Week 2 - 1/28/22

Friday, 21 January 2022 7:55 AM

$Q = -10 \log_{10} P_e$; P_e = P(error in base calling)

eg. $Q = 30$; ASCII (?)

$$30 = -10 \log_{10} P_e$$

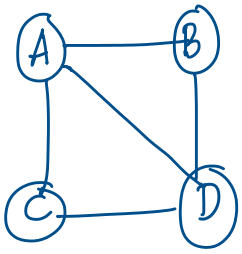
$$-3 = \log_{10} P_e$$

$$\Rightarrow 10^{-3} = P_e = 0.001$$

GRAPH THEORY 'REVIEW'

OBJECT - NODE

RELATIONSHIP - EDGES



NODES = {A, B, C, D}

EDGES = {AB, BD, CD, AC, AD}

Problem 1 : find a path connecting all nodes ; such that every node is visited ONLY once [HAMILTONIAN PATH]

① A → B → D → C

② A → C → D → B

HAMILTONIAN CYCLE :

A → B → D → C → A

Problem 2 : find a path such that every edge is visited ONLY ONCE [EULERIAN PATH]

A → B → D → C → A → D

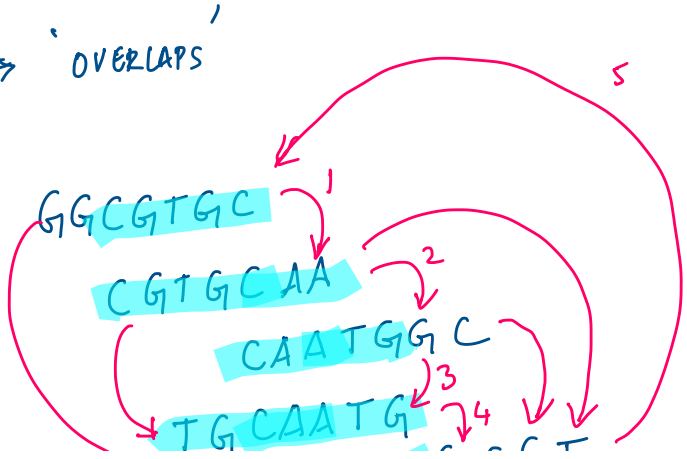
GENOME : ATGGCGTGCA (CIRCULAR)

READS : CGTGCAA, ATGGCGT, CAATGGC,

GGCGTGC, TGCAATG

consider each read as a NODE

edges → 'OVERLAPS'



'k-mer length' → length of overlap between reads

de Bruijn graphs ; Eulerian path

SUPERSTRING PROBLEM

↳ find the shortest superstring that contains all possible substrings of length = k [k-mer] in a given alphabet.

eg. ALPHABET = {0, 1} ; n = 2 [# of letters]

k-mer length = 3

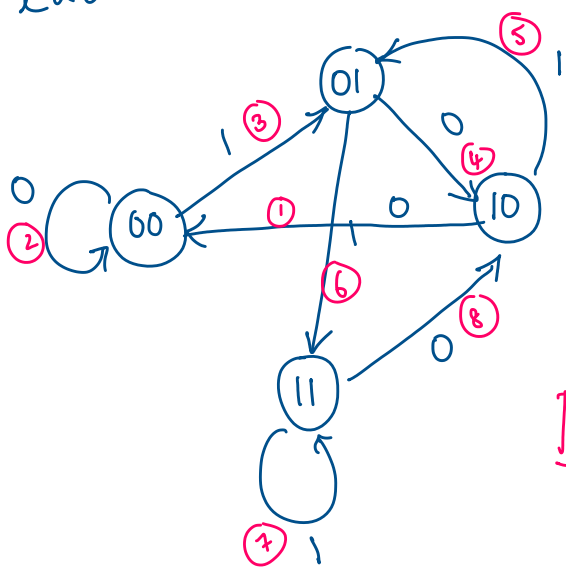
{000, 001, 010, 100, 110, 101, 011, 111}

