Department of Computing and Information Systems
COMP 90016
Assignment 3
Release date: 1st of May 2018
Due date: 24th of May 2018 (11:59pm)
This assignment is worth 15 marks, or 15% of your final score

The topic of our last assignment is copy number changes (CNVs) in DNA and how to detect them from sequencing data.
In the first task, you are going to discuss the application of HMMs to CNV detection in theory. The second task makes you develop an implementation of a CNV detection algorithm using segmentation. In the final task, you will discuss the biology of a particular cancer sample.

Task 1 (4 marks)
The lectures introduced a HMM designed to detect CNVs in a haploid organism. It consisted of three states: One for the normal copy number (CP1) and two alternate copy number states (Cp0, CP2).

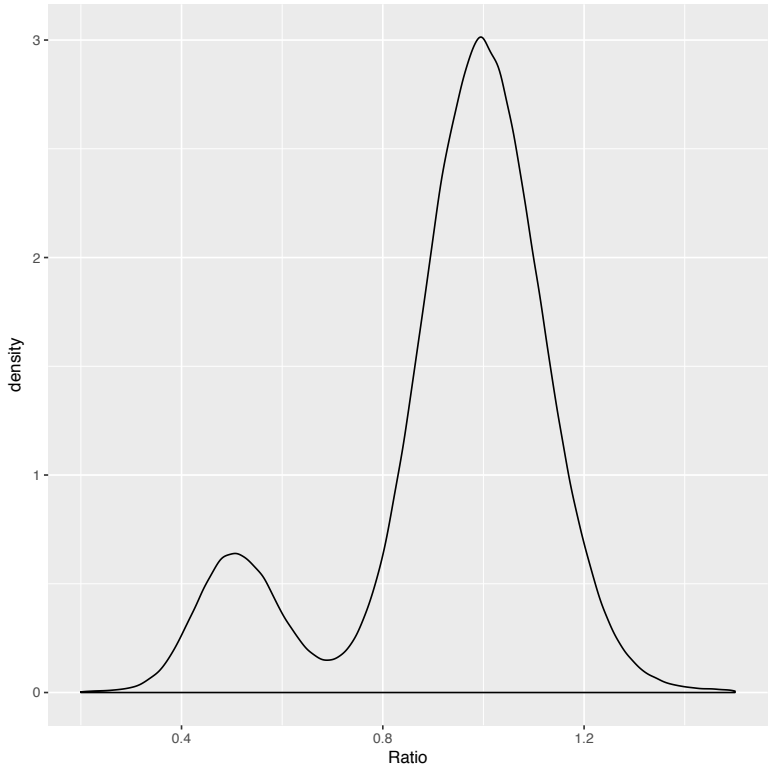

Figure 1: Density plot of ratios of read depth in 5kbp bins over the average read depth in a diploid organism.

1. How would you adapt this approach to a diploid organism, such as human (with states representing different integer copy numbers)?
2. Explain the trade-off between the sensitivity of such a HMM and the computational complexity to solve the Viterbi algorithm for it.
3. Consider the data shown in Figure 1. The plot shows read depth of bins normalized by the average in a diploid organism. How would you use this data to parametrize the emission probabilities of your HMM? Explain what about Figure 1 is general and what is specific to the data that this plot was derived from. How does this affect the HMM in terms of its application to different data sets? Also, how could this data be utilized to derive the transition probabilities for a CNV detection HMM?

4. Describe why a HMM, such as discussed in this task is not very useful for non-clonal data (such as shown in Figure 2 below). Could this shortcoming be alleviated by introducing non-integer CN states?

Task 2 (6 marks)

Implement an algorithm in Python that performs Circular Binary Segmentation (CBS) on a dataset to call copy number variations.

Your task is to take the method proposed by Olshen et al. (see the paper on the LMS) for array data and adapt it for sequencing data. The idea can be transferred directly by using log ratios of binned read depth instead of intensities. The algorithm is to apply the same recursive method and segment the data until no high-scoring segments can be found anymore.

