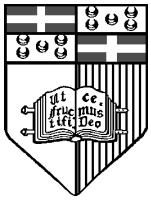
The Malta Human Genome Project

Progress Report

Sara Ann Abdilla (188396M)

Supervisor(s): Dr 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: Human genomes are being globally assembled in order to aid in medical diagnosis, forensic research, genealogy and bioinformatics among other areas. While many countries have already had a reference genome constructed for their population, Malta is still in the process of accomplishing this - a National Maltese Human Reference Genome. In this dissertation, I propose to aid in this project (computation wise) by developing a data visualisation tool along with a genome browser for DNA analysis.

1 Introduction & Motivation

Over the years, more people's genomes are being sequenced in order to identify diseases. This sequencing is stored in a reference genome which is a digital database consisting of DNA sequences; each individual's DNA corresponding to an arrangement of approximately 3 billion bases. DNA (deoxyribonucleic acid) could be described as a model of biological life which encompasses genetic information/genes whereby its 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. DNA variations and mutations may correlate to diseases so discovering any approximately matching alignments between the reference genome and the DNA sequencing reads (parts of DNA produced by the sequencer in question) being analysed could very well help future medical diagnosis and treatments [3]. After all, genes are hereditary from one generation to the next.

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, is a partner in this project.

The problem of aligning genomes comes into play here. In order to accurately and efficiently match alignments between a human reference genome and the respective DNA sequencing reads, different algorithms must be applied. The most common ones use Burrows-Wheeler Transform [4, 5] but there are multiple other variations; all of these being explained further on.

Thus, computation wise, a data visualisation tool along with a genome browser need to be constructed in order to aid in this endeavour. These developments are only possible after the required alignment is completed, as discussed above.

2 Reasoning for Nontriviality of Problem

Globally, large genome projects are being sequenced rapidly. Examples of such projects include the 1000 Genomes Project (now known as the 100,000 Genomes Project in the UK) and the International Cancer Genome Project [4, 6, 7]. Malta should aim to be a part of this endeavour so that further studies can be conducted involving a larger variety of genes (i.e. the Maltese genes).

DNA is analogous to a storage device of genetic information so it aids in multiple research areas such as genealogy, forensics, medical diagnosis and bioinformatics among others. 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.

Hence the problem is non-trivial as Malta has a diverse history and gene pool. By developing the tools required for the Malta Human Genome Project to be a success, Maltese medical research will surely take a step forward in the right direction.

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 (successively). 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 [8]. 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 [6]. Other known algorithm are *Biocompress-2* and *Cfact* [8] along with *DNACompress* and *DNAZip* [6].

3.2 Read Alignment Tools

Next-Generation Sequencing (NGS) technologies are evolving rapidly and to keep up with this 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 [4, 7, 5]. Reads are parts of DNA which are gathered by sequencers from people who wished to contribute to the area. They are given such importance as without them, not only can analysis of genomes not be performed but the actual reference genome would not even exist in the first place.

The majority of read aligners implement Burrows-Wheeler Transform (BWT) but there are also those which use FM-indexing, Smith-Waterman, and Needleman-Wunsch algorithms [4, 7, 5]. 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 more efficient with longer reads [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. These are not as commonly used due to their time complexity but they can still be implemented if reasoned about accordingly.

3.3 Genome Browser Construction

Genome browsers are graphical interfaces which are used in conjuction with genomic databases. They display the information found such that individuals are able to not only browse the stored genomes but also to visualise them, leading to easier data seaching and analysis. Most browsers are web-based applications which allow certain customisations according to the user's requirements, but stand-alone browsers exist as well [14, 15]. An example of a web-based genome browser is *GBrowse* which was constructed by Stein et. al [15].

Genome browser construction is very labour-intensive. Fortunately, many genome browser frameworks have already been constructed such as the most popular *GBrowse* (which was mentioned previously) as well as *Ensembl*, *JBrowse* and *LookSeq* among others. It should also be noted that there are two types of web-based genome browsers, multiple-species and species-specific browsers [14]. As this project deals with

a human genome, a species-specific browser must be implemented. Following the Generic Model Organism Database (GMOD) project, there are multiple open-source tools for this type of browser; *GBrowse*, again, being one of the most used frameworks [14].

3.4 Critique of Missing Studies

Firstly, regarding the genome compression tools discussed above, it is clear that a lot of research has already been done in the area. However specific implementations have not be studied; i.e. using in-built programming language modules as done in this project. Secondly, regarding the read alignment tools also discussed above, it is again clear that a large amount of research has been conducted in the area. However, as mentioned previously, most aligners do not use Hamming or Levenshtein distance so this approach will be one of the methods analysed in this project. Thirdly, regarding alignment visualisation, 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. Finally, regarding genome browser construction, it is again clear that a lot of research has already been conducted in the area. However not a lot has been done in relation to comparing them against each other and so, this project will analyse this aspect as well.

