Regular Expression

In this class, we will talk about the application of regular expressions and develop several exercises for searching patterns with regular expressions in python.

Learning Objectives

- Describe the regular expressions syntax and its applications in pattern search.
- Develop scripts and programs to process text files and search patterns.

1. Review the second part of the slides “Sequence Pattern Finding”.

Quiz

1. Describe a regular expression for the following patterns:

- AATATC, AAATATC, AAATATC
- ACGT, ACCT, ACTT, ACAT
- "She said heed very clearly!", "She Said head very clearly", "She said hood clearly"

2. Which of the following regular expressions are syntactically incorrect? Why? What kind of strings do the valid ones match?

- ?.\b
- [?].\b
- [^\D]{4,1}
- ^[^^]
- ++
- +\+
- \++
- \+\+

3. What kind of strings will be matched by the following regular expressions? In your answer, break them up into parts and explain what each one does.

- ^[+-]?\d+(\.\d+)?
- \b[aeiou][a-z]{,4}\b
- [.?!]\s+([A-Z][A-Za-z]*)

4. Write a regular expression that describes an animal name with three letters and the middle letter is a vowel.

5. "What are blue, red and yellow?". Write a RE that match any of those 6 responses:

- colours
- colors
- they're colours
- they're colors
- they are colours
- they are colors

6. Write an RE to capture the following sentences:

- Red. Blue. Green!
- Red. blue. Green!
- Red. Yellow. Blue. Green.
- Red. Blue. Grey.

7. Which of the following statements are incorrect?

a) ‘pq*’ will match ‘pq’

b) ‘pq?’ matches ‘p’

c) ‘p{4}, q’ does not match ‘pppq’

d) ‘pq+’ matches ‘p’

Quiz

8. *Write a RE to match an IP address.*
9. *Write a RE to match an email address.*
10. *Write a RE to match an entire sequence register in a fasta file (include header and sequence).*
11. *Given an DNA sequence write a RE to detect a putative protein with at least 10 aminoacids.*

Finding patterns in sequences. Develop python programs for the different exercises.

Task 2 – Highlight motif in sequence

Consider the Zinc finger RING-type signature from Prosite (Entry: PS00518), with the following motif:

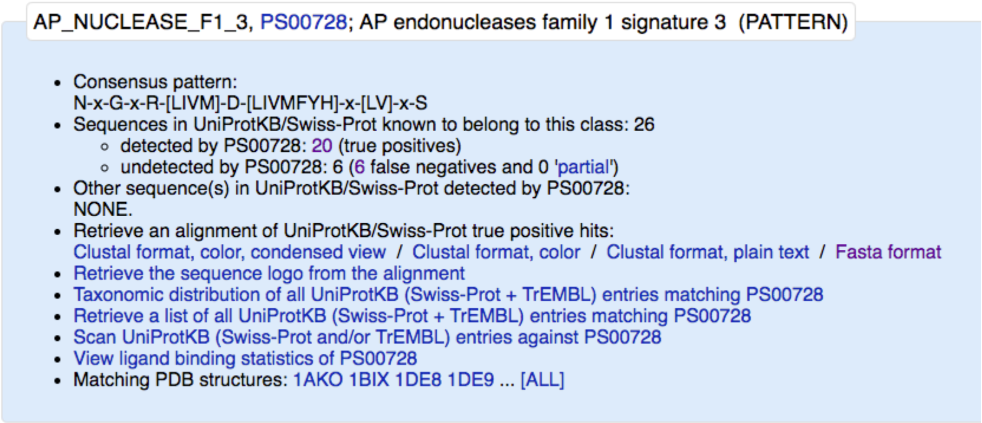
C-x-H-x-[LIVMFY]-C-x(2)-C-[LIVMYA]

Write a script that scans the sequence Q8RXD4.fasta for the occurrence of the motif. Print again the input sequence, where in the positions where the motif occurs it should be printed in upper case letters. The rest of the sequence should be printed in lower case letters.

Task 3 – Classify sequences by motif presence

Consider the following Prosite motif. Sequences from AP endonuclease family should in principle match the PS00728 motif. There are 20 sequences (Tp – true positives) where the pattern is detected and 6 (Fn – False negatives) where it is undetected. Write a program that reads the file PS00727.fasta that contains a subset of sequences from this family and determines which sequences match and those that do not match the pattern.

1) *The script should be called ps00728.py*
2) *It should be run as ps00728.py PS00727.fasta*
3) *For each sequence print its identifier (e.g. sp|Q5XF07|APE1L_ARATH) followed by a space and word MATCH or NOT_MATCH if the sequence matches or not the motif.*



AP_NUCLEASE_F1_3, PS00728; AP endonucleases family 1 signature 3 (PATTERN)

- Consensus pattern:
  N-x-G-x-R-[LIVM]-D-[LIVMFYH]-x-[LV]-x-S
- Sequences in UniProtKB/Swiss-Prot known to belong to this class: 26
  - detected by PS00728: 20 (true positives)
  - undetected by PS00728: 6 (6 false negatives and 0 'partial')
- Other sequence(s) in UniProtKB/Swiss-Prot detected by PS00728:
  NONE.
- Retrieve an alignment of UniProtKB/Swiss-Prot true positive hits:
  Clustal format, color, condensed view / Clustal format, color / Clustal format, plain text / Fasta format
- Retrieve the sequence logo from the alignment
- Taxonomic distribution of all UniProtKB (Swiss-Prot + TrEMBL) entries matching PS00728
- Retrieve a list of all UniProtKB (Swiss-Prot + TrEMBL) entries matching PS00728
- Scan UniProtKB (Swiss-Prot and/or TrEMBL) entries against PS00728
- View ligand binding statistics of PS00728
- Matching PDB structures: 1AKO 1BIX 1DE8 1DE9 ... [ALL]

