

Задача А. FASTA статистика (3 балла)

Имя входного файла: `fasta-statistics.fasta`
Имя выходного файла: `fasta-statistics.out`
Ограничение по времени: 2 секунды
Ограничение по памяти: 256 мегабайт

Вам заданы контиги в формате FASTA.

Требуется посчитать значения метрики N50 и N90, а также среднюю, минимальную и максимальную длины контигов.

Формат входного файла

Входной файл состоит из описания контигов в формате FASTA. Описание каждого контига начинается со строки с комментарием. Строка комментария начинается с символа `>`. В последующих строках следует сам контиг. Длина каждого контига не превышает 10^5 . Сумма длин всех контигов не больше 10^6 . Контиги состоят из символов A, T, G, C.

Входной файл не содержит пустых строк.

Длина каждой строки во входном файле не превышает 80 символов.

Формат выходного файла

В выходной файл выведите 5 чисел, по числу в каждой строке. 4 целых числа: значения метрики N50 и N90, минимальную и максимальную длины контигов. И одно вещественное: среднюю длину контигов. Ответ будет считаться корректным, если абсолютная погрешность не превышает 10^{-5} .

Пример

<code>fasta-statistics.fasta</code>	<code>fasta-statistics.out</code>
<code>>First contig</code>	9
<code>ATGCCGAGC</code>	7
<code>>Second contig</code>	7
<code>AAG</code>	9
<code>GTC</code>	8
<code>A</code>	

Задача В. Обрезание FASTQ (3 балла)

Имя входного файла: `fastqcut.fastq`
Имя выходного файла: `fastqcut.out`
Ограничение по времени: 2 секунды
Ограничение по памяти: 256 мегабайт

Задан файл в формате FASTQ, который содержит чтения геномной последовательности.
Вам надо обрезать чтения по качеству.

Формат входного файла

В первой строке входного файла задано целое число p_0 ($0 \leq p_0 \leq 93$) — пороговая вероятность ошибки.

Вторая строка будет пустой.

Дальше задан файл в формате FASTQ.

Формат выходного файла

Выведите исправленный файл в формате FASTQ.

Пример

<code>fastqcut.fastq</code>	<code>fastqcut.out</code>
67 @This is sample test, length = 7 GATCGCG +Output read length = 4 gfedcba	@This is sample test, length = 7 GATC +Output read length = 4 gfed

Задача С. Подсчет k -меров (4 балла)

Имя входного файла: `count-entries.in`
Имя выходного файла: `count-entries.out`
Ограничение по времени: 2 секунд
Ограничение по памяти: 256 мегабайт

Вам задан набор чтений геномной последовательности.

Требуется для каждого k -мера определить, сколько раз он встречается в чтениях из набора.

Формат входного файла

В первой строке входного файла задано число k ($1 \leq k \leq 20$).

Во второй строке задано число n — количество чтений геномной последовательности в заданном наборе.

В следующих n строках заданы чтения, по одному чтению в строке. Длина каждого из чтений не менее k и не более 100.

Суммарная длина чтений в наборе не превосходит 50000.

Формат выходного файла

В выходной файл выведите m строк, где m — это количество k -меров, найденных во входном файле.

В следующих m строках выведите k -мер и количество вхождений этого k -мера, разделенные пробелом.

k -меры требуется выводить в лексикографическом порядке.

Пример

<code>count-entries.in</code>	<code>count-entries.out</code>
2	CT 1
2	TA 1
СТТ	ТТ 1
ТА	

Задача D. Кратчайшая общая надстрока (5 баллов)

Имя входного файла: `scs.in`
Имя выходного файла: `scs.out`
Ограничение по времени: 2 секунд
Ограничение по памяти: 256 мегабайт

Вам задан набор чтений геномной последовательности.

Требуется найти такую кратчайшую геномную последовательность, которая является надстрокой всей чтений (это означает, что каждое из чтений должно встречаться в ней как подстрока).

Формат входного файла

В первой строке входного файла задано число n ($1 \leq n \leq 20$) — количество чтений геномной последовательности в заданном наборе.

В следующих n строках заданы чтения, по одному чтению в строке. Длина каждого из чтений не более 100.

Формат выходного файла

В выходной файл выведите одну строку, которая является кратчайшей надстрокой всех чтений.

Пример

<code>scs.in</code>	<code>scs.out</code>
2 AGCTA TAC	AGCTAC