Department of Computing and Information Systems
COMP90016
Workshop 6

Circular binary segmentation is a technique for copy number variant (CNV) calling in DNA sequences based upon the change-point problem [2]. In this problem $X_1, X_2, \ldots$ are a sequence of random variables. An index v is called a *change-point* if $X_1, \ldots, X_v$ have a common distribution function F_0 and $X_{v+1}, \ldots$ have a different common distribution function F_1, until the next change point. The sequence to be used for change-point detection are the log ratio of normalised intensities indexed by the corresponding marker locations [1].

Change-point detection [1] Let $X_1, \ldots, X_n$ be a sequence (segment) of log-ratio normalised intensities. Define the sequence $S_i = \sum_{j=1}^{i} X_j, 1 \leq i \leq n$. If we consider joining the segment at both ends, to form a circle, the likelihood ratio test statistic for testing the hypothesis that the arc from i to j and its complement have different means is given by:

$$Z_{ij} = \left(\frac{1}{(j-i)} + \frac{1}{(n-j+i)}\right)^{-1/2} \cdot \left(\frac{(S_j - S_i)}{(j-i)} - \frac{(S_n - S_j + S_i)}{(n-j+i)}\right)$$

The change-point statistic is $Z_C =: \max_{1 \leq i < j \leq n} |Z_{ij}|$. If $Z_C > T$, where T is a critical value, we assert that a change-point exists and is given by the corresponding sub-segment index, i.e., $(i^*, j^*) = \arg\max_{1 \leq i < j \leq n} |Z_{ij}|$.

Task 1 (Computing log-ratio normalised intensities) :

Read in the wiggle file `x.wig`, which contains read counts (intensities) from the e.coli genome, one intensity value per line. The file was generated by parsing the BAM file of aligned reads and counting reads falling into non-overlapping 2000bp bins. Use an appropriate data structure to store the sequence. From this sequence compute the correspond log-ratio normalised intensities. Hint: if $I_1, \ldots, I_n$ are the input intensities, the normalised intensities may be computed as $X_j = \log_2(\frac{I_j}{mean}), 1 \leq j \leq n, mean = \frac{\sum_{i=1}^{n} I_i}{n}$.

Task 2 (CNV calling) :

We are going to perform change-point detection on the sequence of log-ratio normalised intensities from Task 1.

1. Write a function `cbs`, which accepts a list `S`, corresponding to the cumulative sums of $X_1, \ldots, X_n$ (i.e. $S_i = \sum_{j=1}^{i} Xj$), and performs change-point detection (you may use $T = 5$).
2. Use your function from part a) to recursively perform change-point detection until a depth of 5 on the sequence computed in Task 1. The segments in between change-points are then regions corresponding to constant copy number. Hint: if (i, j) is the index of a change-point in the circular segment $X_1, \ldots, X_n$, then $X_{i+1}, \ldots, X_j$ and $X_1, \ldots, X_{i-1}, X_{j+1}, \ldots, X_n$ are inputs to the recursion.

Task 3 (Discussion) :

1. For the most significant pair of change points in the genome: what are the genomic coordinates of the bins involved?
2. Upon inspection of the data at that region in the IGV browser, is the event a duplication or a loss of DNA?

References

[1] Adam B. Olshen, E. S. Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557, 2004.

[2] E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1):100–115, 1954.