I519: Introduction to Bioinformatics, Fall, 2015

October 12th, 2015

Homework 4 (Due via canvas by **11:59pm on Saturday, October 18th, 2015)**

1. [20 pts total] Show all possible topologies of phylogenetic tree (unrooted) for 5 sequences (A–E) (10 pts). Explain briefly why exact tree search is not a clever choice for phylogenetic tree reconstruction for *n* sequences when *n* is large.

2. [20 pts total] Assume you work on a sequencing project for a bacterial genome, of expected size of 4Mb, and you want to achieve 10X sequencing coverage of the genome. You use a Illumina sequencing platform that produces paired-end reads, and the average length of the reads is 100bp. Assembly of the reads results in a collection of contigs with 0.2Mb, 0.1Mb, 0.8Mb, 0.1Mb, 1.0Mb, 0.05Mb and 0.02Mb, respectively, and many more small contigs (of < 1Kb each). Compute the following: a) how many pairs of reads need to be sequenced (10 pts), b) N50 (10 pts), and c) expected number of positions in the genome that are not covered by any read (10 pts).

3. [10 pts] Show the complexity of the below algorithm in Big-O notation (N is the input size). Explain your answer briefly.

```
def myAlgorithm(N):
    something = 0;
    for i in 1 to N:
        for j in i to N:
            for k in j to N:
                something += i * j * k
    for i in 1 to N:
        for j in i to N:
            something += i * j
    return something
```

4. [50 pts total] Consider two sequences v=GTCCCT and w=TCCCCTA, and a scoring function: **-1 for a mismatch and -1 for each indel (insertion or deletion), and 1 for a match**.

   a) Fill out the dynamic programming table for a **local** alignment between v and w (30 pts).

   b) What is the score of the optimal local alignment and what alignment achieves this score? (20 pts)