Problem (1.4)

The problem setup is just like on Homework 1, the only difference is you have an added assumption:

$$P(BB) = 0.$$

Thus, following from homework 1 / lecture:

$$P(E_1, \dots, E_N \mid AB) = \left(\frac{1}{2}\right)^N$$

$$P(E_1, \dots, E_N \mid AA) = \prod_{i:\text{obs } A} P(E_i = 0 \mid AA) \prod_{j:\text{obs } B} P(E_i = 1 \mid AA)$$

$$= P(E_i = 0 \mid AA)^{N_A} P(E_i = 1 \mid AA)^{N_B}$$

$$= (1-\epsilon)^{N_A} \epsilon^{N - N_A}$$

where N_A = # of As we observe, N_B = # of Bs.

So,

$$P(AB \mid E_1, \dots, E_N) = \frac{P(E_1, \dots, E_N \mid AB) P(AB)}{P(E_1, \dots, E_N \mid AB) P(AB) + P(E_1, \dots, E_N \mid AA) P(AA)}$$

$$= \frac{\left(\frac{1}{2}\right)^N P(AB)}{\left(\frac{1}{2}\right)^N P(AB) + (1-\epsilon)^{N_A} \epsilon^{N - N_A} P(AA)}$$

Problem (1.b)

Same as (a), just a different numerator!

$$P(AA \mid E_1, \ldots, E_N) = \frac{P(E_1, \ldots, E_N \mid AA)\, P(AA)}{P(E_1, \ldots, E_N \mid AA) P(AA) + P(E_1, \ldots, E_N \mid AB) P(AB)}$$

$$= \frac{(1-\epsilon)^{N_A} \epsilon^{N - N_a} P(AA)}{\left(\frac{1}{2}\right)^N P(AB) + (1-\epsilon)^{N_a} \epsilon^{N - N_a} P(AA)}$$

Problem (1 c)

$$P(E_1, \dots, E_N \mid AA) = \prod_{i:A} P(E_i = 0 \mid AA) \prod_{i:B} P(E_i = 1 \mid AA)$$

$$= (1-e)^{N_A} e^{N - N_A} = \left(\frac{(1-e)}{e}\right)^{N_A} e^N$$

$$P(AA \mid E_1, \dots, E_N) = \frac{\left(\frac{1-e}{e}\right)^{N_A} e^N \left(\frac{1}{2}\right)}{\left(\frac{1-e}{e}\right)^{N_A} e^N \frac{1}{2} + \left(\frac{1}{2}\right)^N \frac{1}{2}}$$

$$= \frac{\left(\frac{1-e}{e}\right)^{N_A} e^N}{\left(\frac{1-e}{e}\right)^{N_A} e^N + \left(\frac{1}{2}\right)^N} = \alpha \Rightarrow \frac{\left(\frac{1-e}{e}\right)^{N_A} e^N}{\left(\frac{1-e}{e}\right)^{N_A} e^N} + \frac{\left(\frac{1}{2}\right)^N}{\left(\frac{1-e}{e}\right)^{N_A} e^N} = \frac{1}{\alpha}$$

$$\Rightarrow \frac{\frac{1}{2^N}}{\left(\frac{1-e}{e}\right)^{N_A} e^N} = \frac{1}{\alpha} - 1 \Rightarrow \frac{1}{\left(\frac{1-e}{e}\right)^{N_A} e^N 2^N} = \frac{1-\alpha}{\alpha}$$

$$\Rightarrow \frac{\alpha}{(1-\alpha) e^N 2^N} = \left(\frac{1-e}{e}\right)^{N_A} \Rightarrow \frac{\log\left(\frac{\alpha}{(1-\alpha)(2e)^N}\right)}{\log\left(\frac{1-e}{e}\right)} = N_A$$

