Methodology Proposal for De Novo Sequence Assembly Using Seq2Seq Deep Learning Models with a Graph-Theory Based Approach

by

Madala Vikas
Roll No. IMS21106

to

Dr Sabari Sankar Thirupathy
Course: Bioinformatics - BIO326/3202

SCHOOL OF BIOLOGY
INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH
THIRUVANANTHAPURAM - 695 551, INDIA
May 2024

ABSTRACT

Name of the student: **Madala Vikas** Roll No: **IMS21106**

Course for which submitted: **BIO326/3202** Department: **School of Biology**

Project title: **Methodology Proposal for De Novo Sequence Assembly Using Seq2Seq Deep Learning Models with a Graph-Theory Based Approach**

Project supervisor: **Dr Sabari Sankar Thirupathy**

Date of Report submission: **May 2024**

The project aims to provide a methodology to predict the hamilotonian cycles a graph theory approach to generate the assembled genome sequence as a part of the De-Novo genome sequencing problem which plays a crucial role in identifying the complex rearrangements like translocations, deletions and identifying structural variants using Sequence to Sequence(Seq2Seq) Deep Learning models. The main contribution of this work is to provide sufficient information about the required Deep Learning models and contributing to provide a tool for this alignment problem. It considers the time complexity of the proposed model and talks about the methods to overcome the computational cost issue by providing restrictions to predict hamiltonian cycles in a efficient way.

Keywords: De-Novo genome assembly, Seq2Seq(Sequence to Sequence), Deep Learning, LSTM(Long Short Term Memory), Transformers, Hamiltonian Cycles, K-mers

Introduction

De Novo sequencing is a sequence alignment method with no reference genomic sequence. Next generation Sequencing(NGS) techniques like Illumina are cost-effective and are useful to do genome assembly without any prior knowledge on the original genome sequence of an organism.

In our work, we are trying to solve the de novo sequence assembly problem using **graph theory** methods. Particularly, we are focussing on the Hamiltonian cycles that can be considered as the assembled genomic sequences. A hamiltonian path is a cycle in a graph that passes through each and every vertex only once. Also, a K-mer is defined as the sequence of 'k' nucleotide bases ('A', 'T', 'G', 'C') in a given DNA sequence. Figure 1 is a diagrammatic representation of a hamiltonian cycle in a given graph.

To tackle this problem, we are trying to find the Hamiltonian sequence in the graph using the prefixes and suffixes derived from the K-mers, and the predicted hamiltonian sequence will be the required assembly of sequence. Figure 2 is a representation of K-mer geneation for a given sequence. We propose a methodology using Deep

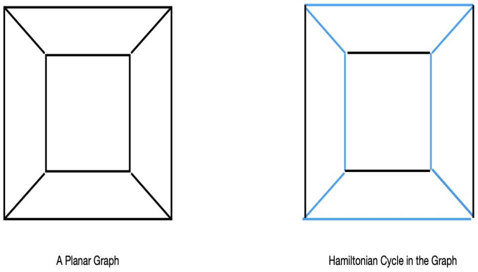

Figure 1: Representation of a Hamiltonian cycle in a graph

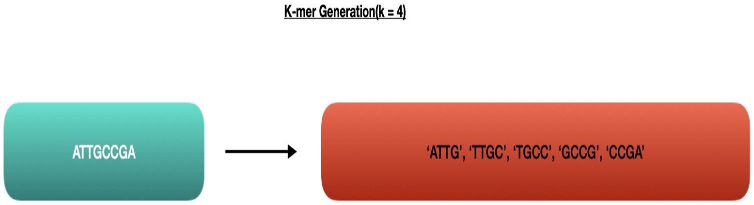

Figure 2: Example of K-mer Generation (K = 4)

Learning architectures for the prediction of Hamiltonian cycles. The current methods application is as depicted in the following Figure 3. The methodology and the problems that arise with this approach are discussed in the following sections.

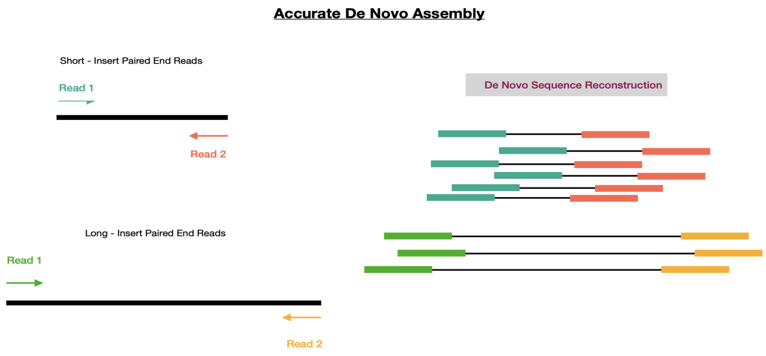
Accurate De Novo Assembly
Short - Insert Paired End Reads
Read 1
Read 2
Long - Insert Paired End Reads
Read 1
Read 2
De Novo Sequence Reconstruction

Figure 3: De novo genome assembly

Methodology

The methodology for De Novo Genome assembly using the Deep Learning architectures in a graph theory based approach is as follows:

- Firstly, we have to train the model on a collection of existing genome sequences by generating the K-mers for the sequence with a preffered 'K' value and shuffling them. Then we create the overlapping graphs using the generated K-mers.
- Then we use a Sequence to Sequence(Seq2Seq) Deep Learning model like LSTM(Long Short Term Memory), Transformers and train the model on this dataset to predict the hamiltonian cycles.

