The Malta Human Genome Project

Progress Report

Sara Ann Abdilla (188396M)

Supervisor(s): Jean Paul Ebejer

Faculty of ICT

University of Malta

December 2016

Submitted in partial fulfillment of the requirements for the degree of B.Sc. Computing Science (Hons.)

Contents

Abstract: The abstract should act as a stand-alone (very) brief description of the whole story: The context, the solution, how effective it was found to be. There is no better way to learn how to write an abstract than by carefully reading the abstracts of good papers. This is usually the last part of the report to be written.

1 Introduction

Over the years, more and more human reference genomes are being assembled. A reference genome is a digital database consisting of DNA sequences, each individual's DNA corresponding to an arrangement of approximately 3 billion bases (made up from four nucleotide bases). These possible bases are adenine, cytosine, guanine and thymine - oftenly referred to by the letters A, C, G and T respectively [Hun93, Les12].

Genome assembly/sequencing technologies are rapidly advancing and their costs are decreasing, both due to the fact that their importance is becoming more widely known. DNA variations and mutations may correlate to diseases so discovering any approximately matching alignments between the reference genome and the DNA reads (sequencing reads) being analysed could very well help future medical diagnosis and treatments [thi16c, thi16a, thi16b].

The University of Malta is developing a National Maltese Human Reference Genome for this reason whereby whole genome sequencing on certain Maltese DNA samples (provided by the Malta BioBank) will be performed.

An amount of challenges need to be tackled, some of which are listed hereunder:

1. A substantial knowledge of bioinformatics - an area where computer technology is applied to biological data;

2. A collections of known algorithms relating to DNA sequencing and their performance comparisons;

3. Compression of reference genomes to improve efficiency, such studies having been already performed by Fritz et al [FLCB11] and Chen et al [CKL00] among others;

4. A data visualisation tool which promotes user-friendliness for even individuals who have no prior knowledge relating to the area in question;

5. A genome browser which also promotes user-friendliness as stated previously.

2 Background and Literature Review

3 Aims and Objectives

The aim of this project is to build tools for the visualization and analysis of genomes sequenced from the Malta Human Genome Project. The research areas studied are bioinformatics, big data, data storage, data visualisation and data analysis among others.

The main objectives of the system are as follows:

1. The alignment of a reference genome against a number of sequencing reads such as that done by Lee et al [LHX$^+$14] and Li et al [LD09] among others;

2. The data visualiation of human genomes;

3. The construction of a genome browser using novel and established components in order to reference DNA for comparitive genomics and to analyse DNA mutations.

4. A comparitive review of existing methods against all the above points.

4 Methods and Techniques Planned

5 Evaluation Stategy

The research question which should be answered is

6 Expected Deliverables

1. The Final Year Project (FYP) which will include: all the relevant background information required to understand said project, a detailed explanation of the system and the evaluation results;

2. The implementation of the designed and developed system;

3. The documentation which will explain to the users how the system should be employed.

References

[BPS14] Bonnie Berger, Jian Peng, and Mona Singh. Computational solutions for omics data. Technical report, USA, 3 2014.

[CKL00] Xin Chen, Sam Kwong, and Ming Li. A compression algorithm for dna sequences and its applications in genome comparison. Technical report, New York, NY, USA, 4 2000.

[FLCB11] Markus Hsi-Yang Fritz, Rasko Leinonen, Guy Cochrane, and Ewan Birney. Efficient storage of high throughput dna sequencing data using reference-based compression. Technical report, United Kingdom, 1 2011.

[HPB13] Lin Huang, Victoria Popic, and Serafim Batzoglou. Short read alignment with populations of genomes. Technical report, 2013.

[Hun93] Lawrence Hunter. *Artificial Intelligence and Molecular Biology.* MIT Press, AAAI Press, 445 Burgess Drive, Menlo Park, California 94025, USA, 1993.

[LD09] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrowswheeler transform. Technical report, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, 5 2009.

[Les12] Arthur M. Lesk. *Introduction to Genomics.* Oxford University Press, Great Claredon Street, Oxford, OX2 6DP, UK, 2 edition, 2012.

[Les14] Arthur M. Lesk. *Introduction to Bioinformatics.* Oxford University Press, Great Claredon Street, Oxford, OX2 6DP, UK, 4 edition, 2014.

[LHX$^+$14] Donghyuk Lee, Farhad Hormozdiari, Hongyi Xin, Faraz Hach, Onur Mutlu, and Can Alkan. Fast and accurate mapping of complete genomics reads. Technical report, 10 2014.

[Mou01] David W. Mount. *Bioinformatics: Sequence and Genome Analysis.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA, 2001.

[RLK11] Matthew Ruffalo, Thomas LaFramboise, and Mehmet Koyutrk. Comparative analysis of algorithms for next-generation sequencing read alignment. Technical report, 7 2011.

[SZV$^+$14] Jing Shang, Fei Zhu, Wanwipa Vongsangnak, Yifei Tang, Wenyu Zhang, , and Bairong Shen. Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. Technical report, 3 20114.

[thi16a] Blood, genes, you. *Think Magazine*, 16:26–31, 4 2016.

[thi16b] Heartbreakers. *Think Magazine*, 16:32–37, 4 2016.

[thi16c] The hidden history of the maltese genome. *Think Magazine*, 16:19–25, 4 2016.