Problem 1 (60pts).

Suppose you are given the following 4-mers or reads of an unknown string.

CGAT
ATCG
GCAG
AGCG
GAGC
GATC
CAGC
GGAG
AGCA
GCGA
TCGG

(a) Create an overlap graph of these reads that may be used to reconstruct the original string. (15 points)

(b) Reconstruct a string that covers the above 4-mers using the overlap graph by constructing the longest path that visits all nodes. (15 points)

(c) Create a De Bruijn graph of these reads that may be used to reconstruct the original string. (15 points)

(d) Reconstruct a string that covers the above 4-mers using the De Bruijn graph by constructing the Eulerian path of this graph. (15 points)

Problem 2 (Extra credits : 20 points)

We consider a greedy algorithm for the shortest superstring problem (See lecture slides): merge a pair of strings with maximum overlap and repeat until only one string left.

Provide an efficient implementation in pseudo-code, runtime analysis, and an example to show that this greedy algorithm cannot find the optimal solution.