UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

MASTER THESIS num. 1572

End-to-End Deep Learning Model for Base Calling of MinION Nanopore Reads

Neven Miculinić

Zagreb, June 2018.

Umjesto ove stranice umetnite izvornik Vašeg rada.
Kako biste uklonili ovu stranicu, obrišite naredbu `\izvornik`.

I would like to thank my mentor, Mile Šikić, for his patient guidance, encouragement and advice provided over the years.

I would also like to thank my family and friends for their continuous support.

In the end, honorable mentions go to Marko Ratković for his help with this thesis.

CONTENTS

LIST OF FIGURES

List of Tables

1. Introduction

Enable TEXT macro for full text rendering

1.1. Objectives

Enable TEXT macro for full text rendering

1.2. Organization

2. Background

Enable TEXT macro for full text rendering

2.1. Oxford Nanopore MinION

The MinION device by Oxford Nanopore Technologies is the first portable DNA sequencing device. Its small weight, low cost, and long read length combined with decent accuracy yield promising results in various applications including full human genome assembly **?** what could potentially lead to personalized genomic medicine.

2.1.1. Technology

As its name says, nanoscaled pores are used to sequence DNA. An electrical potential is applied over a membrane in which a pore is inserted. As the DNA passes through the pore, the sensor detects changes in ionic current caused by different nucleotides present in the pore. Figure **??** shows the change of ionic current as DNA strain is pulled through a nanopore.

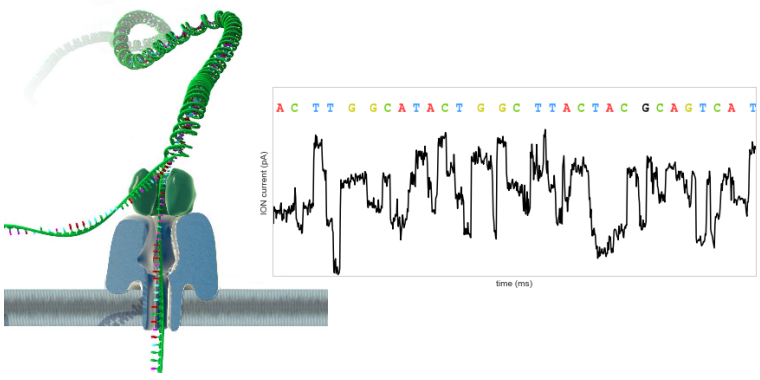

Figure 2.1: DNA strain being pulled through a nanopore [1]

Official software called MinKNOW outputs sequencing data in FAST5 (a variant of the HDF5 standard) file format. It is a hierarchical file format with data arranged in a tree-

[1] Figure adapted from https://nanoporetech.com/how-it-works

structure of groups. Metadata are stored in group and dataset attributes. The same file format is during used different stages of analyses and groups, datasets and attributes are added incrementally. Figure **??** shows raw signal being present in the FAST5 file.

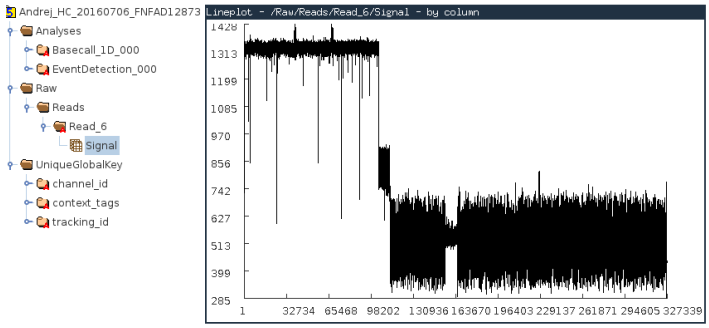

Figure 2.2: Structure of FAST5 file and raw signal line plot show in *HDFView* [2]

Minion offers the possibility of sequencing one or both strands of DNA. Sequencing both strands and combining information results in reads of higher quality. Those reads are called 2D (two-dimensional) reads. Otherwise, if the only single strand is sequenced 1D (one-dimensional) reads are produced.

