Federal State Autonomous Educational Institution
for the Higher Education
National Research University "Higher School of Economics"
Faculty of Mathematics

Grachev Denis Vadimovich

Clustering of Multidimensional Random Variables to Improve HMM Sequence Alignment Accuracy

Project proposal

Scientific supervisor:
Prodanov Timofey Petrovich

Moscow 2022

Contents

1 Introduction

Bioinformatics is an interdisciplinary science that aims to develops methods and software tools for understanding biological data. One of the ways to model haploid genome is to present it as pair of sequences or strings over the alphabet $\{A, C, G, T\}$. Modern technologies of reading genome do not sequence it as one continious string, but a number of random overlaping substrings that are called reads. As within one biological species genomes coincide almost completely, it is convenient to determine one reference genome for one species and identify for every individual deviations from the reference. These single nucleotide variants are called SNVs.

Taking into cosideration these facts and the fact that various errors happen during all stages of the process, a number of problems appears for example:

- **Genome assembly** is process of deciphering genome using reads obtained from it.
- **Sequence alignment** is process of arranging sequences to spot similarities between them.
- **Variant calling** is process of identifying SNVs of an individual based on reads aligned on the reference genome.

The most commonly used technologies nowadays, such as illumina, allow to sequence reads of length 200-500 bp. Sequencing human genome using such short length reads has many limitations. First, due to diploidy of humans, it is important to obtain long-range haplotype information. This might be difficult with short-reads provided by illumina. Secondly 3.6% of human genome consists of long highly repetative duplicated regeions that can not be uniquely aligned, which lowers accuracy of SNVs. Third-generation single-molecule sequencing (SMS) techologies, such as Oxford Nanopore Technologies allow to genereate longer reads of length 10-30 kb. This techology might help overcome limitations that short-reads have. Variant calling tools that were developed for short-reads do not show high accuracy when applied to long-reads sequenced, due to significant difference in error rates and type of errors between these two types of reads. Also these tools process data in short windows of a few hundred bases lengths and not designed to agregate haplotype information present in long-reads, which might be crucial for distinguishing true deviations from errors.

A number of new methods of variant calling were developed to work with long-reads, such as Deep Variant and Longshot that uses deep learning and pair-Hidden Markov model algorithms to effectivly work with such data. Accuracy of these methods can be improved by agrigating information from multiple sequence data sets together. An example of effectiveness of such method is Spades genome assembly tool, which uses short and long read data together.

Longshot works effectivly with one type of reads. To improve it's sencetivity and accuracy various methods of collecting data from diiferent sources can be applied. Previous research show that sequencing reads different profiles and clustering them by their profiles can improve variant calling accuracy.

During this work we are going to extend functionality of Longshot tool to work effectivly with multiple datasets. For that we have to develop clustering algorithm for read profiles and implement such feature into Longshot.

2 Methods

2.1 Clustering

Definition 2.1. Clustering algorithm aims to group points together into predefined number of sets.
Clustering algorithm is a map

$$\text{cluster}(X, m) \to C$$

$$X = \{x_i | x_i \in \mathbb{R}^d, i \in (1 \dots n)\}, m \in \mathbb{N}$$

$$C = \{c_i | c_i \in (1 \dots m), i \in (1 \dots n)\}$$

where m is number of clusters and n is number of points.

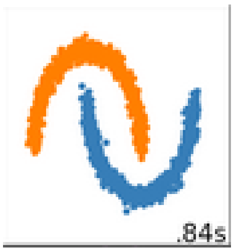

Figure 1: Example of clustering for $d = 2$, $m = 2$, color represents class.

2.2 Strings

Definition 2.2. String is a sequence of letters of finite size.
String of length l over alphabet $A = \{1 \dots m\}$ is a map $s : \{1 \dots l\} \to A$. Usually elements of A are denoted as characters for convenience.

Definition 2.3. Alignment of strings is a way representing them to spot similarities.
Alignment of strings s_1 and s_2 of lengths l_1 and l_2 respectively, over alphabet A is a pair of strings $\hat{s}_1$ and $\hat{s}_2$ of length l over alphabet $A \sqcup \{'-'\}$, such that there exists increasing functions $f_i : \{1 \dots l_i\} \to \{1 \dots l\}, i \in \{1, 2\}$ such that $\hat{s}_i \circ f_i = s_i$ and $\text{Im}(f_i) = \hat{s}_i^{-1}(A)$.

Letter $'-'$ represents gap in string. String $\hat{s}_i$ represents string s_i with inserted letter $'-'$, that was not present in alphabet A, into random places, and function f_i maps indexes of letters in s_i to corresponding indexes in $\hat{s}_i$.

