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 **3**

2 Methods **4**
2.1 Clustering 4
2.2 Strings 4

3 Task **5**

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 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 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 techologies genereate longer reads of length 10-30 kb. This techology might help overcome limitations that short-reads have. Tools that use

2 Methods

2.1 Clustering

Given $X = \{x_i | x_i \in \mathbb{R}^d, i \in (1 \dots n)\}$ and $m \in \mathbb{N}$, where n is the number of points, m - number of clusters.
Clustering algorithm takes X and m and outputs $C = \{c_i | c_i \in (1 \dots m), i \in (1 \dots n)\}$.

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

2.2 Strings

Definition 2.1. 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 2.2. 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 $f_i : \{1 \dots l_i\} \to \{1 \dots l\}$ such that $\hat{s}_i|_{\hat{s}_i^{-1}(A)} \circ f_i = s_i$.

Remark. $\mathrm{Im}(f_i) = \hat{s}_i^{-1}(A)$

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

$$\left\| \begin{array}{c|cccccccc} s_1 & \mathrm{C} & \mathrm{A} & \mathrm{B} & \mathrm{C} & \mathrm{A} & \mathrm{A} & \mathrm{B} & \mathrm{A} & \\ s_2 & \mathrm{A} & \mathrm{B} & \mathrm{A} & \mathrm{D} & \mathrm{B} & \mathrm{B} & \mathrm{A} & & \\ \hat{s}_1 & \mathrm{C} & \mathrm{A} & \mathrm{B} & \mathrm{C} & - & \mathrm{A} & \mathrm{A} & \mathrm{B} & \mathrm{A} \\ \hat{s}_2 & - & \mathrm{A} & \mathrm{B} & - & \mathrm{A} & \mathrm{D} & \mathrm{B} & \mathrm{B} & \mathrm{A} \end{array} \right\|$$

Definition 2.3. For given matrix $G \in \mathbb{R}^{|A| \times |A|}$ and $p \in \mathbb{R}$ score of alignment $\hat{s}_1, \hat{s}_2$ is

$$S(\hat{s}_1, \hat{s}_2) = \sum_{i=1}^{l} \delta_i, \text{ where } \delta_i = \begin{cases} g_{\hat{s}_1(i)\hat{s}_2(i)}, & \hat{s}_1(i) \neq - \text{ and } \hat{s}_2(i) \neq - \\ p, \end{cases}$$

Theorem 1. If G is symmetric and $g_{ij} = \begin{cases} 0, & i = j \\ > 0, \end{cases}$ and $p > 0$, then we can define metric for strings over alphabet A as

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

Proof. □

Definition 2.4. For a string s of length l, sub-string s_s is a string of length l_s, such that there exists an function

$$f : \{1 \dots l_s\} \to \{1 \dots l\}$$
$$f(i) = i + d$$
$$s \circ f = s_s$$

Definition 2.5. For a string s_1 and s_2 of lengths l_1, l_2 correspondingly, define string-sub-string score as

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

and corresponding alignment $\hat{s}_1, \hat{s}_2$ are pair of strings of lengths l over alphabet $A \sqcup \{-\}$ such that there exists increasing functions $f_1 : \{1 \dots l_1\} \to \{1 \dots l\}$

Definition 2.6. 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 minimal.

Definition 2.7. Set of reads R for string s of length l and rate r is

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

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.

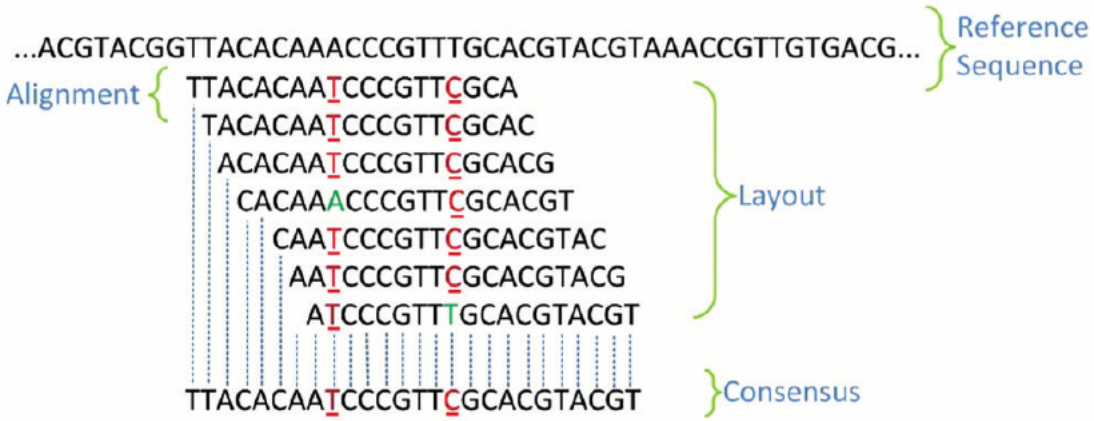

Figure 2: Example of reference string, target string and reads.