GROUP 18

ACCELERATION OF DTW ALGORITHM FOR REAL-TIME NANOPORE SELECTIVE SEQUENCING USING GPUS

SUPERVISORS

Prof. Roshan Ragel

Dr. Hasindu Gamaarachchi

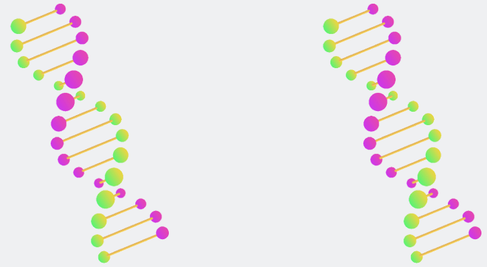
TEAM MEMBERS

E/16/089
Denuke Dissanayake

E/16/313
Maneesha Randeniya

E/16/360
Nipun Dewanarayana

BACKGROUND



- **Nanopore sequencing** is a unique, scalable technology that enables direct, real-time analysis of **long DNA or RNA** fragments.
- Nanopore DNA sequencing data is streamed in **real-time**, providing immediate access to the results.
- The introduction of the **MinION** is one of the breakthroughs in Nanopore Sequencing.
- Modern nanopore sequencers offer selective sequencing capability.

ALGORITHM

- **DTW ALGORITHM**
  - **measures the optimal alignment between signals** with different lengths and different time shifts
  - Earlier, used for **speech recognition applications**.
  - Time and Space complexity - **$O(n^2)$**
  - starts by building the **distance matrix.**
  - Next, **finds the alignment path** that runs through the cost matrix's low-cost areas

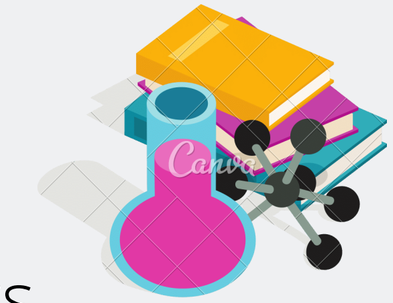
PROBLEM DEFINITION

- DTW algorithm has a computational complexity of **$O(n^2)$** - **High Computational Demand**.

- Because of the high computational demand of the traditional methods, portable MinION sequencers are **required to connect to a large server** to do the analyses.

- Consequently, it will **reduce** the widespread adaptation of selective sequencing in a **portable setting**.

ANALYSING METHODOLOGY



- Mainly use **Google Scholar** to search for research papers

- Main Scope - Dynamic Time Warping
  - DTW Algorithm
  - Variations
  - Acceleration Methods
  - DTW algorithm in Nanopore Selective Sequencing

- Use **Mendeley** as the research paper repository

ANALYSIS - NANOPORE

- Existing sequence analyzing methods
  - **Traditional High-Throughput Sequencing**
  - **Base Calling Method** [Loose et al., 2016]
    - Base callers - Base-callers approximate an electrical signal into a corresponding nucleotide sequence
      - **Slow**
      - **High computational power required**
      - **Less accuracy**

ANALYSIS - NANOPORE

- Existing sequence analyzing methods cont...
  - **Dynamic Time Warping Method** [Loose et al. , 2016]
    - Able to optimize using GPU platform
    - Lot evolved algorithm
    - Match the signal directly to the reference signal
      - **Less computational compared to base-callers**
      - **High accuracy compared to base-callers**

ANALYSIS - NANOPORE

- Existing sequencing methods cont...
  - **Sequencing with Deep Learning** [Danilevsky et al. , 2021]
    - Using a deep learning model, extract the details from raw signals
      - **Reduce rum time complexity**
    - No reference signal is used to train models
      - **Only the required signals are sequenced**
      - **Raw signal can use for accuracy check**

