Multi-Nanopore Sequencing: Modeling, Joint Signal Processing and Coding towards Practical DNA Data Storage

Summary

Nowadays, the world's data storage demand is clearly outstripping the world's storage capacity. DeoxyriboNucleic Acid (DNA), the natural carrier of genetic information, provides an energy-efficient long-term data storage option. To create a practical gateway to access this storage technology, plenty of technical challenges must be addressed. Historically, one of the main technical barriers has been the lack of understanding the statistical nature of the underlying data channel, which is traditionally resolved by deep learning techniques. Moreover, to increase accuracy, PCR amplification is typically applied which increases the cost of storage per stored information bit. In this project, we target high accuracy and ultra-reliable system design through multi-nanopore sequencing with genuine signal processing, detection, and coding schemes for a cost-effective solution. For instance, one of the objectives is to minimize the number of nanopores (effectively the number of reads) during data retrieval. In particular, the project's signal and channel characterization would give better insight into algorithm development for future generations of this storage technology. We will also explore novel trade-offs and propose detection algorithms towards achieving theoretical limits established by past research.

Short Project Description

Introduction

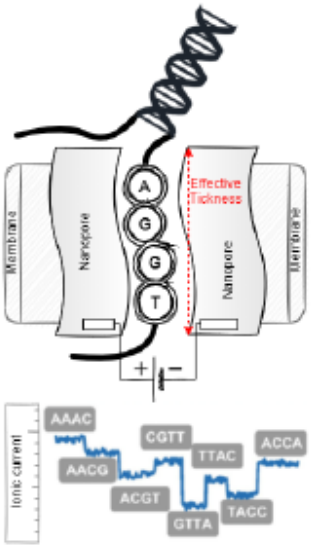

FIGURE 1: **A typical nanopore and a DNA strand passing through it generates an ionic current blokade.**

Nanopore sequencing is a recent method for DeoxyriboNucleic Acid (DNA) data storage and used to read data values chemically embedded in oligonucleotides i.e., a single molecule of DNA can be sequenced without the need for Polymerase Chain Reaction (PCR) amplification or chemical labeling of the sample. A strand of a DNA molecule (consisting of four types of nucleotides: A, T, C, G) passes through a specially designed pore (either biological or solid state) and a voltage is applied across the pore which ends up creating an electrical field across pore ends [1]. As can be seen in Fig. 1, this voltage (the field itself) creates an ionic current to pass through the pore (movement of charges due to the field). Depending on the type of the molecule passing through the pore, different current blockade levels and translocation speeds can be measured and recorded through placing electrodes near the membrane. Based on various factors such as pore geometry, size and chemical composition, the change in the magnitude of the ionic current blockade and the duration of the translocation (so called *dwell time*) will vary over time.

The main objective of the detection process is to be able to differentiate different nucleotides based on the uniquely generated blockade patterns. Due to the sophisticated signal generation process, the classical Artificial Intelligence (AI) approach to base detection (so-called *base calling*) has been based on Neural Networks (NNs) whose main objective is to learn blockade patterns corresponding to different segments of bases to eventually decode the data [2]. However, all NN-based approaches abstract the entire DNA read channel [3]. Though such an approach may work well in practice, it makes it impossible to reason about how the detection works in the data retrieval process. In addition, besides its complexity, current NN-based detection error rates (without any error correction techniques applied) level around 0.10-0.15 [4]. Therefore, for any practical use of DNA as a storage medium, error correction is a must to pull down detection error rates. Due to varying translocation times and molecular differences between the bases, the DNA channel is mostly dominated by **insertions, deletions and substitution errors**. Hence, codes need to be designed to accurately retrieve the data being recorded on DNA molecules. Due to synthesis and sequencing technologies, data is typically stored on a set of short base sequences. Also before retrieval, a given base sequence is replicated multiple times through PCR. Hence one of the fundamental problems

of decoding is to retrieve the data from multiple noisy copies of the sequence known as *sequence reconstruction problem*. Most of the past literature relies on hard and reliable decisions at the end of the detector since they are mostly interested in reconstruction of the original data sequence from their multiple sub/super-sequences. Even in the case of unreliable detection, later stages of concatenated codes are supposed to resolve these errors for a successful data retrieval [5].

Any commercial device with nanopore sequencing capability will come with multiple physical nanopores laid out in a two dimensional grid/membrane that would define independent channels for parallel processing of DNA molecules. For instance, Oxford MinION device [6] has 512 independent channels allowing 512 different DNA molecules to be sequenced all at the same time. Associated with each one of the channels a neural network that processes and detects nucleotides. In fact to do a consensus read (multiple reads of the same data), these networks have to run multiple times (or a separately bigger network designed for consensus) for the same sequence [7]. Finally, in the future generations of such sequencers it is likely to have 10000×10000 nanopores with millions of processing units to be able increase the data access rates for DNA drives/storage devices. However, having 100 million different neural networks inside the device makes it practically infeasible even for testing. At some point, running such a huge number of networks even for testing/classification purposes may be burdensome from an implementation point of view. A resource sharing can be anticipated for practical device development.

Given all that, practical implementation and low complexity algorithm development have to be at the heart of any DNA-based storage system design. In this project, we consider three main work packages. First off, **inspired from tape and more recent racetrack memories** [8], we propose to explore a novel architecture whereby a single DNA strand passes through multiple pores and the translocation times can partially be controlled. Moreover, unlike classical NN-based basecalling, we propose to use NNs to estimate signal shapes corresponding to each base and use signal processing and a specially designed detector architecture (**a new basecaller**). Finally, instead of considering purely soft information at the end of the detector, we introduce "erasures" in addition to deletions, insertions and substutions of a DNA channel. An optimal relability threshold would be determined based on a pore model. Estimated bases which are below the threshold would be labeled as erasures and will be utilized by the subsequent **sequence reconstruction and error correction decoding** stages.