

BINF-5007 Introduction to Python for Bioinformatics

Humber Polytechnic

Final Exam

Fall 2025

- 1.** The hydrophobicity index is a measure of how soluble an amino acid is in water. For a protein (sequence of amino acids), the index can be calculated adding the hydrophobicity of individual amino acids and dividing the result by the length of the sequence.

You must calculate the hydrophobicity index of a given protein sequence. To achieve this, complete the following tasks:

- a.** (20 points) Complete the `calculate_protein_hydrophobicity(...)` function, that takes a protein as a parameter and returns the hydrophobicity index.

Note: Use the `get_aa_hydrophobicity(...)` function to aid you.

- b.** (20 points) Complete the `most_hydrophobic_protein(...)` function, that takes a list of proteins as parameter and returns a tuple with the most hydrophobic protein and its hydrophobicity index. Remember hydrophobicity can be negative.

- c.** (10 points) The `get_aa_hydrophobicity(...)` function could cause a `KeyError` exception if given a non-existent amino acid. Modify the `calculate_protein_hydrophobicity(...)` function to handle this exception, by skipping the non-existent amino acids from the calculation (create a copy of your function and perform the modification over the copy)

- 2.** (25 points) One common problem for the analysis of DNA, RNA and protein sequence variants is extracting relevant information related to the variants from biomedical texts. For example, researchers must locate the variants from lab reports manually, which can be a time-consuming task, and prone to errors. A possible computational aid for variant extraction is the use of regex, that can automatically match and extract the data researchers are looking for.

You must create a Python script that looks for protein insertion variants that comply with the HGVS nomenclature. One example of this type of variant would be `NC_000001.11:p.1234_1235insACGT`, where:

- `NC_000001.11` is the sequence identifier.
- `1234_1235` is the range of the insertion in the sequence
- `ACGT` is the DNA sequence inserted into the sequence.

The general syntax for protein insertion variants is as follows:

```
sequence_id ":p." range "ins" sequence
```

Consider that:

- Sequence_id always start by NC and have 6 digits followed by a dot and the version number (may have 1 or 2 digits)
- The range always contains two numbers (the number of digits can vary) separated by an underscore.
- The sequence must contain one or more nucleotides.

Create a function called `find_prot_ins_variants(...)`, that takes a text as parameter and returns a list of all the protein insertion variants present in the given text.

Then, complete the script by requesting the user to input a text and, using the previously created function, find the protein variants to show them to the user in a message like: “The protein insertion variants found in your text are: `<list_of_variants>`”

3. (25 points) The dna_seq.json file contains multiple DNA sequences that must be translated into proteins. Read the sequences present in the file and translate them into proteins using Biopython. Save the translated protein sequences into a new text file called output.txt