ANALYSIS - NANOPORE

- Existing sequencing methods cont...
  - **Sequencing with Sigmap and Uncalled** [Kovaka et al., 2021, Zhang et al., 2021]
    - Use 'Read Until' API and FM-Index for searching in the reference signal
      - **Suitable for long sequences**
  - **Readfish Sequencing** [Payne et al., 2021]
    - Use nucleotide signals instead of raw signals
    - GPU bases base calling method
      - **Need to convert raw signals into nucleotide signals**

ANALYSIS - DTW

- Existing DTW acceleration methods
  - **Continuous Wavelet DTW** [Han, R., 2018]
    - based on Continous Wavelet Transform
    - Can reduce the details of highly representative signals, which **reduces the noise, and improve the computational time**
  - **Subsequence Pattern Mining** [Tavenard et al., 2018]
    - Identify the similar subsequences in two-time series
    - If two subsequences complete initial requirements, then DTW will calculate for them, else, rejected.

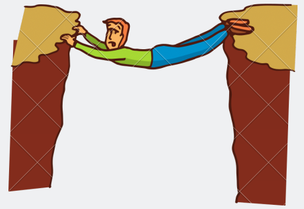
ANALYSIS - DTW

- Existing DTW acceleration methods cont...
  - **Parallelized DTW** [Zhu H et al., 2018, Gustavo et al., 2007]
    - Parallelize the matrix calculation process
    - Each diagonal of the matrix does not depend on each other values
      - **Able to implement parallelly in GPUs**
  - **Subsequence DTW** [Anguera, X., 2013]
    - A subsequence of an input signal compared with a full index sequence
      - **Find the optimum matching subsequence in the index sequence**

ANALYSIS - DTW

- Existing DTW acceleration methods cont...
  - **Lower bounding function** [Tavenard et al., 2018]
    - This can prune unmatching subsequences
    - Because of that, **improve the efficiency**
    - More pruning cause less accuracy
  - **Constraint DTW Paths** [Tavenard et al., 2018]
    - After calculating the cost matrix, calculate the warping path
    - Restrict that path around the diagonal
    - helps to avoid unnecessarily matching

RESEARCH GAPS



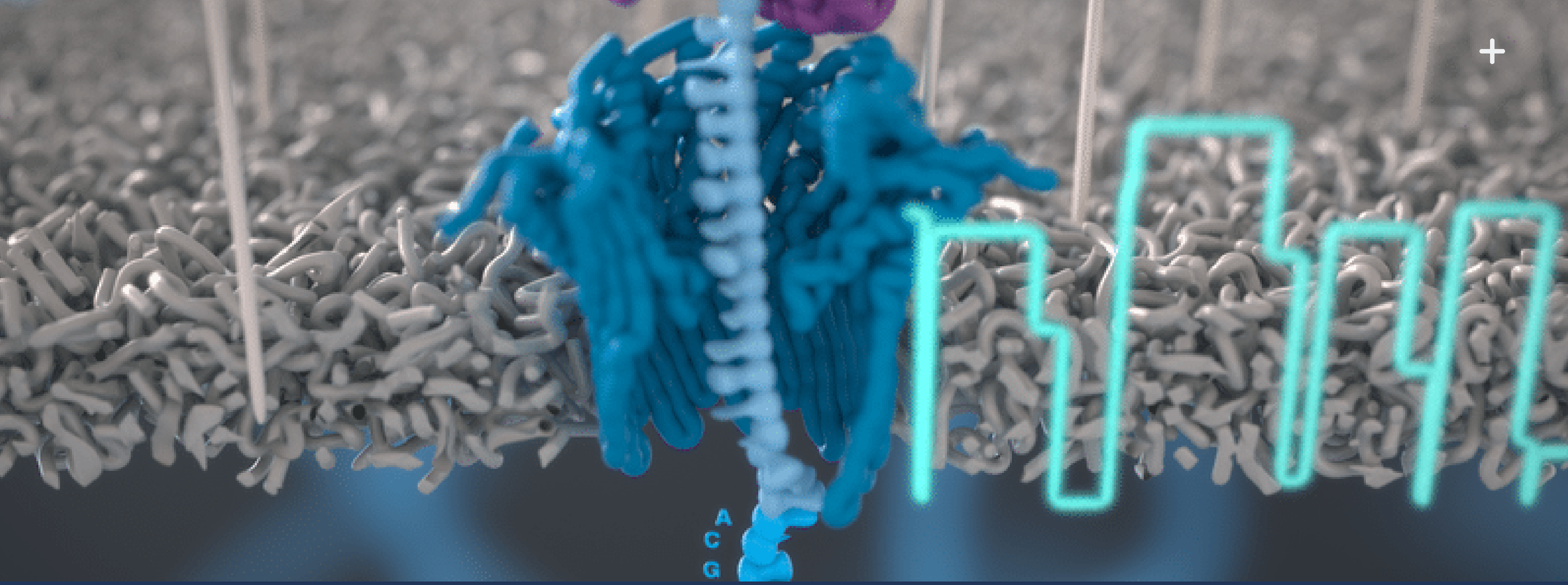
- Existing Nanopore Selective Sequencing using the DTW algorithm is a **computationally intensive** task
  - **The portability of the MinION sequencer will decrease** due to the need for high computational power
  - Unable to analyse DNA in Desktop GPUs
  - **The runtime of the DTW is high** when the sequences are long
  - The **capability of selective sequencing** and **parallel computing** didn't use together

