**Comp 555: BioAlgorithms -- Fall 2013

Problem Set #4

Issued: 11/12/2013      Due: In class 11/21/2013**

**Question 1.**

1. Lookup positions from first half of k-mer. For each such position check the sequence whether k-mer exists starting from that position. Analysis: Hash lookup is constant time. (or O(k) if we also consider hashing time, but this won’t make a difference overall) Then lets say there are m positions in the hash bucket. For each of the m positions O(k) operation to compare k-mer and sequence. Therefore performance is O(km). In the worst case m is equal to the length of the sequence,n so performance is O(kn)
2. We thread the k-mer through the suffix tree. Then we count leaves under in the subtree rooted in the node we reached. Analysis: Threading: O(k). Walking through all over subtree = order of number of leaves which is at most length of sequence n. So overall os O(k+n)
3. 2 binary searches to find first and last occurrences of the string. Number of occurances is just the difference between index numbers of first and last occurances. Analysis: each comparison is O(k). each binary search takes O(logn) time so performance is O(klogn)

**Question 2.**

1. Everybody got this correct

b and c) The trees generated by both methods have the same structure. Only difference is the heights of the brances

d and e) Both initial conditions results in the same clustering. Using first two points converges faster