Mülakat sürecinde paylaşılan belgeler ve materyaller Seven Bridges Genomics Biyoteknoloji A.Ş.'ye aittir ve Seven Bridges Genomics Biyoteknoloji A.Ş.'den izin alınmaksızın kopyalanması veya herhangi bir üçüncü kişiye açıklanması veya paylaşılması yasaktır. Seven Bridges Genomics Biyoteknoloji A.Ş.'ye ait olan belgelerin ve materyallerin gizliliğini ihlal eden herhangi bir aday mülakat sürecinden çıkartılacaktır. Bu koşullar mülakat ve işe alma sürecinin ayrılmaz bir parçası olup, mülakata giren her bir aday tarafından okunmuş, anlaşılmış ve kabul edilmiş sayılacaktır. Seven Bridges Genomics Biyoteknoloji A.Ş.'nin tüm hakları saklıdır.

The documents and materials disclosed during the interview process are the sole properties of Seven Bridges Genomics Biyoteknoloji A.Ş. and they shall not be copied, revealed or disclosed to any third parties without consent of Seven Bridges Genomics Biyoteknoloji A.Ş. Any interviewee violating the confidentiality of the documents and materials of Seven Bridges Genomics Biyoteknoloji A.Ş. shall be dismissed from the interview process. These conditions are an integral part of the interview and recruitment process and shall be deemed read, understood and agreed by all interviewees. All rights of Seven Bridges Genomics Biyoteknoloji A.Ş. are reserved.

Background

FASTQ is a text file format for storing biological sequences as a string of nucleotide codes (e.g., A C T G) along with quality information for each code, and is a common output of DNA sequencing machines. In most cases, the lines are in sets of 4, with the first being a sequence “ID”, the second being the sequence itself, the third containing just a ‘+’ character, and the fourth containing quality information. More information and examples can be found at https://en.wikipedia.org/wiki/FASTQ_format

A “k-mer” is a substring of length k. For example, in the DNA sequence string “GATTACA”, we have the following k-mers of length 3: [“GAT”, “ATT”, “TTA”, “TAC”, “ACA”]. Counting k-mers is a common subtask in many different genomic applications.

Task

Given a FASTQ file, produce a list, sorted by frequency, of the 25 most frequent DNA k-mers (a substring of length k) of length 30 in that file, along with the number of times they appear. You are allowed to use libraries, as long as their purpose is not specifically to count k-mers.

For reference and testing purposes, you can use the following FASTQ files from the 1000 Genomes Project: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/HG01595/sequence_read/. Your submission should complete successfully (not run out of memory or otherwise fail) on consumer grade laptops (e.g., having 4GB of memory or less), even on the files which are multiple gigabytes.

Note that we are only interested in the actual DNA sequences stored in the second line of each entry. The sequences are independent and should not be concatenated.

Requirements

The program must compile and run on Mac or Linux. If you’re on Windows, consider testing using a Linux virtual machine. Virtualbox and Ubuntu are free.

The program should accept inputs via command line arguments only. The user should be able to select the input file, the k-mer size and the number of top k-mers. Exposing other, internal parameters required by the chosen algorithm is at the discretion of the coder.

Okay:

- `./myprogram --filename big.fastq --kmersize 30 --topcount 25`
- `./myprogram big.fastq 30 25`

Not okay:

- `"Please enter a filename" (don’t have anything that waits for user)`
- `ifstream("hardcoded_filepath.fastq");`

Evaluation

Your solution will be evaluated against the following criteria (in order of importance):

1. Correctness
2. Efficiency
3. Code quality

We realize that especially the last point is subjective. Please understand that we do not expect you to meet a certain expectation of ours, but we do expect you to be able defend any choices you’ve made during an interview. There is no time limit, please take as much time as you need and please feel free to ask for clarifications.