Project Suggestion 2:

Remove duplicate sequences. Total 15P

Use the provided FASTA files available on studIP. These files contain sequence data, with some of the sequences appearing multiple times. Develop a python script capable of eliminating all duplicates from each file. The python script should efficiently…

1) Generate new sequence files, with one file for each input file, ensuring that all duplicate sequences are removed from each file. **(6P)**
2) Modify the sequence headers to include information about how frequently the sequence occurred in the original file, such as "*>GeneID_3x*". **(6P)**
3) Provide a brief summary detailing the number of sequences present in the original file and how many of them remain in the new after duplicates have been removed. **(3P)**

Hint: gene sequences could be lower or upper case.