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. [8] and Fritz et al. [6];
2. The alignment of a reference genome against a number of sequencing reads such as that done by Lee et al. [5] and Li et al. [4];
3. The data visualiation of human genomes;
4. The construction of a genome browser using novel and established components (such as that done by Stein et al. [15]) 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 Materials and Methods

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* [8].

5.2 Sequence Alignment

Four possible alignment methods were then taken into consideration - Hamming distance (evaluates the number of mismatches between two sequences with regards to substitutions), Levenshtein/Edit distance (evaluates the number of mismatches between two sequences with regards to substitutions, insertions and deletions), k-mer indexing (evaluates all the possible subsequences of a sequence of length k into an index) and

the most frequently used: FM-indexing (evaluates all the possible subsequences of a sequence using Burrows-Wheeler Transform or BWT into an index). BWT transforms a character sequences into multiple runs having similar characters. 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.

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 Future Plans

As for future development, firstly, the read alignment methods need to be compared in order to deduced the most efficient in terms of the match rate with the corresponding time taken. Secondly, the data visualisation tool needs to be further refined; i.e. become more interactive in terms of genome analysis. This could be accomplished by outputting the complete character match at the point of clicking instead of just the offset, or perhaps even both. Thirdly, the genome browser tool also needs to be constructed using the established technologies (described in section 3) for efficient comparitive genomics. Each method developed will be analysed and compared to deduce the most feasible implementation.

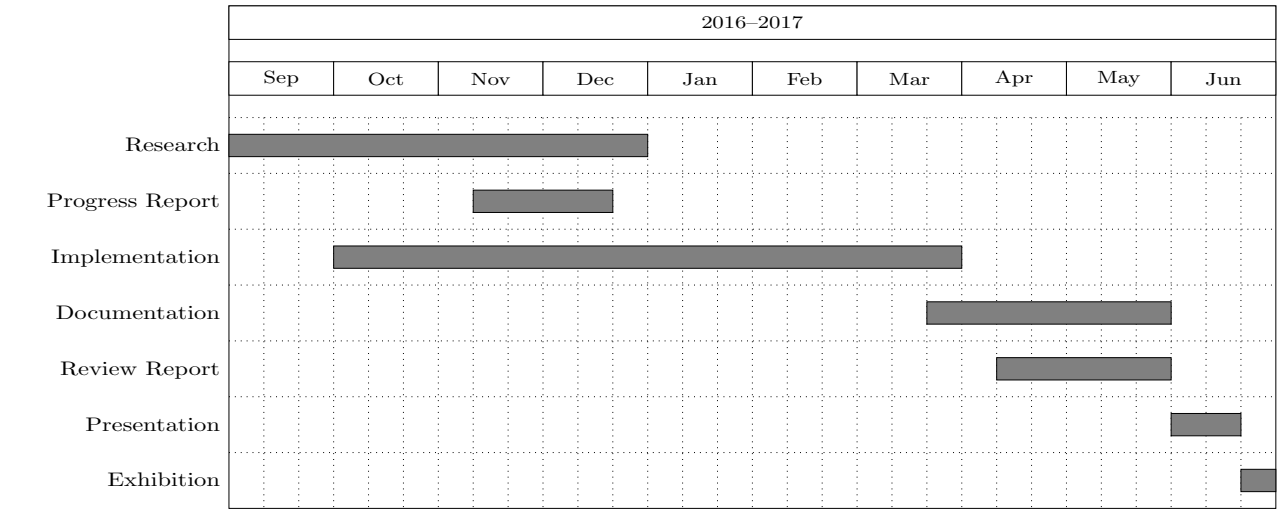
6 Proposed Evaluation Stategy

The main evaluation techniques proposed are as follows:

1. Comparitive review of the developed read aligner with the other developed aligners and with existing methods by comparing match rates with the corresponding time taken, depending on public data with regards to the existing methods;
2. Review of the developed human genome visualisation tool by testing that its interface is easily understood by most users - this could be done, for example, by making sure that the data is output in a clear manner;
3. Comparitive review of the constructed genome browser by referencing DNA and analysing DNA mutations followed by comparison of results with existing methods.

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] 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.

[6] 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.

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

[8] 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.

[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 2014.

[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.

[14] JunWang, L. Kong, G. Gao, and J. Luo, "A brief introduction to web-based genome browsers," tech. rep., 7 2012.

[15] L. D. Stein, C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J. E. Stajich, T. W. Harris, A. Arva, and S. Lewis, "The generic genome browser: A building block for a model organism system database," tech. rep., 12 2002.