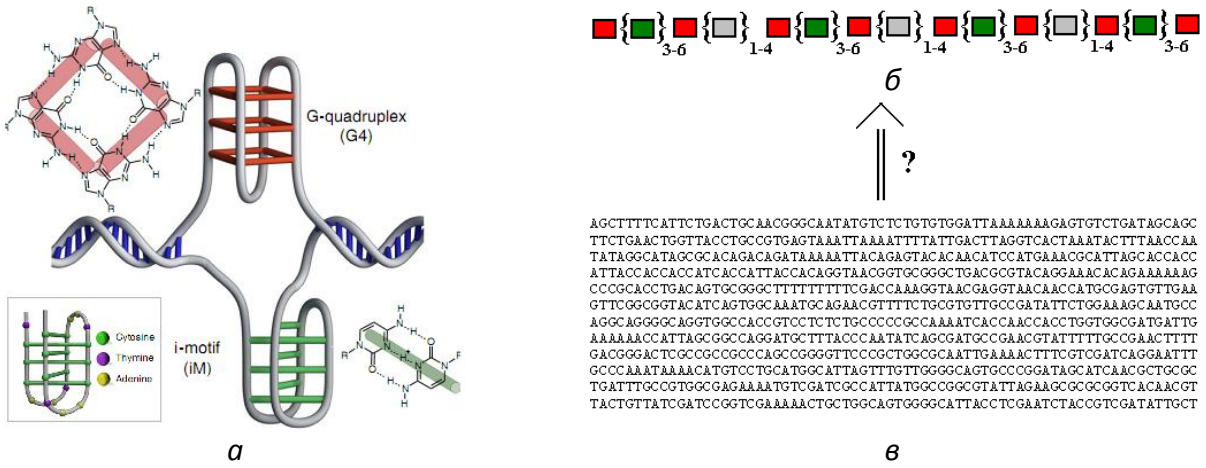
МАТЕМАТИКА

Математика для школьников 7 – 11 класса (заочный тур)
Задача 5. Поиск наномотивов в ДНК *E. Coli*



а *в*

*Рис. 1. а) Наномотивы – неканонические фрагменты структуры ДНК (пояснения см. в тексте задачи). б) Общий шаблон последовательности, отвечающей наномотивам в тексте генома. Здесь: зеленый квадрат отвечает букве **G** для **G**-квадруплексов и **C** для **i**-мотивов, красный квадрат – любой букве кроме **G** и **C**, соответственно, серый – любой из четырех букв. в) Начало файла[2] генома штамма K-12 E. Coli, открытого в текстовом редакторе.*

Единичные нити ДНК[1] с определенным расположением гуанина **G**, способны самопроизвольно сворачиваться в четырехцепочечные спирали – ***G*-квадруплексы** (рис. 1а), которые обладают повышенной устойчивостью. При этом четыре нуклеотида **G** из разных цепей образуют плоскую структуру, называемую G-квартетом. В свою очередь, комплементарные[1] им цепочки, богатые цитозином **C**, также могут образовывать трехмерные ДНК-структуры – ***i*-мотивы** (*i-motif*), в которых нуклеотиды **C** соединены попарно, как показано на рисунке 1а. В начале 2018 года ученым удалось не только впервые зафиксировать ***i*-мотивы *in vivo***, но и исследовать их функции в ядре человеческой клетки. Оказалось, что оба типа структурных наномотивов выполняют регуляторную функцию (входят в состав теломеров и промоторов) и широко представлены во всех известных геномах.

Напишите программу (на любом языке программирования), которая позволит найти, сколько всего ***G*-квадруплексов** и ***i*-мотивов**, соответствующих шаблону (рис. 1б), находится в тексте генома[2] *E. Coli* (рис. 1в). В ответе приведите исходный код программы, а также сами нуклеотидные последовательности и позиции их начала (номер по порядку в геноме) для каждого найденного наномотива.

Подсказка: в программе для упрощения процедуры поиска наномотивов можно использовать регулярные выражения.

[1] 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.

[2] Бактерия *E. Coli* (кишечная палочка) является одним из удобных модельных организмов в биологии, а геном ее лабораторного штамма K-12 был расшифрован одним из первых. Для выполнения этого задания сохраните по указанной ссылке с сайта олимпиады http://enanos.nanometer.ru/uploads/archive/ecoli.zip архив файла (~1.3 Мб) генома штамма K-12 *E. Coli*, который состоит из одной непрерывной строки, содержащей только буквы **A**, **G**, **T**, **C**.

Всего – 8 баллов