Homework 3: Advanced Python

Note I have provided working functions `read_fasta` which can be used as-is to read in a fasta file and return a list of sequences as strings.

1. Write a script `protein_freq.py` which will read in a protein Fasta format file Saccharomyces_cerevisiae.peps.fa
    a. Print out the overall frequency (percentage) of each amino acid observed across all the sequences
    b. Order this output sorted by the frequency of the amino acid, so the most frequent appear first.
2. Using the Fasta files for the genomes 'Ecoli_K-12.fasta' and 'B_subtilis_str_168.fasta'
    a. Print out a table to compute frequency of all di-nucleotide combination (e.g. AA, AC, AG, AT, CA, CC, ...).

    Report should be tab delimited and look like

    ```
    Motif  Ecoli_K-12    B_subtilis_str_168
    AA     7.28     9.85
    ```

3. Process a tabular BLAST report file 'Ecoli-vs-Senterica.BLASTP.tab' which has the following columns for a pairwise sequence alignment report

QUERYNAME SUBJECTNAME PERCENTID LENGTHALN NUMMISMATCHES GAPOPEN QSTART QEND SSTART SEND EVALUE BITSCORE

Only print the lines which match the criteria:

a. SUBJECT ACCESSION between YP_008253351-YP_008253423
b. Percent Identity is >= 25%

Update the report to add 1 more columns after the existing ones.

a. Computed length of the query alignment using QSTART and QEND

print out the new report on the STDOUT