Homework Exercise 4

Use Biopython whenever possible (as opposed to writing your own solution). This applies in general to future exercises – always prefer using the newest tools we've learned to "rolling your own" solution. More often than not, there's already a function that does what you want.

1)

In protein-coding human transcripts, what are the average relative lengths of the 5' UTR, 3' UTR and coding regions? What is the average GC content of each of these regions?

To get high quality results, make sure you look only at reviewed transcripts.

2)

Extract the sequences of all introns on chromosome 11. What is the frequency of each pair of letters at the beginning and end of these introns?

Important: Pay attention to whether each gene is on the positive or negative strand.

NOTE: There are some online tools (e.g. UCSC) that can provide the sequences of selected introns, but for the sake of this exercise, **we want you to extract the sequences from the reference genome by yourself**.

What is the significance of the sequences near the splice junctions? Are your results in agreement with textbooks?

Bonus Questions

1)

Explain function decorators in Python.

Suppose that you have many different functions that get DNA sequences as inputs (other than that, each of them does something completely different). You want to modify all of these functions to also accept RNA sequences. Demonstrate how it can be done.

2)

How do computers generate random numbers? What does Python's ***random.seed*** function do and why is it useful?