Homework assignment #2

CMPB5002

De Bruijn Graph

- Build condensed De Bruijn Graph.

  Write a program that takes fastq-file and k as input and:

  - Builds condensed De Bruijn Graph
  - Outputs edges in fasta-file
  - Counts average kmer coverage for each edge (average kmer coverage is a number of times we see this kmer in reads, so average kmer edge coverage is an average of all its kmer coverages)
  - Outputs result in a .dot file. Each edge should have a label with its length and average coverage

- Graph simplification: implement a graph simplification algorithm for basic erroneous edge removal.

  - Tip removal (remove all edges with short length and small coverage with tip topological structure).
  - Remove all low-covered and short edges

    Assess both approaches. Write a reasonable conclusion (you can add figures of your graphs to illustrate).

- Quality assessment (bonus - 25% extra mark!!)

  Use QUAST to compare contigs that you get with any 2 existing assemblers of your choice. Write a conclusion.

Hints: use k = 55 for illustrations. Add kmer and reverse-complement kmer simultaneously, so your graph will be symmetric.

Data

You can download data from https://drive.google.com/drive/folders/1M1XF4zEKwChssqt501XDZZB661PNIslG?usp=sharing. It contains 3 sets of reads and reference genomes for them.