Hyperion Development

# Python Software Developer

# Task: Bioinformatics

www.hyperiondev.com

# Introduction

**Welcome to the Bioinformatics Task!**

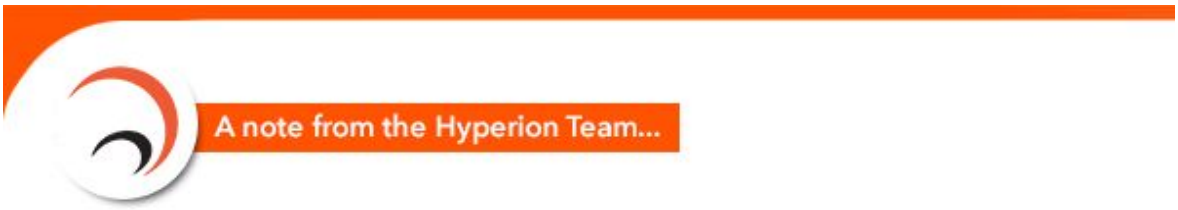
Please feel free to visit www.hyperiondev.com and to view the tools and methods that will help you throughout the course,.

For any queries regarding the course, need help understanding the task or general comments, please contact us at help@hyperiondev.com.

# Overview

Congratulations for making it this far. Now that you are comfortable with Python, we can start working on some applications and interesting fields. This document is an introduction to Bioinformatics. This Task will focus less on teaching you Python and more on showing you applications and writing useful programs in Python.

-The Hyperion Team



**An introduction to Bioinformatics**

Bioinformatics is a type of science which deals with methods for storing, retrieving and analyzing biological data. This includes DNA and protein sequences, structures, functions, pathways and genetic interfaces. It generates new knowledge about drug design and development of new software tools.

Additionally, bioinformatics deals with algorithms, databases and information systems, web technologies, artificial intelligence and soft computing. It also uses information and computation theory, structural biology, software engineering, data mining, image processing and modeling and simulation. Finally, bioinformatics makes use of signal processing, discrete mathematics, control and system theory, circuit theory and statistics. In this task we will focus on using Python to solve a problem known as sequence alignment .


**Sequence Alignment**

In bioinformatics, sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.

Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns. It's nowhere near as complicated as it sounds.

**Let's start with some basic definitions:**

DNA – The structure of DNA was discovered at a lab in the University of Cambridge less than 100 years ago. DNA is a chemical structure found in almost every cell and contains the genetic instructions used in the development and functioning of all known living organisms,

That's pretty impressive! Please read http://en.wikipedia.org/wiki/DNA if you want more information.

DNA is actually very simple. DNA is just a String. Yes, just like from programming. DNA is a long chemical string-like structure. A portion of DNA can be broken down in sub particles called nucleotides. Nucleotides are smaller molecules that when joined up form the complete String of DNA.

**There are only 4 different types of nucleotides in DNA:**
*Adenine* which is just represented by an **A** .
*Cytosine* which is just represented by a **C** .
*Guanine* which is just represented by a **G** .
*Thymine* which is just represented by a **T** .

DNA is just a long String of **A**'s, **C**'s, **G**'s and **T**'s aka nucleotides. We can even forget that these actually represent chemicals and are held together by sugar and hydrogen bonds. You will never have to read those names again. They are just different chemical elements. The entire DNA is a PATTERN of the different combinations of these nucleotides. They are held together by a structure that makes the whole DNA structure form a double helix and the iconic DNA 'shape'.

What is important to understand is that DNA is said to be a 'code' because it IS a code. It is the 'programming language' of all living things. Within a cell, all instructions are read from the DNA. Portions of the DNA are also directly responsible for genes. Genes control eye colour, hair colour, and many of traits that we don't even understand yet. The DNA of a human is an enormously long string of nucleotides. It has been decoded in a massive project called the human genome project. For more information on this project, you can view the details here: http://en.wikipedia/wiki/Human_genome_project. The act of determining the exact pattern of a DNA is known as **sequencing** and the entire string is known as the **genome**. A fully sequenced human genome would look something like this:

ACGTAAAAGGTCATACGGATCA..... (essentially the longest string you could ever imagine).

Just like all computers run at their most basic level from binary patterns of 101000010..., all living things use exactly the same pattern method except that instead of two items (0 and 1),

DNA uses four (A, C, G, T). This is where computing, combinations, and informatics directly overlap with life as we know it at the most basic level.

**How does DNA work?**
The DNA literally opens up and a 'messenger' body reads strands of the DNA. The combination of **A**'s, **C**'s, **G**'s, and **T**'s on the strand read determines the type of **amino acid** produced.

