Central Dogma Script

Created by **Georgios I. Iatropoulos**
City: **Piraeus**
Country: **Greece**

A project for Harvard's CS50 python 2023

Description

- The Central Dogma script is a console application crafted to handle the input of a DNA sequence 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 the double stranded molecule of each DNA sequence.

Usage

```
python3 project.py [data.file]
```

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.

Purpose

- Emulating the Central Dogma of Biology (CDB) procedures.
- Aiding ***secondary education biology students*** in understanding and applying the CDB procedural concepts.
- Assisting ***biology teachers*** in writing their own practical exercises.

Operation modes

1. Manual input
2. Data file input

Manual mode output

1. On Screen
2. Output file **dogma_manual.txt**

Input is a Coding DNA Sequence

```
PS C:\Users\Georgios\iCloudDrive\Programming_Projects\central_dogma> python project.py
Existing output files may be overwritten.
Continue? (Y/N)y
If you want to load a data file for multiple coding sequences run this script as:
python3 central_dogma.py /path/name_of_file.extension
Enter your coding DNA sequence:
ATGCTACAATTATAG
Enter the direction of your sequence:
1 for 5->3
2 for 3->5
Enter your direction: 1
DNA Coding Strand:         5' ATG-CTA-CAA-TTA-TAG 3'
DNA Complementary Strand:  3' TAC-GAT-GTT-AAT-ATC 5'
mRNA sequence:             5' AUG-CUA-CAA-UUA-UAG 3'
Peptide sequence:         H₂N-Met-Leu-Gln-Leu-COOH
Hydrogen bonds = 34
Phosphodiester bonds = 28
GC content = 0.26666666666666666
The output was added to the dogma_manual.txt file.
Do you want to continue? (Y/N)y
```

Input is a Non Coding DNA Sequence - Inverted Orientation

```
Enter your coding DNA sequence:
ATGCCCATAGCGTGA
Enter the direction of your sequence:
1 for 5->3
2 for 3->5
Enter your direction: 2
DNA Coding Strand:          5' AGT-GCG-ATA-CCC-GTA 3'
DNA Complementary Strand:   3' TCA-CGC-TAT-GGG-CAT 5'
mRNA sequence:              5' AGU-GCG-AUA-CCC-GUA 3'
Peptide sequence:          H₂N-Non Coding-COOH
Hydrogen bonds = 38
Phosphodiester bonds = 28
GC content = 0.5333333333333333
The output was added to the dogma_manual.txt file.
Do you want to continue? (Y/N)n
PS C:\Users\Georgios\iCloudDrive\Programming_Projects\central_dogma>
```

dogma_manual.txt output file sample

dogma_manual.txt

C: > Users > Georgios > iCloudDrive > Programming_Projects > central_dogma > dogma_manual.txt

```
DNA Coding Strand:          5' ATG-CTA-CAA-TTA-TAG 3'
DNA Complementary Strand:   3' TAC-GAT-GTT-AAT-ATC 5'
mRNA sequence:              5' AUG-CUA-CAA-UUA-UAG 3'
Peptide sequence:          H₂N-Met-Leu-Gln-Leu-COOH
Hydrogen bonds = 34
Phosphodiester bonds = 28
gc content = 0.26666666666666666

DNA Coding Strand:          5' AGT-GCG-ATA-CCC-GTA 3'
DNA Complementary Strand:   3' TCA-CGC-TAT-GGG-CAT 5'
mRNA sequence:              5' AGU-GCG-AUA-CCC-GUA 3'
Peptide sequence:          H₂N-Non Coding-COOH
Hydrogen bonds = 38
Phosphodiester bonds = 28
gc content = 0.5333333333333333
```

Invalid DNA sequence input

```
Windows PowerShell
PS C:\Users\Georgios\iCloudDrive\Programming_Projects\central_dogma> python project.py
Existing output files may be overwritten.
Continue? (Y/N)y
If you want to load a data file for multiple coding sequences run this script as:
python3 central_dogma.py /path/name_of_file.extension
Enter your coding DNA sequence:
atgfgtagatga3agatagacg
Enter the direction of your sequence:
1 for 5->3
2 for 3->5
Enter your direction: 2
Traceback (most recent call last):
  File "C:\Users\Georgios\iCloudDrive\Programming_Projects\central_dogma\project.py", line 175, in <module>
    main()
  File "C:\Users\Georgios\iCloudDrive\Programming_Projects\central_dogma\project.py", line 60, in main
    dna_object = set_coding_sequence(d_seq, d)
                 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "C:\Users\Georgios\iCloudDrive\Programming_Projects\central_dogma\project.py", line 76, in set_coding_sequence
    dna_obj = DNA_obj(c_seq, direct)
              ^^^^^^^^^^^^^^^^^^^^^^
  File "C:\Users\Georgios\iCloudDrive\Programming_Projects\central_dogma\bio.py", line 73, in __init__
    self.seq = seq.strip().upper()
    ^^^^^^^^
  File "C:\Users\Georgios\iCloudDrive\Programming_Projects\central_dogma\bio.py", line 92, in seq
    raise ValueError("Invalid Sequence!")
ValueError: Invalid Sequence!
PS C:\Users\Georgios\iCloudDrive\Programming_Projects\central_dogma>
```

