Abstract Book

Fourth International Workshop on Data Science

Zagreb, Croatia, October 15, 2019

Organiser

Centre of Research Excellence for Data Science and Cooperative Systems
Research Unit for Data Science

Member Institutions

University of Zagreb Faculty of Electrical Engineering and Computing

Ruđer Bošković Institute

University of Zagreb Faculty of Science

University of Zagreb Faculty of Transport and Traffic Sciences

Catholic University of Croatia

University of Split Faculty of Electrical Engineering, Mechanical Engineering and Naval Architeture

University of Rijeka Centre for Advanced Computing and Modelling

University of Rijeka University of Rijeka Faculty of Civil Engineering

Josip Juraj Strossmayer University of Osijek Faculty of Electrical Engineering, Computer Science and Information Technology

Sponsors

Ministry of Science and Education, Republic of Croatia
University of Zagreb Faculty of Electrical Engineering and Computing, Croatia

Organising Committee

General Co-Chairs

Sven Lončarić, University of Zagreb, Croatia
Tomislav Šmuc, Ruđer Bošković Institute, Croatia

Program Committee

Ana Babić, Croatia
Bojan Basrak, Croatia
Vuko Brigljević, Croatia
Zlatan Car, Croatia
Tonči Carić, Croatia
Bojana Dalbelo Bašić, Croatia
Davor Davidović, Croatia
Mirjana Domazet Lošo, Croatia
Tomislav Domazet Lošo, Croatia
Ante Đerek, Croatia
Neven Elezović, Croatia
Joao Gama, Portugal
Nikola Godinović, Croatia
Hrvoje Gold, Croatia
Sonja Grgić, Croatia
Edouard Ivanjko, Croatia
Zoran Kalafatić, Croatia
Wolfgang Ketter, The Netherlands
Mladen Kolar, USA
Ivica Kopriva, Croatia
Nada Lavrač, Slovenia
Damir Lelas, Croatia
Robert Manger, Croatia
Goran Martinović, Croatia
Branka Medved Rogina, Croatia
Igor Mekterović, Croatia
Marie-Francine Moens, Belgium
Niranjan Nagarajan, Singapore
Davor Petrinović, Croatia
Boris Podobnik, Croatia
Tomislav Pribanić, Croatia
Krešimir Pripužić, Croatia
Ivica Puljak, Croatia
Strahil Ristov, Croatia
Damir Seršić, Croatia
Karolj Skala, Croatia
Vernesa Smolčić, Croatia
Siniša Srbljić, Croatia
Marko Subašić, Croatia
Mile Šikić, Croatia
Jan Šnajder, Croatia
Hrvoje Štefančić, Croatia
Kristian Vlahoviček, Croatia
Mladen Vouk, USA
Boris Vrdoljak, Croatia
Mladen Vučić, Croatia
Vinko Zlatić, Croatia

[3] Shabestary, S. M. A., Baher, A.: Deep Learning vs. Discrete Reinforcement Learning for Adaptive Traffic Signal Control, 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, Hawaii, USA, November. 4-7, 2018, pp. 286-293

Supervised learning approach to long read classification

L. Vrček, M. Šikić

University of Zagreb, Faculty of Electrical Engineering and Computing

Determining the complete genetic material of an organism is a central task of genomics as it would enable various applications in medicine and biotechnology. In order to make that possible, several techniques for DNA sequencing have been developed. The most recent approach, called third generation sequencing technology, provides us with DNA sequences of length that can surpass a 100,000 bases, but that comes at a cost of high sequencing error which is often greater than 15%. The most popular approach for assembling the genome using the mentioned sequences, also called reads, is based on the OLC paradigm which consists of three steps: overlap, layout and consensus. In the overlap phase all the reads are mutually aligned which provides an overlap graph. In the assembly step such graph has to be simplified in order to obtain the Hamiltonian path which defines the assembly genome. In the final phase, consensus, the polishing of the assembly genome is performed by comparing the assembly with the reads obtained from the sequencer.