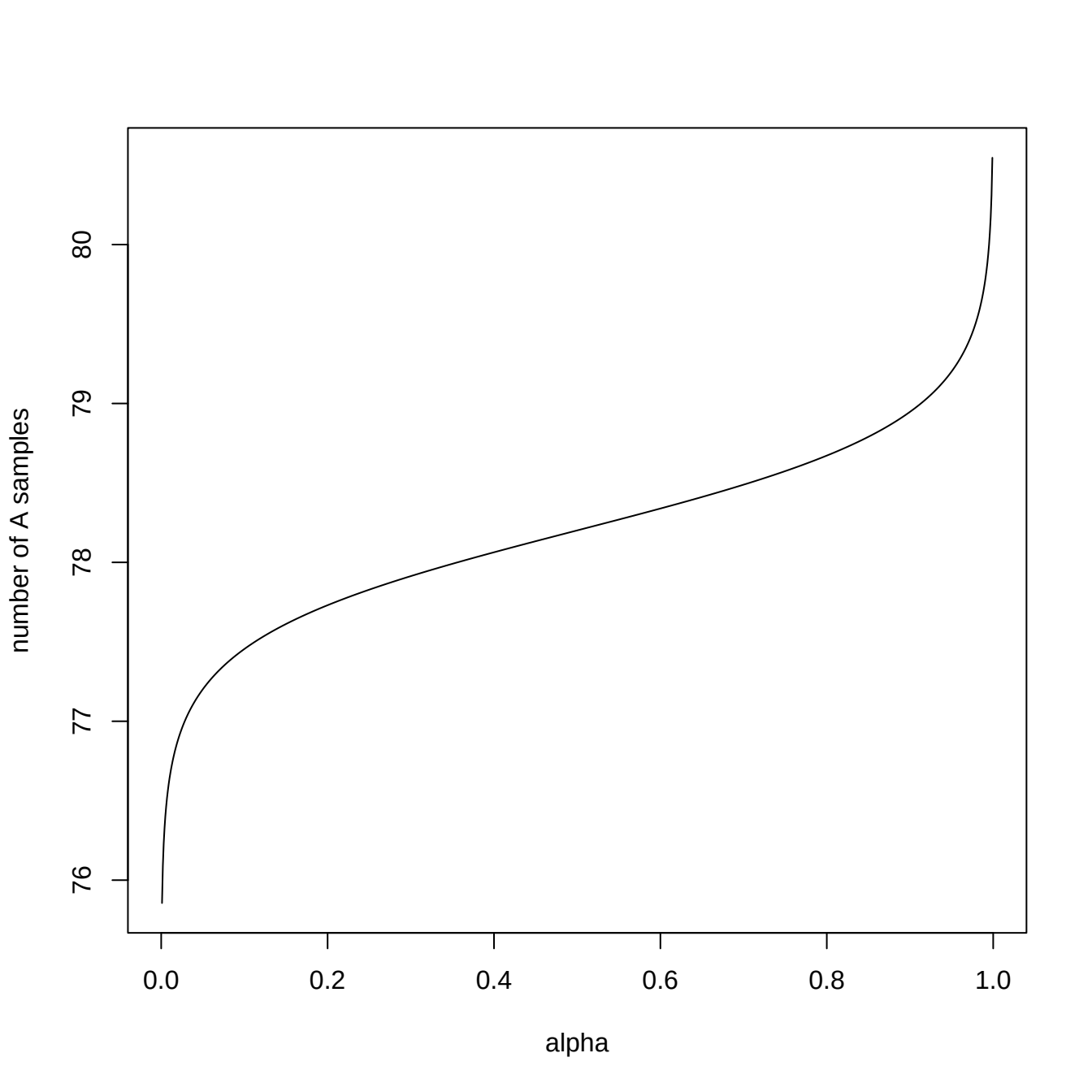
$$\Rightarrow \boxed{N_A = \frac{\log(\alpha) - \log(1-\alpha) - N\log(2e)}{\log(1-e) - \log e}}$$

Problem (1.d)

Our solution from (1c) is a function of α, N, and e:

$$N_A = \frac{\log(\alpha) - \log(1-\alpha) - N\log(2e)}{\log(1-e) - \log(e)}.$$

Since N and e are given, we can simply plot N_A as a function of α:


number of A samples
76
77
78
79
80
0.0
0.2
0.4
0.6
0.8
1.0
alpha

Problem (1e)

We need to account for case BB, which from class / homework 1:

$$P(E_1, \ldots, E_N \mid BB) = \prod_{i \in A} P(E_i = 1 \mid BB) \prod_{j \in B} P(E_j = 0 \mid BB)$$

$$= (1-\epsilon)^{N-N_a} \epsilon^{N_a} = (1-\epsilon)^{N-N_a} \epsilon^{N_a}$$

Note, given the description, we know $N_A = 0$, so,

$$P(E_1, \ldots, E_N \mid BB) = (1-\epsilon)^N$$

$$P(BB \mid E_1, \ldots, E_N) = \frac{(1-\epsilon)^N P(BB)}{(1-\epsilon)^N P(BB) + \left(\frac{1}{2}\right)^N P(AB) + \left(\frac{1-\epsilon}{\epsilon}\right)^{N_a} \epsilon^N P(AA)}$$

