Homework: peptide detection from tandem mass spectra

In particular, you should implement the scoring function described in Equation 7 and the e-value algorithm described in Figure 1. In broad strokes, the program should parse the MS/MS spectrum in “peptide_msms.txt” (presented as mass/intensity pairs) and match it against every peptide sequence in “peptide_database.txt” to determine the best scoring peptide-spectrum match. Afterwards the e-value should be calculated.

I expect you to simplify the scoring function in a few ways:
• Only use b- and y-type ions without any a-type ions or neutral losses. Calculate b-type ions as \((1.008 + \sum \text{residue masses})\) and y-type ions as \((19.018 + \sum \text{residue masses})\). Use monoisotopic residue masses.
• Round all fragment masses to the nearest integer. This will allow you to perform “=” operations to find matching fragment masses, rather than check them within a mass tolerance. I have already rounded the masses in “peptide_msms.txt”. Please note: do not round individual amino acid masses, just the final fragment masses.
• \(\log_{10}\) the scores before computing the e-value. The final score equation should be:

\[
Hyperscore = \log_{10}\left(\left(\sum_{i=0}^{n} I_i \cdot P_i\right) \cdot n_b! \cdot n_y!\right)
\]

When computing the e-value histogram, use bin sizes of 0.1. Calculate the e-value for the top scoring peptide using linear regression applied to the \(\log_{10}\) (count) in all bins from the bin with the highest count (the distribution mode) to the bin \(j-1\) greater than the mode right before bin \(j\) where the count first equals 0.

What to hand in:
• All the code used to perform the calculations
• The peptide sequence and the Hyperscore for the top scoring peptide
• The slope (m) and intercept (b) of the linear regression used to calculate the e-value
• The \(\log_{10}\) e-value for the top scoring peptide
• A figure showing the histogram points used to calculate the linear regression, the line, and the estimated e-value for the top scoring match (see right)

Please contact me with any questions! (searleb@uw.edu)