More specifically:

1. The inputs to the algorithm (*cbs.py*) should be provided on the command line: *python cbs.py input_bins Z_threshold output_file*, specifying a file for input, a floating-point number of the absolute termination condition (see below) and an output filename.
2. Transform the input read depth into log ratios. That is, each bin b is to be transformed to log2(b/m) with m a good representative of a "normal" bin in the data. Often, the median of the data is a useful value to this extent. In the specific data used here and its highly read depleted nature you are to take the median from **the first third of the input data.**

   Any log2-ratios that are larger than 2 or smaller than -5 should be set to 0. This is a preventative measure towards making the data from Task 3 easier to analyse. This filtering step is going to remove some extreme values around the centromere of the chromosome (centromeres are notorious for influencing read-depth). Your algorithm is going to have to do less computation due to this hard-coded filter.
3. Perform CBS on the log ratio data: Identify the maximum segmentation score Z for the current range of bins, that is z := |Z(S, i, j)| is maximal for the cumulative scores S. If z is as high or higher than the user-specified Z-value, segment the range between (i,j] and the rest of the bins, and save both segments for further analysis. Otherwise, the current interval cannot be segmented further and the algorithm can move on to the next, or terminate if all intervals have been analysed.
4. Write an output file in the BED format containing of four tab-separated columns (*CHR START END RATIO*) and one row per CNV. The file should report the start and the end of any copy number segment that has been identified during step 2 and its average log ratio. Only segments with an absolute average log-ratio of 0.1 or more should be reported here (all other segments are to be assumed normal (that is, having a copy number of two).

Your discussion file should use about half a page to describe your algorithm and the design choices.

Further, discuss the theoretical complexity of your algorithm: What variables in the input does it scale with and in what relationship?

To make development of this method easier, there is a sample input set provided on the LMS (*test.txt*). This small data set provides a simple example of 50 lines of input. Instead of read

depths, it directly shows the copy number within a made-up diploid sample. If you use CBS on this with a Z-threshold of 1, it should provide you with two CNVs:

seq 21 31 0.58
seq 36 41 -1

Task 3 (5 marks)

Run your CBS copy number algorithm from Task 2 on the data from the LMS (*data.txt*). The input data for this task is a sequence of binned read depths from a single chromosome of a human cancer sample. The aligned read data is summarised into over 3000 non-overlapping bins of length 50kbp. See Figure 1 for a plot of the log2 ratios of the bins.

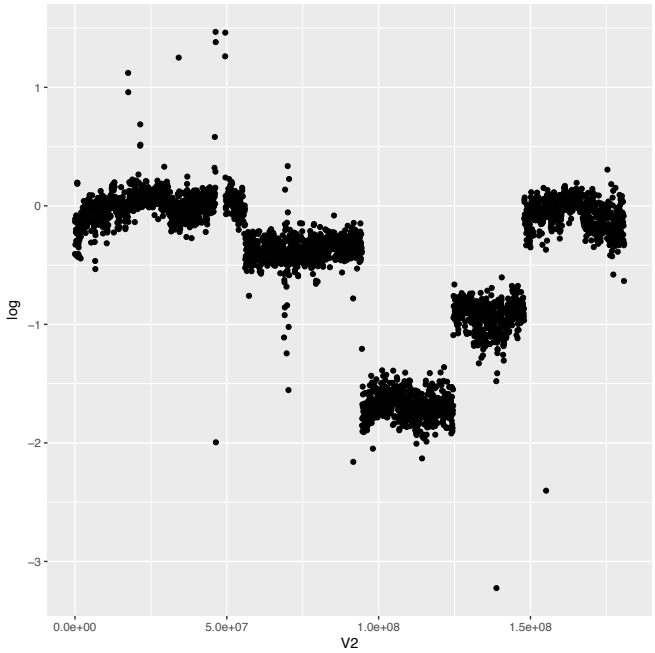

Figure 2: Log ratio of binned read depth on chr5. Normalisation was performed with the average from all chromsomes.

Your task is to run your algorithm with a Z-cut-off of 10 followed by:

1. Visualise the reported CNVs alongside the log-ratios. Create a plot similar to Figure 2, but highlighting CNV segments. If it is too difficult to draw the segments on top of the data points, it is fine to draw them beneath, all on the same height.
2. Discuss the results using about half a page. Are all of the identified CNVs real? Are there any false negatives? How would you parametrise the algorithm differently to increase sensitivity?
3. Discuss the biology of the analysed sample. Focusing on the large and obvious changes in the data, we can generalise to three large CNVs in the data: 50-90Mbp -0.3, 90-125Mbp -1.3, and 125-140Mbp -0.8. None of these log ratios translate directly to a clonal CNV of losing an entire copy of chromosome 5.

Discuss what the clonal structure of the sequenced sample could be. Consider if the three changes are single or double copy losses of the DNA, and whether they are present in the same cells or different cells. Present some mathematics of clonalities and allele counts to back up your theory.

Also discuss what additional information could be obtained from the original BAM file to substantiate your theory.

Finally, draw a diagram of the clonal structure that you have identified showing the different haplotypes and how they are represented in the overall population of cells. This part of the task should be addressed with a page of writing.