1. Given a proteomics database DB, and a spectrum S, implement an algorithm to find the best match between this spectrum and the database. You can use the following simplistic assumptions:

(i) Only consider b-ions and y-ions.
(ii) Only consider charge +1,
(iii) The best match is defined as the match that explains maximum number of b-ions and y-ions in the peptide.
(iv) Use a peptide mass tolerance and ion mass tolerance of 0.02 Da.
(v) Assume fully tryptic peptides (trypsin cuts from K and R residues).

A - Formulate this problem as a computational problem, and state the input, output and the goal.

B - Design an algorithm to find the best match.

C - implement the algorithm in programming language of your choice, and test it on part_1.peaklist and sequence.fasta .

Consider proton-mass = 1.00728 Da and water-mass = 18.01056 Da. Lets consider the peptide DEFG. The second b-ion mass is mass(D) + mass(E) + 1.00728 = 245.075. The second y-ion mass is mass(F) + mass(G) + 18.00728 + 1.00728 = 223.1044 . Find the list of amino acid masses below.

2. Now consider each amino acid can go through one or more known post-translational modifications. For example, if the native peptide is TGST, and we allow PTMs T-18 and S-18, we can have 8 possible modified peptides:

T,G,S,T (no modification)
T,G,S,T-18
T,G,S-18,T
T,G,S-18,T-18
T-18,G,S,T
T-18,G,S,T-18
T-18,G,S-18,T
T-18,G,S-18,T-18

Given a peptide P and a spectrum S, the goal of modification discovery is to find a modification of peptide P that (i) has the same mass as S, and (ii) is the best match in terms of number of b-ions and y-ions explained.

A - Formulate this problem as a computational problem, and state the input, output and the goal.

B - Design an algorithm to find the best match using sequence.fasta and part_2.peaklist

C - implement the algorithm in programming language of your choice, and test it on the data given, assuming T-18 and S-18 modifications.

D - For a peptide with *n* S and T residues, what is the complexity of your search ? Is it possible to do the search in time polynomial with *n* ?

List of amino acid masses :

D=115.026943031
E=129.042593095
F=147.068413915
G=57.021463723
A=71.037113787
C=103.009184477
L=113.084063979
M=131.040484605
N=114.042927446
H=137.058911861
I=113.084063979
K=128.094963016
T=101.047678473
W=186.079312952
V=99.068413915
Q=128.058577510
P=97.052763851
S=87.032028409
R=156.101111026
Y=163.063328537
T-18=83.0371184
S-18= 69.021468409