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

Молекулы ДНК обладают одной из самых больших плотностей хранения информации. Недавно группа ученых предложила способ кодирования информации с использованием адресной записи в короткие последовательности нуклеотидов. Например, ученые смогли закодировать в ДНК, а затем успешно прочесть разнообразные файлы с данными, включая 3 изображения (рис. 1) и даже операционную систему. Такой способ позволяет быстро находить и считывать только нужные фрагменты данных, не требуя технически сложно реализуемых чтения и записи длинных молекул ДНК.

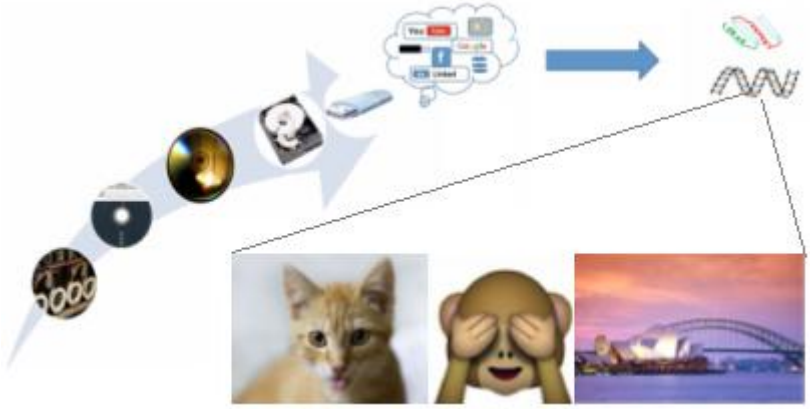

Рис. 1

Рассмотрим пример использования такого способа кодирования информации (рис. 2). Файл, состоящий из логических нулей и единиц, кодируется последовательностью нуклеотидов, записанной 4 буквами (A, C G, T), которая содержит такое же количество информации. Эта последовательность затем разбивается на строки, содержащие не более 192 нуклеотидов (блок II, рис. 2). Порядковый номер строк (начиная с 0) называется адресом и кодируется в последовательности из 8 нуклеотидов (блок I, рис. 2), для этого он записывается в двоичном виде и кодируется тем же кодом, что и остальные данные.

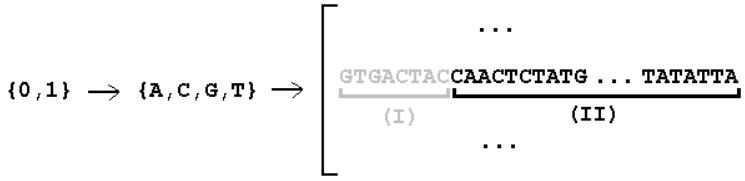

Рис. 2

1. Найдите, какой максимальный объем файла (в мегабайтах) можно закодировать таким способом. **(2 балла)**

Далее приведены все прочтенные ДНК-цепочки (расположенные в случайном порядке), отвечающие некоторому файлу-изображению.

2. Сколько строк и символов нуклеотидов содержит такая запись файла, **(1 балл)** рассчитайте размер (в байтах) исходного файла изображения. **(1 балл)**

3. Сколько возможных вариантов нуклеотидного кода существует для такой записи файла? **(1 балл)** Установите, каким вариантом кода было закодировано изображение. **(3,5 балла)**

4. Напишите программу*, с помощью которой можно будет раскодировать эту ДНК-запись изображения, и опишите вкратце алгоритм ее работы. Что изображено на закодированной картинке? **(8,5 баллов)**

* Программу можно написать на любом языке программирования, но обязательно приведите ее исходный код и полученный файл **image.png**.

Учтите, что в файл изображения **image.png** по мере раскодирования необходимо записывать не *текстовые символы* «0» «1», а *двоичные коды:* «0» «1» (при этом размер файла должен соответствовать рассчитанному в п.2). Если нет ответа на вопрос 3, то при раскодировании файла переберите все возможные варианты кодировок: только верный вариант кода позволит открыть и увидеть картинку.

Приведенные ниже последовательности можно скачать в виде текстового файла **image.txt** (архив: http://enanos.nanometer.ru/uploads/archive/image.zip).

