Central Dogma Script

A python script created by Georgios I. Iatropoulos

City: Piraeus, Country: Greece

Video Demo: https://youtu.be/aTOSJQ3SC_c

Description

The Central Dogma script is a console application crafted to handle the input of a coding DNA sequence, *without introns (non coding regions)*, in plain text format, along with its orientation, according to the principles of the Central Dogma of Biology. It outputs the corresponding complementary DNA sequence, the equivalent mRNA sequence, and, if applicable, the corresponding polypeptide sequence. In addition, the script calculates the number of phosphodiester and hydrogen bonds, as well as the GC content for each DNA sequence.

Usage and Input

`python3 project.py [data.file]`

The **Central Dogma script** can be executed with an optional data file in plain text format. Each line of the data file should contain a DNA sequence written in the 5' to 3' orientation.

1. If a data file is not provided, the script will prompt the user for manual input. Initially, it will request a coding DNA sequence, followed by its orientation. After processing and presenting the output, the script will inquire whether the user wishes to continue with another sequence. If not, the application will exit."

2. If a data file is given as a command line argument, the script will interpret each line as a coding DNA sequence and automatically assign the 5' to 3' orientation. After processing all lines in the data file, the program will exit.

Script files

1. **`project.py`**: It contains the **main()** function, and the rest of the custom functions that compose the procedural part of the script.
2. **`bio.py`**: It contains the `translation_table` dictionary and the `DNA_obj` custom class, which is imported in `project.py`.

3. `test_project.py`: It contains the unit tests for the functions (procedural part) and the methods (object oriented part) of the script.

Input file

`data.txt`: It contains some DNA sequences, in plain text format, for script testing.

Purpose

The **Central Dogma of Biology** describes how Genetic Information is inherited to the next generations of cells and how it is manipulated and decoded for polypeptide synthesis.

The purpose of the **Central Dogma script** is emulating the Central Dogma of Biology procedures, aiding ***secondary education biology students*** in understanding and applying their concepts. For example, one of the primary objectives is to help students acquire how to read the **Genetic Code Table**. It will also assist ***biology teachers*** in writing their own practical exercises. These procedures are:

1. **Replication**
2. **Transcription**
3. **Translation**

1. Replication

Replication creates the complementary and antiparallel DNA strand of the coding DNA strand, following **Chargaff's rule**. Its purpose is to create two copies of a DNA molecule, that they will be inherited by the daughter cells of a parent cell.

DNA strands are composed of subunits that are called **deoxyribonucleotides**. Deoxyribonucleotides are linked together by **phosphodiester bonds**, forming a long strand. **DNA polymerase** is the enzyme responsible for synthesizing the **daughter** DNA strands. It achieves this by pairing deoxyribonucleotides with the complementary deoxyribonucleotides of the **parent** (old) strand in an antiparallel orientation.

Each deoxyribonucleotide can have one of the following nitrogenous bases:

Table 1A: DNA Nitrogenous bases

	Name	**Category**
1.	Adenine (A)	Purine
2.	Thymine (T)	Pyrimidine
3.	Guanine (G)	Purine
4.	Cytosine (C)	Pyrimidine

According to Chargaff's rule a Purine can pair only with a Pyrimidine. Specifically:

1. A pairs with T via 2 Hydrogen bonds.
2. G pairs with C via 3 Hydrogen bonds.

The formation of **double stranded DNA** occurs, when two DNA strands pair. Pairing prerequisites for the pairing strands are that they consist of nucleotides whose nitrogenous bases sequences are **complementary**, and their orientations are **antiparallel**.

Example 1

For the input DNA starnd sequence:

5' ATGGAGCTCTAA 3'

The complementary and antiparallel DNA strand sequence, after replication, will be:

3' TACCTCGAGATT 5'

Finally, the double-stranded DNA will be:

5' ATGGAGCTCTAA 3'
3' TACCTCGAGATT 5'

2. Transcription

Transcription produces a **messenger RNA** molecule (**mRNA**), which has the same nitrogenous bases sequence as the **coding DNA strand** in the same orientation. **RNA strands** are made of **ribonucleotides**, which can have the same nitrogenous bases as those in Table 1A, except that **Thymine (T)** is replaced by **Uracil (U)**. mRNAs are expendable molecules that will be translated into polypetide chains in the cytoplasm. mRNA molecules are suitable for this operation, because mRNAs carry the information of a gene that is currently needed, and cells can produce, or destroy them, depending on various environmental factors, or signals. DNA molecules serve only as the library that always keep safe the total genetic information of the cell during its life cycle.