- With this, we can obtain the Contigs which are defined as the contiguos DNA sequences obtained using smaller overlapping DNA sequences. After this, Scaffolding methods can be applied along with Gap-filling inorder to obtain the Assembled genome sequence. The model's performance is then evaluated and to improve, hyperparameter tuning may be applied accordingly.

The main contibution of our work is to explain how the Deep learning models can be beneficial for this task. The detailed overview of our methodology is depicted in the Figure 4

Why Deep Learning Models?

The genome assembly is a Seq2Seq task. Also, hamiltonian cycle prediction for larger graphs is quite expensive and is NP-Hard problem i.e, it is not a polynomial time solvable problem. Since, our task is similar to this, by tansitivity, our task is also an NP-Hard problem. There are no direct solutions to this problem yet. The recent advancements in Deep Learning architectures include LSTM, Transformer architectures. These architectures have shown promising results in the field of Seq2Seq tasks.

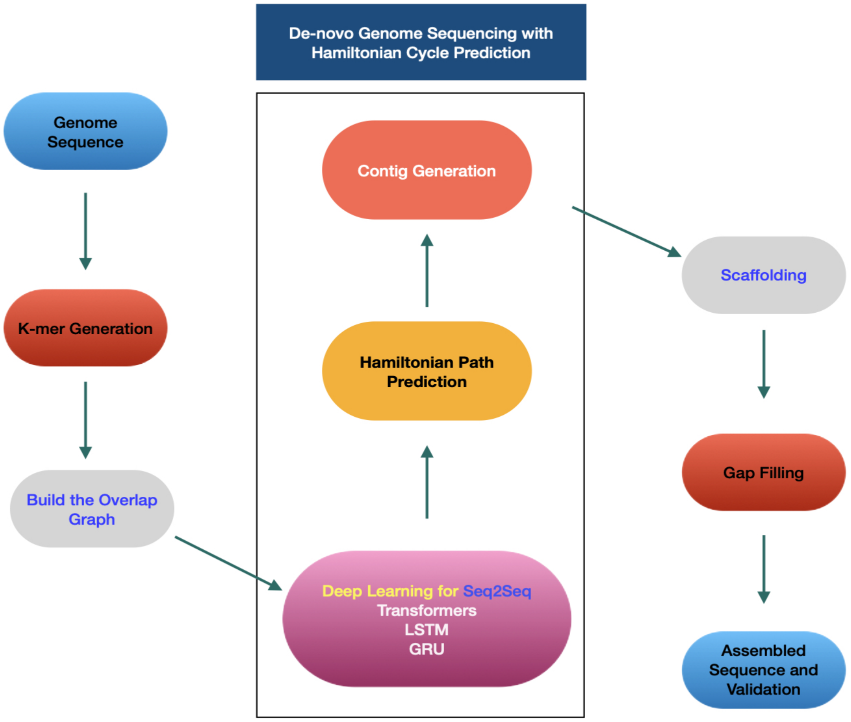
De-novo Genome Sequencing with Hamiltonian Cycle Prediction
Genome Sequence
K-mer Generation
Build the Overlap Graph
Deep Learning for Seq2Seq
Transformers
LSTM
GRU
Hamiltonian Path Prediction
Contig Generation
Scaffolding
Gap Filling
Assembled Sequence and Validation

Figure 4: Overview of our Methodology

Researchers have been working these problems for decades and recently there have been proposal on finding hamiltonian cycles using Graph Neural Networks. The other related working on De Novo genome sequence reconstruction includes the applications of Convolutional Neural Networks(CNNs) and LSTM architectures for mass spectroscopy but not on the hamiltonian cycle prediction. The current best time complexity is $\mathcal{O}(1.657^n)$ for undirected graphs with n vertices using monte-carlo approach.

While training a transformer model, we aim to reduce this time complexity by adding some restrictions to find the hamiltonian cycles such as considering the condition that if some point becomes an interior point while selecting the faces, that is not going to be a hamiltonian cycle. And provide a tool with a better performance compared to the exsiting models.