NEXT STEPS

- Implement the DTW algorithm for GPUs.
  - The **Compute Unified Device Architecture (CUDA)** developed by NVIDIA will be used to parallelize the DTW algorithm between multiple cores.

- Research **different optimizing techniques** for DTW algorithm for Nanopore Selective Sequencing

- **Implement and Observe** the results with each optimizing technique and select the best method



THANK YOU
Q & A

FM Index

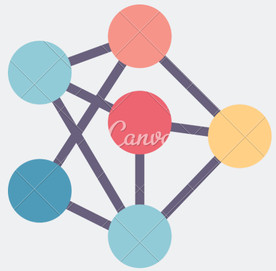
The FM-index algorithm check for k-mers to represent the raw signal.

It converts raw signals to events and calculates the possibilities of event matches to k-mers using the ONT probabilistic model.

To develop the FM-index search algorithm high-probability, k-mers are used, considering all possible sequences and locations.

Then to filter out false-positive locations seed-clustering algorithm is used by grouping seeds together.

BACKGROUND CONT.



- In principle, the sequencer can **reject individual sequences** to enable selective sequencing.

- One approach is to **compare decoded bases to reference base signals**. This method is **slow** and **less accurate**.

- Using the **DTW** algorithm, signals can **directly match** the reference signal.

ALGORITHM

- **DTW ALGORITHM**
  - **measures the optimal alignment between signals** with different lengths and different time shifts
  - Earlier, used for **speech recognition applications**.
  - **O(n^2)** time and space complexity.
  - starts by building the **distance matrix** representing all pairwise distances between two sequences.
  - Next, **finds the alignment path** that runs through the cost matrix's low-cost areas

NANOPORE DNA SEQUENCING

- Using long nanopore DNA sequencing reads, researchers can:
- Resolve complex structural variants and repetitive regions
- Simplify de novo genome assembly and improve existing reference genomes
- Study linkage and phasing
- Enhance metagenomic identification of closely related species and distinguish plasmid from genome
- Sequence entire microbes in single reads – in real-time
- Explore epigenetic modifications using direct, long-read DNA sequencing

- Using long nanopore DNA sequencing reads researchers can:
- Resolve complex structural variants and repetitive regions
- Simplify de novo genome assembly and improve existing reference genomes
- Study linkage and phasing
- Enhance metagenomic identification of closely related species and distinguish plasmid from genome
- Sequence entire microbes in single reads – in real-time
- Explore epigenetic modifications using direct, long-read DNA sequencing