Assignment 3. Due date (Mar 26, 11:59pm)

Write a program to identify peptides from their tandem mass spectra.

Input:

The input will have two files: An mgf file provides a list of tandem mass spectra in MGF format, and a fasta file provides a list of proteins in FASTA format.

The command line used to call your program will be:

```
run.sh file1.mgf file2.fasta
```

Note that the file names used in test may be replaced when we test your program.

Output:

For each spectrum in the mgf file, print the identified peptide from the proteins in the fasta file, as well as a score. A higher score should indicate a better confidence of your software's identification. Each spectrum occupies one line in the following format:

```
Id m/z z peptide protein score1 score2
```

The columns are:

- id is the index of the spectrum in the mgf file. The first spectrum has id=0. Then the id number increase by 1 for each additional spectrum.
- m/z is the mass to charge ratio of the spectrum, which is read from the mgf file as "PEPMASS=400.6561". Despite the tag name "PEPMASS", the value is actually mass to charge ratio.
- z is the charge state, which is read from the data as "CHARGE=3+".
- peptide is whatever peptide your program identified.
- protein is the first 10 characters of the header line (including the greater sign) of the protein that contains the identified peptide. If a peptide appears in multiple proteins, output one of them arbitrarily.
- score1 and score2 are two scores your program assigns to this peptide spectrum match. It will be used to filter your results. Score1 is described later. You can implement score2 as you wish.

Test:

We will use a pair of mgf and fasta files to test your program. These files will be very similar to the sample files provided with the assignment. So, any reasonable efforts in your program to read the sample files should read the test files correctly.

The testing fasta file will contain both correct (target) and random (decoy) proteins. The following criterion is used to test the performance of each scoring function. Given a score threshold T, we will first filter your program's results to keep only those with score $\geq T$. Then we will count the number of remaining peptides

appearing in target and decoy proteins, respectively. The FDR at score threshold T is defined as

$$\text{FDR(T)} = \frac{\#decoy\ matches\ with\ score \geq T}{\#target\ matches\ with\ score \geq T}$$

We will test different T to find the least threshold T_0 such that $FDR(T) \leq 0.05$. The number of target peptides with score $\geq T_0$ will be used as the performance of your program.

More specifications:

Overall procedure:

1. For each spectrum, you calculate the mass of the peptide by mz × z − 1.0073 × z. Here mz is the mass to charge ratio of the precursor and z is the charge state. The −1.0073 ×z in the formula is to subtract the mass of the extra protons due to the charge.
2. Each protein can be digested with trypsin rule: after R or K, and not before P. You should only consider the peptides respecting the digestion rule. Do not use other peptides.
3. For each tryptic peptide, the peptide mass is calculated as the total residue mass +18.0105. The +18.0105 is because of the extra water on the peptide.
4. If a peptide mass matches the spectrum's peptide mass within error $\pm$0.1Da, then evaluate the peptide-spectrum match with your scoring function.
5. After all proteins are evaluated, output the peptide with the highest matching score to the spectrum.

The first scoring function:

1. Calculate the y-ion mass of a peptide as the total residue mass of the suffix, plus 19.0178.
2. Find the tallest peak within $\pm$0.5Da of the calculated y-ion mass. Suppose it's relative intensity (defined as the ratio between the peak's intensity and the intensity of the tallest peak in the spectrum) is x, add $\max\{0, \log_{10}(100 \cdot x)\}$ to the total score.

Amino acid residue mass:

1. The amino acid residue's mass can be found at the following webpage. Use the monoisotopic mass in that table. http://education.expasy.org/student_projects/isotopident/htdocs/aa-list.html
2. A special case is the residue Cystein (Cys, or C). It is purposefully modified during the sample preparation before mass spec. The mass of the (modified) residue should be 160.03065 instead of the one given in the above table.
3. The following tool can be used as a reference to check if your mass is calculated correctly: https://www.rapidnovor.com/mass-calculator/

Hints:

1. The description of the procedure is for reference only. You may need to adjust it in order to implement it in a program.
2. Some suggestions to improve the score include, but not limited to the following: the use of b-ions, converting the intensity to a score with a different formula, and the use of the mass errors as a feature in the scoring function.

Data file supplied:

1. Ups.fasta: A fasta file containing a list of proteins.
2. Test.mgf: A mgf file containing a list of MS/MS spectra. This is for your development purpose only. The actual marking will use a different mgf file.
3. A3-test.peaksdb.csv: A list of identified peptides by a commercial tool. This is for reference only to assist your coding. The answers given in the file may not be completely correct, but should be mostly correct. The commercial tool also identified peptides with PTM and do not follow the trypsin rule completely. Your program needs not to (and should not) look for these. Also notice that the file format is different from our specification. The "Scan" column in that file can be used to find the corresponding spectrum in test.mgf by matching the "SCANS=???" lines.