MinION devices can produce long reads, usually tens of thousand base pairs (with reported reads lengths of 100 thousand **?** and even recently above 800 thousand base pairs **?**), but with high sequencing error than older generations of sequencing technologies. Switch from older R7.3 to R9 chemistry in 2016 increased accuracy of produced data. With this change, the accuracy of 1D data increased from 70% to 85% and the accuracy of 2D reads from 88% to 94% **?**. This increase of accuracy makes 1D reads usable for analysis with benefits over 2D reads being faster sample preparation and faster sequencing. Developed tool in this thesis focuses on base calling 1D reads.

2.2. Existing basecallers

2.2.1. Official

Oxford Nanopore has, with the R9 version of the platform, introduced a variety of base calling options. Some of those are production ready and some experimental. The majority of information regarding differences, specifications and similar is only available through Nanoporetech Community [3].

Metrichor is an Oxford Nanopore company that offers cloud-based platform *EPI2ME* for analysis of nanopore data. Initially, base calling was only available by uploading data to the

[2] https://support.hdfgroup.org/products/java/hdfview/

[3] https://community.nanoporetech.com/

platform - that being the reason why this basecaller is often called Metrichor even though it is a name of the company.

The older version of Metrichor relied on *hidden Markov models* (HMM) to find the biological sequence corresponding to the signal. Preprocess included segmentation of the signal into smaller chunks called events defined by start location of the chunk, length, mean value and variance of the signal in the chunk. Metrichor than assumed that each event usually corresponds to a context of 6 bases being present in the pore and that the context is typically shifted by one base in each step. The states of HMM are modeled as a context present in the pore and transition correspond to change of bases in the pore. During the transition from one state to another, an event is emitted. Base calling is performed using the Viterbi algorithm which determines the most likely sequence of states for the observed sequence of events. This approach showed poor results when calling long homopolymer stretches as the context in the pore remains the same **??**.

With the release of R9 chemistry, this model was replaced by a more accurate recurrent neural network (RNN) implementation. Currently, Oxford Nanopore offers several RNN-based local basecaller versions under different names: Albacore, Nanonet and basecaller integrated into MinKNOW **?**.

Albacore is basecaller by Oxford Nanopore Technologies ready for production and actively supported. It is available to the Nanopore Community served as a binary. The source code of Albacore was not provided and is only available through the ONT Developer Channel. Tool supports only R9.4 and future R9.5 version of the chemistry.

Nanonet[4] uses the same neural network that is used in Albacore but it is continually under development and does contain features such as error handling or logging needed for production use. It uses *CURRENNT* library for running neural networks. It supportes basecalling of both R9 and R9.4 chemistry versions.

Scrappie[5] is another basecaller by Oxford Nanopore Technologies. Similar to Nanonet, it is the platform for ongoing development. Scrappie is reported to be the first basecaller that specifically address homopolymer base calling. It became publicly available just recently in June, 2017 and supports R9.4 and future R9.5 data.

[4] `https://github.com/nanoporetech/nanonet/`

[5] `https://github.com/nanoporetech/scrappie`

2.2.2. Third-party basecallers

Nanocall **?** was the first third-party open source basecaller for nanopore data. It uses HMM approach like the original R7 Metrichor. Nanocall does not support newer chemistries after R7.3.

DeepNano **?** was the first open-source basecaller based on neural networks. It uses bidirectional recurrent neural networks implemented in Python, using the Theano library. When released, originally only supported R7 chemistry, but support for R9 and R9.4 was added recently.

3. Methods

In this chapter all key deep learning concepts shall be described.

4. System architecture

4.1. Data Preparation

DEBUG was off

title Data has been downloaded from `https://data.genomicsresearch.org/Projects/online_dataset/train_set_all/`. The following species were provided there by the Chiron team:

- Human
- E. Coli
- Lambda Phage

The raw dataset is transformed using my repo, minion-data [1]. It defines common dataset training structure in the protobuf (**?**) interface description language (IDL). The whole definition can be seen in figure **??**.

For the concrete Chiron dataset, the re-squiggled preparation method was used. (The data was re-squiggled, that is after aligning the read on the reference, the read data is improved and each base pairs place on the raw signal is calculated.)