Table 1B: RNA Nitrogenous bases

	Name	**Category**
1.	Adenine (A)	Purine
2.	Uracil (U)	Pyrimidine
3.	Guanine (G)	Purine
4.	Cytosine (C)	Pyrimidine

RNA polymerase is the enzyme that synthesizes the mRNA strands by pairing **ribonucleotides** to the complementary **deoxyribonucleotides** of the non coding strand in an antiparallel orientation.

According to Chargaff's rule the complementary pairings are:

1. A pairs with U via 2 Hydrogen bonds.
2. A pairs with T via 2 Hydrogen bonds.
3. G pairs with C via 3 Hydrogen bonds.

Example 2

For the DNA coding sequence:

5' ATGGAGCTCTAA 3'

the mRNA will be:

5' AUGGAGCUCUAA 3'

The script accepts that introns does not exist in the DNA sequence.

3. Translation

Proteins are the operational powerhouses of the cells, as well as some of their most important building blocks. Usually, proteins are made up of one or more than one polypeptide chains.

Translation connects **amino acids** forming a **polypeptide chain**, using the information that is stored in **DNA coding regions**. In living organisms 20 different amino acids exist, and polypeptides are formed by various combinations of these amino acids. The sequence in a polypeptide chain is not random, and occurs when cells' structures or organelles called **ribosomes**, read the nitrogenous bases sequence in the mRNA molecule and build the corresponding polypeptide chains accordingly.

Ribosomes are enzyme complexes that "read" the mRNA sequence from 5' to 3' orientation, codon by codon in a continuous fashion. **Codons** are groups of 3 ribonucleotides, thus 3 nitrogenous bases, starting at the **starting codon** 5' AUG 3' (5' ATG 3' for the DNA coding sequence), and ending at the termination codons 5' UGA 3', or 5' UAA 3', or 5' UAG 3' (5' TGA 3', or 5' TAA 3', or 5' TAG 3' for the DNA coding sequence). Then ribosomes, whith the help of special RNA molecules called **transfer RNA** (**tRNA**) "translate" those codons to amino acids, according to the **genetic code**, except the termination codons. The first amino acid of the polypeptide sequence, **Methionine**, often serving as the initiator amino acid, corresponds to the 5' AUG 3' starting codon and it is located at the **amino (N-terminus) end (H_2N-)**, while the last amino acid is located at the **carboxyl (C-terminus) end (-COOH)** of the polypeptide chain.

Thus, the starting and ending codons define the reading frame of a coding sequence, whose length must be a multiple of three. Central Dogma script checks if a coding DNA sequence starts with 5' ATG 3' and ends with a termination codon, while reading the coding sequence by steps of 3 bases (codons). If the coding sequence fulfills the criteria above, it will be translated according to the **Standard Genetic Code Table**. If not, the sequence will not be translated.

Example 3

For the mRNA of the previous example:

5' AUGGAGCUCUAA 3'

the polypeptide chain will be:

H_2N-Met-Glu-Leu-COOH

Example 4: Summary of Central Dogma

Table 2: DNA Coding Sequence Replication, Transcription and Translation (Central Dogma)

Procedure	Macromolecule	Starting Edge	Starting Codon	Codon 1	Codon 2	Termination Codon	Ending Edge
	DNA coding	5'	ATG	GAG	CTC	TAA	3'
Replication	DNA complementary	3'	TAC	CTC	GAG	ATT	5'
Transcription	mRNA	5'	AUG	GAG	CUC	UAA	3'
Translation	Polypeptide	H_2N	Met	Glu	Leu	-	COOH

Various Calculations

The Central Dogma script expands its operation in calculating the number of:

1. The number of ***hydrogen bonds*** between the nitrogenous bases of the DNA complementary strands.
2. The number of ***phosphodiester bonds*** between the nucleotides of a linear double stranded DNA molecule.

Finally, the script calculates the ***GC content*** for each double stranded DNA molecule. This metric evaluates the magnitude of the ***Melting Temperature*** of the double stranded DNA molecule.

Output files and on screen console output

When script is used with a DNA sequences data file

1. "***dogma.txt***" file: Contains the results of Replication, Transcription and Translation of the DNA sequences in the data file. It also contains the phosphodiester and hydrogen bonds calculations, plus the GC content for each DNA sequence. (see the "Various Calculations" paragraph)
2. "***peptides.csv***" file: Contains a comma separated table that corresponds each DNA sequence to its peptide.
3. "***peptides_tab.txt***" Contains the same output as the peptides.csv file, but in a tabular format for better readability.
4. On screen warning for overwriting the previous data files.
5. On screen notification that the data file exists.
6. On screen notifications about the names and the contents of the output files.

