De novo genome assembly based on graph convolutional neural networks

UNIVERSITY OF ZAGREB
Faculty of Electrical Engineering and Computing

Lovro Vrček, MSc

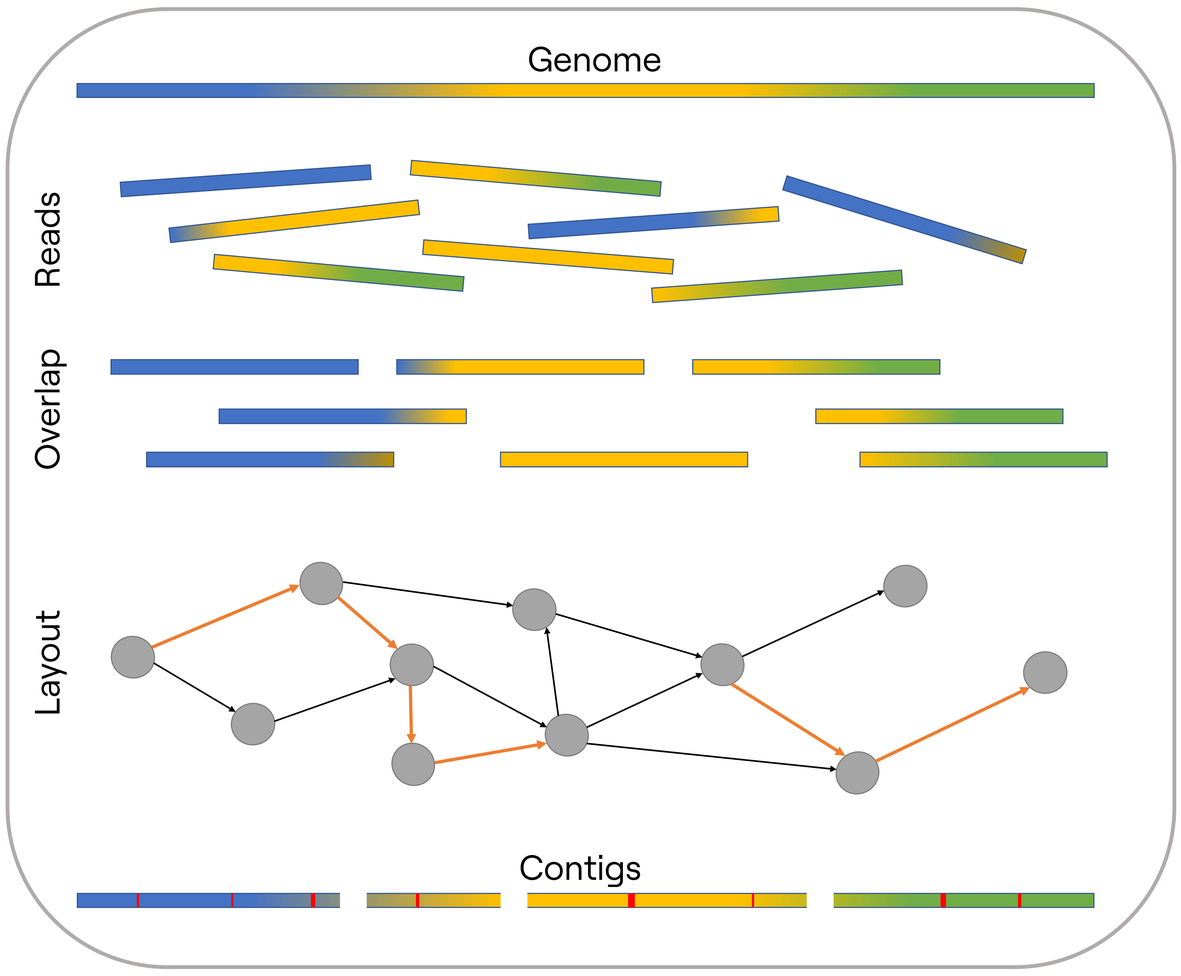
mentors: Prof. Mile Šikić, PhD; Assoc. Prof. Xavier Bresson, PhD (National University of Singapore)
University of Zagreb Faculty of Electrical Engineering and Computing

1. Introduction

De novo genome assembly is a process of reconstructing the original genomic sequence from short DNA fragments, called reads. The problem can be described as finding a Hamiltonian walk over an assembly graph, with the added requirement of avoiding a certain set of nodes. As this is an NP-complete problem, there is no exact solution for any given graph, and the developed heuristics often result in a fragmented assembly genome. Here, we present a novel method based on graph neural networks (GNNs) that produces more contiguous assembly genomes.

