Verona Yue

Prof. Anna Ritz

BIO131

Final Project Report

05/14/2020

How Do We Sequence Antibiotics?

My code: https://repl.it/join/yfxrfcvg-yuev

This project is designed according to the chapter "How Do We Sequence Antibiotics" in the textbook. Originally, we discovered that antibiotics are composed by amino-acid sequences. It is critical to be able to predict the sequence of the antibiotics, so that the antibiotics could be massively generated. Now, we have this mass spectrometer to shatter the molecules to pieces, and then weight fragments' masses (measured by Daltons). Then, our problem becomes how can we predict the peptide sequence according to a mass specturm. For turning this problem into a computational problem, we assume that we have this mass specturm (input), and we need to return a list of possible cyclic peptides that contains the pieces with same masses in the given specturm (output, in their mass forms). This function is called "Cyclopeptide Sequencing".

First, I need to generate a theoretical specturm from a peptide (string). Notice that I need two functions for this step—generating a cyclic theoretical specturm and a linear theoretical specturm. I am more likely to get cyclic specturm from the mass spectrometer, and I need to check if each possible mers' theoretical specturm is in the given specturm by using linear theoretical specturm function. Second, I need to find the possible one mers from the given specturm. Third, I add each one mer to each possible mer and remove this one mer from the candidate list. If this new mer (with adding this one mer) is consistent with specturm, then I add this mer to the possible mer list. Then, I try to add one and check each mer again to expand the peptide. Once, the mass of the possible mers is equal to the biggest mass in the given specturm, the addition ends. From these possible peptide strings, I made a function to read the mass of each amino-acid in each string, and convert them into a list of lists of masses. Some small functions are helpful in my case, such as reading an amino-acid mass and make it into lists, getting the entire peptide mass, or reversing a list.

I also have several functions to make my input and output have the correct format for testing on Rosalind. I ran my code for the data from Rosalind, and it says I got the correct peptides from given spectrum. My code was slow earlier, because it does not consider that the linear theoretical specturm for each possible mers needs to be in the given spectrum as well. Also, one amino-acid could appear twice in a peptide, and I had some trouble with the remove function. However, once I had my linear theoretical specturm function and remove one mers if they are used in the current mer, the problem was solved. For future exploration, I can implement this method with the leaderboard, and come up with a better strategy to determine the sequence of an antibiotic.

It's my pleasure to complete BIO131 course with Anna and my classmates. I'm willing to share my code and my report on the course webpage. Thank you.