Genome regions

Background

Consider we have a number of regions, each region corresponding to a portion of the human genome. Each region is specified by a start and end position, both of which are integers and within the range [1,N].

We have provided two text files in Windows format which give example data.

Regions_Small.txt

Regions_Large.txt

Each line of the file specifies a region and contains a pair of integers separated by a tab. These are the start and end coordinates of each region. Please use these region files in the following problems. Please provide your C++ code, any instructions for compilation and running the code and all your output files. You may use standard template library containers, but should not use other 3rd party libraries (e.g. Boost).

Part 1

We have a set of regions as read from the text file. We would like to draw the regions in the following style where overlapping regions stack up.

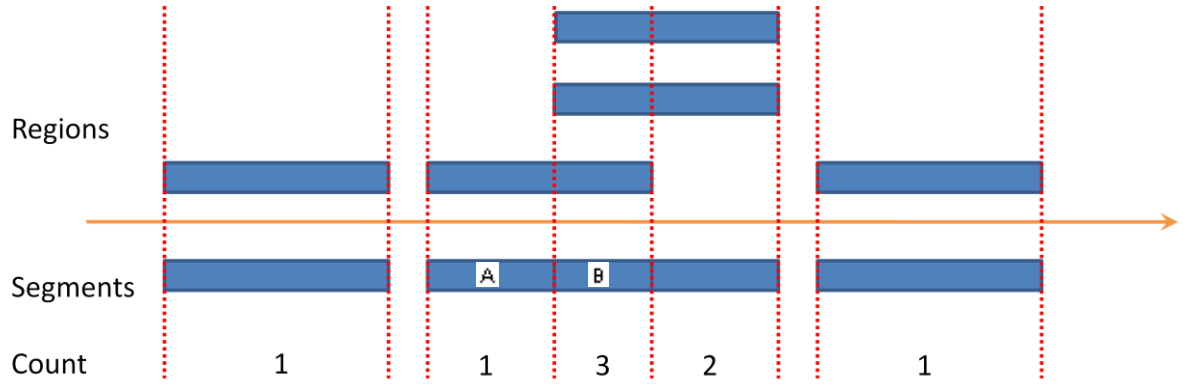
To do this, each region must be assigned to a drawing row. Because on screen real estate is limited we want to draw the stack of regions using the smallest number of drawing rows.

Please write code which takes as input a regions file. The code should produce as output a text file. The text file should contain the same number of rows as the input file. On each line of the output file there should be an integer which indicates the drawing row to which the corresponding region in the input file was assigned.

Part 2

When there are a large number of overlapping regions, the stack of regions can become too deep to display in limited space. Instead we can summarise the data and display it in the style of a histogram.

To summarise the regions data, the regions must be resolved into non overlapping segments. Together the segments cover the regions. Finally associated with each segment is a count of the number of regions which overlap that segment. This idea is illustrated below.



In the above, if segment B has start position X, then segment A has end position X-1. The segments do not overlap.

Please write code which takes as input a regions file. The code should produce as output a text file. The text file should contain one row per segment. Each segment must be described by three tab separated integers, start, end and count.