STEM Games 2018 - Technology Arena

Filip Šuste, Ivan Krpelnik, Robert Vaser, Marin Petričević, Petar Šegina,
Stanko Krtalić Rusendić, Luka Abramušić

Main task

1.1 Introduction

Knowing the actual genome sequence can aid many biological and medical studies. For example, being able to detect genetic variations (which include duplications, translocations, inversions, insertions and deletions) has a great impact on detection of diseases and genetic disorders. To be able to obtain that information, the order of nucleotide bases within a DNA sequence has to be determined through a process called genome sequencing. Sequencing technologies have emerged in the late 1970s and have gone through a lot of improvements but they all share the disability to read the whole genome at once. Today is the era of third generation of sequencing, which produce the longest fragments so far but with the flaw of relatively high error rates.

Once the fragments are obtained and if there exists a reference genome, the reference can be used as a backbone to reconstruct the sequenced genome. If a reference genome is not available, the fragments need to be stitched together with a tool called assembler in a process which is similar to assembling a jigsaw puzzles. Due to sequencing errors, many DNA molecules are sequenced in order to increase the confidence of the reconstruction. Number of molecules sequenced is called sequencing depth, and represents the average number of fragments covering a base in the sequenced genome.

This assignment will simulate a part of the process which determines whether an individual is prone to some of the known human diseases in the scenario where the reference genome is known.

1.2 Assignment

For a given reference genome (of a healthy individual) and a set of fragments obtained with third generation of sequencing, find all mutations in the genome of interest. Report their positions in the reference genome and visualize them.

Note: Using or copying publicly available tools is prohibited!

1.3 Data

The data was obtained with NanoSim (Yang *et al.*, 2017), a tool which simulates fragments produced by Oxford Nanopore Technologies. Reference genomes were first introduced with random mutations which include substitutions, insertions and deletions. Afterwards, fragments were generated with a R9 2D Escherichia coli profile with length boundaries between 1k and 100k nucleotide bases and sequencing depth around 50. The resulting fragments have a median read length around 7.5k and a median error rate of 14%.

The training set was created from the Escherichia coli K-12 genome and will be provided at the start of this assignment.

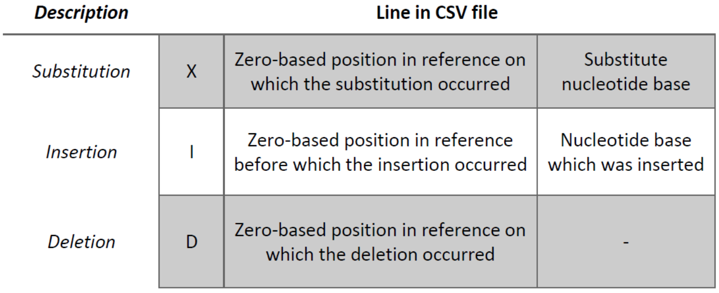
The test sets (five in total) were obtained from parts of the human chromosome 17 and will be used to evaluate this assignment's solutions.

Every fragment set and every reference genome are provided in FASTA[1] format.

[1] https://en.wikipedia.org/wiki/FASTA_format

1.4 Evaluation

The output of the solution should be in CSV[2] format. Each line should have three comma separated values which indicate the type of the mutation, at which position in the reference genome the mutation occurred and the actual mutation. Formal definition of each column can be seen in the following table:



Description	Line in CSV file		
Substitution	X	Zero-based position in reference on which the substitution occurred	Substitute nucleotide base
Insertion	I	Zero-based position in reference before which the insertion occurred	Nucleotide base which was inserted
Deletion	D	Zero-based position in reference on which the deletion occurred	-

Measuring the accuracy of the solution can be done with the provided *jaccard.py* script. It calculates the Jaccard index which is a similarity measure between two finite sets. The script takes two arguments as input, the output of the solution (set of reported mutations A) and the actual set of mutations (B), and computes the following formula:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

To see the usage of the script run the following command: *python jaccard.py –help*.

1.5 Literature

Following links should provide hints towards the solution:

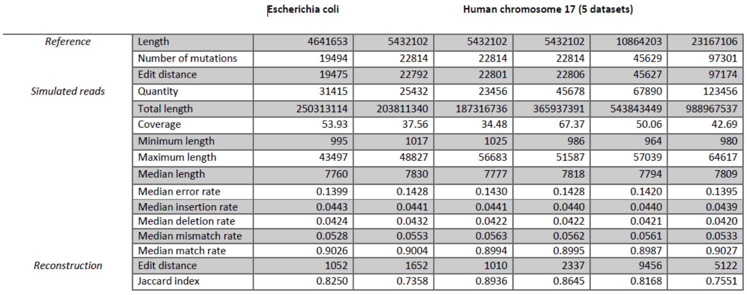
- https://en.wikipedia.org/wiki/Sequence_alignment
- Emphasis on global alignment
- https://academic.oup.com/bioinformatics/article/32/14/2103/1742895
- Only the part about Minimap

1.6 Bibliography

Yang,C. et al. (2017) NanoSim: Nanopore sequence read simulator based on statistical characterization. *Gigascience*, 6.

[2] https://en.wikipedia.org/wiki/Comma-separated_values

Statistics



		Escherichia coli	Human chromosome 17 (5 datasets)				
Reference	Length	4641653	5432102	5432102	5432102	10864203	23167106
	Number of mutations	19494	22814	22814	22814	45629	97301
	Edit distance	19475	22792	22801	22806	45627	97174
Simulated reads	Quantity	31415	25432	23456	45678	67890	123456
	Total length	250313114	203811340	187316736	365937391	543843449	988967537
	Coverage	53.93	37.56	34.48	67.37	50.06	42.69
	Minimum length	995	1017	1025	986	964	980
	Maximum length	43497	48827	56683	51587	57039	64617
	Median length	7760	7830	7777	7818	7794	7809
	Median error rate	0.1399	0.1428	0.1430	0.1428	0.1420	0.1395
	Median insertion rate	0.0443	0.0441	0.0441	0.0440	0.0440	0.0439
	Median deletion rate	0.0424	0.0432	0.0422	0.0422	0.0421	0.0420
	Median mismatch rate	0.0528	0.0553	0.0563	0.0562	0.0561	0.0533
	Median match rate	0.9026	0.9004	0.8994	0.8995	0.8987	0.9027
Reconstruction	Edit distance	1052	1652	1010	2337	9456	5122
	Jaccard index	0.8250	0.7358	0.8936	0.8645	0.8168	0.7551