Organize genomic data for random access

Questions

1. Provide a brief explanation of your approach.
2. Write a Java implementation for this assignment, without the use of a database.

Details

Input is a tab-delimited file containing points on the genome. Each point is defined by chromosome, start and end location. Organize this data in persisted form (not in memory) for fast random access, such that given a specific genome region(s) returns all points that fall within it, including partially overlapping ones. Provide an interface to execute queries against the data and display the results. Keep in mind that for each genome we can have gigabytes worth of data and the system needs to hold data for many thousands of such genomes. However, the query here is just for one genome at a time.

Sample queries:

1. chr18:0-60000000
   This should return all points in chromosome 18 within 0 to 60000000

2. chr3:5000-chr5:8000 (This query spans multiple chromosomes starting at chromosome 3 base position 5000 ending at chromosome 5 base position 8000)

The tab-delimited input file is available in zipped format (Note: Please ignore the last column called “Array” in the tab-delimited file.)