to 8, the run time decreases from 4.3ms to 3.7ms, indicating no significant gain in using higher k values for short and medium-length sequences.

For longer string lengths such as 10,000 and 30,000, the benefits of LCSkPOA (Parallel) become more evident. vanilla POA exhibits a drastic increase in both run time and memory usage, reaching 4.6s and 5.3GB for 10,000 strings, and 41.3s and 55GB for 30,000 strings. Conversely, LCSkPOA (Parallel) with higher k values significantly reduces run time and memory usage. For example, at a string length of 10,000 and k=12, the run time is 64.5ms, reduced from 807ms at k=8, while maintaining a high score of 99.1%. At 30,000 strings and k=12, LCSkPOA (Parallel) achieves a run time of 301ms, reduced from 1,720ms at k=8, showcasing a stark contrast to vanilla POA.

For a real sequencing dataset, PacBio HiFi (10,000~30,000), LCSkPOA (Parallel) with k=12 achieves a run time of 400ms, reduced from 3,960ms at k=8, maintaining a perfect score of 100%. This comparison clearly highlights the significant impact of increasing k in LCSkPOA (Parallel) for long sequences, consistently improving run time and memory efficiency while maintaining near perfect alignment scores, making it highly effective for large-scale sequence alignment tasks.

Conclusion

In this study, we introduced an innovative adaptation of the Longest Common Subsequence with k-mer matches (LCSk++) algorithm, tailored specifically for graph structures and particularly for Partial Order Alignment (POA) graphs. POA graphs, which are directed acyclic graphs, effectively represent multiple sequence alignments and capture the relationships between sequences with more flexibility than traditional alignment methods, accommodating variations such as insertions, deletions, and substitutions.

Our approach extends the LCSk++ algorithm, originally designed for linear string sequences, to manage the complexities of graph data structures by integrating dynamic programming and graph traversal techniques. This extension allows for the detection of conserved regions within POA graphs, referred to as the LCSk++ backbone, which facilitates banding of the POA matrix for accurate local and semi-global alignment. This method significantly

improves the construction of consensus sequences with high accuracy and efficiency and supports other critical applications of POA, such as multiple sequence alignment (MSA) for phylogeny construction and graph-based reference alignment.

The results of our study demonstrate a substantial enhancement in performance, with a 37-fold increase in speed and a 98% reduction in memory usage for sufficiently long sequences. This reduction in memory usage is particularly beneficial, allowing up to 150x speed improvements on conventional PCs through parallel processing without compromising alignment quality, a feat not achievable with vanilla unbanded POA. These advancements enable fast and efficient local and semi-global alignment on the banded POA graph, significantly enhancing its utility in bioinformatics.

Overall, the extended LCSk++ algorithm for graph structures represents a significant advancement in sequence alignment technology, offering improved performance and accuracy for the analysis of complex biological datasets, thereby addressing some of the longstanding challenges in multiple sequence alignment.

Availability and requirements

Project name: lcskgraph

Project home page: https://github.com/lorewar2/lcskgraph

Operating system(s): Linux, MacOS

Programming language: Rust

Other requirements: Rust 1.65.0+

License MIT license

Any restrictions to use by non-academics: None

Abbreviations

POA: partial order alignment

LCSK: Longest Common Subsequence K-mer

Declarations

Ethics approval and consent to participate

We acquired a cord blood sample (ID: PD47269d) from a newborn female patient through the Cambridge Blood and Stem Cell Biobank (CBSB). This sample was collected at Addenbrooke's Hospital with informed consent and approved by the Cambridge East Ethics Committee under reference 18/EE/0199.

Consent for publication

Not applicable

Availability of data and materials

The datasets analysed during the current study are available in the github repository, https://github.com/lorewar2/lcskgraph and can also be accessed by SRA repository, SRA biosample accession PRJNA1128051.

Competing interests

Not applicable

Funding

Not applicable

Authors' contributions

Methodology, MW, HH; Conceptualization, MW, CS; Resources, HH; Supervision, HH, CS; Writing original draft, MW, CS, HH; Validation, MW;

All authors reviewed the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

References

1. Rautiainen, Mikko, and Tobias Marschall. “Graphaligner: Rapid and Versatile Sequence-to-Graph Alignment.” *Genome Biology* 21, no. 1 (September 24, 2020). https://doi.org/10.1186/s13059-020-02157-2.

2. Li, Heng. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics* 34, no. 18 (May 10, 2018): 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

3. Zhang, Yun, Chanhee Park, Christopher Bennett, Micah Thornton, and Daehwan Kim. “Rapid and Accurate Alignment of Nucleotide Conversion Sequencing Reads with HISAT-3N.” *Genome Research* 31, no. 7 (June 8, 2021): 1290–95. https://doi.org/10.1101/gr.275193.120.

4. Li, Heng, and Richard Durbin. “Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform.” *Bioinformatics* 25, no. 14 (May 18, 2009): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.

5. Lee, Christopher, Catherine Grasso, and Mark F. Sharlow. “Multiple Sequence Alignment Using Partial Order Graphs.” *Bioinformatics* 18, no. 3 (March 1, 2002): 452–64. https://doi.org/10.1093/bioinformatics/18.3.452.

6. Gao, Yan, Yongzhuang Liu, Yanmei Ma, Bo Liu, Yadong Wang, and Yi Xing. “ABPOA: An SIMD-Based C Library for Fast Partial Order Alignment Using Adaptive Band.” *Bioinformatics* 37, no. 15 (November 9, 2020): 2209–11. https://doi.org/10.1093/bioinformatics/btaa963.

7. Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. “Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads.” *Genome Research* 27, no. 5 (January 18, 2017): 737–46. https://doi.org/10.1101/gr.214270.116.

8. Pavetić, Filip, Goran Žužić, and Mile Šikić. *LCSk++: Practical similarity metric for long strings.* https://doi.org/https://doi.org/10.48550/arXiv.1407.2407.

9. Pavetić, Filip, Ivan Katanić, Gustav Matula, Goran Žužić, and Mile Šikić. *Fast and simple algorithms for computing both LCSk and LCSk+ Filip Pavetić, Ivan Katanić, Gustav Matula, Goran Žužić, Mile Šikić*, n.d. https://doi.org/https://doi.org/10.48550/arXiv.1705.07279.

10. Benson, Gary, Avivit Levy, and B. Riva Shalom. “Longest Common Subsequence in k Length Substrings.” *Similarity Search and Applications*, 2013, 257–65. https://doi.org/10.1007/978-3-642-41062-8_26.

11. Köster, Johannes. “Rust-Bio: A Fast and Safe Bioinformatics Library.” *Bioinformatics* 32, no. 3 (October 6, 2015): 444–46. https://doi.org/10.1093/bioinformatics/btv573.

12. Weerakoon, Minindu, Sangjin Lee, Emily Mitchell, and Haynes Heaton. *TopoQual polishes circular consensus sequencing data and accurately predicts quality scores*, February 12, 2024. https://doi.org/10.1101/2024.02.08.579541.