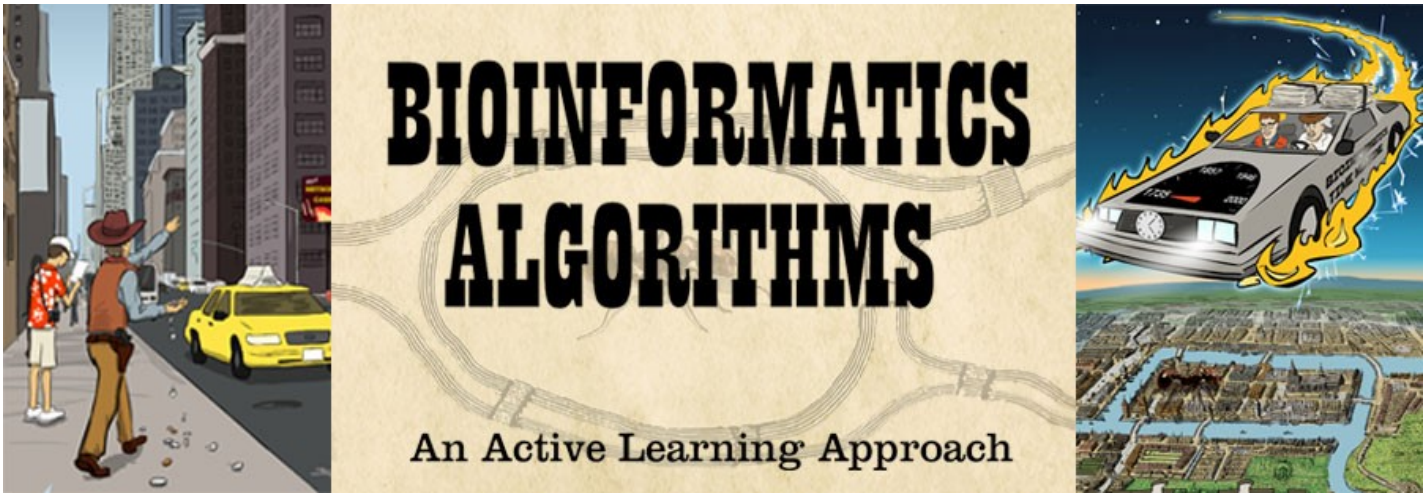

Chapter 4: How Do We Sequence Antibiotics?

(Coursera Week 3)

How do I expand the list of candidate peptides in **CyclopeptideSequencing**?

One of the students in the first session of our class, John Cloutier, provided the following example that illustrates how **CyclopeptideSequencing** works. Consider a strange amino acid alphabet consisting of just two amino acids with masses 1 and 3. The figure below shows the peptides generated at each step by **CyclopeptideSequencing** with respect to a sample experimental spectrum {0, 1, 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 5, 6}. Consistent peptides are shown in black, and inconsistent peptides are shown in red. In step 4, **CyclopeptideSequencing** produces the blue peptides 1-1-1-3, 1-1-3-1, 1-3-1-1, and 3-1-1-1; these four linear peptides all represent the same cyclic peptide, whose spectrum is equal to the experimental spectrum.

Step	Candidate peptide	Linear Spectrum
1	(1)	(0, 1)
	(3)	(0, 3)
2	(1, 1)	(0, 1, 1, 2)
	(1, 3)	(0, 1, 3, 4)
	(3, 1)	(0, 1, 3, 4)
	(3, 3)	(0, 3, 3, 6)
3	(1, 1, 1)	(0, 1, 1, 1, 2, 2, 3)
	(1, 1, 3)	(0, 1, 1, 2, 3, 4, 5)
	(1, 3, 1)	(0, 1, 1, 3, 4, 4, 5)
	(1, 3, 3)	(0, 1, 3, 3, 4, 6, 7) - remove
	(3, 1, 1)	(0, 1, 1, 2, 3, 4, 5)
	(3, 1, 3)	(0, 1, 3, 3, 4, 4, 7) - remove
	(3, 3, 1)	(0, 1, 3, 3, 4, 6, 7) - remove
	(3, 3, 3)	(0, 3, 3, 3, 6, 6, 9) - remove
4	(1, 1, 1, 1)	(0, 1, 1, 1, 1, 2, 2, 2, 3, 3, 4) - remove
	(1, 1, 1, 3)	(0, 1, 1, 1, 2, 2, 3, 3, 4, 5, 6) – output
	(1,1, 3, 1)	(0, 1, 1, 1, 2, 3, 4, 4, 5, 5, 6) – output
	(1, 1, 3, 3)	(0, 1, 1, 2, 3, 3, 4, 6, 5, 7, 8) - remove