2. Problem Description

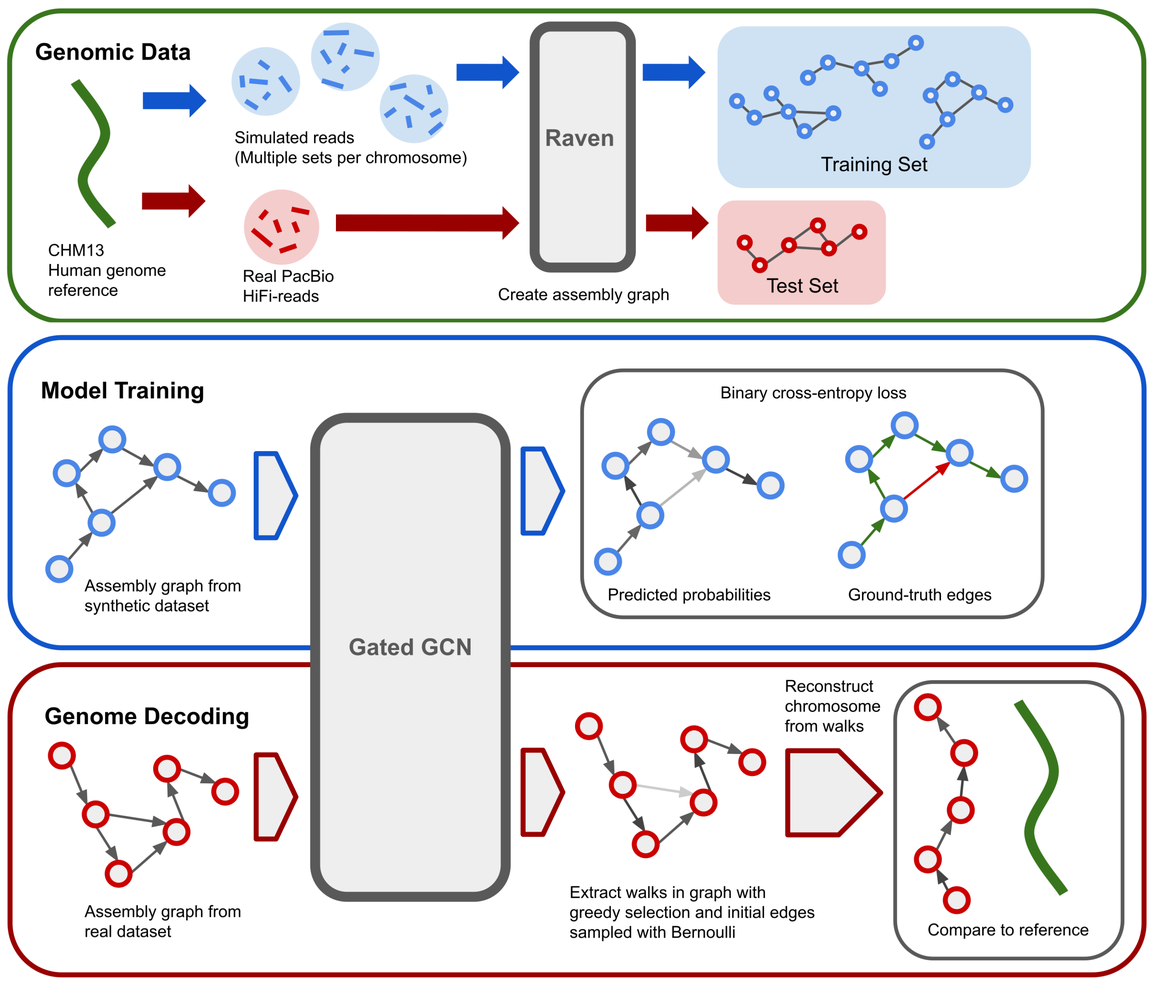
In de novo genome assembly, the first stage focuses on aligning all the reads in the sample. From this, an assembly graph is built, in which nodes represent the reads and edges represent the overlaps between them. The second stage focuses on simplifying this graph into a collection of path graphs, which can be translated into long genomic sequences called contigs.



