Curing Staph Project

Hunter Casillas 12/1/19

Abstract

Staphylococcus or staph infection is an infection caused by bacteria commonly found on the skin or in the nose. Staph can be spread person-to-person and is exceptionally contagious. Once the bacterium has been identified as the cause of the illness, treatment is often in the form of antibiotics. However, many strains have become antibiotic resistant, such as methicillin-resistant staphylococcus aureus or MRSA. As a result, MRSA causes over twenty-thousand deaths a year in the United States. To diagnose the infection, it's possible to differentiate the antibiotic resistant staphylococcus from non-resistant infections with the help of mass spectrometry. Mass spectrometry is an analytical technique that measures the mass-to-charge ratio of ions. It can be used to identify unknown compounds via molecular weight determination, to quantify known compounds, and to determine structure and chemical properties of molecules. In the following report, I used the spectrum generated from a sample of MRSA that was killed by an antibiotic to find the sequence of the cyclic peptide. Then I found the most likely sequence of linear peptide from the available sequence data.

Methods/Algorithms

In order to calculate the sequence for the cyclic peptide, I used the convolution cyclopeptide sequencing algorithm from Rosalind. This algorithm first computes the convolution of a given experimental spectrum. It then selects the M most frequent elements (with ties) between 57 and 200 in the convolution to form an extended alphabet of amino acid masses. Finally it runs the algorithm leaderboard cyclopeptide sequencing, where amino acids are restricted to this alphabet. The leaderboard cyclopeptide sequencing algorithm uses a leaderboard which holds the N highest scoring candidate peptides for further extension. At each step, it expands all candidate peptides found in the leaderboard by adding every possible amino acid to the end. Then, it eliminates those peptides whose newly calculated scores are not high enough to keep them on the leaderboard. The algorithm also uses a trim function to reduce the leaderboard to the "N highest-scoring peptides including ties".

In order to calculate the sequence for the linear peptide from the available sequence data, I used the algorithms described above and used the gene sequence to check against the spectrum of masses. In order to deal with the real value masses, I rounded them to the nearest sum of amino acid masses in order to stay consistent. I ran a sliding window across the gene, generated the spectrum for the amino acids in the sliding window, and scored the spectrum found against the given spectrum. In the end, the highest scoring section of the gene correlated to the section of the gene that generated the given spectrum.

Results

The sequence for the cyclic peptide I found was 114-130-164-97-128-163-66-147-161. Although there could be missing amino acids (or masses that don't correlate to the standard masses of the amino acids) there's not really much that I can do to justify my answer because there wasn't any pertinent biological information provided. I used the convolution algorithm as opposed to the regular leaderboard algorithm because it is more successful at finding sequences. In the textbook the regular leaderboard algorithm failed to reconstruct Tyrocidine B1 from Spectrum10 when using the extended alphabet of amino acids, however the convolution algorithm was able to do so successfully.

The sequence for the linear peptide I found was SELDGRCSLRRGSSFTFLTPGP or 87-129-113-115-57-156-103-87-113-156-156-57-87-87-147-101-147-113-101-97-57-97 with a score of 18. I was able to generate the spectrum for the amino acids in the sliding window, and score the window spectrum against the given spectrum. The sequence I found in the end is in fact a substring of the gene sequence provided to us, which is justification for my answer being biologically accurate. The highest scoring section of the gene correlated to the section of the gene that generated the given spectrum.

Discussion

The majority of my issues were code related. Trying to get my algorithms working properly with the provided files, how to handle the pesky real valued masses, making sure my linear peptide sequence was a substring of the sequence provided in the lab specifications. I decided to test both rounding the real valued masses and truncating them. Even after implementing multiple methods for dealing with this, my results were the same. This is convincing that my functions are doing what they are supposed to be. I also had an issue where converting my masses back into a peptide was returning the incorrect sequence and therefore wasn't a substring in the provided gene. I realized I just had to tweak a few aspects of that function in order to get the correct answer. Other than those minor things, I didn't have any issues with the lab.

Recommendations

I think for the most part this project is structured well the way it is and it's a great learning experience for our bioinformatics course. I appreciated the hints and clarification section that was added to the project description. It gave me better direction and helped when I got confused. One thing that may help for future semesters would be to verify that the algorithms students have worked on for Rosalind are correct and finished before starting the project. This could save time reworking broken code and emphasize the results in the report.