

NAME: _____

BIO/CMPSC 300**Activity 3****Chapter 9 Overlap in Sequence Assembly****Fall 2019**

1. Suppose two sequence reads give GGCAGCC and GCCCCAGG.

a) What are the possible overlaps? Which one is the largest?

 GCCCCAGG ←-largest
GGCAGCC

GCCCCAGG
 GGCAGCC

b) Describe an algorithm that finds the largest overlap. Use the legos to help you with the development of the algorithm and with the run through your algorithm, where the first sequence are represented by the rectangles and the second one by squares. A - red, C - yellow/orange, G - blue. Remember that the algorithm is a step-by-step procedure, thus your algorithm should outline the sequence of steps required for finding the largest overlap given two sequences.

First overlay the two sequences at the ends and check for matches.

If they overlap perfectly then that is the best overlap.

If not then repeat the process, overlapping at the other end, repeatedly offsetting by a base.

(Stopping of course when there isn't any overlap left to check)

Ex;

GCCCCAGG
GGCAGCC

GCCCCAGG
 GGCAGCC

 GCCCCAGG
GGCAGCC

 GCCCCAGG
 GGCAGCC

 GCCCCAGG
 GGCAGCC

 GCCCCAGG
 GGCAGCC

 GCCCCAGG
 GGCAGCC

 GCCCCAGG
 GGCAGCC

c) What would be the sequence of the contig produced using your algorithm?

GGCAGCCCCAGG

2. Now suppose you would like your algorithm to account for the possibility that the sequences could come from either strand of the DNA. How would you modify the algorithm to accomplish this? Would the contig resulting from the two sequences in 1 change as a result?

Also check the inverse complement of one of the sequences against the unmodified other sequence. (or simply pipe in the sequences from the DNA manipulation algorithm if the details are actually known by the user)