	(1, 3, 1, 1)	(0, 1, 1, 1, 2, 3, 4, 4, 5, 5, 6) - output
	(1, 3, 1, 3)	(0, 1, 1, 2, 3, 3, 4, 6, 5, 7, 8) - remove
	(3, 1, 1, 1)	(0, 1, 1, 1, 2, 2, 3, 3, 4, 5, 6) - output
	(3, 1, 1, 3)	(0, 1, 1, 2, 3, 3, 4, 6, 5, 7, 8) - remove
5	Empty list.	

I've noticed a discrepancy between the mass of an amino acid cited in the book and in other sources. Why is this?

For example book suggests that glycine has elemental composition C_2H_3ON (integer mass 24+3+16+14=57 Da), whereas Wikipedia suggests that it is $C_2H_5ON_2$ (integer mass 24+5+16+28=75 Da). We should use the former formula in analyzing mass spectra, since when an amino acid forms a peptide bond, it loses a water molecule (H_2O).

(Coursera Week 4)

How can I improve the performance of **LeaderboardCyclopeptideSequencing**?

You should not need to optimize your implementation for **LeaderboardCyclopeptideSequencing** in order to pass its Code Challenge. However, various optimization approaches can be applied. To take one example, if the leaderboard has a peptide with mass smaller than *ParentMass*(*Spectrum*) but exceeding *ParentMass*(*Spectrum*) - 57 (recall that 57 is the mass of the lightest amino acid, glycine), this peptide can be safely removed from the leaderboard.

How can I trim the peptide leaderboard without sorting?

To trim a peptide leaderboard without using sorting, we will first compute an array *ScoreHistogram,* where *ScoreHistogram*(*i*) holds the number of peptides in *Leaderboard* with score *i*. For example, if we are trimming the leaderboard from Charging Station: Trimming the Peptide Leaderboard to $N = 5$ peptides (including ties), then *ScoreHistogram* = *ScoreHistogram* = (0, 0, 2, 1, 3, 2, 2). As a result, 2 + 2 + 3 = 7 peptides will be retained and the remaining 0 + 0 + 2 + 1 = 3 peptides will be trimmed. Here, the minimum score that a peptide can have without being cut is denoted $ScoreThreshold_N(Spectrum)$.

Assuming that *N* is smaller than the number of elements on *Leaderboard*, note that the number of peptides cut is at most |*Leaderboard*| - *N*. In order to compute $ScoreThreshold_N(Spectrum)$, we need to find the index *i* such that the sum of the first *i* elements in *ScoreHistogram* is at most |*Leaderboard*| - *N* and the sum of the first *i* + 1 elements in *ScoreHistogram* exceeds |*Leaderboard*| - *N*. To find this index, we will compute *CumulativeHistogram*, where *CumulativeHistogram*(*i*) holds the number of peptides in *Leaderboard* with score below *i*. For our ongoing example, *CumulativeHistogram* = (0, 0, 2, 3, 6, 8, 10). This leads us to the following pseudocode.

```
AnotherTrim(Leaderboard, Spectrum, N, AminoAcid, AminoAcidMass)
    for i ← 0 to |Spectrum|
        ScoreHistogram(i) ← 0
    for j ← 1 to |Leaderboard|
        Peptide ← j-th peptide in Leaderboard
        LinearSpectrum  ← LinearSpectrum(Peptide, AminoAcid, AminoAcidMass)
        LinearScore ← Score(Peptide, LinearSpectrum)
        LinearScore(j)  ← LinearScore
        ScoreHistogram(LinearScore) ← ScoreHistogram(LinearScore) + 1
        sum ← 0
        for i ← 0 to |Spectrum|
            sum ←  sum + ScoreHistogram(i)
            if sum > |Leaderboard| - N
                ScoreThreshold ←  i - 1
                for j ←  1 to |Leaderboard|
                    Peptide ←  j-th peptide in Leaderboard
                    if LinearScores(j) < ScoreThreshold
                        remove Peptide from Leaderboard
    return Leaderboard
```

© 2018 by Phillip Compeau & Pavel Pevzner | All Rights Reserved
ISBN: 978-0-9903746-3-3