02-425/725 HW4

March 13, 2024

1 Database Search

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:

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

1.1 Problem Statement

Formulate this problem as a computational problem. State the input, output, and the goal.

1.2 Algorithm Design

Design an algorithm to find the best match.

1.3 Test Implementation

Implement the algorithm in the programming language of your choice and test it on `part_1.peaklist` and `sequence.fasta`.

2 Modification Discovery

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 $mass(F) + mass(G) + 18.00728 + 1.00728 = 223.1044$.

The list of amino acid masses can be found below.
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:

1. `T`,`G`,`S`,`T` (no modification)

2. T,G,S,T-18
3. T,G,S-18,T
4. T,G,S-18,T-18
5. T-18,G,S,T
6. T-18,G,S,T-18
7. T-18,G,S-18,T
8. 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

1. has the same mass as S, and
2. is the best match in terms of number of b-ions and y-ions explained.

2.1 Problem Statement

Formulate this problem as a computational problem. State the input, output and the goal.

2.2 Algorithm Design

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

2.3 Test Implementation

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

The parent-mass (PM) on the first row of the peak-list file, is the sum of residual-masses of all the amino-acids, plus mass of water and mass of proton-charge:

$$PM = sum\text{-}of\text{-}residual\text{-}mass\text{-}of\text{-}amino\text{-}acids + mass(water) + mass(proton) =$$

$$sum\text{-}of\text{-}residual\text{-}mass\text{-}of\text{-}amino\text{-}acids + 19.01$$