Example 2.1. Alignment of strings $s_1 = CABCAABA$ and $s_2 = ABADBBAD$ over alphabet $\{A, B, C, D\}$.

Initial strings.

s_1	C	A	B	C	A	A	B	A
s_2	A	B	A	D	B	B	A	

Aligned strings.

$\hat{s}_1$	C	A	B	C	-	A	A	B	A
$\hat{s}_2$	-	A	B	-	A	D	B	B	A

Definition 2.4. For given matrix $G \in \mathbb{R}^{|A \sqcup \{'-'\}| \times |A \sqcup \{'-'\}|}$ and $p \in \mathbb{R}$ score of alignment $\hat{s}_1, \hat{s}_2$ is

$$S(\hat{s}_1, \hat{s}_2) = \sum_{i=1}^{l} g_{\hat{s}_1(i), \hat{s}_2(i)}.$$

Matrix G stores predefined penalties for mismatches and gaps and encouragement for matches.

There are other popular ways to define score of alignment, that penalty one longer gap less than several small gaps.

Definition 2.5. We are interested in the alignment with the biggest possible score. So we define score between two strings as

$$S(s_1, s_2) = \max_{\hat{s}_1, \hat{s}_2} S(\hat{s}_1, \hat{s}_2)$$

Definition 2.6. Substring is continious piece of a given string.
For a string s of length l, sub-string s_s is a string of length l_s, such that there exists an function $f : \{1 \dots l_s\} \to \{1 \dots l\}$ such that

$$f(i) = i + d$$

$$s \circ f = s_s.$$

Definition 2.7. For a string s_1 and s_2 of lengths l_1, l_2 correspondingly, define string-sub-string score as

$$S_s(s_1, s_2) = \max\{S(s_s, s_2) | s_s \text{ is a sub-string of } s_1\}$$

and corresponding alignment $\hat{s}_1, \hat{s}_2$.

Definition 2.8. For a string s of length l and set of strings $R = \{s_1 \dots s_n\}$ of lengths $\{l_1 \dots l_n\}$ correspondingly, multiple alignment is tuple $\hat{s}, \hat{s}_1 \dots \hat{s}_n$, of strings of length l over alphabet $A \sqcup \{'-'\}$, such that $\sum_{i=1}^{n} S(\hat{s}, \hat{s}_i)$ is maximal.

Definition 2.9. Set of reads R for string s of length l and rate r is

$$R = \{s_s | \text{length of } s_s > l, S_s(s, s_s) < r\}$$

2.3 Task

Given reference string s_r and reads R for an unknown target string s_t, we know that $S(s_r, s_t) < D$ and whant to find s_t.

Plan:

1. Make multiple alignment of R over s_r.
2. Estimate most likely difference between s_r and s_t.

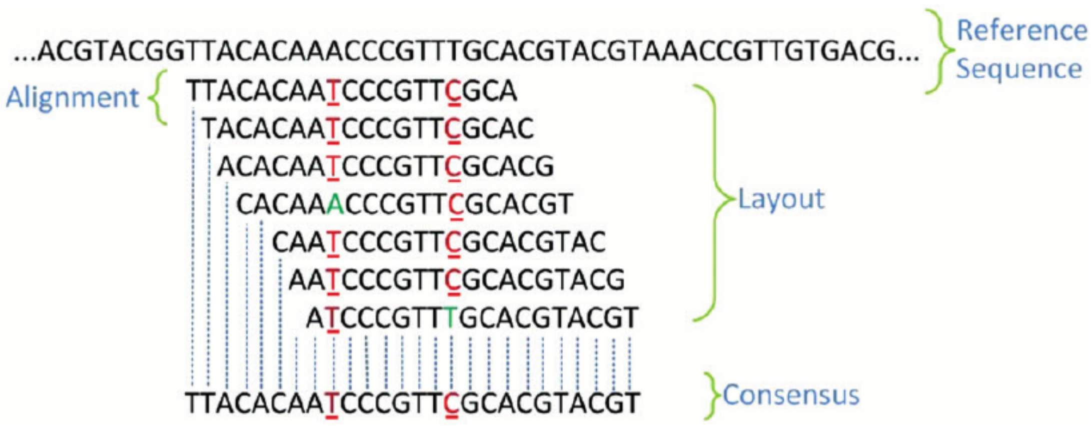

Figure 2: Example of reference string, target string and reads.

2.4 Pair Hidden Markov Model

Each step of pairwise alignment can be assigned to one of the three states $\{M, X, Y\}$, where M is a match, X is a gap in s_1, Y is a gap in s_2.

References

[1] Longshot is variant calling tool for long reads based on pair-Hidden Markov Model. `https://www.nature.com/articles/s41467-019-12493-y`