# Lab Exercise 6

1)

The file sequence.txt contains a protein sequence in a corrupted format (with line numbers). Use regex in order to extract the clean amino-acid sequence of that protein.

2)

Search for the **zf-C2H2** pfam domain HMM logo and write a regex for it. Find the occurrences of this domain in the sequence you extracted in (1).

**It should be noted, however, that regex is NOT a good way to deal with HMM profiles.**

3)

Many functionally important cellular peptides and proteins, including hormones, neuropeptides, and growth factors, are synthesized as inactive precursor polypeptides, which require post-translational proteolytic processing to become biologically active. This is achieved by the action of a relatively small number of proteases that belong to a family of seven proprotein convertases (PCs), which tend to cleave peptides precursors at sites with multiple (usually two) positively-charged amino acids (i.e. Lysine and Arginine). These cleavage sites are often referred to as “dibasic cleavage sites”.

The known motifs of dibasic cleavage are:

1. R-R
2. K-[K or R]
3. R-X-X-[K or R]

Where **X** denotes any amino-acid.

For example, the following sequence:

*MFGYRSLLVLLVTLSLCLLLQSSHCSAVRTYGNDLDARARREIISLAARLIKLSMYGPEDDSFVKRNGGTADALYNLPDLEKIGKR*

Would be cleaved into the following products:

1. *MFGYRSLLVLLVTLSLCLLLQSSHCSAVRTYGNDLDARARR*
2. *EIISLAARLIK*
3. *LSMYGPEDDSFVKR*
4. *NGGTADALYNLPDLEKIGKR*

Read the sequences in **neuropeptides.fasta** and write the cleavage products into a new FASTA file.

4)

A)

Read **viperdb.csv** into a dictionary of columns. The keys of the dictionary should be the column names, and the values should be NumPy arrays with the values of the records. The arrays should have of a proper type, depending on whether they contain numeric or textual data. For example, the “Inner Radius” column should be a numeric array with the inner radius values of the 419 records. Make sure to parse N/A values into *np.nan*.

B)

What are the average inner/outer/average radiuses of the viral records? What are the standard deviations?

C)

We define outlier records to be records that are at least 2.5 standard deviations away from the mean with respect to any of the three radius types.

How many outliers are there? To what viral families do they belong?

What is the average number of subunits after filtering out the outliers?