Exercises – DNA

Introduction

RNA is a nucleic acid, consisting of many nucleotides. In RNA there are only four types of nucleotides called adenine (coded with a capital **A**), guanine (**G**), cytosine (**C**) and uracil (**U**).

Three nucleotides together are called a codon. Each of these triplets of nucleotides in a nucleic acid sequence specifies (codes for) a single amino acid. The amino acids are chained together to form proteins.

For instance, the RNA sequence CGACUCUGACUG contains the four codons CGA, CUC, UGA and CUG. The codon CGA codes for the amino acid Arginine. The codon CUC codes for Leucine. And so on.

Because every codon contains a triplet of four possible nucleotides there are 4^3=64 different types of codons. Some codons code for the same amino acid. For instance the codons UUA, UUG, CUU, CUC, CUA and CUG all code for the same amino acid Leucine. In total, there are only 20 different types of amino acids.

In addition to the codes for amino acids there are a number of special codons that indicate the beginning and the end of a piece of RNA sequence that translates into a specific protein. These codons are called START and STOP codons. There is only one type of START codon (AUG), but there are three types of STOP codons (UAG, UGA and UAA).

An RNA sequence is represented by a Python string called rna_string.

The string rna_string is initialized as follows:

```
rna_string = 'CUUCGGAUGAAGCUGUGGGCAAGUUGGGAUGAAUCGUGAUGGGUC'
```

Exercise 1 - Expressions

a) give an expression for the number of nucleotides in the RNA string rna_string.

b) give an expression for the number of codons in an RNA string rna_string.

c) give an expression that gives the n-th codon from an RNA string rna_string.

Exercise 2 – Iteration with a while loop

We want to print the codons on a separate line. Fill in the missing expressions in the following while loop:

```
i = ...
while i < ...           :
   print ...
   i = ...
```

Exercise 3 – represent RNA sequences as Python lists

The format of the RNA sequence has to be changed. We need the RNA sequence in the form of a list of codons to look like this:

rna_list = ['CUU', 'CGG', 'AUG', 'AAG', 'CUG', 'UGG', 'GCA', ...]

a) What is the data type of the elements of the list rna_list?

b) Fill in the missing expressions in the following while loop that converts the RNA string rna_string into the RNA list called rna_list:

```
rna_list = ...
i = ...
while i < ...                  :
   rna_list = rna_list +  ...
   i = ...
```

Exercise 4 - Expressions

a) give an expression for the number of codons in the RNA list rna_list.

b) give an expression that gives the n-th codon from the RNA list rna_list.

c) create a function for a) and b)

Exercise 5 – finding START and STOP codons

a) give an expression that gives the index of the first START codon in an RNA list called rna_list. Call this index start_idx.

b) give the expression or statement(s) that determine the index of the first STOP codon in the RNA list called rna_list. Call this index stop_idx.

Exercise 6 – find

In some sequences there may be STOP codons *before* the first START codon. We want to make sure that the STOP codon we found under 5b) is a STOP codon that comes *after* a START codon.

a) Give the Boolean expression that gives True as a result if the STOP codon with index stop_idx comes after the START codon with index start_idx.

b) Give an expression that determines the index of the first STOP codon *after* the START codon with index start_idx.

c) create a function for finding a specific codon, starting at a given index.

Exercise 7 – slices

We want to make a slice of the RNA list rna_list that includes the START codon with index start_idx and the STOP codon with index stop_idx and all the codons in between.

a) Give the expression that gives the slice of rna_list from index start_idx up to and including stop_idx.

b) create a function that finds the next sequence that starts with a START codon and ends with a STOP codon, starting at a given position.

Exercise 8 – Iteration with a for loop

The codon UGG codes for the amino acid Tryptophan. We are interested in the number of UGG codons in the RNA sequence called rna_list.

a) Fill in the expressions on the dots in the following statements to determine how many codons of the type UGG there are in the RNA sequence rna_string.

```
count = ...
for codon in rna_list:
  if ...    == ...    :
    count = ...
print count
```

b) create a function that counts the number of codons for a specific amino acid in an RNA sequence.

Challenge

Write a program that takes any RNA sequence, cuts it into parts that start with a START codon and end with a STOP codon, and translates the codons into amino acids.

You can generate a random RNA sequence with the following function.

```
import random

# list of nucleotides
RNA_NUCLEOTIDES = ('A', 'C', 'G', 'U')

# This function generates a random sequence of n nucleotides
# The result is returned in a (potentially long) string
def random_sequence(n):
    result = ''
    while n > 0:
        result = result + random.choice(RNA_NUCLEOTIDES)
        n = n - 1
    return result
```

See also:

http://en.wikipedia.org/wiki/Codon