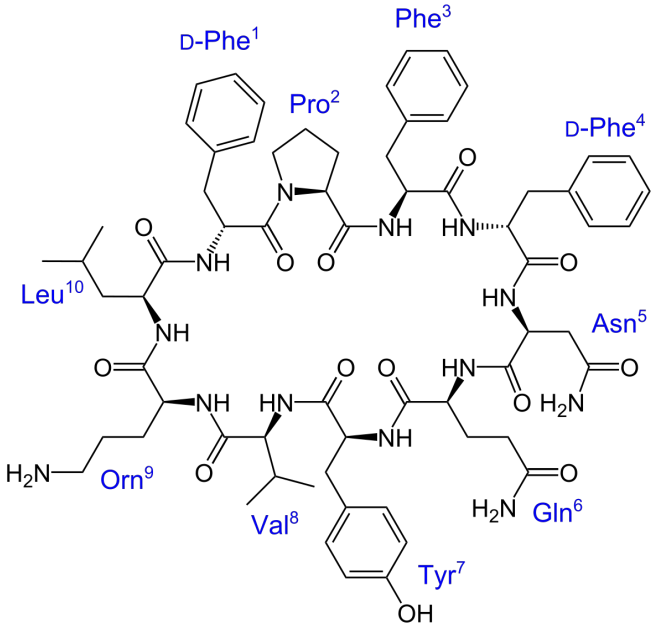
Cyclopeptide Sequencing

Sol Taylor-Brill

The Problem:

How can we determine the amino acid sequence of a peptide from its mass spectrometry spectrum?

(Described in Chapter Four of *Bioinformatics Algorithms*)



Tyrocidine B. Source: https://en.wikipedia.org/wiki/Tyrocidine

Cyclopeptide Sequencing

- Relies on a “branching and bounding algorithm”
  - Every time the while loop runs, each peptide grows in length by one amino acid
  - Then the spectrum each peptide in the list is be compared to the spectrum of the mystery peptide. If it contains a peak value that is not in the spectrum, it is removed from the list
  - This continues until the peptide whose spectrum exactly matches the given spectrum is identified
- Pros:
  - Faster than a brute force algorithm
- Cons:
  - Still very slow
  - Only worked if the given spectrum was perfect and didn’t have any missing or extraneous peaks.
  - Certain peptides take a lot longer to run than others.

Leaderboard

- Follows a branching-bounding algorithm, but instead of looking for a peptide that perfectly fits the spectrum, each peptide is scored based on how closely it matches the mystery spectrum.
- Every iteration, the candidate peptides with the highest scores are kept and the rest are discarded. The program should return the highest scoring peptide not a “perfect” peptide.
- Pros:
  - Works with imperfect spectra
- Cons:
  - Still very slow.
  - Doesn’t always accurately reproduce the correct peptide.

Leaderboard w/ Spectral Convolution

- Spectral convolution narrows down the number of amino acids that are added to the peptides each time.
  - Calculate the difference between each peak in the spectrum
  - Based on the numbers between 57 and 186 that appear most frequently, a list of probable amino acids is formed
- Every iteration Leaderboard grows peptides using only those amino acids - not all 20.
- Pros:
  - Faster (Still pretty slow on peptides greater than 5aa long)
  - Hypothetically more accurate

Results

- While Leaderboard conceptually is an improvement over regular Cyclopeptide sequencing, it is still slow for certain peptides and often does not produce accurate results.
  - However this is likely due to me implementing the algorithm incorrectly
    - I have not been able to make Leaderboard run as described in the book, and the program does not currently self-limit correctly.
  - Was unable to test whether it accurately reproduced Tyrocidine B because the program ran too long.

References

1. Compeau, P., & Pevzner, P. (2015). *BIOINFORMATICS ALGORITHMS: An active learning approach* (Vol. 1). S.l.: ACTIVE LEARNING.
2. “Tyrocidine.” *Wikipedia*, Wikimedia Foundation, 30 Apr. 2018, en.wikipedia.org/wiki/Tyrocidine. [Source of Image on 2nd Slide]