Different combinations of amino acids create different proteins which are responsible for absolutely everything in not only the human body, but all other living creatures. DNA is translated into amino acids, which is then translated into **proteins**. Every group of three nucleotides (**A**, **C**, **G**, or **T**) is known as a **codon** and one codon corresponds exactly to one amino acid. We know all amino acids, e.g codon **ATT** translates to the amino acid Isoleucine. A table of all codons and their corresponding amino acids can be seen at: http://www.cbs.dtu.dk/courses/27619/codon.html.

The amino acids are thus formed in an order depending on the order of the nucleotides in the base DNA sequence that was read. The order of the amino acids influence what type of protein is created. There are many proteins, some consisting of twenty amino acids, some consisting of less. The protein leaves the cell and does it's required job. An example of these proteins are Insulin and proteins that are used to form red blood cells.

# Instructions

First read **example.py**, open it using Notepad++ (Right click the file and select 'Edit with Notepad++').

- example.py should help you understand some more Python. Every task will have example code to help you get started. Make sure you read all of example.py and try your best to understand.

- You may run example.py to see the output. The instructions on how to do this are inside the file. Feel free to write and run your own example code before doing this

task to become more comfortable with Python.

- You are not required to read the entirety of Additional Reading.pdf, it is purely for extra reference.

***Note:*** You have reached a milestone in your Python learning, as you're now beginning to create some genuinely useful programs.

# Compulsory Task 1

**Follow these steps:**

- Visit the website: http://www.cbs.dtu.dk/courses/27619/codon.html
- Note the 'SLC' code for each Amino Acid.
- Create a program called SickleCellDisease.py . You will simulate the effects of the Single Nucleotide Polymorphism that leads to this genetic disease.
- Write a function called 'translate' that when given a DNA sequence of arbitrary length, the program identifies returns the amino acid sequence of the DNA using the amino acid SLC code found in that table.
  E.g DNA Input: ATTATTATT
  Output: III
- There are many different amino acids so this may get a bit repetitive. Just do the first five Amino Acids (i.e I L V F M) and make any other codon be printed as the amino acid 'X' . So basically, you would use an if - elif - elif .... else structure to translate each codon of DNA into the correct Amino Acid.
- Note that the program must be able to handle DNA sequences that are not of a length divisible by 3.
- Hint:
  len(DNA) - (Will return the length of a String)

DNA[0:3] - (Will get the first 3 characters of the string stored in DNA num = 3)
DNA[0:num] - (This will work too!)

# Compulsory Task 2

**Follow these steps:**

- Add another function to the program SickleCellDisease.py called 'mutate'. This function must read in the contents of the text file named 'DNA.txt'. It must then identify the first occurrence of the lowercase letter 'a' in 'DNA.txt'.
- You must then write two new text files, one named normalDNA.txt and the other named mutatedDNA.txt.
- The normalDNA.txt must have the same DNA sequence as DNA.txt with the 'a' changed to an 'A'.
- The mutatedDNA.txt must have the same DNA sequence as DNA.txt with the 'a' changed to a 'T'.
- Now create a new function, 'txtTranslate', that calls the translate function that you wrote in Task 1, to take in textfile input.
- Call it on both mutatedDNA.txt and normalDNA.txt, and output both Amino Acid sequences to the user.

# Compulsory Task 3

**Follow these steps:**

- Create a new program called AminoAlign.py with a function named 'align'.
- This function must take in any two amino acid sequences , and find all characters in which they differ.
- Run this function on the output of Task 2 (The AA of the mutatedDNA.txt file vs the AA of the normalDNA.txt file) and identify the mutated amino acid.
- Remember: print "a" == "b" (You can compare strings like this)

# Optional Task 1

**Follow these steps:**

- The strings must be formatted in such a way:

  AA Sequence 1: ILLVFMCAAGPT
  AA Sequence 2: ILLCFMCAAGPS
  ***_*******_
  Alignment: 83%

- Where there is a match in amino acids, a * should be printed beneath the pair. Where there isn't a match, a '_' should be printed beneath the pair.
- The percentage of aligned (matching) characters should also be printed.
- In bioinformatics, the alignment % of the DNA of two different animals can tell us how closely related they are to each other. More closely related animals with a higher alignment share common physical features.

- You should now have a program named SickleCellDisease.py that has functions mutate, translate and txtTranslate as well as a program called AminoAlign.py with the function align.

### Things to look out for:

1. Make sure that you have installed and setup all programs correctly. You have setup **Dropbox** correctly if you are reading this, but **Python or Notepad++** may not be installed correctly.
2. If you are not using Windows, please ask your tutor for alternative instructions.

**Still need help?**

Just write your queries in your comments.txt file and your tutor will respond. Alternatively you can email us on help@hyperiondev.com.

# Task Statistics

Last update to task: 21/05/2016.
Author: Riaz Moola
Main Tutor: Umar Randeree
Task Feedback link: Hyperion Development Feedback.