Department of Computing and Information Systems
COMP 90016
Group Assignment
Release date: 7th of March 2018
Due date: 19th of March 2018 (11:59pm)

Introduction:

Alignment is the task of matching two sequences to each other. In this assignment you are going to write an alignment program with the specific task of matching sequencing reads to a reference genome.
The program is to be written in python.
You will need to check that you have a working account on the Melbourne School of Engineering UNIX servers. See the instructions on the LMS for accessing these servers. In particular, you will need to use the Melbourne School of Engineering UNIX machine *digitalis.eng.unimelb.edu.au*. The MSE Linux machines can be accessed from the Engineering student labs, using *MobaXterm*, which is found on the lab machines under Start → MobaXterm Personal Edition → MobaXterm Personal Edition. To access these machines from home, you will have to use the university's Virtual Private Network. For instructions and support see the documents on the LMS in the section *LabDocuments*. The instructions for using *MobaXterm*, both from the laboratories and from your own machine, are found in the first pages of the Unix document.

Tasks:

You are going to analyse simulated sequencing data from a short reference genome. The data are accessible on the LMS and in the comp90016 folder on the MSE unix servers (nutmeg, digitalis, dimefox – in /home/subjects/comp90016/assignments/group_assignment/). You are free to work with the data in our lab environment or download the sequencing data to your own computer. The data contains the following files: 1) *reference.fa*, 2) *reads.fa*. All the files are in FASTA format. You can assume that the reference contains only one sequence. Only align to the first sequence in the reference file (not to be confused with the first line of the first sequence!)

Write a python program (*aligner.py*) that takes the following inputs from the command line:

a. A reference filename (expecting a FASTA file).
b. A read filename (expecting a FASTA file).

The program should perform the following actions:

c. Align each read to the reference sequence. The alignment should only consider **perfect** matches to either the forward or reverse strand of the reference. The position of an alignment is the coordinate closest to the **start of the reference**. For a read on the reverse strand the position is therefore its last base. Count the number of possible alignments for each read and their positions within the genome.

d. Write the alignments to an output file in the same order as the reads present in the input file. The output file should be called *alignment.txt*. There should be one line per read containing the following fields: *READ_NAME, REF_NAME, POS, STRAND, NUMBER_OF_ALIGNMENTS*
The fields should be tab delimited. The *POS* field should contain the "left-most" position out of all possible alignments, that is, the position that is closest to the **5' end of the reference sequence**, regardless of strand. The strand should indicate "+" if the specific alignment is to the forward strand ("-" otherwise).
Finally, the last field should specify how many possible alignments there are for the read.
If no alignments are found for a read, the output should read: "*READ_NAME* * 0 * 0" (with the according read name).

e. Write alignment statistics to the command line. Specifically, how many reads were aligned to 0 positions, exactly 1 position, and more than one position (in absolute numbers and percent).

The program should perform sensible error checking regarding the existence and readability of the provided filenames.

Discuss the results. In particular, what could be the reasons for reads not aligning to the reference? Are the numbers higher or lower than you expected? Back your discussion up with theoretical considerations and practical examples from the data.

Notes:

- The alignments should be reported as 1-based coordinates. That is, a read that align to the first base of the reference should have position 1.
- All alignments are reported with respect to the forward strand of the reference. That is, if a read aligns to the reverse strand, the position should still reflect how far away the read is from the 5' end of the reference.
- A read can be palindromic. A palindromic read has the same sequence as its reverse complement. For example, "ACGT". To report an alignment for such a read, default to the forward strand of the reference.
- If you develop the program on your own computer, make sure that it also runs on the MSE machines, as this will be the environment we test in.
- The assignment is to be submitted via Turnitin. Please submit .zip file with all the relevant documents: The alignment program and the discussion.
  As this is a group assignment, only one student per group should submit a solution to the assignment. Indicate all member of the group within the submission files.