Data file input

data.txt input file sample

data.txt

C: > Users > Georgios > iCloudDrive > Programming_Projects > central_dogma > data.txt

```
atgagtcgccacgaggggtga
agtcccacagcttgtcgctag
atgcgagagtgtgcaccacactattattaa
```

Data file mode output

1. **On Screen**
2. **Output files:**
   - dogma.txt
   - peptides.csv
   - peptides_tab.txt

Invalid data filename output

```
PS C:\Users\Georgios\iCloudDrive\Programming_Projects\central_dogma> python project.py wrong_filename.txt
Existing output files may be overwritten.
Continue? (Y/N)y
Coding Sequence file not found.
Check that the path and the name of your sequence file are correct,
or run the program without arguments and input a coding DNA sequence manually.
PS C:\Users\Georgios\iCloudDrive\Programming_Projects\central_dogma>
```

Valid data filename output

```
PS C:\Users\Georgios\iCloudDrive\Programming_Projects\central_dogma> python project.py data.txt
Existing output files may be overwritten.
Continue? (Y/N)y
Your data file exists!
Read the dogma.txt file for Replication, Transcription and Translation data., plus various calculations for each DNA sequence.
Your translation data is written in the peptides.csv file.
For tabular format the peptides_tab.txt was created.
PS C:\Users\Georgios\iCloudDrive\Programming_Projects\central_dogma>
```

dogma.txt output file sample

dogma.txt

C: > Users > Georgios > iCloudDrive > Programming_Projects > central_dogma > dogma.txt

```
DNA Coding Strand:          5' ATG-AGT-CGC-CAC-GAG-GGG-TGA 3'
DNA Complementary Strand:   3' TAC-TCA-GCG-GTG-CTC-CCC-ACT 5'
mRNA sequence:              5' AUG-AGU-CGC-CAC-GAG-GGG-UGA 3'
Peptide sequence:          H₂N-Met-Ser-Arg-His-Glu-Gly-COOH
Hydrogen bonds = 55
Phosphodiester bonds = 40
gc content = 0.6190476190476191

DNA Coding Strand:          5' AGT-CCC-ACA-GCT-TGT-CGC-TAG 3'
DNA Complementary Strand:   3' TCA-GGG-TGT-CGA-ACA-GCG-ATC 5'
mRNA sequence:              5' AGU-CCC-ACA-GCU-UGU-CGC-UAG 3'
Peptide sequence:          H₂N-Non Coding-COOH
Hydrogen bonds = 54
Phosphodiester bonds = 40
gc content = 0.5714285714285714

DNA Coding Strand:          5' ATG-CGA-GAG-TGT-GCA-CCA-CAC-TAT-TAT-TAA 3'
DNA Complementary Strand:   3' TAC-GCT-CTC-ACA-CGT-GGT-GTG-ATA-ATA-ATT 5'
mRNA sequence:              5' AUG-CGA-GAG-UGU-GCA-CCA-CAC-UAU-UAU-UAA 3'
Peptide sequence:          H₂N-Met-Arg-Glu-Cys-Ala-Pro-His-Tyr-Tyr-COOH
Hydrogen bonds = 72
Phosphodiester bonds = 58
gc content = 0.4
```

peptides.csv output file sample

peptides.csv

C: > Users > Georgios > iCloudDrive > Programming_Projects > central_dogma > peptides.csv

```csv
DNA Coding Sequence 5' -> 3',Peptide Sequence H₂N -> COOH
ATG-AGT-CGC-CAC-GAG-GGG-TGA,Met-Ser-Arg-His-Glu-Gly
AGT-CCC-ACA-GCT-TGT-CGC-TAG,Non Coding
ATG-CGA-GAG-TGT-GCA-CCA-CAC-TAT-TAT-TAA,Met-Arg-Glu-Cys-Ala-Pro-His-Tyr-Tyr
```

peptides_tab.txt output file sample

peptides_tab.txt

C: > Users > Georgios > iCloudDrive > Programming_Projects > central_dogma > peptides_tab.txt

```
+----------------------------------------+-------------------------------------+
| DNA Coding Sequence 5' -> 3'           | Peptide Sequence H₂N -> COOH        |
+========================================+=====================================+
| ATG-AGT-CGC-CAC-GAG-GGG-TGA            | Met-Ser-Arg-His-Glu-Gly             |
+----------------------------------------+-------------------------------------+
| AGT-CCC-ACA-GCT-TGT-CGC-TAG            | Non Coding                          |
+----------------------------------------+-------------------------------------+
| ATG-CGA-GAG-TGT-GCA-CCA-CAC-TAT-TAT-TAA | Met-Arg-Glu-Cys-Ala-Pro-His-Tyr-Tyr |
+----------------------------------------+-------------------------------------+
```

Central Dogma of Biology Summary Table (Central Dogma Overview)

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

Future plans

- Adding non standard genetic tables.
- Introducing a new feature for translating a DNA coding sequence with 2 different genetic tables. Then the script will align the 2 corresponding peptides, counting their differences in amino acids.

