Assignment 2: Read Mapping

VUB 4018221FNR / ULB INFO-F439 - Matthieu Defrance

Project Objective

The overall objective of this project is to implement a read mapping software that is able to align a large set of reads to a given genome. This reap mapping software will be applied to the provided ChIP-seq data.

Data set

The data set has been generated using an ChIP-seq experiment performed on the Drosophila melanogaster "Oregon-R S2" strain. The sequencing data have been produced using the Illumina tehcnologie using single-end sequencing. The data are available here:

https://owncloud.ulb.ac.be/index.php/s/ceP7Vcu039z2Aw7

The reads and the dm6 version of the Drosophila melanogaster reference genome are provided:

- all_reads.fastq.gz, all the ChIP-seq reads
- dm6.fa.gz, the full D melanogaster reference genome

To facilitate the development of the software, reduced data files, resulting from down sampling the full data set are also provided:

- dm6_chr2L.fa.gz, only chromosome chr2L extracted from the dm6.fa.gz genome
- 10k_reads.fastq.gz, only 10,000 reads extracted from all_reads.fastq.gz

Work to be done

The work consists in the production of a functional 'read mapper' and its application to the previously described data set. The description of method implemented and the summarized performances when applied to the data should be presented in a report. Try to first run your method on the reduced dataset. You should provide at least the following information:

- The description of the mapping software algorithm and its pseudo-code (generic code).

- A summary of the alignment output when applied to D melanogaster data set (e.g. percentage of reads mapped).
- Extract of the implementation code with comments (e.g. python, R or C code).

A PDF version of the report named 'assignment2_lastname_firstname.pdf' should be emailed to matthieu.dc.defrance@ulb.ac.be before the deadline (May 8, 2021).

Important note

This project is an individual work. Plagiarism, copying is forbidden.