Elements of Computational Biology

Subject 6: Searching amino acid sequence with PROSITE regular expressions

Agata Radys, Paweł Cejrowski, Łukasz Myśliński

18.01.2016

1. *Problem describtion*

Our program in the input has a sequence of amino acids and regular expression pattern written with PROSITE notation. It gives section of the sequence matching with the pattern as an output.

PROSITE pattern notation:

- – – separator between the pattern's elements,
- V – any letter, one letter amino acid code,
- x – any amino acid,
- […] – one amino acid from bracket,
- {…} – one amino acid, but not from bracket,
- e(i) – for element e and number i: repetition of e exactly i times,
- e(i,j) – repetition of e exactly k times, where k≥i and k≤j.

2. *Technology*

Our program is created in C# in Visual Studio without using any extra libraries.

It includes 2 classes:

- AAlist – class used to define output from sequence with given PROSITE pattern notation
- Match – class which has amino acid sequence, given pattern, initializes them and has a method to return a matching pattern while using AAlist class

3. *Usage*

To launch it, one has to clone the source from Github repository:

```
git clone https://github.com/MiSSLab/Bioinf6.git
```

and open and launch the solution in Visual Studio.

4. Test data modification

In file Program.cs there are 3 testing sequences of amino acids taken from lecture slides and 1 PROSITE regular expression to match in every sequence. After launching the program in the console window we can see given 3 sequences and 3 matched fragments of them, as below:

Pattern:
`[RK]-G-{EDRKHPCG}-[AGSCI]-[FY]-[LIVA]-x-[FYM]`

Sequences:
```
SRSLKMRGQAFVIFKEVSSAT
RGQAFVIF

KLTGRPRGVAFVRYNKREEAQ
RGVAFVRY

VGCSVHKGFAFVQYVNERNAR
KGFAFVQY
```

To modify the test data, one has to change the strings in Program.cs file and then lanuch the program again.