UiO : Department of Informatics
University of Oslo

# Alignment against a graph-based reference genome

## Fuzzy searching in large and complex structures

Esten Høyland Leonardsen

Master's Thesis Spring 2016

# Alignment against a graph-based reference genome

Esten Høyland Leonardsen

2nd February 2016

# Abstract

# Contents

# List of Figures

# List of Tables

# Preface

# Part I

# Introduction

# Chapter 1

# Background

## 1.1 DNA

*Deoxyribonucleic acid* (DNA) is a molecule which allows organisms to store and pass on genetic information. The molecule is stored in the cells and encodes proteins which regulates the vital functions of the organism. DNA is able to replicate itself and can thus live on for generations through reproduction.

The DNA of an individual is made up by two complementary strands of nucleotides bound together in a double helix. The nucleotides of DNA can contain the bases Adenine, Cytosine, Guanine or Thymine, typically denoted A, C, G and T. Complementary in this context means that instead of one singular sequence of bases DNA is made up by a sequence of paired bases, A's with T's and C's with G's, called base pairs (bp). Due to the chemical structure of the molecules making up a single strand of DNA each strand can be said to have a direction, upstream towards the 5′ end or downstream towards the 3′ end. Two complementary strands have opposing directions and are thus called the reverse complements of eachother.

The size of DNA varies across species, from a couple of thousand basepairs (kb) in some viruses to several hundred billion basepairs (gb) in larger, more complex organisms. The human genome comes in at the higher end of this range, with a length of roughly 3 gb. A continuous sequence of bases is called a contig, several contigs combined is a scaffold which again sits together into chromosomes, the building block of the genome of an organism.

### 1.1.1 Genomic variations

Over the span of time DNA is subject to change. Through random mutations and recombination a genetic sequence can be changed either within an individual or as a product of reproduction. The fact that these changes are able to survive and propagate through generations leads to a genepool where even though the DNA comes from a common ancestor, different in-

dividuals will have different variants of the original sequence. These variations form the basis for the division into species, but even within species alot of natural variation will occur.

The least complex of these variations are Single Nucleotide Polymorphisms (SNPs), where a single nucleotide has changed between two individuals, and insertions and deletions (Indels) where either one or a short sequence of bases have appeared or disappeared from the DNA of an individual. Longer and more complex structural variations can also occur when a larger part of a chromosome breaks free and disappears completely or inserts itself in a different place or the opposite direction.

### 1.1.2 Alignment

The fact that a genome is built by discrete entities, the bases A, C, G and T, means that any DNA sequence can be represented by a text string. The double-stranded nature of DNA could be encoded into the string, but as one side can easily be derived from the other representing one of the strands is usually expressive enough. The process of determining genetic variation between two separate individuals can then be seen as the problem of finding the similarities and differences in the two corresponding text strings. Motivation?

#### String comparison in computer science

There exists several ways of determining the difference, also called the edit distance, between two strings mathematically. The main difference between the approaches are which operations are possible on the two strings and how the result is scored. The most common technique is called Levenshtein distance, which allows the operations deletion, insertion, and substitution. All of the preceding operations works on a single character, and one operation by itself yields a distance of 1. Finding the optimal Levenshtein distance can be done by dynamic programming in linear time.

#### DNA sequence comparison

### 1.1.3 Reference genomes

To say that an individual of a species deviates from the normal in any way there has to be a standard to compare it to, a reference genome. For humans the current reference genome is the GRCh38 developed and maintained by the Genome Reference Consortium. Something about the represented variation

## 1.2 Graph-based reference genomes

As sequencing technology improves and cost decreases we are able to sequence an increasing number of genomes. A natural consequence of

sequencing more genomes is discovering a larger number of variations. When mapping reads against a reference genome, knowledge of possible variants will provide a better basis for completing a correct mapping**??**. A natural way to represent this variation is through a graph, where different sequences share regions they have in common and branch out wherever they diverge from eachother. Motivation

### 1.2.1 De Bruijn graphs

### 1.2.2 Mapping/Coordinate system

When a reference genome is represented by a single sequence of letters, each letter can be uniquely identified by its position. Variations can then be stored based on the position they occur in, or the position they start in if they span several nodes. When using a graph-based reference the nodes are not a part of any naturally occuring coordinate system. If one also imagines the reference should be able to merge in new data at any given point the relationship between nodes might change and yield a fixed mapping scheme generated for the old graph stale and incorrect. In the article *Mapping to a reference genome structure***??**, Benedict Paten et al. introduce the concept of context-driven mapping. They define a context for a node N in a reference as a triplet (L, B, R) where B is the base contained in the node, L is a set of sequences generated by traversing the paths leading in to N and similarily R is a set of sequences generated by traversal of the paths going out from N. A unique context is a context mapping to exactly one node in the reference genome graph.

Further the authors go on to describe different mapping schemas using different variants of contexts. A general context-driven mapping scheme defines contexts as the sequences achieved by traversing the left context, the base of the node and the right context. A left-right exact mapping scheme requires either the left or the right context to be unique.

### 1.2.3 Alignment

## 1.3 Searching graphs

# Chapter 2

# Method

# Bibliography

[1] Deanna M. Church et al. 'Extending reference assembly models'. In: *Genome Biology* (2015).