GGGGGGTGCTGTGAGCGCGCGGACCGGCCGAGTATCCCCTATTCTAAATAAGAACTCTTTTGTCACCATTATACTAACT
GAATATTCCCGCGAGGGCGTTTCGTACCCTTCGAAGGATCACTCATTCATCCGGATTTAAGCAAGACGTGAACAGTTCTG
TCAGACGTCTCTCGACCTGTCAGTCTCTTCTTGGCGGTGGA

GGGGGGATCGGCGGCTTCACTACTCTACAACCAACAACGGACGGCTACCAAGGGTGTAGAGCGATAGAGGCAACAACCC
CTAGACCTAGCCACGCAACAACCTGACCTTGACAGGCCCGCGCGGTAGAAGGCTTTATAATTCGAGCTTGCCTAGTAGAC
TGGTAGATTAGCGCAGTACGGGGCCGCAATAAGTTCCGGTA

GGGGGCGTGGGGGGGGGGGGGGGGCGTCCGCCCGATCGCGTTATCGGTCTGGTGGT

GGGGGGAGAATCGTAAAGACGGTGCCTCATCACAAACTTCCCAGAAGAGTCGCCGAGCTCGGCTTCAGTCGGAGTTTAG
GTGCGGGTGTTTCACAGGCTGGCCCCCGTGCTCGAGTGTTTCGTAGCTTCCTCACGTTGTGGGTGTTTCGTGTCGGTTTT
TAGTGGACGTTCGACCTCACCCTACGACCCAAGCTACTAGA

GGGGGGGAGGTTTTTCGGAGCATCACACAGTCTCGTCGCTTCATCAATCGCCGACTAACCAACGCGTATCCAATCAAAG
AACTCGAACAATCCACCTAACGGCTGACCCGGCGCACCCAATGCATGTGGAATCATGCGTATACGGGGCTAATACGCCAG
CGGGAATCTCGGCCGAGTTTATGGCTGATAATGTAAGAGTA

GGGGGGCGTCCTTTTCTTATGACCGGTGCCTGGTTCCCTGCTCGCCGCACCTTACAATGGCACAATGCGGTTAACAATC
TCTTGCAATTCGGCAATTGTGAAAGGCGCGTGACACCGGGAATCCGACATCACAATATTTCTGGATGGGTTTGTCTCTAT
GCTAGAGACTGTGCGGATAGCCGATCATTACCTGCCTCGTC

GGGGGGCTTAGGATCGTGGTATCACCGATTCTAAGACTGTCCCCGCTCATGGCATTGAGTCGCCAACTTGGAGGTGCCC
TCCCCCCCCCCCCCACCCATTCCACGGCTTGCTTCACGCGTATTCTAGATCCCAGGATCGTATCTTAACGGGCAAGATAAT
CCTCATGGCCCGCTAAGAGATGGGGCGGTGAGCCGGAGCAT

GGGGGGCAAAAGGGCATTGCAAGAGCTTGTCGGTAGGAACTTAGTTTCCTAACTTCCAAGAAGAGGCGGATGCCCCGCG
GGTCAGACTAATTGTTTCACTTTAGTCAAACGCGAGACTTAGTGTTTAAAGCGAATCGCCCTCAGCGCGGATCTCCAACA
ATATGCCGGTCACCTCAGAGTTCATGTGCCCAATAGAAGCG

GGGGGGAACAGCTCGGGGGACATTGCACACGGATCTGGATGCCTTCTAGTGGATCTGTTTGACCCCCTACTGTATCTGC
AGGTGGAACTATGGCTCCTATTGGTGTGTTCGCGCCGGAAGCTTCGGCCGCGGATATGGATGGCCCCGTAGCGTATAGCG
TTGAGGACCCCAAAATCGGCCGCCCGTTGATGCTTAGTAGC

GGGGGCGCTTGTGTGACGGCACCATCGGTCGACTGGTCGTCCGCTGTTTGGCAAATCACTTAGAACGATTAGACTACAA
AGGCTATACCGCCTCCTTCGGAGTAATCCGGGTTACTGTTCAATTCGGCTCGCGCACGTCCCCTCTTCATGTTCTAGCGG
CCAATGGGCCCATTTGACTATAGTGCCGGCATGATTCTGGA

