CS 8, Winter 2015
Homework Assignment #? (draft)

Assignment Overview

This assignment is for you to practice using basic control structures such as loop and conditional statements and data structures such as character string.

Background Assignment Specifications

DNA (deoxyribonucleic acid) is made up of molecules called nucleotides. Each nucleotide contains a phosphate group, a sugar group and a nitrogen base. The four types of nitrogen bases are adenine (A), thymine (T), guanine (G) and cytosine (C). The order of these bases is what determines DNA's instructions, or genetic code.

DNA encodes genes. A gene is the basic physical and functional unit of heredity. Genes, which are made up of DNA, act as instructions to make molecules called proteins. In humans, genes vary in size from a few hundred DNA bases to more than 2 million bases.

Most genes contain the information needed to make functional molecules called proteins. (A few genes produce other molecules that help the cell assemble proteins.) The journey from gene to protein is complex and tightly controlled within each cell. It consists of two major steps: transcription and translation. Together, transcription and translation are known as gene expression.

During the process of transcription, the information stored in a gene's DNA is transferred to a similar molecule called RNA (ribonucleic acid) in the cell nucleus. Both RNA and DNA are made up of a chain of nucleotide bases, but they have slightly different chemical properties. The type of RNA that contains the information for making a protein is called messenger RNA (mRNA) because it carries the information, or message, from the DNA out of the nucleus into the cytoplasm.

Translation, the second step in getting from a gene to a protein, takes place in the cytoplasm. The mRNA interacts with a specialized complex called a ribosome, which "reads" the sequence of mRNA bases. Each sequence of three bases, called a codon, usually codes for one particular amino acid. (Amino acids are the building blocks of proteins.) A type of RNA called transfer RNA (tRNA) assembles the protein, one amino acid at a time. Protein assembly continues until the ribosome encounters a "stop" codon (a sequence of three bases that does not code for an amino acid).

Assignment Specifications

We will simulate the DNA to protein transcription and translation processes in this homework. **Caveat**: Be warned that this is not a course in biology. The transcription and translation processes that we simulate in this assignment can be quite complex and our simulation can be biologically incorrect, but is sufficient for a Python programming assignment.

The transcription process "reads" a DNA sequence and transcribe it into an mRNA sequence of the same length. RNA is again made of four nucleotides, adenine (A), cytosine (C), guanine (G), or uracil (U). The transcription process replaces C with G, G with C, A with U, and T with A (no T in RNA).

The translation process "reads" an mRNA sequence and search for a particular 3-nucleotide sequence (AUG) that corresponds to a "start" codon. Once the start codon is located, every subsequent 3-nucleotide subsequence is used to transcribe a protein. This translation process ends when a "stop" codon is encountered.

Your program must comprise (at least) the following function definitions:

- DNA2Protein(ifname, cfname)
  This is the entry point of your program. It takes two files: The first is a file containing the input DNA sequence and the second is a file containing the mRNA to protein translation table. All parameters must be specified.

- DNA2mRNA(ifname,ofname="mRNA.txt")
  This is a function that transcribes between a DNA sequence and an mRNA sequence. It takes two files: The first is a file that contains the input DNA sequence and the second is an output file that contains the transcribed mRNA sequence. The mRNA sequence must be written out in the same format as the input DNA sequence.

- mRNA2Protein(cfname,ifname="mRNA.txt")
  This is a function that translates between an input mRNA sequence and the resulting proteins. It takes two arguments: the first is the mapping table, and the optional second argument is the file storing the input mRNA sequence. The protein is written out on the console.

Assignment Deliverables

The deliverable for this assignment is the following file:

dna.py – the source code for your Python program

Be sure to use the specified file name and submit it for grading via the **turnin** system before the project deadline.

Assignment Notes

1. DNA files have a standard recording format: Each line contains 60 nucleotides assembled in six groups of 10 each with a space in between each group. The first entry in each line is the sequence number of the first nucleotide in the line (1 for first line, 61 for second line, 121 for third line, etc.).

