Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 110 (2017) 440–447

Procedia Computer Science

www.elsevier.com/locate/procedia

The 1st International Workshop on Algorithms, Tools and new Frontiers on the use of Networks in Biology and Clinical Science (BioNet 2017)

Efficient data structures for mobile de novo genome assembly by third-generation sequencing

Franco Milicchio[a], Mattia Prosperi[b] *

[a]*Department of Engineering, Section of Informatics & Automation, Roma Tre University, Rome, Italy*
[b]*Department of Epidemiology, College of Public Health and Health Professions, College of Medicine, University of Florida, Gainesville, Florida, USA*

Abstract

Mobile/portable (third-generation) sequencing technologies, including Oxford Nanopore's MinION and SmidgION, are revolutionizing once again –after the advent of high-throughput sequencing– biomedical sciences. They combine an increase in sequence length (up to hundred thousands of bases) with extreme portability. While a sequencer now fits the palm of a hand and needs only a USB outlet or a mobile phone/tablet to work, the data analysis phases are bound to an available Internet connection and cloud computing. This somehow hampers the portability paradigm, especially if the technology is used in resource-limited settings or remote areas with limited connectivity. In this work, we introduce efficient data structures to effectively enable portable data analytics by means of third-generation sequencing. Specifically, we show how sequence overlap graphs (fixed length *k*-mers, with an extension on variable lengths) can be built and stored on a mobile phone, thereby allowing the execution of de novo genome assembly algorithms (along with ad-hoc strategies for error correction) without the need of transfer data over the Internet nor execution on a desktop.

Peer-review under responsibility of the Conference Program Chairs.

Keywords: DNA sequencing, nanopore, genome assembly, de Bruijn graphs, mobile computing, cache-oblivious

* Corresponding author. Tel.: +1-352-273-5860; fax: +1-352-273-5365.
E-mail address: m.prosperi@ufl.edu

1877-0509
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.06.115

1. Introduction

Portable (third-generation) sequencing technologies have literally brought high-throughput sequencing to the palm of a hand [1]. The chemistry behind is 'nanopore' sequencing, which exploits 1-nanometer holes on a membrane through which a DNA molecule is passed and recognized [2]. The advantage of nanopore sequencing is that very long DNA strands can be retrieved (up to hundred thousands of bases), in opposition to shorter fragments usually output by high-throughput sequencers, e.g. Illumina (hundreds of bases) or even PacBio (thousands). Currently, Oxford Nanopore Technologies ltd (Oxford Science Park, Oxford OX4 4GA, UK) is the sole relevant market supplier with three machineries: the MinION sequencer (on sale since 2015), PromethION (early access), and SmidgION (under development). The MinION sequencer weighs 90g and measures 10×3×2cm, making it the smallest sequencing device currently available on the market. The maximum throughput is within the Gigabyte range; read (i.e. sequence) lengths have a median of a few Kilobases, but can reach a hundred of Kilobases [3]. Error rates estimates for MinION are in the order of 10%-30% which is consistently higher than other high-throughput technologies, yet promising given that they have been decreasing consistently over time [4, 5].

The MinION output format is called FAST5, which is based on the hierarchical data format 5 (HDF5, http://www.hdfgroup.org/HDF5/) and allows for metadata content. Currently, FAST5 preprocessing, such as base calling, is done by transferring files over the Internet on to a dedicated service, the Metrichor (https://metrichor.com). Given the high error rates and the peculiar read length distribution, many ad-hoc software for data analytics have been developed and released, also to overcome the dependence on an entirely web-based data processing pipeline [6]. De novo assembly methods are still based on the *k*-mer (fixed string length) and read overlap (variable string length) graph paradigms; for instance the popular Celera Assembler has been adapted to nanopore data, in a suite named Canu [7].

While a MinION and SmidgION need only a USB outlet or a mobile phone/tablet to work, the data analysis phases are bound to an available Internet connection and cloud computing. This somehow hampers the portability paradigm, especially if the technology is used in resource-limited settings or remote areas with limited connectivity. Most of MinION applications have been 'on the field', and included rapid identification of pathogens such as Ebola [8], and environmental/food safety monitoring [9].

To date, there is no software purposely built to run on directly portable devices such as tablets or phones. While this may seem an impossible objective for very high throughput machines like Illumina, it can be feasible for the MinION. It is true that absence of Internet does not imply absence of a desktop/laptop on which MinION or PromethION data can be processed; nonetheless, the upcoming SmidgION directly plugs into a smartphone and opens up a vast perspective for development in this direction, since smartphone and tablet CPUs have a very different architecture than personal computers or servers. We have seen how gaming, computer-aided design, and other leisure/working activities have been migrating into portable devices, and this could become also the case for biotechnology analytics.

Therefore, in this work we introduce original data structures to enable portable, on-chip assembly of genomes using third-generation sequencing. In detail, we show how a de Bruijn graph (discussing the extension for a generic read overlap graph, used by most nanopore assemblers) can be computed and stored efficiently on mobile phone chips such as A10 mounted on iPhone 7 (Apple Inc.), thereby allowing the execution of de novo genome assembly algorithms (along with embedded strategies for error correction) without the need of transfer data over the Internet.

2. Methods

Smartphones and tablets have limited amount of RAM (but similar levels of cache) as personal computers or servers. Likewise, a mobile CPU can be less performant than that of a standard desktop, often to reduce battery usage and excessive heating, although the gap is narrowing every day. For instance the iPhone 7 has only 3GB of RAM, three cache levels, and the A10 CPU at 2.3GHz has benchmarks comparable to a 2013 MacBook Pro or even a more recent MacBook Air (https://tinyurl.com/jgkafcf). Therefore, the main problems relative to make de novo assembly feasible on a mobile device are: (1) to parse, process and store MinION read files into opportune data structures such that they fit the RAM; (2) to implement efficient data structures for de Bruijn / read overlap graphs that allow also error correction (e.g. storing quality scores) besides generic Eulerian / Hamiltonian path discovery;