GENOMIC PROJECT

December 8, 2016

1 Introduction

We were given the objective to build a program with an application in actual genomic problems. Several inputs can be given by the user, such as a genome sequence, a position-weight-matrix (PWM), binding scores along a specific sequence ... The idea is to offer different outputs that can help the user to solve his problem. The output can be a list of sites, affinity score for a given sequence, a PWM (or a PSSM), or a graphic representation of the PMW (logo showing the different nucleotides A, T, C and G at different sizes, among their probability).

Figure 1: An example of PWM

$$\begin{pmatrix} 0.9913 & 0.0019 & 0.0044 & 0.0024 \\ 0.0045 & 0.0031 & 0.9611 & 0.0313 \\ 0.0032 & 0.9915 & 0.0019 & 0.0034 \\ 0.8917 & 0.0019 & 0.0089 & 0.0975 \\ 0.6241 & 0.2510 & 0.0059 & 0.1187 \\ 0.7289 & 0.1045 & 0.0116 & 0.1550 \\ 0.3048 & 0.4911 & 0.0045 & 0.1996 \end{pmatrix}$$

A C G T

Figure 2: A sequence logo

2 Download the program

To download the program, go to : https://github.com/EPFL-SV-cpp-projects/genom-1. Open a terminal, and type "git clone : **https://github.com/EPFL-SV-cpp-projects/genom-1".** The genom-1 folder and all its content will be copied in your directory, you can then compile and execute the program.

3 Compilation and execution

To compile and execute the program, a fews steps are required. Make sure you are in the genom-1 folder and then :

rm –rf build and **mkdir build** to make sure an empty build folder is crated

cmake../

make

And then you have several options

If you want to see the documentation with the doxyfile, describing more precisely the different classes and fonctions involved in the program you will then have to write **make doc**

If you want to execute the program, then do **../src/Main**

If you want to run the tests of the program, then do **make test**

4 The functionalities of the program

When you execute the program, a menu will appear, proposing you a list of outputs designed by their numbers. You can choose the task the program must do by hitting '1', '2' or '3' or '4'. Then you will be asked to provide the inputs (if the program needs a file you can write its name like "$example.fasta$"). Here are the mains functionalities of the program :

1.- Being able to read a DNA sequence and a PWM (or/and it's logarithmic version) and give as output the list of site along the genome where the protein is gonna attach.

2.- Being able to read a DNA sequence, a list of sites and their respective binding score (the product of the probabilities of each nucleotide along the sequence) and output a PWM (or/and it's logarithmic version).

3.- Based on the matrix or based on the binding scores and list of sites, being able produce a sequence logo.