Task 5:

Welcome to Task 5 and 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.

Bioinformatics:

Bioinformatics is a type of science which deals with methods for storing, retrieving and analyzing biological data, such as DNA and protein sequences, structures, functions, pathways and genetic interfaces.

It generates new knowledge about drug design and development of new software tools. Bioinformatics also deals with algorithms, databases and information systems, web technologies, artificial intelligence and soft computing, information and computation theory, structural biology, software engineering, data mining, image processing, modeling and simulation, 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**.

dafuq did i just read

Sequence alignment:

In bioinformatics, a 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:

Definitions in Biology:

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:

Adenosine which is just represented by a A.
Cytidine which is just represented by a C.
Guanosine which is just represented by a G.
Thymidine 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'.

Why is DNA a code?

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 tons of traits we don't even understand yet.

The DNA of a human is an **enormously** long string of nucleotides. It has recently been decoded in a massive project called the human genome project: http://en.wikipedia.org/wiki/Human_genome_project. The act of determining the exact pattern of a DNA is know as '**sequencing**' and the entire String is known as the **'Genome'**.

As fully sequenced human genome will look something like:
ACGTAAAAGGTCACTACGGGATCA.....the longest string you can ever imagine.

Just like all computers run at their most basic level from binary patterns of 10100000101010101... , all living things use exactly the same pattern method except that instead of two items (0 and 1), DNA uses 4 (A , C , G , T). This is where computing, combinations and informatics directly overlaps with life as we know it at a most basic level.

How does DNA work?

The DNA literally opens up, 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 ➡ Translated into Amino Acids ➡ 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 known Amino Acid. We know all Amino Acids.

EG the Codon ATT (3 nucleotides) translates exactly 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 tons of proteins, some of length 20 amino acids, some of less. The protein leaves the cell and does it's job. An example of these proteins are Insulin and proteins that are used to form red blood cells.

EXAMPLE:

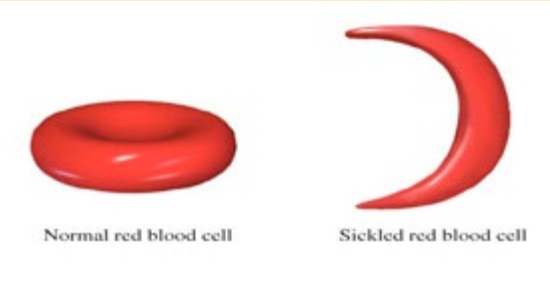
If the DNA sequence ACG TTT ACG TAT is read, it forms the Amino Acid sequence: Threonine, Phenylalanine, Threonine, Tyrosine. Go to http://webusers.physics.illinois.edu/~esham2/dna/dna.php and copy and paste in ACG TTT ACG TAT into the textbox, then press 'Convert' to see this! (You will need to select the "This DNA is the coding strand" option in the drop down box to get the correct answer.)

Mutations and their effects:

Sometimes, errors occur in the DNA. There are known as mutations. If JUST ONE OF THE NUCLEOTIDES (A, C, G OR T) CHANGE THAT IS KNOWN AS A MUTATION. Mutations can occur while reading the DNA, copying the DNA, at birth, through environmental causes such as gamma rays released by radioactive particles (explaining birth deformities after the Chernobyl disaster etc). Look at what a drastic effect one single mutation can have:

Say there was a mutation and the original DNA sequence ACG TTT ACG TAT changes to ACG TAT ACG TAT.

Now the amino acid sequence changes from Cysteine Lysine Cysteine Isoleucine.



To Cysteine Isoleucine Cysteine Isoleucine.

The resulting Protein is not formed correctly because of the one incorrect Isoleucine in the pattern. In this particular case, the protein that is made here is one used in Haemoglobin – the red blood cells of the body. Because of this one permanent error in DNA, the Haemoglobin protein is not formed correctly in every single cell of the body and **ALL OF A** persons blood cells look like a 'sickle', as seen in the diagram above.

This is the single cause of the genetic disease – Sickle Cell Anemia. Red Blood cells carry oxygen throughout the body. Because a sickled red blood cell has less volume than a normal red blood cell, as seen above, a person with sickle cell disease **gets less oxygen around their body even though they breath the same amount of oxygen as a normal person**. As a result they feel **tired all the time, and have a much shorter life expectancy.**

Single Nucleotide Polymorphisms:

When a single A C G or T gets mutated in the DNA of a person, it is known as a SINGLE NUCLEOTIDE POLYMORPHISM or SNP. There is an extremely interesting wiki of SNPs seen here: http://www.snpedia.com/index.php/SNPedia

Example: Here is an entry http://www.snpedia.com/index.php/Rs1805007. This small change influences the development of red hair in people and also has a link to an increased resistance to anaesthetics!

Summary:

The bottom line: DNA is really important. Sequences which are just simple strings are extremely important and identifying slight differences is a huge market and field. We can even take entire DNA sequences of different animals and find the % of matching nucleotides and this is shown in animals more closely related to humans. We can take specific genes for eye colour etc and compare their DNA sequences with those found in humans and see 99% match.

Instructions:

- Read example.py and see some examples of problems involving matching DNA sequences to other sequences using Strings and Lists in Python.
- Find the instructions on the compulsory task in the example.py file. Follow the instructions in the comments to complete the task.

You have reached a milestone in your Python learning and you're reading to write some genuinely useful programs.

Feedback:

Please email all feedback about this task and the Bioinformatics explanation above to students@hyperiondev.com – especially if you feel the explanation is not clear! Also go to http://www.surveymonkey.com/s/PSLGN6F to fill out an online survey about the course so far.

Need some help?

Firstly, make sure you have installed and setup all programs correctly.

Please refer to the pdf file **PythonReference.pdf** if you would like more examples of Python coding and explanations.

If you having problems understanding example.py or how to complete Task 5, please contact students@hyperiondev.com. One on one help sessions are available over the internet or in person in Westville, Durban or UKZN (Westville Campus) and these can be arranged by contacting us. **We employ paid teachers who are here to help you!**

Alternatively, simply create a text file in this folder and call it 'Comments'. Inside it, type any problems you're having and one of our teachers will reply as quickly as possible.

If there are any specific areas that are unclear or areas that require additional information:

Please add to 'What do you want to learn.txt' and one of our teachers will assist you once they read your request.

A peek ahead:

Task 6: An exciting introduction to Artificial Intelligence. This task was developed jointly with the University of Edinburgh's Artificial Intelligence research department. Find out about Natural Language Processing and how:

- **Ironman uses it,**
- **iPhone 4's 'Siri' uses it**
- **Google uses it**

Write your own fully featured AI program to automatically classify tweets using the same tools Gmail uses to identify spam!"