БиоИнформатика и *E. Coli* *(16 баллов)*

Без компьютеров и электронных методов хранения и обработки информации современная наука, включая нанотехнологии, немыслима. Все больше открытий совершается путем компьютерной обработки уже имеющихся данных, например, геномов, поскольку в наши дни уже расшифрованы и доступны для всех желающих геномы множества организмов.

Бактерия *E. Coli* (кишечная палочка) является одним из удобных модельных организмов в биологии, а геном ее лабораторного штамма K-12 был в числе самых первых расшифрованных. Для выполнения этого задания сохраните по ссылке[1] (<5 MB) с сайта Национального центра биотехнологической информации США (NCBI) к себе на компьютер файл генома *E. Coli*.

Часть 1. ДНК и данные

Рис. 1.

1. Посмотрите свойства сохраненного вами файла: сколько байт составляет его размер? (**1 балл**) Рассчитайте точное число пар оснований[2,3] в геноме *E. Coli*. (**2 балла**)

Длина генома измеряется, как правило, в мегабазах (Mbp, от «mega base pairs») – т.е. в миллионах пар оснований.

2. Оцените, сколько примерно мегабайт (MB) будет занимать файл генома, имеющий длину в одну мегабазу (1Mbp), при таком же[3] способе электронной записи файла. (**1 балл**) Поместится ли на обычный DVD диск файл с геном человека, имеющим длину 3 234.83 Mb? (**1 балл**)

3. На любом языке программирования напишите программу (к решению приложите ее код), которая прочитает скачанный вами файл и сосчитает суммарное число нуклеотидов в геноме *E. Coli* и процентное содержание каждого из них по отдельности, чему они равны? (**4 балла**)

[1] Рекомендуется сразу сохранить файл (или переименовать скачанный sequence.fasta) в coli.txt Также вы можете скачать этот файл в архиве с сайта Nanometer по ссылке со страницы задачи (~1.4 MB).

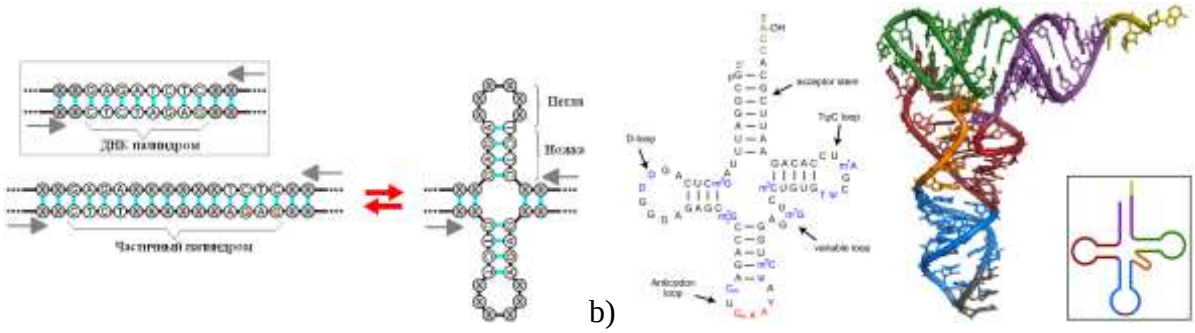
[2] Наследственная информация в молекуле ДНК хранится в виде текста записанного всего четырьмя буквами – **A**, **G**, **T**, **C**. Каждой букве из одной ДНК цепочки соответствует строго определенная (комплементарная: **A** напротив **T**, **C** напротив **G**, а также наоборот) буква второй цепочки. Поэтому для описания генома достаточно записать буквами только одну из них, что и сделано в скачиваемом вами файле, поэтому число пар оснований равно числу символов нуклеотидов в этом файле.

[3] Учтите, что каждый символ в файле кодируется 8 битами информации, а через каждые 70 букв нуклеотидов стоит символ переноса строки; первая строка, содержащая описание файла, вместе с пустой последней строкой содержат суммарно 87 символов, включая символы переноса строк.

Часть 2. Поиск настоящих ДНК-палиндромов

ДНК-палиндромом называется такая последовательность ДНК, прочтение которой совпадает с прочтением в обратном направлении по комплементарной цепочке. Например, последовательность **ATTA** – «обычный» палиндром, а последовательность **AATT** – ДНК-палиндром. Важным биологическим свойством ДНК-палиндрома является то, что если его цепочку сложить пополам, она будет сама себе комплементарна (см. рис. 2a).

Палиндромы могут задавать особенности наноструктуры молекулы ДНК, которые отвечают за выполнение тех или иных важных функций во внутриклеточных процессах и могут принимать весьма сложные формы. Поэтому поиск таких ДНК-палиндромов может требовать применения достаточно сложных алгоритмов.



a) b)

Рис. 2. Примеры нуклеотидных палиндромов:
b) транспортная РНК дрожжей (нуклеотидная последовательность и трехмерная модель);

4. На любом языке программирования напишите программу для поиска ДНК-палиндромов в файле генома *E. Coli* K-12 и найдите все палиндромы с длиной ножки от 17 до 25 пар оснований (максимальную длину петли при поиске ограничьте 8 нуклеотидами). Приложите к решению текст программы и найденные палиндромы (достаточно привести номер первого нуклеотида палиндрома, число пар нуклеотидов в ножке и число нуклеотидов в петле). **(7 баллов)**

- рекомендуется читать файл последовательно, используя строки (а не массивы) для хранения прочитанных символов. Для быстрого поиска палиндромов удобно использовать функции поиска подстроки в строке (например, функцию pos в Turbo Pascal); для упрощения алгоритма можно не искать палиндромы в начале и конце файла.
- Если вы не можете выполнить задание полностью, найдите палиндромы с ножкой от 14 до 25 пар нуклеотидов с размером петли 0 нуклеотидов.