Assignment: It is assembly-based. A DNA is a long molecule that can be sequenced in one shot. One way of sequencing DNA is - split the DNA into shorter segments randomly and sequence them. These random sequences are then assembled into a larger DNA sequence to get to the pre-segmentation sequence. This process is called “DNA sequence assembly.”

For example, if we assume our DNA sequence is “ATGAGGAATTT” that is randomly segmented into 3 segments, where individual segment’s sequences are:

> Segment_1

ATGAG

>Segment_2

AGGAA

>Segment_3

AATTT

Notice that the last 2-bases of Segment 1 overlaps with the first 2-bases of the Segment 2. Similarly, the last 2 bases of Segment 2 overlap with the first 2-bases of Segment 3. If we assume that all these segments are from the same contiguous DNA, then we can assemble these segments into the following sequence (where overlaps are highlighted)

>Output

ATGAGGAATTT

This process is called DNA assembly (or genome assembly) and the software/tool that achieves this goal is called DNA assembler.

Write our own DNA assembler using your preferred language. [i.e., write a script that can perform DNA assembly based on the sequence overlaps.]

Concepts tested: DNA assembly principles, Algorithm for sequence assembly