$$= \frac{(1-\epsilon)^N P(BB)}{(1-\epsilon)^N P(BB) + \left(\frac{1}{2}\right)^N P(AB) + \epsilon^N P(AA)} = \alpha$$

flip both sides:

$$1 + \left(\frac{\frac{1}{2}}{1-\epsilon}\right)^N \frac{P(AB)}{P(BB)} + \left(\frac{\epsilon}{1-\epsilon}\right)^N \frac{P(AA)}{P(BB)} = \frac{1}{\alpha}$$

$$\Rightarrow \left(\frac{1}{(1-e)2}\right)^N \frac{P(AB)}{P(BB)} + \left(\frac{e}{1-e}\right)^N \frac{P(AA)}{P(BB)} = \frac{1-\alpha}{\alpha}$$

Log both sides

$$\Rightarrow N\left[\log 1 - \log(2(1-e))\right] + \log\frac{P(AB)}{P(BB)} + N\log\left(\frac{e}{1-e}\right)$$

$$+\log\frac{P(AA)}{P(BB)} = \log\left(\frac{(1-\alpha)}{\alpha}\right)$$

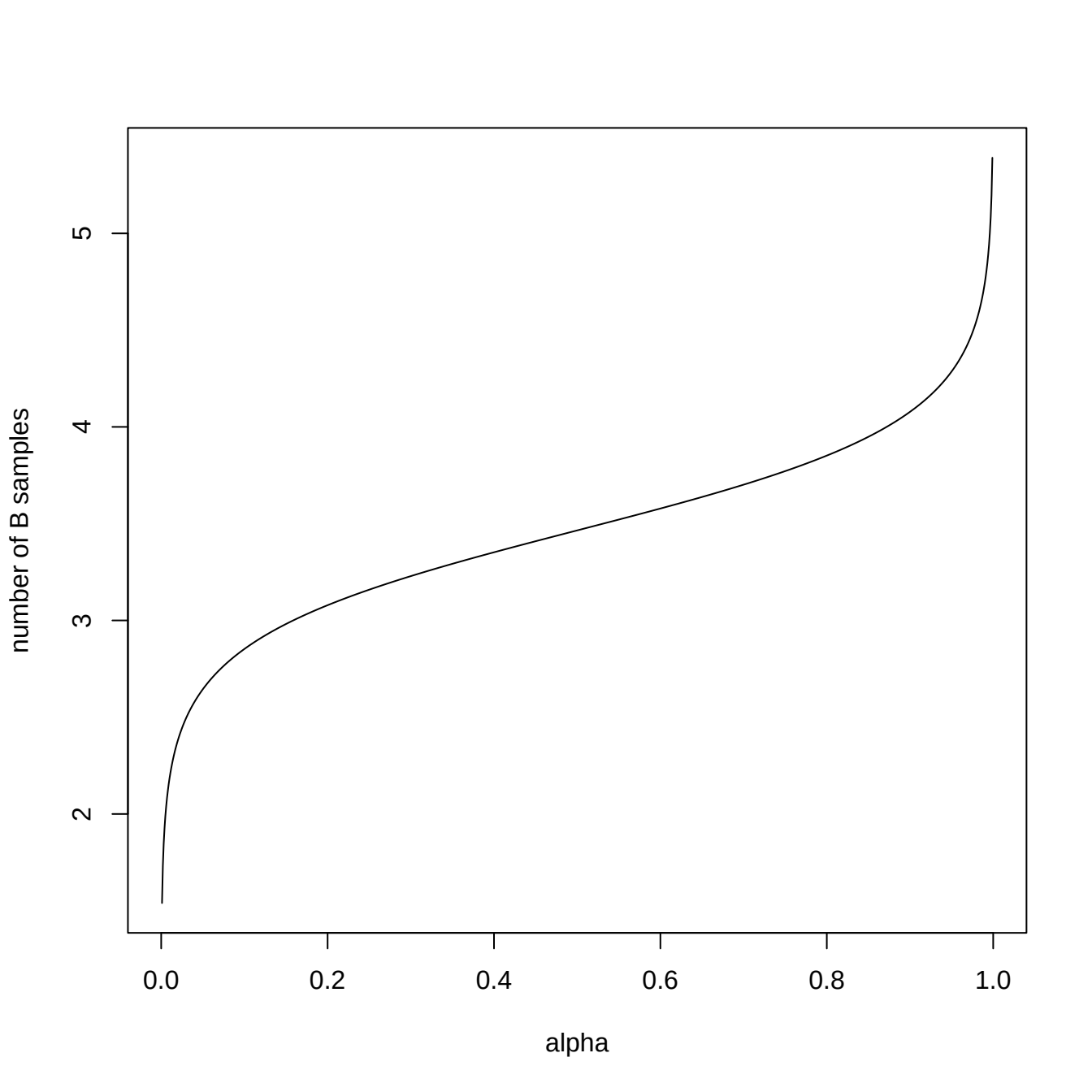
$$\Rightarrow N\left[\log\left(\frac{e}{1-e}\right) - \log(2(1-e))\right] = \log\left(\frac{(1-\alpha)}{\alpha}\right)$$