Standard approach to *de novo* genome assembly.

3. Methodology

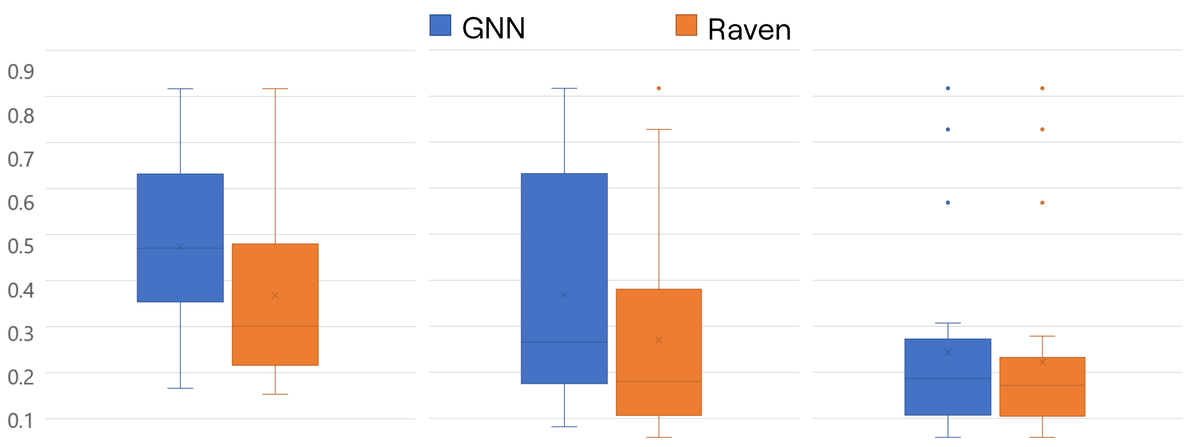
In our approach, instead of simplifying the assembly graph into path graphs, we find paths directly through the whole assembly graph. For this, we train a neural network based on GatedGCN layer [1] to correctly predict which edges in the assembly graph should be traversed in order to get the optimal assembly. We simulate a synthetic dataset of reads—allowing us to get a label for each edge—and generate graphs with Raven assembler [2]. Once trained, the network will output probability that an edge is valid. In inference, we decode the network output with a greedy search over the probabilities and obtain a collection of walks. These walks can be translated into contigs which we can then evaluate as any other assembly.



Framework based on obtaining scores with a graph neural network and decoding with a search algorithm.

4. Results

We trained the model on 15 instances of synthetic reads of chromosome 19, the shortest non-acrocentric chromosome, and assembled the real human chromosomes. The graphs on which we performed the decoding were constructed with Raven assembler, and we compare the performance of GNN against Raven's simplification algorithms. The metrics we show are the length of the longest contig, NG50, and NGA50, shown as a fraction of the length of the chromosome (higher is better in all three cases). We show that GNN approach outperforms classical heuristics on all three metrics, while creating fewer contigs—on average, GNN created 13 contigs per chromosome, while Raven created 71.



5. Conclusion

We present a novel approach to de novo genome assembly, based on finding a path through the graph with the help of graph neural networks instead of resorting to heuristics and cutting of the entangled regions. With this approach, we achieve less fragmented assemblies of human chromosomes when compared to standard heuristics-based methods.

Acknowledgments

This project has been supported by the Croatian Science Foundation under the project Single genome and metagenome assembly (IP-2018-01-5886).

References

[1] Joshi et al. An efficient graph convolutional network technique for the travelling salesman problem. arXiv preprint
[2] Vaser and Šikić. Time-and memory-efficient genome assembly with raven. Nature Comp. Sci., 2021

Contact

Lovro Vrček, MSc
Lovro.Vrcek@fer.hr