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 aim of this report is to detail the groundwork accomplished on developing the visualisation tool and the genome browser required for the Malta Human Genome Project. The problem is introduced with the motivation behind it along with the reasoning for its nontriviality. The reader is then provided with the required background knowledge in order to properly discern the problem, followed by the project's objectives. The current and intended techniques and evaluation strategy are explained. Finally, the expected deliverables are listed with an accompanying Gantt chart.

1 Introduction & Motivation

Over the years, more and more human reference genomes are being globally 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. The possible nucleotide bases are adenine, cytosine, guanine and thymine - oftenly referred to by the letters A, C, G and T respectively [1, 2].

Genome assembly/sequencing technologies are rapidly advancing and their costs are decreasing; both factors being 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 sequencing reads being analysed could very well help future medical diagnosis and treatments [3].

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. The American human genome sequencing facility Complete Genomics, which was founded in 2006, will be a partner in this project.

Thus, computation wise, a data visualisation tool along with a genome browser need to be constructed in order to aid in this endeavour.

2 Reasoning for Nontriviality of Problem

Globally, large genome projects are being sequenced rapidly. Examples of such projects include the 1000 Genomes Project and the International Cancer Genome Project [4, 5, 6]. Malta should aim to be a part of this endevaour so that further studies can be conducted involving a larger variety of genes (i.e. the Maltese genes). After all, while the reading of an individual's DNA shows the likeliness of that person developing a disease, the reading of a nations DNA shows why that population is more likely to develop a disease [3]. Such analyses can therefore prove to not only medically aid global research but also nation-wide studies.

3 Background Research and Literature Review

The development of a genome assembly technology consists of multiple stages; the main ones being genome compression, sequence alignment, alignment visualisation and genome browser construction;

all the steps following one another. The following points detail some research which has already been conducted in these areas.

3.1 Genome Compression Tools

Genomes, particularly human ones, consist of a large amount of data. For this reason, in order to efficiently analyse them, said genomes are compressed using various methods. For example, Chen et al devised the lossless *GenCompress* algorithm which implements certain established compression algorithms such as Lempel-Ziv. By analysing approximate matches based on the evaluated edit distances, it was found that such a method not only achieves the best compression ratio but also finds common sections in DNA sequences [7]. Another example would be the lossless reference-based compression algorithm devised by Fritz et al. It implements established components such as Golomb codes and De Bruijn graphs and was found to be quite efficient for read alignments similar to the reference genome [5]. Other known algorithm are *Biocompress-2* and *Cfact* [7] along with *DNACompress* and *DNAZip* [5].

3.2 Read Alignment Tools

Next-Generation Sequencing (NGS) technologies are evolving rapidly and to keep up with evolution, multiple read aligners are being produced; the most known being BWA, Bowtie, Soapv2, MAQ, BOAT, SHRiMP2 for the NGS platforms Illumina, Roche454 and ABI SOLiD. The majority implement Burrows-Wheeler Transform (BWT) but there are also those which use FM-indexing, Smith-Waterman, and Needleman-Wunsch algorithms [4, 6, 8]. All these mentioned aligners are quite efficient relative to the task they are given; for example, while some aligners may be more efficient with short reads, others may be the opposite [9, 10]. Other known algorithms which assist in approximate read mapping are the Hamming distance as well as the Levenshtein distance, both of which return a measure of the similarity between two sequences.

3.3 Critique of Missing Studies

Regarding alignment visualisation and genome browser construction, there do not seem to be any concrete algorithms for such cases. This is a fundamental area in genome assembly technology which should be researched extensively; i.e. part of this project's goals.

4 Aims and Objectives

The aim of this project is to build tools for the analysis of sequenced genomes 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 compression of a reference genome such as that done by Chen et al. [7] and Fritz et al. [5] among others;
2. The alignment of a reference genome against a number of sequencing reads such as that done by Lee et al. [8] and Li et al. [4] among others;
3. The data visualiation of human genomes;
4. The construction of a genome browser using novel and established components in order to reference DNA for comparitive genomics and to analyse DNA mutations;
5. A comparitive review of existing methods against all the above points.

5 Methods and Techniques Used or Planned

This section details the components which have been implemented along with a description of future plans.

5.1 Genome Compression

The genome compression method implemented makes use of certain in-built Python modules in order to compress the genome into a binary file by means of integers (each integer being of length 4 bytes). Each nucleotide base is assigned two bits to represent it as there are only 4 possible bases. Chen et al. used the same deduction when developing *GenCompress* [7].

5.2 Sequence Alignment

Four possible alignment methods were then taken into consideration - Hamming distance, Levenshtein/Edit distance, k-mer indexing and the most frequently used: FM-indexing with BWT. These were first implemented and tested on string inputs; i.e. on an uncompressed genome of marginally less size than the human genome to confirm their applicability. These same functions were then converted in order to instead support integer inputs due to the previously mentioned genome compression. Analysing each algorithms' resulting match rate with the corresponding time taken led to the deduction that the ... algorithm performed best with ... matches done in ... seconds.

5.3 Alignment Visualisation

The data visualisation tool's initial construction is based on *Tkinter*, Python's standard Graphical User Interface (GUI) package. The result is a depiction of the reads aligned with the genome, with lines repsresnting the position and length of each alignment/match. As of now, its interactivity is in the form of clicking a line at a certain position to output the offset of the match at said point of clicking.