$$- \log\frac{P(AB)}{P(BB)} - \log\frac{P(AA)}{P(BB)}$$

$$\Rightarrow N = \frac{\log\left(\frac{(1-\alpha)}{\alpha}\right) - \log\left(\frac{P(AB)}{P(BB)}\right) - \log\left(\frac{P(AA)}{P(BB)}\right)}{\log\left(\frac{e}{1-e}\right) - \log(2(1-e))}$$

Fixing $e = 0.05$, $P(AA) = P(AB) = 0.4995$, $P(BB) = 0.001$,

you see the plot:


number of B samples
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
alpha

Problem 2

Problem (2.a)

t_1 = AAAACGCGAAA

t_2 = TTTTCGCGTTT

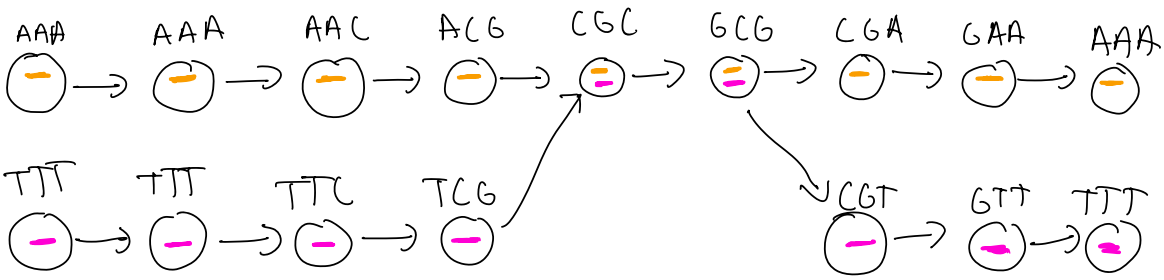
t_1 3-mers

AAA
AAA
AAC
ACG
CGC
GCG
CGA
GAA
AAA

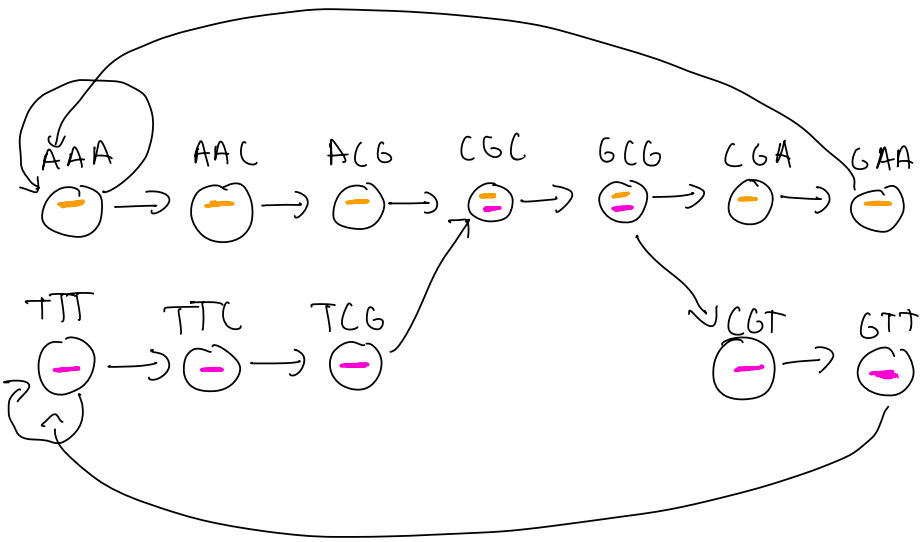