However, not only is finding a Hamiltonian path in a graph an NP-hard problem, but due to high error rate of the third generation sequencers the overlap graph can also be overly complex. In order to avoid that, reads are first analyzed and divided into three groups: regular, chimeric and repeat. When mapped to the reference genome, regular reads have uniform coverage since they have a unique position in the genome. Chimeric reads are usually created as a flaw of sequencer which connects two distant regions into a single read which is characterized as a sudden drop in coverage. Repeat reads have a significantly higher coverage at either end of a read, originating from overlap of bases from that end with multiple positions on a reference genome. The first assembly tool to perform such analysis was HINGE [1]. It observes pile-o-grams of each read, which are plots of coverage versus base index. Another tool that utilizes similar method is Ra [2]. It stores signals from pile-o-grams into vectors of unsigned short integers and calculates the coverage slope at each position by keeping the sliding window on both sides of the observed position. The importance of identifying reads as chimeric and repeats is that they can produce complex overlap graph or add errors into the assembly. Therefore, each type is dealt in its own manner: chimeric

reads are cut and only the longest non-chimeric region is retained and ridges are removed from repeat.

In this work, we introduce a supervised learning approach for solving this problem. Since the main goal is to improve de-novo genome assembly reads are not mapped onto the reference, but onto each other. This introduces noise into the pile-o-grams and makes the classification more difficult. To deal with that, reads are also mapped onto the reference using the ratlesnake tool [3] which helps classify dubious reads during the training. The training is performed on grayscale images of pile-o-grams on a dataset of 5.000 reads of each class, using the convolutional neural networks. Accuracy obtained for regular reads is 88.93 %, for chimeric 88.17 %, while for repeat reads it is 86.99 %. Code can be found at https://github.com/lvrcek/LongReadClassification under the MIT license.

References:

[1] G. M. Kamath, I. Shomorony, F. Xia, T. A. Courtade, and D. N. Tse, "HINGE: Long-read assembly achieves optimal repeat resolution", Genome Research, 2017.
[2] R. Vaser and M. Šikić, "Yet another de novo genome assembler", 2019.
[3] ratlesnake, https://github.com/lbcb-sci/ratlesnake

RNA Splice Aware Mapper

J. Marić, K. Križanović, M. Šikić

Faculty of Electrical Engineering and Computing and University of Zagreb, Faculty of Electrical Engineering and Computing

Advances in sequencing technology achieved by companies such as Oxford Nanopore technologies (ONT) and Pacific Biosciences (PacBio) have resulted in production of long reads that are over 10 kbp in length. Initially, such long reads had high error rate which has steadily improved and the latest generation of PacBio protocols produce reads comparable in accuracy to Illumina short reads. However, most of long-read technologies in use still produce error rate up to 10 %. Although, short reads are still predominantly used in the field of RNA-seq analysis, longer reads help in detection and quantification of isoforms and better annotation of new genomes. Algorithmically, mapping RNA-seq reads to known transcripts is equivalent to mapping DNA reads. Yet, mapping these reads to eukaryotic genomes is more complex due to RNA splicing.

In this paper we present a new splice-aware mapping method, built upon our previously developed DNA mapping method, which is tailored for long reads produced by Pacific Biosciences and Oxford Nanopore devices. It uses several

Supervised learning approach to long read classification

Lovro Vrček[1], Mile Šikić[1,2]

[1]University of Zagreb, Faculty of Electrical Engineering and Computing
[2]Bioinformatics Institute, Singapore 138671, Singapore

Centre of Research Excellence for Data Science and Advanced Cooperative Systems

INTRODUCTION

Determining the complete genetic material of an organism is a central task of genomics, as it would enable various applications in medicine and biotechnology. The most recent approach, called third generation sequencing, provides us with reads that can surpass 100,000 bases, but have error rate often greater than 15%. That can result in faulty and overly complicated overlap graphs obtained from the first phase of the OLC assembly process.

PROBLEM DESCRIPTION

Main problem in the genome assembly is finding a Hamiltonian path in the overlap graph. To make that possible, reads first have to be classified into one of the three classes which helps determine to what extent they are useful in the assembly. Those classes are chimeric, repeat and regular, whose pile-o-grams can be seen on the Figure 1. Pile-o-gram is a plot of coverage versus base index obtained by mapping read onto a reference genome.

METHODOLOGY

In this work a supervised learning approach for classifying reads is introduced. Since the main goal is to improve *de-novo* assembly in which the reference genome is not given, the reads are first mapped onto each other. This introduces noise and makes the classification of pile-o-grams more difficult. Thus, for the training process reads have also been mapped to reference genome in order to more precisely determine to which class each pile-o-gram belongs. Reference mapping was not used in the validation and testing phase in order not to overfit the model. Training was performed on grayscale images of pile-o-grams and the used models were AlexNet and ResNet50, both convolutional neural networks proven to be useful in computer vision problems.

