Βιοπληροφορική

Παπαδομανωλάκης Ελευθέριος 1634

Άσκηση :

Open the file (yersinia_genome.fasta) with the complete Yersinia genome and find the possible start and end positions of its genes.

In the code given we do the following steps:

- Read from file everything except the first line that contains some info. In the end we have a string containing the whole sequence
- Then we also find its reverse complement and keep it
- We have a fuction that takes a string and finds the sequence(s) we are searching with the following logic
  1. Every match finds the Shine-Dalgarno sequence or any of its variants ,the start codon ,the sequence we are looking for and the stop codon. From its start/end position we make a string.
     - after Shine-Dalgarno there are {4,8} bases

GGC GGG CAG TAA GGA GGT GCG GTT ATG CGC CTC CTC CTG TTT TTT ACT TAA TCA GCA

MATCH

TAA GGA GGT GCG GTT ATG CGC CTC CTC CTG TTT TTT ACT TAA

  2. Then we discard everything before the start codon and the two codons to extract the sequence we search for.

CGC CTC CTC CTG TTT TTT ACT

- We call the above fuction for original sequence **and** its reverse complement.
- The regular expression that follows the rules is:

([TA][AC]AGGA[GA][GA])([ATCG]{4,10})(ATG)([ATCG]{3})*?(TGA|TAA|TAG)

*? **Quantifier**: Matches between **zero** and **unlimited** times, as few times as possible, expanding as needed

The sequence we are searching for is between the green and red color

Output:

```
>gi|22123922|ref|NC_004088.1| Yersinia pestis KIM, complete genome

------Analysing normal sequence------

------Analysing reverse complement sequence------

        Total triplets: 70128 (210384 bases) ,in 346 matches
```