t_2 3-mers

TTT
TTT
TTC
TCG
CGC
GCG
CGT
GTT
TTT

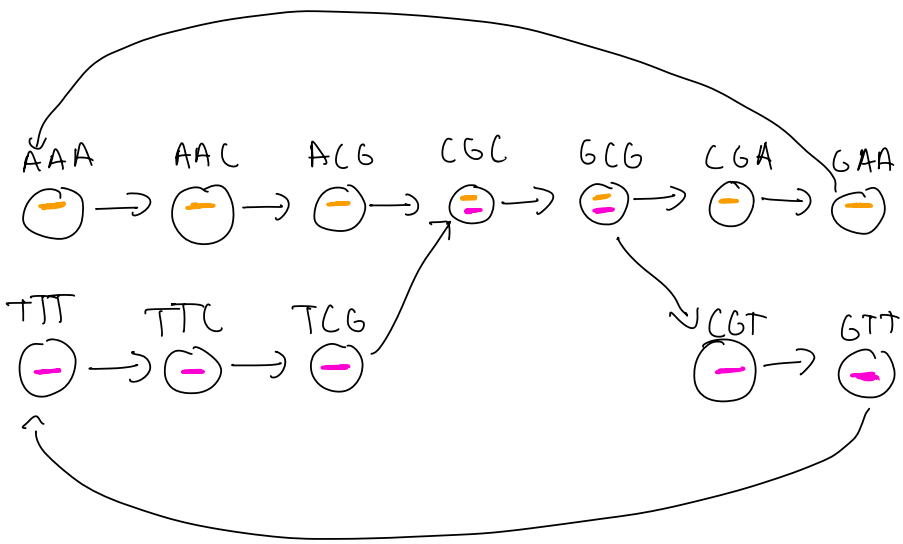
Given the k-mers, I can draw the "naive" graph w/o collapsing.



Best practice is to "compress" the graph. Let's just use multi-edges



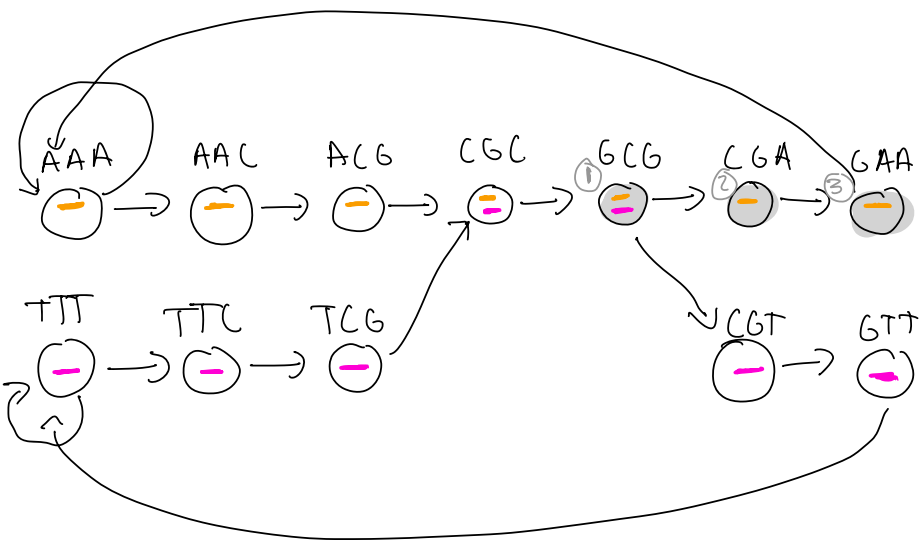
Alternatively:



Problem (2.b)

We will use the multi-edge version from (a) to be a bit more explicit

There are a few possibilities. 1.) There is an error. 2.) Your prof. screwed up & meant to give you the sequence GCGAA. Let's work under ass. (2). This case is straightforward as there is one path. Lets number & color the nodes we visit.