5.4 Planned Methods

As for future development, firstly, the data visualisation tool needs to be further refined; i.e. become more interactive in terms of genome analysis. Secondly, the genome browser tool also needs to be constructed using established technologies for efficient comparitive genomics. Each method developed will be analysed and discussed to deduce the most feasible implementation.

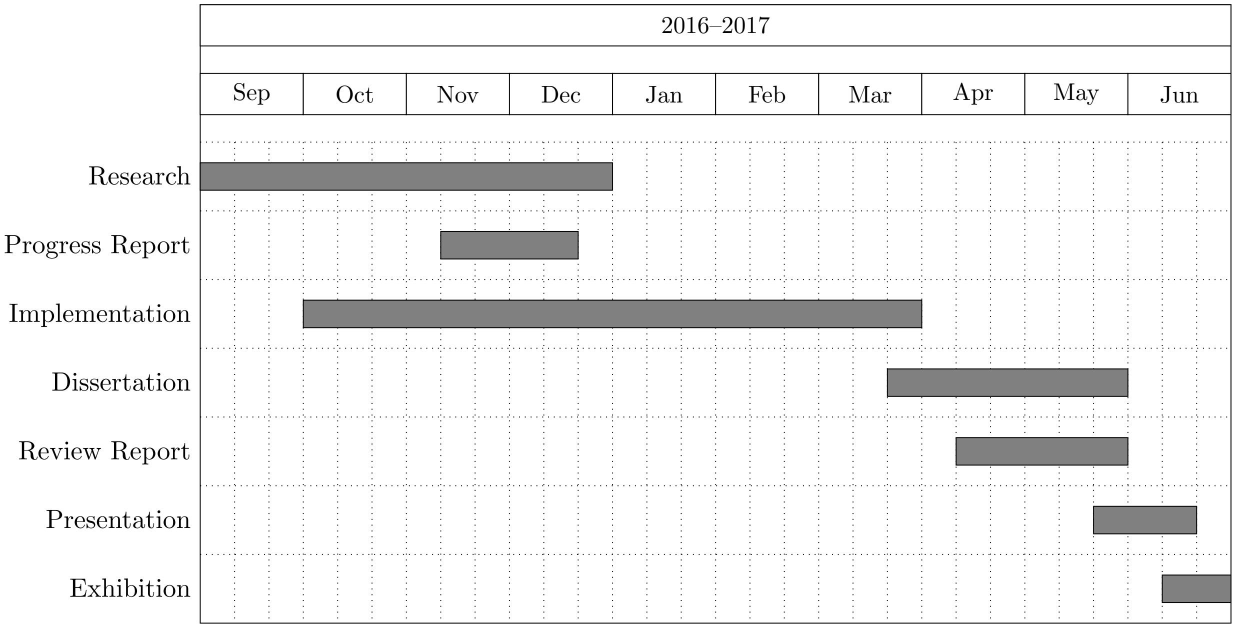
6 Proposed Evaluation Stategy

The main evaluation techniques proposed are as follows:

1. Comparitive review of the developed human genome visualisation tool with existing methods;
2. Reference DNA and analyse DNA mutations using the constructed genome browser.

7 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.
   The milestone schedule is represented in the following Gantt chart:



References

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

[2] A. M. Lesk, *Introduction to Genomics.* Great Claredon Street, Oxford, OX2 6DP, UK: Oxford University Press, 2 ed., 2012.

[3] "The hidden history of the maltese genome," *Think Magazine*, vol. 16, pp. 19–25, 4 2016.

[4] H. Li and R. Durbin, "Fast and accurate short read alignment with burrowswheeler transform," tech. rep., Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, 5 2009.

[5] M. H.-Y. Fritz, R. Leinonen, G. Cochrane, and E. Birney, "Efficient storage of high throughput dna sequencing data using reference-based compression," tech. rep., United Kingdom, 1 2011.

[6] L. Huang, V. Popic, and S. Batzoglou, "Short read alignment with populations of genomes," tech. rep., 2013.

[7] X. Chen, S. Kwong, and M. Li, "A compression algorithm for dna sequences and its applications in genome comparison," tech. rep., New York, NY, USA, 4 2000.

[8] D. Lee, F. Hormozdiari, H. Xin, F. Hach, O. Mutlu, and C. Alkan, "Fast and accurate mapping of complete genomics reads," tech. rep., 10 2014.

[9] M. Ruffalo, T. LaFramboise, and M. Koyutrk, "Comparative analysis of algorithms for next-generation sequencing read alignment," tech. rep., 7 2011.

[10] J. Shang, F. Zhu, W. Vongsangnak, Y. Tang, W. Zhang, , and B. Shen, "Evaluation and comparison of multiple aligners for next-generation sequencing data analysis," tech. rep., 3 20114.

[11] B. Berger, J. Peng, and M. Singh, "Computational solutions for omics data," tech. rep., USA, 3 2014.

[12] A. M. Lesk, *Introduction to Bioinformatics.* Great Claredon Street, Oxford, OX2 6DP, UK: Oxford University Press, 4 ed., 2014.

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