Task 4 – Match Sequence info

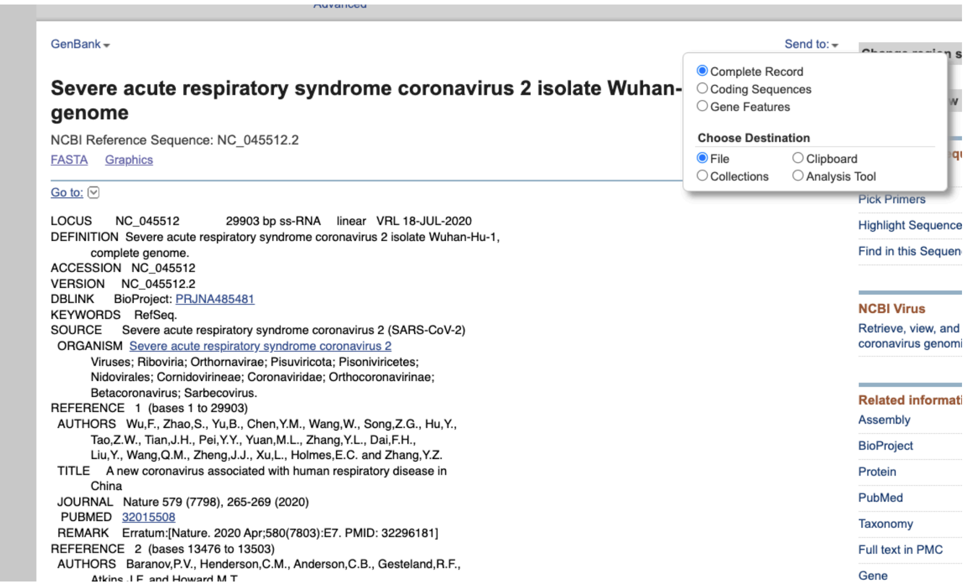
For each of the following tasks write a function that applies regular expression and returns the required information as specified. Use the PS00727.fasta to test you code.

1) *Function that scans the input file and returns a list with all the sequence identifiers.*
2) *Function that returns the most frequent species in the file. The species is identified by the keyword OS.*
3) *A function that returns a dictionary, with sequence ids as the key a tuple as the value. The tuple should be the corresponding information of the sequence (OS, OX, GN, PE, SV).*
4) *A script called sequence_info.py that tests the above functions.*

Task 5 – Parse the Genbank file for SarsCov2

In this exercise, we will parse the information contained in a Genbank file about the SarsCov2 genome. For this you should apply regular expression to extract the requested information.

- *Retrieve the Genbank file for SarsCov2 genome file. Download the Genome sequence from Genbank:*
- *Goto: https://www.ncbi.nlm.nih.gov/genome/?term=MT072688*
- *Click on the RefSeq link https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2*
- *Dowload the Genbank file (see figure below).*



Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-genome

NCBI Reference Sequence: NC_045512.2

Send to:

● Complete Record
○ Coding Sequences
○ Gene Features

Choose Destination

● File ○ Clipboard
○ Collections ○ Analysis Tool

```
LOCUS       NC_045512            29903 bp ss-RNA     linear   VRL 18-JUL-2020
DEFINITION  Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1,
            complete genome.
ACCESSION   NC_045512
VERSION     NC_045512.2
DBLINK      BioProject: PRJNA485481
KEYWORDS    RefSeq.
SOURCE      Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)
  ORGANISM  Severe acute respiratory syndrome coronavirus 2
            Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes;
            Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae;
            Betacoronavirus; Sarbecovirus.
REFERENCE   1  (bases 1 to 29903)
  AUTHORS   Wu,F., Zhao,S., Yu,B., Chen,Y.M., Wang,W., Song,Z.G., Hu,Y.,
            Tao,Z.W., Tian,J.H., Pei,Y.Y., Yuan,M.L., Zhang,Y.L., Dai,F.H.,
            Liu,Y., Wang,Q.M., Zheng,J.J., Xu,L., Holmes,E.C. and Zhang,Y.Z.
  TITLE     A new coronavirus associated with human respiratory disease in
            China
  JOURNAL   Nature 579 (7798), 265-269 (2020)
   PUBMED   32015508
  REMARK    Erratum:[Nature. 2020 Apr;580(7803):E7. PMID: 32296181]
REFERENCE   2  (bases 13476 to 13503)
  AUTHORS   Baranov,P.V., Henderson,C.M., Anderson,C.B., Gesteland,R.F.,
            Atkins,J.F. and Howard,M.T.
```

1) *Develop a script implementing different functions to retrieve the following items of information:*
   a. *Retrieve the list of PUBMED identifiers. Pubmed is a database of biomedical literature. Each entry corresponds to a scientific article.*
   b. *Retrieve the list of genes (ORFs).*
   c. *Retrieve the list of protein identifiers.*
   d. *Obtain the translated protein sequence for each gene. Return a dictionary with the key = gene and value = protein_sequence.*

Task 6 - Cleaning Up a Sequence

It is frequent than a sequence appears in format different than those in the fasta format. See for instance the ORIGIN in the Genbank file.

```
ORIGIN

        1 attaaaggtt tataccttcc caggtaacaa accaaccaac tttcgatctc ttgtagatct

       61 gttctctaaa cgaactttaa aatctgtgtg gctgtcactc ggctgcatgc ttagtgcact

      121 cacgcagtat aattaataac taattactgt cgttgacagg acacgagtaa ctcgtctatc
```

Write a script that returns a contiguous sequence without spaces, new lines and numbers. Apply to the file names genome_unformatted.fas

Hint: use the sub function from regular expressions.