GGGGGGTCACAGTACCACAACTTAGGATGTTCAAGAAACGGAGTTAGATTCAGATCGAGATCCCGAATGCAAAGCGATA
CACGAAGATTAGGGCAGAGTTAGTGCTTGCACGTTTTTTCGCCGAAGATATACACCCACTACCGTTAAGACGTGCTCTTC
CCAGGTCCTAGTGCTTAAGGCGGGAACGAAACGAGCATATT

GGGGGGTTTTATATTATATTTTATTATAACACCTATCGGACTGAACCGTAATCCACGTGTTTGAACTCTACATCTCGCC
TACTGAATATTCCATGAGCCCTACAAAAGTGACTCTAACACTGTCTCTATAACACTCTTTAATATATGGCCGCACATTCC
CCTTAGAGCCGGAGCATGAGGTTTGACGTGAATAAACCCGC

GGGGGGTATGAGTGATACACTACCTGAATTCTTATCATGGGGCGAAACTTTGGCTAATTCTCACCGAAAGCGATTGCAC
GCGCCTATGCCTGTCAGTGAGCCGCAGTTCATCGCCAGGGACCAAACCACTTAAACGCGATCTAGGATTTTGAACGATCC
CGACGAATCGACCGTGCCGGATCGCCCAAGCGAGAGTGCCG

GGGGGGGCGAGGGGAGCCCCGGAGGGAAGCTCCAGATGGGGCAGTTGACGTACGCTGGCAGCACCATATGACAATGGCG
TCGCCGGGCGGGACGCCATAACCGGAATGGTAGCGCCGAAGATCTGTGTGAACCGGGAGCTCGAGCACCGGCAGGGATGG
TGTTTCCGCGCCGGTACAGTCGAAATCCTCGCGGCGGAGTG

GGGGGGACGAGCTGTGCATACTTGTACCTTACCTAACTGTGTCAAGGCGTGCAGAGTTATCGGGAATACGACATGACAA
CATCTGCGCCGCAGAGCGGCAGAGTTCCCAGGCGCATGTTGACCTCCTTGTGATATTTAATTATGGACAGTGTAAGGCCG
TGAGATACCTTATATTATACTCTACCGGCTGAGAACGACCC

GGGGGCGGGACTCCGAGATCGGTATACCTCCTCGTAAATGGTGCGTCCTAGCAGGGTTTACTGGTCGTTATCGCAGAAT
GCGATTCTTACTCTGAAGCCATCGTGTGGGTCTCTGGTTCCTAGCGCAGGTTCTGGACGTCTGGGCGCGCCGGTAGGCCT
GATGCTGTCATGTAAGAGCTCCGGCCTCGTTGTGTCAGGTA

GGGGGGGGTGTCCCGGCGATCGCAGGACGGTTGCTTGGTTGGGGGGGGGGGGGGACCGTCCGTGCGCGCCGTGGGGGGG
GGGGGTCAGGGGGGGGGGGGGGAGGGGTGGGGTGGGGGGGGGGGGCTTAGGCGTGCGTGAAGGGGGGGGGGGAGAGACGT
CCGCGCGGCCCCGCATGACTTATACCCTAGATACTATAGGA

GGGGGGGTGACTGTACTCGCATAATCGCTCCGGTCCGTCATATATAATATAATCCCGGTGGTAAGTTCGGCGGGGTGTG
CCCCCCTCGGGGACCGTATTTACCTTAACGATCGGTTGCAGTATGGCAGTCTTCTAAAAGACAGGGTCCTGTGCCTCCCC
TCGTCTTCTCAGTGCGGGACATACTTGCGCCCGTGCTTAAG

GGGGGGCCTTGGTAATTATAATTTCGCACATGGCACCCTAAATCCCGATGTTCAAATTTCCATGAGTCAAGAAATCGCA
GTGCAAGCCATTAACTATCTACCGTCTTTTAAAACAAGAAAGCATGGAATTCCACCGAGCAAATAGATAATCCTTATCGG
AAAGACTACGCGCCATCCTAATGATGTATACTCTCTTGTCG

Всего – 17 баллов