Heidi Lischer
University of Zurich

URPP Evolution
in Action

17 April 2015

Exercises Python basics Tutorial – part 2

Download the files:

https://www.dropbox.com/s/ekgw6npd9utuxxs/Python2.zip?dl=0

Work locally on your laptop

The following instructions have been tested under Python 2.7.3:

1. Copy the zipped data to your computer (Ubuntu):
   ```
   wget
   https://www.dropbox.com/s/ekgw6npd9utuxxs/Python2.zip?dl=0
   ```

2. Unzip the data:
   ```
   unzip Python.zip
   ```

3. Download Biopython under Ubuntu:
   ```
   sudo apt-get install python-biopython
   ```

4. Start Python by typing `python` in the terminal

Exercise1:
Create a DNA sequence object of following 3 sequences:

```
>seq1
CTTTGCTCGTCTGATGCGCATTATTCCGCACTCGCTTGCGGCGGCAATGCTTtGGCGGGATTTTATTACGC
TTTGGATTACAGGCGTTTGCCAGTCTGGAC
>seq2
GCTGAAAGGCGCATGGGCGGCGCGTACCATCCaGATGAAAGCTCAGGTGAAGCGTCAGGAAgAGGTGGCGA
AAGCCATCTACGACCGCGGGATGAACAGCATTGAGCGGGCG
>seq3
ATCATAGCCTGCAAGTGGCCGGAGAGCGAAGGGcTATCCGGCCAGGGTGAAATTATCGCCGCGAACGCACA
ATTTGATATCGACGaGTAAAGTACTCAAACGGCGCGCTCCACACATGCAC
```

a) Extract a sub-sequence from base 9 to 30 of the first sequence

b) Concatenate the 3 sequences to one and make sure that they are in upper case

c) Estimate the GC content of the concatenated sequence

Exercise 2: translation

In the Mouse_genes.fa file you can find a few sequences of mouse genes. Generate a new FASTA file containing their protein sequences. Thus, take each nucleotide sequence from the original file, and translate it (use SeqIO to read the file into memory). The key point is that for each nucleotide `SeqRecord`, we need to create a protein `SeqRecord`. You can write your own function to do this.

Exercise 3: Quality filtering of FASTQ files

The FASTQ file Example_800_R1.fastq contains Illumina reads with PHRED scale quality scores. One common task is taking a large set of sequencing reads and filtering them (or cropping them) based on their quality scores. The following exercise is very simplistic, but should help understanding the basics of working with quality data in a `SeqRecord` object. FASTQ files are an excellent example of per-letter-annotation, because for each nucleotide in the sequence there is an associated quality score. Any per-letter-annotation is held in a SeqRecord in the `letter_annotations` dictionary as a list, tuple or string. The quality values can be accessed by `rec.letter_annotations["phred_quality"]`

a) First count the number of reads in the file (use SeqIO to read the file)

b) Filter the file for reads with a minimum PHRED quality of 10 and write it into a new FASTQ file. How many reads pass the filter?

Exercise 4: Exercises without Biopython

a) Print out the names of the amino acids that would be produced by the DNA sequence "GTT GCA CCA CAA CCG". Split this string into the individual codons and then use a dictionary to map between codon sequences and the amino acids they encode (GTT=Val, GCA=Ala, CCA=Pro, CAA=Glu, CCG=Pro)

b) Write a function that takes a list of numbers and returns the mean of all the numbers in the list. Estimate the mean of the even numbers under 20.

c) Write a function that takes a single DNA sequence as an argument and estimates the molecular weight of the sequence (A=331, C=307, G=347, T=306 g/mol, N=mean weight of the other bases). What is the molecular weight of `AAGGACTGTCNCGTNNCGTAGGATNATAGNN`

d) Write a program that reads in a tab delimited file () with 4 columns: gene, chromosome, start and end coordinates. Compute the length of each gene and print the name of each gene and its corresponding length, seperated by a space, to a new file.

e) Write a script that runs the UNIX command ls to get a list the files in the current directory, then prints out all upper case versions of all the file names.

f) The file RNAseq_Mus_foldChange.txt contains the result of a differential gene expression analysis between males and females from a laboratory strain of mouse (*Mus musculus*).
   i. Print out gene IDs for genes located on mitochondrion (MT).
   ii. Print out rows of genes significantly differential expressed (padj < 0.05), a minimum log2-fold-change of 5 and located on the Y chromosome.

Solutions:
You can find solutions to all exercises within the downloaded folder (solutions.py).

Sources:

- http://biopython.org/DIST/docs/tutorial/Tutorial.html
- http://pycam.github.io/