Master's Thesis Project

Genome sequencing, the process of determining the DNA sequence of an organism's genome, can be used to study the genetic basis of various diseases and traits, to understand evolutionary relationships between organisms, and to develop new treatments and therapies. It has become increasingly accessible and affordable in recent years, enabling scientists and researchers to analyze large amounts of genetic data and make new discoveries about the genetic underpinnings of life.

At the core of the sequencing pipeline is the **basecalling** process, or converting sensed electrical signals into the string of bases. Current state-of-the-art algorithms use **deep neural networks** that account for **40-80% of the sequencing pipeline**, and require expensive GPUs to run, presenting a bottleneck in the sequencing pipeline.

To overcome this bottleneck, a new **in-memory computing** device has been proposed, which can perform basecalling in realtime on the portable DNA sensing device, eliminating the need for a costly basecalling workstation and enabling sequencing in even more remote and off-the-grid scenarios. Alongside the device, a new DNN basecalling algorithm has been proposed, which maintains >90% inference accuracy.

Despite these advances, there are still many improvements to the proposed HW/SW system that can be achieved. This project will focus on one of the following three topics:

- **Exploring DNN implementations** that take advantage of the in-memory computing device's unique and special capabilities to improve accuracy while maintaining a small area and power footprint (Fig.1-a).
- **Developing new decoding hardware in HDL** for converting the output of the DNN, which represents probabilities that the DNA strand contains a base at a given timestep, into the actual inferred base string (Fig.1-b).
- **Developing downstream analysis algorithms** that run on the in-memory computing device after basecalling to extend the utility of the device beyond basecalling (Fig.1-c).

We are looking for enthusiastic students interested in applying new in-memory computing paradigms to bioinformatics to enable genome sequencing in off-grid environments.

Requirements

- Outstanding programming skills (Python/C++ for Project 1/3, also HDL for Project 2)
- Understanding of CNN/LSTM DNNs.
- Independent learning/working abilities
- Interest in bioinformatics
- Strong work ethic

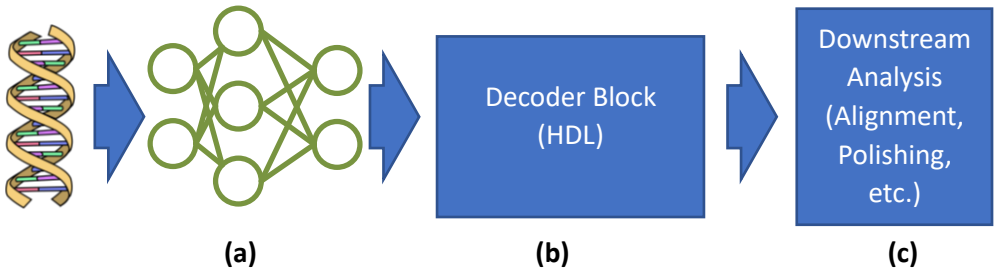

Figure 1: Basecalling pipeline with DNN (a), decoder block (b), and downstream analysis (c).

For background and past works, see:

- Nanopore Basecalling
  - Nanopore website explanation
- In-memory computing
  - Helix: SW/HW Co-design for Accelerating Nanopore Genome Base-calling
  - A Heterogeneous and Programmable Compute-In-Memory Accelerator Architecture for Analog-AI Using Dense 2-D Mesh
- Decoding algorithms
  - CTC decoding
  - CRF decoding Pt1, Pt2

If you are interested, please contact **Professor Onur Mutlu, Dr. Mohammed Alser** and **Dr. William Simon:**

omutlu@gmail.com

william.simon1@ibm.com

mealser@gmail.com

https://safari.ethz.ch

https://people.inf.ethz.ch/omutlu/