COMP90016 Workshop 3

Take a closer look at read file reads.fa, which is the same data from the group assignment. You can access it from /home/subjects/comp90016/assignments/group_assignment/.

Q1. If you were to assemble the reads by overlapping them with each other and extending as described in the lecture, what length of the overlapping region would be a good choice?
Even if we don't know the length of the reference genome, we can make some arguments in this regard.

A1. In other to find such reads that both overlap with each other and extend the sequence, the reads needs to have at least 1 base outside the overlapping region. Furthermore, a good overlap should add to the confidence of the assembly instead of introducing confusion, that is, the length of the overlapping region should be as long as possible. Given above reasons, a possible good choice would be:

$$length(overlapping\ region) = \ length(read) - 1$$

Q2. Write a program that counts the number of occurrences of each k-mer for k in {3, 4, 5}.
Should the reverse complement of each k-mer be included as well? Discuss in groups in the tutorial.
Given the number of occurrences, which of these k might be used to establish significant overlap? Explain why.

A2. Program is in /home/subjects/comp90016/tutorials/week3/
The reverse complement should be included in the search. We know that if a certain sequence exists in the genome, then the reverse complement of such sequence must exist as well. However, the reads data doesn't necessarily contain sequences from both strands for each read. In order to not miss possible overlaps, we need to consider not only the reads in the data, but the reverse complement of them.
For example, we are searching for a sequence of 'ATCTG' in the reads data, but there is no match. However, there does exist 'CAGAT' among the reads, which is the reverse complement of 'ATCTG' – exactly what we are looking for! We would've missed it if the reverse complement is not considered.

The 'significant overlap' in here is the same as 'a good choice' in Q1. We want to find overlapping reads that we are confident to assemble with and extend the sequence. Thus, what we are looking for is sub-sequence of reads with exactly one match among all other reads. According to the output of the program, it's clear that when k equals 5, the number of sub-sequence with exactly one match is significantly higher than when k equals 3 or 4.

Q3. Write a program that calculates the overlap of length 5 between any two reads. Note that overlaps can happen on both strands. Print all the possible overlaps of all reads.
Discuss if this set of reads could be assembled by overlaps based on your evidence. In what ways could the data be improved?

A3. Program is in /home/subjects/comp90016/tutorials/week3/

The number of pairs is not enough to connect all 218 reads in the data according to the output. Two ways that could improve the data would be having longer reads, which should increase the number of significant overlap, as well as having more reads.