Bioinformatics, computer lab IV

University of Potsdam WS 2019/2020, S. Hartmann

Today's exercise will help you to understand and practice using regular expressions by searching for protein motifs in amio acid sequences. You will explore the online resource Pfam, and optionally you can also search DNA sequences for restriction enzyme patterns.

1 Regular expressions

Download the two data files `proteins.fasta` and `dna.fasta` from Moodle. They have been compressed into a single file. Navigate to the downloaded file and extract it, using the contextual menu (right-click!).

Open a terminal and navigate to the directory in which you stored these files. Briefly look at them with `less` to inspect their content.

1.1 Patterns in protein sequence

1.1.1 Thionins

Thionins are small plant defense proteins that occur in seeds and cell walls of leaves. These proteins are relatively short (45-50 amino acids), and they always contain conserved Cysteines (C) that are involved in forming disulfide bonds. The consensus pattern used to identify thionins is described below:

- Two C, *followed by*
- any five amino acids *followed by*
- R, *followed by*
- any two amino acids *followed by*
- either an F or Y *followed by*
- any two amino acids *followed by*
- C.

✎ Write a regular expression that will find this pattern. Use it to search for this motif in the sequence stored in the test file `proteins.fasta`. How many times does this motif occur in this file?

✎ Search again for thionins in the test file. This time, however, use a regular expression that searches for the word 'thionin'. How many times does this term occur in the test file?

✎ Are the two numbers you get for the previous two exercises the same? How do you explain the difference? You may look at the content of the file with `less` to answer this question.

1.1.2 Zinc fingers

✎ Zinc finger motifs were discussed in class. Write a regular expression that matches the following Prosite zinc finger motif:

C - x(2,4) - C - x(3) - [LIVMFYWC] - x(8) - H - x(3,5) - H.

✎ Use your regular expression to search for occurrences of the zinc finger motif in the file `proteins.fasta`. How many times does this motif occur in the test file?

✎ Search again for zinc fingers in the test file. This time, however, use a regular expression that searches for the term 'zinc finger'. For this, do a case-insensitive search that doesn't distinguish between upper- and lower-case letters. How many times does the search-term occur in the test file?

✎ Are the numbers you got for previous two questions identical? Do you have an explanation for this? You may look at the content of the file with `less` to answer this question.

2 A database of families of functional domains and motifs: Pfam

Two weeks ago you searched for globins on GenBank and Uniprot. To find the corresponding Pfam entry, you could first go to the UniProt entry for the human hemoglobin sequence with the ID P68871. From there, you could follow the link to the corresponding Pfam entry. Alternatively, direct your browser to the Pfam site at `http://pfam.sanger.ac.uk`. You can diretly go to the Pfam globin entry by entering the ID PF00042 in the form under "JUMP TO enter any accession or ID" and click on "GO". You will be taken to the Pfam entry for the globin domain.

Note: Keep in mind that the globin domain may only be a part of the proteins in which it occurs, and that other domains may also be present in these proteins. Some of the information you will look up pertains only to the globin domain(s) in the different proteins, while other information tells you something about the entire protein sequences with globin domains.

☞ Skim the summary information to see what kind of information it contains. Does it contain more or less detailed information about the function of globins, as compared to UniProt? How is the information related to Wikipedia?

To the left there is a menu that allows you to explore different aspects of this domain family: Click on the first link, "Domain organization" to view the different domain architectures of proteins that contain the globin domain.

A one-sentence description of telling you about the kind of information that is shown on this page is given at the top. Click on the blue link "More..." to read a more detailed description. Click on the blue link "Less..." to get back to the one-sentence description.

☞ How many protein sequences have no other domain except the globin domain?

☞ Skim through the entries to see in what kinds of proteins the globin domain occurs. Do these proteins have similar functions?

Next, click on the menu item "Species" to get to a list of species in which this domain is found. The information is given in the form of a sunburst or a species tree, depicting the relationships of organisms. Sunburst is the default display. For domains that are found in fewer species, you can click on the tab "Tree" to generate the species tree. In this case, however, this option is not available.

☞ Is the globin domain found in all domains of life, or is it restricted to some domains? Which domain of life has the most sequences with the globin domain?

Later in the semester we cover alignments and structure prediction, and you may want to come back to this section then. But feel free to explore the menu items “Alignments” and “Structures” now, if you like. Follow any of the other links that interest you. Explore!

2.1 OPTIONAL: Patterns in DNA sequence

2.1.1 Restriction enzyme patterns

Restriction enzymes are enzymes that are isolated from bacteria and used frequently by molecular biologists: restriction enzymes recognize specific sequences double-stranded DNA and then cut the DNA into fragments. The specific recognition sequences for three different restriction enzymes are given below:

AgeI ACCGGT
BanI GG(T/C)(A/G)CC
PsrI GAACNNNNNNTAC

✎ Write a regular expression that finds the recognition sequence for the restriction enzyme AgeI. How many times (if at all) does it occur in the sequences within the file `dna.fasta`?

✎ Write a different regular expression that finds the recognition sequence for the restriction enzyme AgeI. How many times (if at all) does it occur in the sequences within the file `dna.fasta`?

✎ Write a regular expression that finds the recognition sequence for the restriction enzyme BanI. How many times (if at all) does it occur in the sequences within the file `dna.fasta`?

✎ Write a regular expression that finds the recognition sequence for the restriction enzyme PsrI. Did you check whether the sequences contain only valid nucleotides? How many times (if at all) does it occur in the sequences within the file `dna.fasta`?