The End :)

Central Dogma of Biology

Procedures

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

Replication

1. **Purpose**: Replication produces two identical copies of a DNA double-stranded molecule, so that they can be inherited by the two daughter cells of a parent cell.
2. **Procedure**: Each strand serves as a template for the synthesis of its complementary strand in an antiparallel orientation, following Chargaff's rule.

Transcription

1. **Purpose**: Transcription produces a messenger RNA (mRNA) copy of the gene's DNA coding strand. mRNAs are expendable molecules that will be translated into polypetide chains in the cytoplasm.
2. **Procedure**: The non-coding DNA strand of the gene serves as a template for the synthesis of mRNA, according to Chargaff's rule.

Translation

1. **Purpose**: Translation connects amino acids, forming a polypeptide chain, using the information that is stored in DNA coding regions.
2. **Procedure**: Ribosomes facilitate the formation of the peptide bonds between amino acids. Transfer RNAs (tRNAs) read the mRNA's sequence, 3 nitrogenous bases (codons) at a time. They pair their anti-codon with the next available codon, ensuring the proper positioning of the corresponding amino acid. The Codon - Anti-codon pairing, in anti-parallel orientation, follows the Chargaff's rule.

DNA Nitrogenous Bases Table

Table 1A: DNA Nitrogenous bases

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

RNA Nitrogenous Bases Table

Table 1B: RNA Nitrogenous bases

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

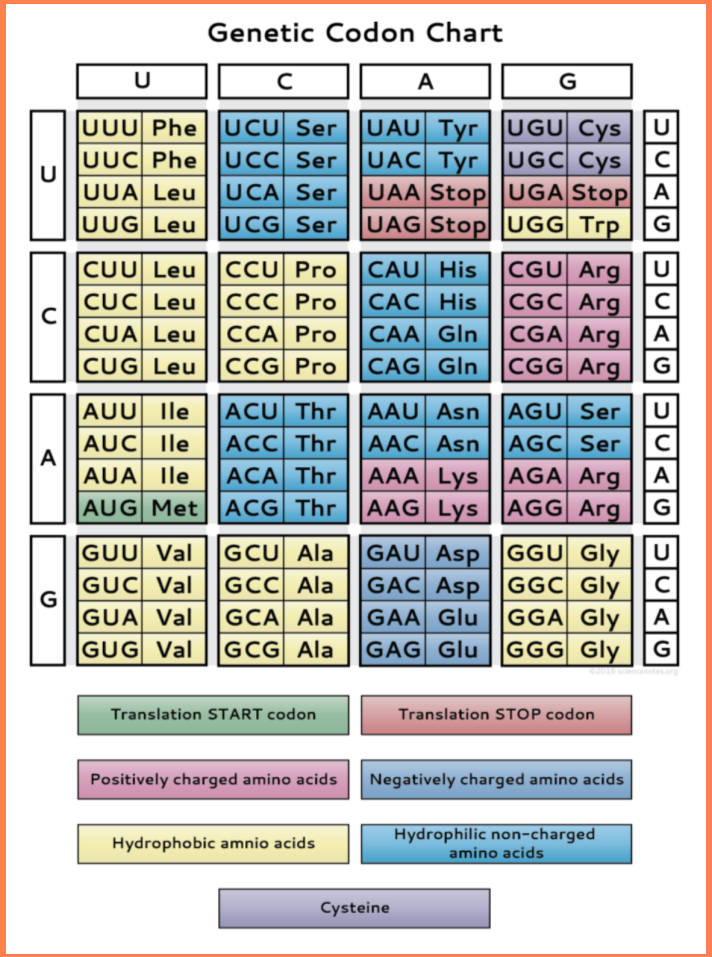
Chargaff's Rule of Complementary Nitrogenous Bases Pairing

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.

Standard Genetic Code Table

Image taken from: sciencenotes.org



Genetic Codon Chart

	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
	UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop	A
	UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

Translation START codon

Translation STOP codon

Positively charged amino acids

Negatively charged amino acids

Hydrophobic amnio acids

Hydrophilic non-charged amino acids

Cysteine

Central Dogma of Biology Summary Table

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

The End

- For more information check your eclass!