Transformers Architecture

In this section, we will provide a brief introduction to the transformer architecture. The transformer model consists mainly of two components:

1. Encoder part
2. Decoder part.

The input is first embedded into machine readable form by using 'input embedding' and then there is something called 'Positional embedding' which has given transformers an edge over the LSTM models which were previously the state of the art for Seq2Seq tasks. The embedded input is then passed onto the encoder layer, where, the multihead attention layers are used followed by a normalization and finally through a feedforward neural network to extract the features.

The Key, Value vectors generated in the encoder layer are then passed in to the decodern layer. The decoder layer also similar input as in encoder layer with an additional layer called as 'masked attention layer'. Pictorial representation of the transformer's architecture is depicted in the Figure 5.

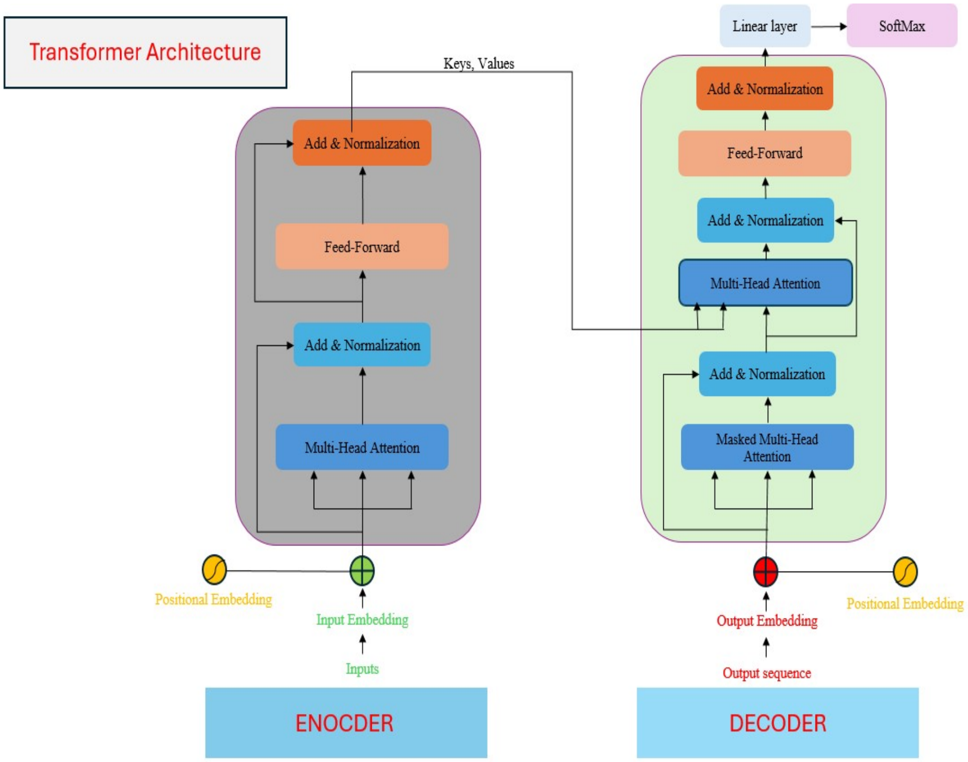

Figure 5: Transformer Architecture

Drawbacks

The problems that arise are the computational cost of transformers, which is very high. A deeper understanding of the architecture along with implementation is required for creating a fully functional model. Our approach with the code is not yet completed and we are currently working on it.

Future Research

Our future interests include working on real-life sequence data with the proposed methodology and creating a tool for the researcher to handle this task efficiently.

Images and Code Availability

The images used in this report are our own geneated work and the code will be uploaded in the github for the public use once we complete the project.

Bibliography

Björklund, Andreas. *Determinant Sums for Undirected Hamiltonicity*. 2010. arXiv: 1008.0541 [cs.DS].

Bosnić, Filip and Mile Šikić. *Finding Hamiltonian cycles with graph neural networks*. 2023. arXiv: 2306.06523 [cs.LG].

Bresson, Xavier and Thomas Laurent. *The Transformer Network for the Traveling Salesman Problem*. 2021. arXiv: 2103.03012 [cs.LG].

Ebrahimi, Shiva and Xuan Guo. *Transformer-based de novo peptide sequencing for data-independent acquisition mass spectrometry*. 2024. arXiv: 2402.11363 [q-bio.QM].

— "Transformer-based de novo peptide sequencing for data-independent acquisition mass spectrometry". In: *ArXiv* (2024). URL: https://api.semanticscholar.org/CorpusID:267750788.

Pandey, Ambrish Kr. and Shriya Kanchan. "Usage of Eulerian and Hamiltonian Graph in Pandemic Situation". In: *Journal of Applied Science and Education (JASE)* 1.1 (Nov. 2021), pp. 1–8. DOI: 10.54060/JASE/001.01.002. URL: https://jase.a2zjournals.com/index.php/ase/article/view/2.