Lecturer: Jan Baumbach Term: Winter 2015/2016

Class: Introduction to Bioinformatics

Exercise sheet – ChIP-seq data analysis

1. Explain/define: "bi-modal enrichment pattern". How will the pattern change if you use single-end instead of paired-end sequencing

2. Write a JAVA program CHIP that performs very simple/basic peak detections: You are given a set of read mappings, which can be downloaded from the course webpage (URL1, see below). Each line in the dataset contains two numbers, the start position and the end position of the read as well as a symbol that denotes whether the read is on the forward (+) or reverse strand (-). We don't care about the actual sequence. Each field is separated by a blank space. An example might look like this:

```
132 432 +
203 345 +
940 245 -
```

The size of the genome is 1000bp. Proceed as follows:

(a) Read the file from URL1 and compute the signal maps (histograms) for each position in the genome on the forward and backward strand. You should maintain two suitable data-structures (e.g. arrays).

(b) In the next step you should smooth both signal maps by replacing each entry with the mean over the 2k+1 neighbors. Use a value of k=5, i.e. each count is replaced by the mean over its 5 left and right neighbors and its own value.

(c) Now, identify all local maxima on the forward and backward strand. You can use the following definition for a local maxima at position i:

```
array[i-1] <= array[i] > array[i+1]
```

(d) As a simple filtering criterion, remove all identified maxima on both strands if the peak height is <100 read counts.

(e) Finally, for each peak on the forward strand find the largest peak on the reverse strand (if it exists) within a window of 120 to 200bp downstream and report each pair of positions as a binding site region. Write the list of start/end positions to a user-specified file on harddisk.

(f) Add a plotting functionality to your program that plots the two signal histograms and highlights the binding site regions.

3. Use pen and paper to compute the Burrows Wheeler Transformation (BWT) of the string "attgtg". Introduce the $-symbol to denote the end of the string. Treat "$" as lexicographic smaller than any other character.

(a) Match the strings "ttg" and "ctt" against the BWT by utilizing the EXACT matching algorithm from the lecture. Draw all intermediate steps.

(b) Why do we need the $-symbol for the BWT?

URL1:

http://www.imada.sdu.dk/~jbaumbac/download/teaching/ws15-16/DM847/exercises/bioinformatics_intro_class_chip_seq_data_analysis_readmapping.txt

Please send the JAVA program as well as the source code and the input file via email to your TA. Also email the names of all group members and a short tutorial on how to execute the program with the input file.