# A Progressive Alignment Algorithm for Rotation and Orientation of Circular Sequences

André Rosengaard Jørgensen, Alexandra Weronika Balicki,
Joshua Daniel Rubin, Kristine Rosted Petersen, Peter Wad Sackett,
Gabriel Renaud

November 11, 2025

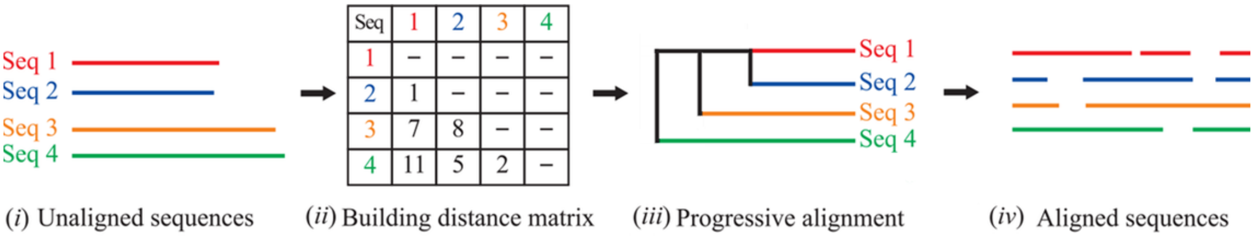
## 1 Abstract

Current bioinformatics tools often struggle with circular genomes, primarily due to challenges in linearizing them for analysis. We introduce `vgOrient`, a progressive, graph-based tool designed to linearize circular mtDNA sequences to facilitate downstream analyses. By detecting and correcting misoriented or rotated genomes and identifying a common anchor region, `vgOrient` produces consistently oriented, linearized sequences. In comparisons with MARS across three datasets of varying complexity, `vgOrient` achieved equal or better alignment quality for reversed sequences and comparable performance on non-flipped data. These findings underscore `vgOrient`'s potential as a powerful solution for handling circular genomic data.

# 2 Methodology

## 2.1 Preliminary Orientation and Graph Construction

vgOrient makes use of progressive alignment methods widely employed for MSA (see Fig. 1.



| Seq | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | – | – | – | – |
| 2 | 1 | – | – | – |
| 3 | 7 | 8 | – | – |
| 4 | 11 | 5 | 2 | – |

Figure 1: An overview of the steps of vgOrient. Figure adapted from [1]

Before building any graphs, vgOrient uses a k-mer–based approach to detect reverse-complemented sequences and generate a distance matrix that dictates the processing order of genomes. After orienting the sequences, an initial circular genome variation graph (see Fig. 2) is constructed from the first sequence in this order. Because genome variation graphs naturally model loops, they avoid arbitrary linear start points typical of circular DNA alignments. Subsequent mtDNA sequences are then progressively integrated into the initial graph by mapping their nodes and edges onto the existing structure until all sequences are represented in the graph.

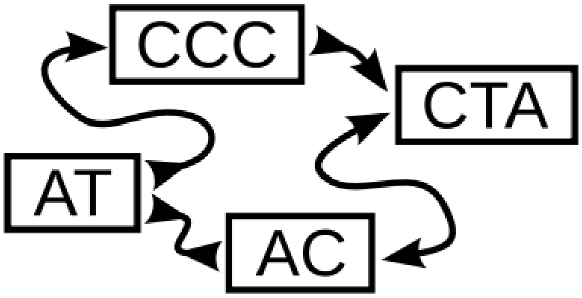


Figure 2: A bidirectional genome graph. The nodes can be identified as $N = n_1(AT), n_2 : (CCC), n_3 : (CTA), n_4 : (AC)$ and paths are $P = p_1 : (e_{12}, e_{23}), p_2 : (e^*_{14}, e_{43})$ Figure from [2]

## 2.2 Anchor Selection and Linearization

Once the final graph is built, `vgOrient` identifies a shared "anchor" region common to all sequences. Each sequence is then rotated and cut so that this anchor appears at the start, producing a set of linearized genomes suitable for conventional alignment tools. By using consistent cut points, `vgOrient` mitigates issues encountered when standard linear methods are applied directly to circular data.

# 3 Results

We compared `vgOrient` to MARS using three circular mtDNA benchmark datasets of increasing complexity (*Easy*, *Medium*, and *Hard*). Each dataset was transformed into two variants: one containing randomly rotated sequences (`no-flip`) and another incorporating both random rotation and a 50% chance of reverse-complementing each sequence (`flip`).

## 3.1 Alignment Quality (Wentropy)

Following rotation/orientation by either `vgOrient` or MARS, the resulting sequences were aligned with MAFFT and evaluated using Wentropy (a weighted Shannon entropy variant). Lower Wentropy values indicate higher alignment conservation. On flipped datasets, vgOrient produced distinctly lower Wentropy scores than MARS, reflecting its strong ability to correct reversed input. For non-flipped data, vgOrient and MARS showed comparable performance on the Medium and Hard datasets, while vgOrient held a small advantage on the Easy dataset.

### 3.2 Performance and Reliability

vgOrient ran faster and used fewer resources than MARS on smaller datasets but was sometimes slower and less stable on the Hard dataset. Despite occasional errors, vgOrient's significantly lower Wentropy on flipped data indicates its suitability for handling misoriented circular genomes, where traditional linear methods struggle.

## 4 Discussion

vgOrient addresses a key challenge in circular mtDNA analysis: producing consistently oriented, linearized sequences that avoid the pitfalls of arbitrary start points. Its primary strength lies in handling reversed genomes more effectively than MARS, offering superior alignment quality in flipped scenarios. By anchoring all sequences at a common region, vgOrient delivers linearized output that integrates smoothly with standard alignment pipelines. While vgOrient is more computationally intensive on large and divergent datasets, refining its graph-based algorithm or pruning strategy may improve scalability and reduce errors.

## 5 Conclusion

vgOrient provides a robust alternative to MARS for linearizing circular mtDNA sequences, particularly when dealing with reversed or arbitrarily rotated data. Its progressive, graph-based workflow identifies a shared anchor region, enabling clearer alignments and more reliable downstream analyses. Although further optimization is needed to enhance performance on challenging datasets, `vgOrient` demonstrates that leveraging circular-aware graph methods can substantially improve the accuracy and efficiency of circular genome analysis.

# References

[1] Soniya Lalwani, Harish Sharma, Abhay Verma, and Rajesh Kumar. An efficient discrete firefly algorithm for ctrie based caching of multiple sequence alignment on optimally scheduled parallel machines. *CAAI Transactions on Intelligence Technology*, 4, 2019.

[2] Novak A. M. Eizenga J. M. & Garrison E. Paten, B. Genome graphs and the evolution of genome inference. *Genome Research*, 27(5):665–676, 2017.