Python Unit 1 programming problem

September 13, 2016

These are the exercises for this week. Solve exercises 1-6 in an interactive python shell and write an script to solve problem 7. YOU ONLY HAVE TO UPLOAD A TEXT FILE CONTAINING THE SCRIPT TO SOLVE PROBLEM 7, problems 1-6 are optional and won’t be marked.

As we already mentioned, string methods are extremely useful to work with sequences. Since we do not have time to try all of them in class, we encourage you to explore them yourself. To see all the methods and variables associated to a string type *help(str)* in a python prompt (python interactive shell). Although the displayed information may seem confusing, do not worry about understanding everything, just go over it and focus on the methods (e.g. capitalize (...), etc).

Now create a string variable such as MySeq=”MHKPASSTAG” and observe the output of different methods[1]:

Problem 1. MySeq.isdigit(), MySeq.isalpha(), MySeq.startswith(’MHK’) (try other parameters ’M’, ’HHK’,...), MySeq.enswith(’TAG’),...

1. compare the result of MySeq.upper() with MySeq.isupper() same with MySeq.lower() and MySeq.islower(),...

2. compare the results you get in these tests with the information reported by help(str)

3. assign different strings to MySeq and repeat steps above

...

Problem 2. The nucleotide frequency in the genome of a newly discovered specie is given in the following dictionary: NucFrq={’A’:0.35,’C’:0.25,’G’:0.3,’T’:0.2}. Using the variable NucFrq, what would be the probability of finding the EcoRI restriction site (’GAATTC’)? How many EcoRI sites would you expect if the genome is 2.7*10^5bases long? Solve it using the interactive python shell and the provided variable (dictionary).

...

Problem 3. The nucleotide frequency in the genome of a newly discovered specie is given in the following array: NucFrq=[0.4,0.3,0.2,0.1], where the values correspond to the frequencies of “A”, “C”, “G” and “T” respectively. Using the variable NucFrq, what would be the probability of finding the HindIII restriction site (’AAGCTT’)? How many HindIII sites would you expect if the genome is 3.4*10^9bases long? Solve it using the interactive python shell and the provided variable (array).

Given that MySeq=”ACGTGG” is a DNA fragment written in the standard 5->3’ orientation:

Problem 4. How would you compute its reverse (i.e the output should be the string “GGTGCA”). Tip: take a look at list slicing.

...

Problem 5. The method *str*.translate(), coverts (i.e. map or “translate”) symbols in *str* to another symbols using a translation table. For example:

```
>>> Original_symbols="abcdefghijklmnopqrstuvwxyz"
>>> Coded_message=Message.translate(TransTable)
```

[1] dir(MySeq) will display all variables and methods associated to the variable MySeq

```
>>> TransTable=str.maketrans(Original_symbols,Secret_code) #makes a translation table
>>> Message="python rocks!"
>>> Coded_message=Message.translate(TransTable)
>>> Coded_message
'(E_8/& =/3#?!'
>>>
```

Use these string methods to produce the complementary strand of MySeq=”ACGTGG” (i.e. the ouput should be “TGCACC”

...

Problem 6. Combine your code from problems 4 and 5 to generate the reverse complement of MySeq=”ACGTGG”.

THIS IS THE (only) ONE YOU MUST SOLVE:

Problem 7. Write a program that ask the user to type (keyboard input) a DNA sequence and the name of an enzyme and returns:

1. composition of the DNA sequence (% of each base)

2. indicate if the DNA sequence will be digested by the selected enzyme

3. indicate the (first) cutting position of the selected restriction enzyme, if applicable.

To make things simple the user can only choose one of these four enzymes: EcoRI, BamHI, HindIII and NotI.