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

In recent years, deep learning methods significantly improved the state-of-the-art in multiple domains such as computer vision, speech recognition, and natural language processing LeCun and Bengio (1998)Krizhevsky et al. (2012). In this thesis, we present application of deep learning in the field of Bioinformatics for analysis of DNA sequencing data.

DNA is a molecule that makes up the genetic material of a cell, and it is responsible for carrying the information needed for survival, growth, and reproduction of an organism. DNA is a long polymer of simple blocks called nucleotides connected together forming two spiraling strands to a structure called a double helix. Possible nucleotide bases of a DNA strand are adenine, cytosine, guanine, thymine usually represented with letters A, C, G, and T. The order of these bases is what defines genetic code.

DNA sequencing is the process of determining this sequence of nucleotides. Originally sequencing was an expensive process, but during the last couple of decades, the price of sequencing has drastically decreased. A significant breakthrough occurred in May 2015 with the release of MinION sequencer by Oxford Nanopore making DNA sequencing inexpensive and more available, even for small research teams.

Base calling is a process assigning sequence of nucleotides (letters) to the raw data generated by the sequencing device. Simply put, it is a process of decoding the output from the sequencer.

1.1. Objectives

The objective of this thesis is try out novel approach in basecalling the raw sequence. We had good results with earlier R9 chemisty (Miculinić et al., 2017) and we're experimenting with new approaches.

1.2. Organization

2. Background

2.1. Related work

3. Methods

4. Implementation

5. Results

6. Conclusion

BIBLIOGRAPHY

Yann LeCun and Yoshua Bengio. The handbook of brain theory and neural networks. chapter Convolutional Networks for Images, Speech, and Time Series, pages 255–258. MIT Press, Cambridge, MA, USA, 1998. ISBN 0-262-51102-9. URL http://dl.acm.org/citation.cfm?id=303568.303704.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. URL https://goo.gl/UpFBv8.

Neven Miculinić, Marko Ratković, and Mile Šikić. Mincall| minion end2end convolutional deep learning basecaller. In *2nd International workshop on deep learning for precision medicine, ECML-PKDD 2017*, 2017.

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