CME 211 HW1 Bowei Ma SUNetID: boweima

Part 2

The test data is designed so that the only read with two alignments has the initial alignment in the latter half of the reference. And thus the designed algorithm is guaranteed to work properly on the test data set.

The algorithm would not work on the reads with multiple alignments but the first alignment occurs in the first 50% portion of the reference.

No. The portion is not exact due to the mechanism of random generator and the simulation of uniform distribution.

To implement and debug part 2, it took me about 90 minutes to complete the task.

Part 3

Yes. The numbers of 0, 1 and 2 alignments reads are exactly the same. Because the algorithm I used to recognize the number of alignments of a read is by its starting position, which is the same method when I generated the dataset.

Suppose the length of the reference is m, the number of reads is n and the length of each read as r. Assuming that the find function in the string class would take O(mr) to complete in worst case, we know that our algorithm would take O(nmr) to complete. Taking account of cache capacity and increased time of buffering caches, to have 30x cover of the 3 billion gene sequence it would probably take (3*10^9/10^3)^2 * 0.01 = 3 * 10^10 seconds ~ 1000 years to complete calculation. So it is not feasible probably.

It took me about 60 minutes to complete part 3.