Computational biology: project

January 2020-2021

A FASTA file contains a number of protein sequences. Write a Python script that, given the name of the FASTA file, writes the sequence identifier and the molecular weight for each sequence in the file.  
Note: do not take the ends of the molecule into account.

*Example: for the sequence “DHPFWKQTACKHV” the molecular weight would be 1578.8.*

A FASTA file contains a single protein sequence. A second FASTA file contains sequences that are fragments of the sequence in the first file. Compute the molecular weight of each of the sequences, and using those, determine whether there is a combination of fragments that may cover the complete protein sequence, without those fragments overlapping

*Example: suppose the sequence has a molecular weight of 1623.5, than fragments with weights 326.2, 487.4, and 819.9 might cover it without overlap.*

Write a script, that given the same input as before, i.e., a sequence file containing a single protein sequence, and a file containing fragments, checks whether the fragments with that weight indeed cover the protein without overlap. For bonus points, take into account that there may be multiple candidate sets of fragments.

*Example: the sequence “QYLCRNI” with molecular weight 891.0 is covered by the fragments “CR” (259.3), “QYL” (404.5), and “NI” (227.3).*

Write a script that, given the name of a FASTA file and a protein sequence identifier, writes all possible tRNA sequences from which the protein sequence may have been translated, but only if there are less than 5000. If the given sequence identifier does not occur in the FASTA file, or if there are more than 5000 possible sequences, an appropriate error message is printed. Note that there are potentially very many tRNA sequences, so write a function to count the total number, without generating the sequences first.

*Example: for the sequence “CH” the script’s output would be “UGUCAU”, “UGUCAC”, “UGCCAU”, and “UGCCAC”.*

G**eneral remarks**: the FASTA files can have an arbitrary number of sequences, unless explicitly mentioned otherwise. Sequence length can be arbitrary. Remember that floating point computations are subject to round off errors. Also remember that magic is *not* allowed: ensure that you know what the functions you use actually do!

If you have questions, feel free to ask: [geertjan.bex@uhasselt.be](mailto:geertjan.bex@uhasselt.be)