CS202: PROGRAMMING PARADIGMS & PRAGMATICS

Semester II, 2022 – 2023

Lab 9: Regular Expressions Exercise

- **AIM -** Using Regular Expressions in Perl to find Patterns in Biological Sequence Data

- **Introduction**
  - Use the sample data files provided for these exercises
  - First you should study the files and notice the structure of the data
  - In all exercises you will have to parse the files using regular expressions

- **Exercise 1: Extract Details (`Details.pl`)**
  - Extract the accession number, the definition and the organism (and print it)
  - **Example:** For **`data1.gb`** file output should be

```
J00265
Human insulin gene, complete cds.
Homo sapiens
```

  - Should work for all similar files like **`data1-4.gb`**

- **Exercise 2: Extract MELINE Article Number (`Medline.pl`)**
  - Extract and print all MEDLINE article numbers which are mentioned in the entries.
  - **Example:** For **`data1.gb`** file output should be

```
80054779
80120725
80147417
…
```

  - Should work for all similar files like **`data1-4.gb`**

- **Exercise 3: Extract Translated Gene (`Translated.pl`)**
  - Look for the line starting with **`/translation=`**
  - An amino acid sequence can be short, i.e. only one line, or long, i.e. more than one line
  - If more than one line, concatenate lines to form the whole/complete sequence
  - **Example:** For **`data1.gb`** file output should be

```
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLAL
EGSLQKRGIVEQCCTSICSLYQLENYCN
```

  - Should work for all similar files like **`data1-4.gb`**

- **Exercise 4: Extract DNA Sequence (`DNA.pl`)**
  - This is the whole base sequence in the end of the file
  - Remove indexing number, & spaces and concatenate all the lines into one long sequence
  - **Example:** For **`data1.gb`** file output should be

```
gccagggtgtccccttcctaccttggagagagcagccccagggcatcctgcaggggtgc
tgggacaccagctggccttcaaggtctctgcctccctccagccaccccactacacgctgc
tgggatcctggatctcagctccctggccgacaacactggcaaactcctactcatccacga
...
tggagtccccagagaccttgttcaggaaagggaatgagaacattccagcaattttccccc
cacctagccctcccaggttctatttttagagttatttctgatggagtccctgtggaggga
ggaggctgggctgagggaggggggt
```

- Should work for all similar files like **`data1-4.gb`**

- **Exercise 5: Extract only Coding DNA Sequence (`CodingDNA.pl`)**
  - This is described in **`FEATURES`** - **`CDS`** (Coding DNA Sequence)
  - **Example:** For **`data1.gb`** file says **`'join(2424..2610,3397..3542`**)' and means that the coding sequence are bases 2424-2610 followed by bases 3397-3542 of the whole DNA sequence (found in Exercise 4). Concatenate all the bases to form the whole coding DNA sequence.
  - Remember to generalize; there can be more (or less) than two bases, and the 'join' line can continue on the next line.
  - Should work for all similar files like **`data1-4.gb`**

- **Submitting your work:**
  - All source files and class files as one tar-gzipped archive.
    - When unzipped, it should create a directory with your ID. Example: **2008CSB1001** (NO OTHER FORMAT IS ACCEPTABLE!!! Case sensitive!!!)
  - Source files should include the following: (Case-Sensitive file names!!)
    - **`Details.pl`** [3 Points]
    - **`Medline.pl`** [2 Points]
    - **`Translated.pl`** [3 Points]
    - **`DNA.pl`** [5 Points]
    - **`CodingDNA.pl`** [7 Points]
  - ***Negative marks for any problems/errors in running your programs***
  - Submit/Upload it to Google Classroom