PRELIMINARY RESULTS

Data was obtained from the Zymo dataset consisting genomes of eight bacterial and two yeast species. Five thousand pile-o-grams of each class were generated and horizontal mirroring was performed on each image in order to expand the dataset. When classifying with AlexNet accuracy was 84.10% for regular reads, 84.11% for chimeric and 81.34% for repeats. When ResNet50 was used, accuracies were 88.93%, 88.17% and 86.99% respectively. Confusion matrices for both cases can be seen Table 1 and 2, respectively. It is worth mentioning that, while these results fall short of heuristic methods, dubious pile-o-grams which can't be classified by heuristics have not yet been inserted into the dataset.

Table 1: Confusion matrix for AlexNet model

		Predicted class		
		Regular	Chimeric	Repeat
Actual class	Regular	1742	121	207
	Chimeric	179	1656	135
	Repeat	147	218	1595

Table 2: Confusion matrix for ResNet50 model

		Predicted class		
		Regular	Chimeric	Repeat
Actual class	Regular	1841	103	126
	Chimeric	102	1737	131
	Repeat	108	147	1705

CONCLUSIONS

We have developed a method based on supervised learning and convolutional neural networks for classifying pile-o-grams into three classes. The accuracy almost reaches that of heuristic methods, but by introducing difficult pile-o-grams that heuristics are unable to classify, we believe this method could surpass heuristic approach and become dominant approach for classifying reads in order to improve de-novo genome assembly.

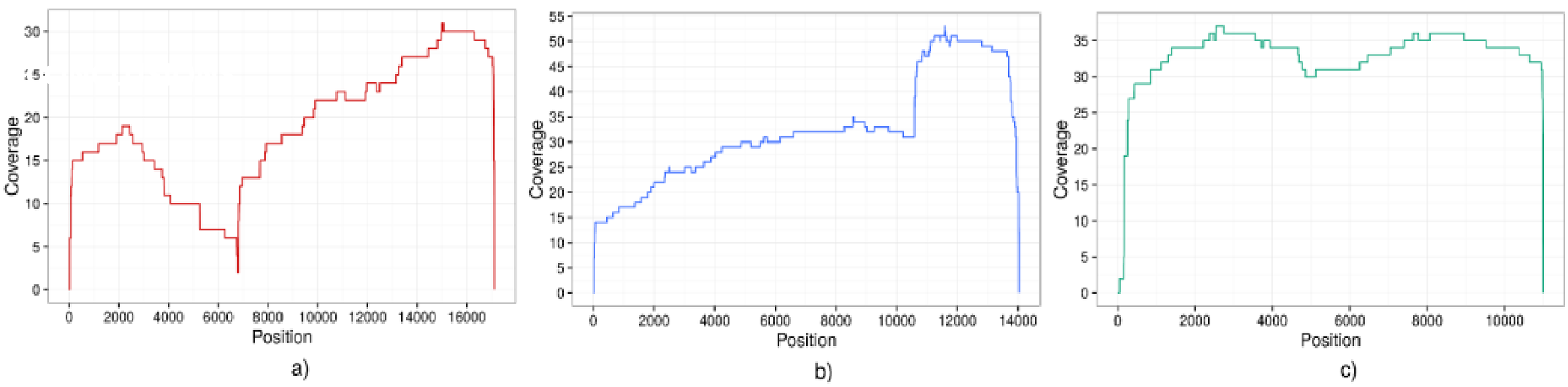

Figure 1. Coverage graph of a) chimeric read, b) repeat read and c) regular read. Chimeric reads are characterised by the sudden drop in coverage (also called rift), repeat reads by a higher coverage on one end of the read (called ridge), while regular reads have approximately uniform coverage.

ACKNOWLEDGMENT

This research was funded by the European Union through the European Regional Development Fund, under the grant KK.01.1.1.01.0009 (DATACROSS) and has been supported in part by the Croatian Science Foundation under the project Single genome and metagenome assembly (IP-2018-01-5886) and „Young Researchers" Career Development Program.

Europska unija
Zajedno do fondova EU

Ministry of Science and Education

EUROPSKI STRUKTURNI I INVESTICIJSKI FONDOVI

Operativni program KONKURENTNOST I KOHEZIJA

This project was supported by European Union's European Regional Development Fund

HRZZ Hrvatska zaklada za znanost

The contents of this poster are the sole responsibility of the University of Zagreb Faculty of Electrical Engineering and Computing and do not necessary reflect the views of the European Union.