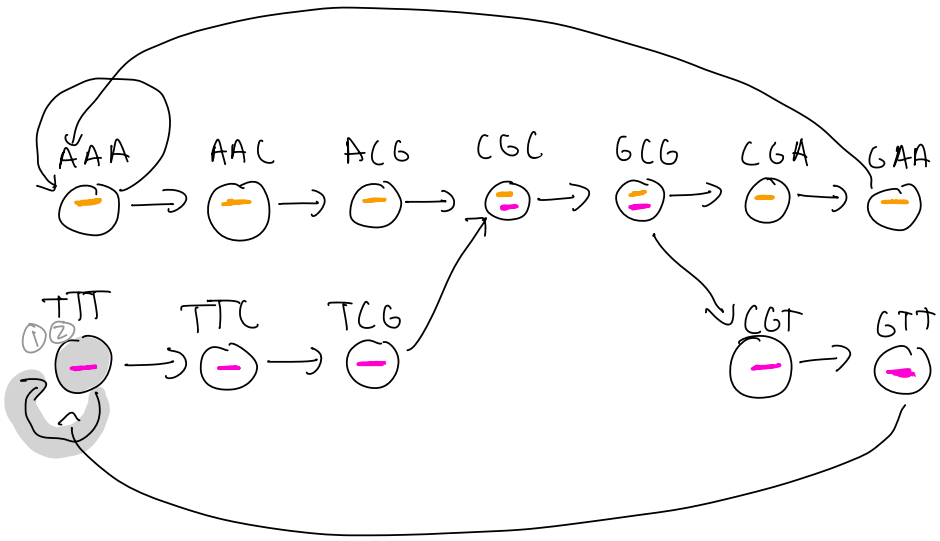
Pseudoalgorithm:

$\textcircled{1} \wedge \textcircled{2} \wedge \textcircled{3} = -$, so clearly t_1.

The problem does not state forward/reverse, so we can pretend they are in the forward strand only as long as we are explicit about this assumption.

Problem (2.c)

Because we know all reads come from the forward strand, there is only one possibility.



Pseudoalignment:

1 ∩ 2 = —, so t_2.

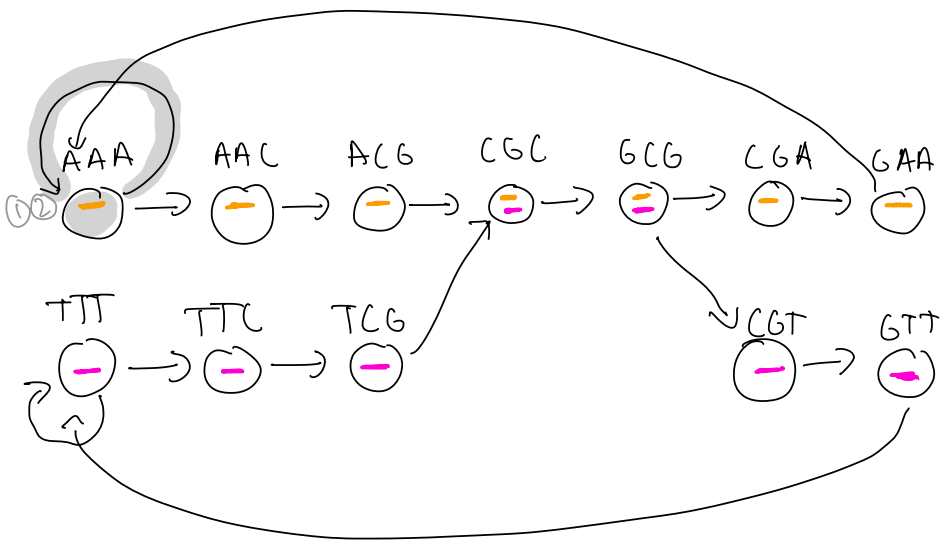
Problem (2d)

There are a few ways to deal with this. See disc. week 6 for graph alternatives. The easiest (& slowest) is to take the reverse complent of the read:

Read: TTTT

Rev. comp: AAAA

The pseudoalignment for the direct observation is in (2c). Below is the PA for the rev. complement.



Pseudoalignment:

① ∧ ② = —

Now, we need to "combine" the information somehow. One way is the union:

— ∪ — = = Final answer.

We see that this read is ambiguous given the protocol.

Problem (2e)

Again, a few different ways to do this problem.

The key here is to realize that the sequences share many portions because of the reverse complementarity.

In fact, no 4-mer exists that will allow us to uniquely resolve the read:

t_1: AAAA CGCG AAA

t_1 r.c.: TTTCGCGTTTT

t_2: TTTTCGCGTTT

t_2 r.c.: AAACGCGAAAA

By taking the reverse complement of both transcripts, we see that every 4-mer of t_1 is in the reverse complement of t_2. So, we have shown that there is no position of length 4 that enables us to uniquely identify the transcript.