About your main assignment

Background: For many diseases with known causative mutations, screening methods have been developed to detect whether people have a high risk of becoming sick, even before the onset of the actual disease.

Over the last few years, the cost of full genome sequencing has gone down so that, in some cases, it might be cheaper to collect the complete genome sequence of patients with a high risk of carrying variants associated with the disease, rather than using targeted screening procedures.

Cystic fibrosis is a complex disease, where patients often manifest the following symptoms: problems with lung functions, diabetes and infertility. From a genetic point of view, there are several mutations associated with this disease. In particular, the CFTR gene (short for Cystic Fibrosis Transmembrane Conductance Regulator) encodes an ion channel protein acting in epithelial cells, and carries several non-synonymous genetic variants, with alterations leading to premature stop codons, that are known to cause the disease.

Goal: In this assignment, you have access to the human reference genome as well as the genome annotation. In addition, you have full genome sequence data from five individuals from a family at risk of carrying mutations related to the disease.

Your task is to write a Python program that will extract the CFTR gene, translate the gene sequence to its corresponding amino-acid sequence and based on the inferred amino-acid sequence determine whether any of the five given individuals is affected.

» Fetch the appropriate files

The main task is divided in several steps. The first step is to fetch the sequence file (in `fasta` format) and the appropriate annotation file (in `GTF` format) from the Ensembl database.

The CTFR gene is chromosome `7`.

» Warmup

1. What is the length of the chosen DNA sequence?
   ▸ Tip
2. How many genes are annotated in the GTF file?
   ▸ Note
3. What fraction of the chromosome is annotated as genes?

» Architect a method

All the following tasks are now related to the CTFR gene.

In the annotation file (from the Ensembl database), that gene has the id `ENSG00000001626` on chromosome `7`.

» Course Content

During this course, you will learn about:

- Core concepts about Python syntax: Data types, blocks and indentation, variable scoping, iteration, functions, methods and arguments
- Different ways to control program flow using loops and conditional tests
- Regular expressions and pattern matching
- Writing functions and best-practice ways of making them usable
- Reading from and writing to files
- Code packaging and Python libraries
- How to work with biological data using external libraries (if time allows).

» Learning Outcomes

After this course you should be able to:

- Edit and run Python code
- Write file-processing python programs that produce output to the terminal and/or external files.
- Create stand-alone python programs to process biological data
- Know how to develop your skills in Python after the course (including debugging)

Learning objectives (ie goals for the teachers)

- Increase the student's toolbelt for better quality and performance at work
- Make students understand that there is more to programming than only *knowing* the syntax of a language. This expertise is precisely what NBIS provides.