МАТЕМАТИКА

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

Алгоритм поиска: проходим по всему файлу скользящим окном, в котором содержится 44 нуклеотида, внутри которого ищем ***G**-квадруплексы* и ***i**-мотивы* по указанному шаблону.

Текст программы (*на PascalABC Net*):

```
var
  f: Text;
  str, regC, regG: String;
  char: Char;
  len, n, nPos: Longint;

begin
  len := 44; {размер "скользящего окна", максимальная длина последовательности рис. 1б условия}

{шаблон поиска для i-мотива*}
  regC := '^[^C]C{3,7}[^C].{1,4}[^C]C{3,7}[^C].{1,4}[^C]C{3,7}[^C].{1,4}[^C]C{3,7}[^C]';
{шаблон поиска для G-квадруплекса}
  regG := '^[^G]G{3,7}[^G].{1,4}[^G]G{3,7}[^G].{1,4}[^G]G{3,7}[^G].{1,4}[^G]G{3,7}[^G]';

{посимвольное чтение файла}
 Assign(f, 'ecoli.txt');
  Reset(f);
  while (not Eof(f)) do {пока не достигнут конец файла}
  begin
    Read(f, char);
    n := n + 1; {счетчик прочитанных нуклеотидов}
    str := str + char; {добавляем прочитанный нуклеотид в строку}
    if n >= len then
    begin
      nPos := n - len;  {позиция первого символа последовательности str}
      if (length(str.MatchValue(regC, RegexOptions.None)) > 0) then
        writeln('C: ', nPos, ' ', str.MatchValue(regC, RegexOptions.None));
      if (length(str.MatchValue(regG, RegexOptions.None)) > 0) then
        writeln('G: ', nPos, ' ', str.MatchValue(regG, RegexOptions.None));
      str := copy(str, 2, length(str)); {отбрасываем первый символ str, чтобы начало строки на следующем шаге приходилось на следующий нуклеотид}
    end;
  end;
        end.
```

* Пояснение к шаблону поиска на примере i-мотивов:

- ^ в начале строки шаблона означает, что начало шаблона должно совпадать с началом строки, по которой будет вестись поиск;
- [^C] – любой символ кроме «C»;
- C{3,7} – от 3 до 7 символов «C» подряд;
- .{1,4} – от 1 до 4 любых символов.

Всего найдено 11 ***G**-квадруплексов* и 12 ***i**-мотивов*:

G: 53224	AGGGGAGTTGGGGGAATAAGGGCGGAGGGT
C: 164596	ACCCTCACCCTAACCCTCTCCCT
G: 171663	TGGGCGCGGGTCTGGGGCTGGTGGGC
G: 388675	TGGGGAGAGGGTTAGGGTGAGGGGGC
C: 425039	GCCCGAATCCCTGATTGCCCACTATCCCA

G: 497854	CGGGGAGAGGGTTAGGGTGAGGGGA
G: 624590	TGGGGAGAGGGTTAGGGTGAGGGGA
C: 632040	GCCCAGGGTTCCCTCTCACCCTAACCCT
C: 632049	TCCCTCTCACCCTAACCCTCTCCCCG
C: 1351202	TCCCCTCACCCTAACCCTCTCCCCA
C: 3046050	TCCCCTCACCCTAACCCTCTCCCCA
C: 3239660	TCCCCTCACCCTAACCCTCTCCCCA
C: 3390492	TCCCCTCACCCTAACCCTCACCCCA
C: 3504855	TCCCCTCACCCTAACCCTCTCCCCA
G: 3592474	TGGGTGAGGGAAAATGGGAGATGGGGC
G: 3608695	TGGGGAGAGGGTTAGGGTGAGGGGA
C: 3695942	TCCCCACGCCTCCCCGCACCCCCTGCTATCCCA
C: 3781025	GCCCCCTCACCCTAACCCTCTCCCT
G: 3908506	TGGGGAGAGGGTTAGGGTGAGGGGA
G: 4070463	TGGGGAGAGGGTTAGGGAGAGGGGA
G: 4231285	CGGGAAAAGGGTTAGGGTGAGGGGA
G: 4314296	TGGGGAGAGGGTTAGGGTGAGGGGGC
C: 4549846	TCCCCTCACCCTAACCCTCTCCCCG

Любопытно, что среди найденных наномотивов много одинаковых или близких последовательностей, что может свидетельствовать об их важной роли в жизнедеятельности клетки.