COMP 364: Homework 1

Fundamental of Python

Assigned: Jan. 21, 2013
Due: Feb. 6, 2013 (in class)

Submit all solutions in the form of python script files that the TA can run. Zip your solutions up into a compressed file called `<your name>_hw1.zip` and email it to `jerome.waldispuhl@mcgill.ca` (indicate COMP364 in the subject).

1 Concatenation (2 pts)

Write a script called `concat.py` that concatenates the command line arguments of your script together, with spaces between them. Put a period at the end. Your script should print the string out. As an example, if your program was called like this:

```
./concat.py The quick fox jumped over the lazy dog
```

Then it would print out this:

```
The quick fox jumped over the lazy dog.
```

2 Functions (2 pts)

The function `int(x)` will take a string value and convert it into its integer equivalent. For example, `int('3')` returns an integer value `3`. Write a script called `add2.py` that takes two numbers as arguments and prints their sum.

3 Functions 2 (2 pts)

Create a new script `addN.py` that does the same thing as `add2.py`, except that it will add an arbitrary number of integer arguments. As an example, if your program was called like this:

```
python addN.py 1 1 2 5 10
```

Then it would print: `19`

4 Input/Output (2 pts)

Write a script `fileconcat.py` that takes two filenames as arguments, read these files, concatenate their content and store the result in a new file named `outputQ4.txt`.

5 Input/Output (2 pts)

Download the file `http://www.cs.mcgill.ca/~jeromew/comp364/data/HW1Q4.fasta` and store in on your computer. This is a file storing a sequence using the FASTA format. Lines starting by '>' are identifier or comments about the sequence stored.

Write a script that reads a file passed as an argument in the command line. Remove the comments and print it in the terminal.

Bonus (1 pt): Perform a check on the sequence to determine if it is a DNA, RNA or protein sequence. Print the result in the terminal as well. (Indication: You can use regular expressions).

6 Gene structure (5 pts)

Write a regular expression that describes a mature mRNA surrounded by START and STOP codons. Then, write a script that:

- identify if an input sequence given stored in a file contains a mature mRNA, The name of the input file should be passed through the command line (3 pt),
- returns the coding sequence without the START and STOP codons found in the sequence stored at `http://www.cs.mcgill.ca/~jeromew/comp364/data/HW1Q6.fasta` (1pt).

N.B.: Store the file `HW1Q6.fasta` on your computer. Your script must take the filename as an argument in the command line and read the file.

7 PROSITE patterns (5 pts)

PROSITE (`http://prosite.expasy.org/`) is a protein database. It consists of entries describing the protein families, domains and functional sites as well as amino acid patterns, signatures, and profiles in them. We saw the syntax of PROSITE patterns in class. You can also find additional information at `http://prosite.expasy.org/scanprosite/scanprosite-doc.html#pattern_syntax`.

Go to `http://prosite.expasy.org/PDOC50135.txt` and extract from this file the PROSITE consensus pattern. Translate this pattern into a Python regular expression (1 pt).

Write a program that extracts the subsequence matching this pattern in the sequence stored at `http://www.cs.mcgill.ca/~jeromew/comp364/data/HW1Q7.fasta`, print it in the terminal and returns the index where the pattern starts.

N.B.: Store the file `HW1Q7.fasta` on your computer. Your script must take the filename as an argument in the command line and read the file.

8 Translation (5 pts)

Re-start from the end of question 6. You have extracted a coding sequence from `http://www.cs.mcgill.ca/~je`
We ask now to translate this sequence of codons into an amino acid sequence. Write a script that takes a RNA sequence as an argument and translate it into its protein sequence.