When script is used without a DNA sequences data file (manual input mode)

1. "***dogma_manual.txt***" file: Contains the results of Replication, Transcription and Translation of the DNA sequences that are input manually, by using the script's terminal text interface. It also contains the various calculations for each DNA sequence. Manual input mode is designed for experimenting with only one or few DNA sequences. It doesn't create the "***peptides.csv***" and "***peptides_tab.txt***" files which provide an overview of DNA sequences that can be translated and the corresponding peptides.
2. On screen input notifications.
3. On screen output for each DNA sequence's Replication, Transcription and Translation products. In addition, it contains the number of phosphodiester and hydrogen bonds, plus the GC content calculations.
4. On screen notification about the name and the contents of the output file.

Design

Central Dogma Script uses a dictionary that stores the ***Standard Genetic Code Table*** and a custom class called **`DNA_obj`**. The "**bio.py**" file contains both data structures, because the script in its current state uses only the Standard Genetic Code Table. However, for utilizing more Genetic Code Tables than just the Standard, it would be better if the corresponding dictionaries were written in a separate file.

This class represents a single-stranded DNA molecule and must be initialized with user-input values for the following two **attributes**:

1. ***seq***: The nitrogenous bases sequence of the coding DNA strand.
2. ***direction***: The orientation of the coding DNA strand ("1" for 5'->3' or "2" for 3'->5').

In case that ***direction*** is "2", the class method ***set_seq_dir(self)*** reverses the orientation of the DNA sequence in ***seq***.

The rest of the **`DNA_obj`** **class attributes** are calculated by the corresponding **class methods**. These are:

1. ***`complementary`***: It is calculated by the ***`comp(self)`*** class function, and holds the complementary and antiparallel DNA strand.
2. ***`rna_seq`***: It is calculated by the ***`transcription(self)`*** class function, and holds the mRNA sequence which corresponds to the coding DNA sequence.
3. ***`peptide_seq`***: It is calculated by the ***`translation(self)`*** class function, and holds the amino acid sequence of the peptide chain that corresponds to the given coding DNA sequence.

`DNA_obj` class also uses a function named `justify_nucleotide_seq(self, seq)` to format the DNA and mRNA sequences as triplets of nitrogenous bases (codons) using dashes as separators. Thus, it is shown in the formatted DNA and mRNA sequences which codon is translated to which amino acid, and the final output has better readability.

The script also consists of a procedural part that contains a series of functions. Below a **functions** overview is given:

1. **`set_coding_sequence(c_seq, direct)`**: This function creates a DNA object using the DNA_obj class, and executes the in class functions in order to calculate, or modify, all the class attributes. Its parameters are: ***c_seq*** the nucleotide sequence and ***direct*** the orientation of the sequence.
2. **`create_dogma_file(out_filename, d_object, write_mode)`**: This function creates the "***dogma.txt***" output file when a data file is provided or the "***dogma_manual.txt***" output file, when the script is used for manual input. It has 3 parameters: ***out_filename*** accepts the name of the output file, ***d_object*** holds a ***DNA_obj*** object and ***write_mode*** is used to define the writing mode for the output file.
3. **`create_csv_dogma_files(in_filename, out_filename1, out_filename2)`**: This function creates the ***peptides.csv*** (*out_filename1*) output file and the ***dogma.txt*** (*out_filename2*) output file, when multiple DNA sequences are given via a ***data input*** (*in_filename*) file. This function utilizes the `create_dogma_file` function for writing in the ***dogma.txt*** file. This design choice was made because the two output files are simultaneously created in the same loop, thus using the `create_dogma_file` function improves the readability and simplifies the code in the `create_csv_dogma_files` function.
4. **`create_tab_table(in_filename, out_filename)`**: This function creates the ***peptides_tab.txt*** (*out_filename*) output file. It utilizes the `create_table_list(file)` function that creates a **list of rows** from the lines of ***peptides.csv*** output file, which is now used as input (*in_filename*). This list is used as input by the `tabulate` function from the `tabulate` library. `tabulate` function is also utilized by the `create_tab_table` function, in order to write the tabular formatted file ***peptides_tab.txt***.
5. **`count_bases(dna_seq)`**: This function takes a single stranded DNA sequence as input (*dna_seq*) and counts its nitrogenous bases. It returns a dictionary that contains the nitrogenous bases symbols as keys and their corresponding number as values.
6. **`compute_hydro_bonds(b_dict)`**: This function takes as input (*b_dict*) the dictionary that is returned form the `count_bases` function and returns the **number of hydrogen bonds** in the double stranded DNA molecule.
7. **`compute_gc_content`**: This function takes as input (*b_dict*) the dictionary that is returned from the `count_bases` function and returns the **GC content ratio** of the single or double stranded DNA molecule, which are equal.
8. **`compute_phospho_bonds(dna_seq)`**: This function gets a DNA sequence as input and outputs the **number of phosphodiester bonds** for the corresponding linear double stranded molecule.

Contact

For any questions, or suggestions about the Central Dogma script, contact me at: iatropoul@gmail.com