

Математика для школьников 7 – 11 класса (заочный тур) Решение задачи 6. ДНК для хранения информации: от теории к практике

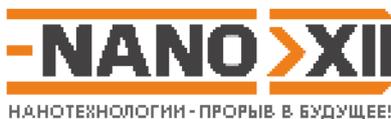


Рис. 1

1. Каждый нуклеотид произвольной последовательности ДНК может принимать 4 значения – т.е. кодирует 2 бита информации. Следовательно, 192 нуклеотида кодируют $192 \cdot 2 = 384$ бит информации.

Сколько этих строк можно закодировать таким способом? Столько же, сколько чисел можно закодировать $8 \cdot 2 = 16$ битами, т.е. суммарно $2^{16} = 65536$ строк (или по-другому: максимальный номер строки будет $1111111111111111_2 = 65535$; всего строк, с учетом нулевой, будет 65536). Следовательно, объем информации (в МБ) составит $384 \cdot 65536 / 8 / 1024 / 1024 = \underline{\underline{3 \text{ Мегабайта}}}$.

2. Файл содержит **19** строк, из которых 18 полных (содержат 200 символов нуклеотидов) и одна неполная из 76 нуклеотидов, следовательно, запись файла состоит из $200 \cdot 18 + 56 = 3656$ символов нуклеотидов, из которых информацию кодируют $3656 - 19 \cdot 8 = 3504$. Поскольку каждый символ кодирует 2 бита информации, то исходный файл имеет размер $3504 \cdot 2 / 8 = \underline{\underline{876 \text{ байт}}}$.
3. Расшифруем код, зная, что в блоке адреса закодированы цифры от нуля до 18 (19 строк). Бросается в глаза, что стоящий слева нуклеотид, кодирующий номер строки и его соседи, не меняются. Значит – этот нуклеотид кодирует 00. Меняются только 3 последних нуклеотида адреса, очевидно, что адрес, состоящий из одинаковых нуклеотидов – G – нулевой (т.е. G = 00). Самая короткая последовательность (56 символов) – это, очевидно, самая последняя строка. Поскольку всего 19 строк, то ее адрес 18, следовательно:

$$18 = 10010_2 = 00\ 00\ 00\ 00\ 00\ 00\ \mathbf{01\ 00\ 10}_2 \Leftrightarrow \mathbf{G\ G\ G\ G\ G\ C\ G\ T}$$

Таким образом расшифровываются коды нуклеотидов C = 01 и T = 10 (остается A = 11).

4. Алгоритм: создаем в памяти массив строк, читаем последовательно строки из файла **image.txt**, раскодируем в первую переменную первые 8 нуклеотидов и во вторую переменную – оставшиеся 192 нуклеотида. Переводим из двоичной системы в десятичную адрес строки (первая переменная) и заносим в соответствующий адресу элемент массива вторую строку. После прочтения всех строк перебираем в цикле строки массива, дописываем байты в новый файл **image.png**. В конце работы программы у нас будет готовый раскодированный файл, который достаточно открыть в любой системе (программой умеющий читать распространенные графические файлы), чтобы увидеть, что картинка – это логотип 12-й олимпиады (рис. 1).