The re-squiggled basecalled data is located at `/Analyses/RawGenomeCorrected_000/BaseCa`
The interesting code fragments are in function processDataPoint of file `minion_data/preperation/_` from minion-data python package.

After the gzipped dataset is prepared, it goes into the training pipeline. The whole training testing pipeline is available open source on `https://github.com/nmiculinic/minion-basecaller`

4.2. Training pipeline

4.3. Hyperparameter optimization

[1] `https://github.com/nmiculinic/minion-data`

```
syntax = "proto3";

package dataset;

enum BasePair {
    A = 0;
    C = 1;
    G = 2;
    T = 3;
    BLANK = 4;
}

enum Cigar {
    MATCH = 0;
    MISMATCH = 1;
    INSERTION = 2; // Insertion, soft clip, hard clip
    DELETION = 3;  // Deletion, N, P
}



message DataPoint {
    message BPConfidenceInterval {
        uint64 lower = 1;
        uint64 upper = 2;
        BasePair pair = 3;
    }
    repeated float signal = 1;
    repeated BasePair basecalled = 2; // What we basecalled
    repeated BPConfidenceInterval labels = 3; // labels describe corrected basecalled signal for training

    // Aligment data:
    repeated Cigar cigar = 8;
    repeated BasePair aligned_ref = 9; // Which is the reference string for this read after aligning

    // squiggled are the same length with blanks inserted for space filling. Matches/mismatched are aligned with BLANKS
    repeated BasePair aligned_ref_squiggle = 10; // Which is the reference string for this read after aligning, BLANKS inserted for aligments
    repeated BasePair basecalled_squiggle = 11; // What the  BLANKS inserted for aligments

    reserved 12 to 100; // Further assigement
    reserved 101 to 110; // For further Mincall use
}
```


Figure 4.1: dataset protobuf description

5. Results

6. Conclusion

End-to-End Deep Learning Model for Base Calling of MinION Nanopore Reads

Abstract

In the MinION device, single-stranded DNA fragments move through nanopores, which causes drops in the electric current. The electric current is measured at each pore several thousand times per second. Each event is described by the mean and variance of the current and by event duration. This sequence of events is then translated into a DNA sequence by a base caller. Develop a base-caller for MinION nanopore sequencing platform using a deep learning architecture such as convolutional neural networks and recurrent neural networks. Instead of events, use current waveform at the input. Compare the accuracy with the state-of-the-art basecallers. For testing purposes use publicly, available datasets and Graphmap or Minimap 2 tools for aligning called reads on reference genomes. Implement method using TensorFlow or similar library. The code should be documented and hosted on a publicly available Github repository.

Keywords: base calling, Oxford Nanopore Technologies, MinION, deep learning, seq2seq, convolutional neural network, residual network, CTC loss

S kraja na kraj model dubokog učenja za određivanje očitanih baza dobivenih uređajem za sekvenciranje MinION

Sažetak

Unutar uređaja MinION, fragmenti jednostruke DNA prolaze kroz nanopore, što uzrokuje promjene u električnoj struji. Struja proizvedena na svakoj nanopori mjeri se nekoliko tisuća puta u sekundi. Svaki događaj opisan je srednjom vrijednosti i varijancom struje te svojim trajanjem. Postupak kojim se takav slijed događaja prevodi u niz nukleotida naziva se određivanje očitanih baza. Razviti alat za prozivanje baza za uređaj za sekvenciranje MinION koristeći modele dubokog učenje kao što su konvolucijske i povratne neuronske mreže. Umjesto događaja na ulazu koristi valni oblik struje. Usporediti dobivenu točnost s postojećim rješenjima. U svrhu testiranja koristiti javno dostupne skupove podataka i alate GraphMap ili Minimap 2 za poravnanje očitanja na referentni genom. Alat implementirati koristeći programsku biblioteku TensorFlow (ili neku sličnu). Programski kod treba biti dokumentiran i javno dostupan preko repozitorija GitHub. **Ključne riječi:** određivanje

baza, Oxford Nanopore Technologies, MinION, duboko učenje, prevođenje, konvolucijske neuronske mreže, rezidualne mreže, CTC gubitak