n^k possible k-mers

$\binom{3}{2} = 8$

de Bruijn graph

Construct a de Bruijn graph such that each node = 'k-1' mer



EULERIAN PATH

10 → 00 → 00 → 01 → 10 → 01 → 11 → 11 → 10

10001011 0

ALPHABET = {A, C, G, T} $n = 4$ k-mer = 3

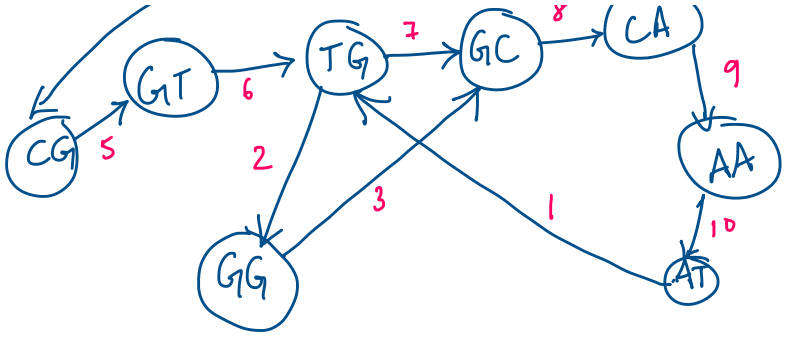
de Bruijn graph has to have nodes of length = $k-1 = 3-1 = 2$

possible nodes = {AA, AC, AG, AT, ..., TT} 4^2

OBSERVED READS: CGTGCAA, TGCAATG, ATGGCGT, GGCGTGC, CAATGGC

nodes = {CG, GT, TG, GC, CA, AA, AT, GG}

4



EULERIAN CYCLE:

ATGGCGTGCA

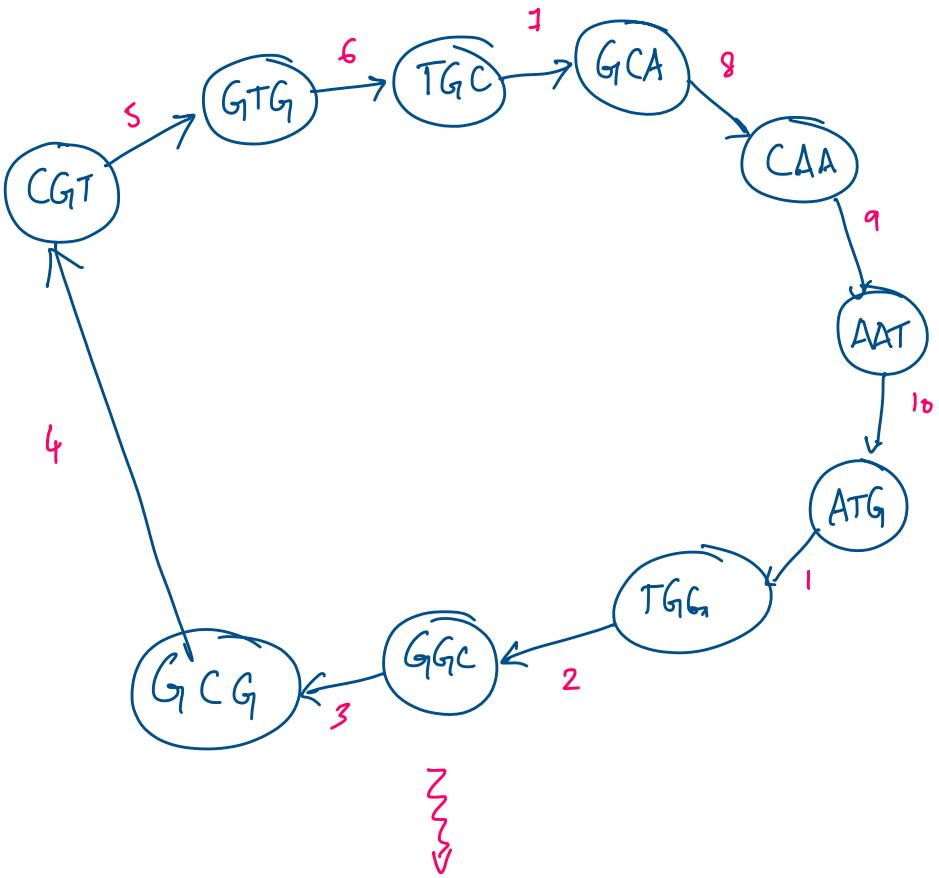
ASSEMBLED GENOME

---

k = 4

nodes : { CGT, GTG, TGC, GCA, CAA, ATG, TGG, GGC, GCG }

AAT :




ATGGCGTGCA