# A6: It's in your Genes

* Assignment A6 should be complete individually or with one partner.
* If you work with a partner, be sure to follow good pair programming practices.
* Be sure to read the entire prompt and understand the problem before beginning coding.

## Learning Objectives

* More practice breaking a larger problem down into smaller pieces using functions
* Gain practice manipulating strings
* Introduce some concepts about DNA

A DNA researcher is interested in some code to help him with his research project dealing with bioinformatics, which is an interesting blend of math, computing, and biology.

## Deoxyribonucleic Acid (DNA)http://www.calabriadna.com/wp-content/uploads/2013/06/dna.jpg

DNA, or deoxyribonucleic acid, is the hereditary material in humans and almost all other organisms. In the nucleus of each cell, the DNA molecule is packaged into thread-like structures called chromosomes. Nearly every cell in a person’s body has the same DNA, and an organism's complete set of DNA is called their genome. Gene sequencing tries to determine the exact sequence of nucleotide bases in a strand of DNA to better understand the behavior of a gene.

DNA is a double-stranded entity consisting of a sequence pair of nucleotides (also referred to as bases). The picture to the right shows an example of a sequence.

There are four possible nucleotides in DNA: **A, C, G, and T**

Each strand contains the complementary sequence of the other, where:

* + T pairs with A.
  + A pairs with T.
  + C pairs with G.
  + G pairs with C.

## Your task

In this assignment, we will create all the necessary functions to turn a DNA sequence into its resultant Amino Acid sequence. Most of the structure of this program has been provided for you; you need to complete each function, as well as create the tests to prove your functions are correct.

|  |  |
| --- | --- |
| is\_nucleotide()  Tests a string to see if it is a valid DNA strand. A valid strand consists solely of the 4 nucleotides: **A, C, G, and T**  What do the following calls to is\_nucleotide() return?   1. is\_nucleotide(“TAG”) 2. is\_nucleotide(“CAB”) | 1.a.  1.b. |
| num\_times()  Determines how often a particular nucleotide occurs in a particular DNA sequence.  What do the following calls to num\_times() return?   1. num\_times("CACAGTGT",'T') 2. num\_times("ACGGGTTT",'C') | 2.a.  2.b. |
| complement\_strand()  Determines the complementary sequence of a valid DNA strand given that:   * + T pairs with A.   + A pairs with T.   + C pairs with G.   + G pairs with C.   Example: the strand **TACG** would produce a **ATGC**.  What do the following calls to complement\_strand() return?   1. complement\_strand("CAT") 2. complement\_strand("GTG") | 3.a.  3.b. |
| mRNA()  Though this is a bit of a simplification, we will assume that the mRNA of a DNA sequence will take the complement strand of the DNA sequence as input and will replace each occurrence of the nucleotide **T** with the nucleotide **U**. Therefore, the function mRNA()will compute this mRNA, given the complement of any valid DNA sequence.  Example: the complement strand **TACG** would produce **UACG**.  What do the following calls to mRNA()return?   * mRNA("CAT") * mRNA("TTA") | 4.a.  4.b. |
| chunk\_amino\_acid()  Each mRNA will then be translated to a sequence of Amino Acids by "breaking" the mRNA sequence into groups of three nucleotides. The function chunk\_amino\_acid() should accomplish this. Though you may not assume the mRNA input has length equally divisible by three, any "extra" nucleotides which extend beyond chunks of three, should simply be discarded.  This function is already written for you. If your above functions are working correctly, this function should return the correct amino acid sequence for any valid input DNA sequence. | |
| sequence\_gene()  Takes a sequence of nucleotides such as **CGTAGGCAT** and will utilize all of the above functions to return the corresponding amino acid sequence, such as **ASV**.  This function is already written for you. If your above functions are working correctly, this function should return the correct amino acid sequence for any valid input DNA sequence. | |

## Suggestions and Requirements:

* Begin by downloading [a6\_genes.py](https://drive.google.com/file/d/0B0J8Yj0B6KRSTWNkYmpMMVdWZms/view?usp=sharing)
* I HIGHLY recommend you begin by writing some more test cases. By writing test cases first, you are ensuring you understand each and every function in this rather large starter code.
* Then, complete each of the functions described above.
* After writing each function, check your unit tests. Are the tests correct? Is your function correct? If either answer is no, fix them before moving on to the next function!
* Be sure to include good comments throughout your code.
* Be sure to add the appropriate docstrings for each function.
* Be sure to modify the standard header at the top of your program with name, username, assignment number, purpose and acknowledgements.

## Checking for Understanding

Answer the following questions after you’ve completed the code for this assignment.

|  |  |
| --- | --- |
| One way of thinking about functions is to imagine them as a black box, where you can only see the inputs and outputs, but not what is happening inside the function. I tried to make you think this way about your functions by having you write the test cases before writing the functions.  Which did you write first, the **functions** or the **test cases**?  Had you used the alternative approach, do you think writing the functions would have been easier or harder? | 5. |
| This is easily the largest code you’ve encountered and written in this course.  What was challenging about this assignment? How could you mitigate those challenges in future assignments? | 6. |

## Submission Instructions

1. Review the requirements above to ensure you completed everything that was expected of you.
2. Save your code as **A6\_genes\_*username*.py**. Replace *username* with your Berea usernames. For example, the TA Bianca Marrero’s file would be **A6\_genes\_marrerob.py.**
3. Save this document at **A6\_genes\_*username*.pdf.** Replace *username* with your Berea usernames.  **NOTE:** Incorrect filenames will automatically reduce your grade by 1 point. Fortunately, the format is always the same no matter what the assignment.
4. Zip the two files together.
5. Upload the Zip file to Moodle by the due date listed on the course website: <https://trello.com/b/w7bIrLoV/>.
6. If you worked with a partner, your partner should upload a file named **a6\_genes\_*usernames*.txt** and include both partner’s name in the document.