2. For each protein assembled, you must give the information on the starting and end nucleotide positions, length of the protein, and the protein sequence itself. Note that the starting and end nucleotide positions are according to the DNA recoding format (the first nucleotide is position 1), not Python array index (the first entry is 0)!

3. DNA symbols are case insensitive, or A and a mean the same thing.

4. Your program needs to handle proper end condition: If protein transcription ends before a stop codon is encountered (end of file error), you should output the protein sequence generated so far and print out a warning message.

Sample Outputs

```
>>> dna.DNA2Protein('dna.txt', 'rna_codon.txt')
protein found from base   15 to   20 of length    2: LP
protein found from base   96 to  161 of length   22: FDFARHQSRRRLRRLQDDSHLL
protein found from base  223 to  279 of length   19: ELLLFGGVTVIILIFVLRF
protein found from base  396 to  503 of length   36: SSGQSSYIIMGSIHTNFYRRGVGVSRNGSQREETAH
protein found from base  546 to  551 of length    2: RV
protein found from base  609 to  665 of length   19: NYLFNIEGALLKDEGCRCI
protein found from base  686 to  706 of length    7: VLCRSLK
protein found from base  725 to  727 of length    1: I
protein found from base  736 to  738 of length    1: R
protein found from base  789 to  827 of length   13: GGHRSQLLSKCKV
protein found from base  842 to  847 of length    2: DI
protein found from base  882 to 1073 of length   64: LTKLNGSTESKLRSRSCKSPLGRRLNDRLRLWCNIKLHYELPCLRRLSCRNLLCMVKQQCLAGR
protein found from base 1130 to 1150 of length    7: IAVFARL
protein found from base 1230 to 1298 of length   23: PARVNILRGNGLTDKKLRPLNFK
protein found from base 1350 to 1358 of length    3: SKQ
protein found from base 1370 to 1381 of length    4: SVTS
protein found from base 1439 to 1546 of length   36: EISFIKLLVAMTVSIAKYTEWRFDTNRATARIKKTI
protein found from base 1604 to 1627 of length    8: VKAQTGST
protein found from base 1682 to 1837 of length   52: NTTTLNKVEASTEVLPKQTVIKRRVIIAMFPTYQEDDKKRKSQMSDALMFAK
protein found from base 1850 to 1879 of length   10: IKFGSGTDPL
protein found from base 2088 to 2096 of length    3: CRR
protein found from base 2123 to 2179 of length   19: EATTMKKTRSSRRSSVILK
protein found from base 2311 to 2331 of length    7: AYDNHLD
protein found from base 2375 to 2437 of length   21: WKLVGETTTTKEHATMKLSFF
protein found from base 2540 to 2557 of length    6: YVTSQG
protein found from base 2599 to 2772 of length   58: SEVSRARNWVSYPEEDTYTYHLVVIYLGPLYDRWTVVRDYNSLCQCLVFFDNYVFLKS
protein found from base 2804 to 2827 of length    8: QFPTVIQK
protein found from base 2926 to 2937 of length    4: FYKF
protein found from base 3165 to 3182 of length    6: RCYKTN
protein found from base 3221 to 3259 of length   13: SMGYKINQKYESL
protein found from base 4008 to 4031 of length    8: GVGEEMGR
protein found from base 4058 to 4192 of length   45: IKKKEEPRVSSVFAYGRKRLEKKSKDLFKYNQFVPDQNHTFDHQS
protein found from base 4277 to 4285 of length    3: HEN
protein found from base 4312 to 4455 of length   48: YDKKPFPLSYYIFDLILPALFLSESSRKKVFRGLLLLILLFLNGRQSL
protein found from base 4679 to 4801 of length   41: GKRVSSLAGESQIGNEGSQTEVTEDRQQDLGPKRTTSRVII
stop codon not found, the last protein transcription may be incomplete
```