Tayla Isensee

Dr. Anna Ritz

Introduction to Computational Biology

7 May 2018

Final Project Report

Motivation

In this project, I wanted to implement a program that sequences cyclic peptides like antibiotics. Because of their unique structure and their absence from the coding genome of the organisms that synthesize them, sequencing cyclic peptides requires the use of mass spectrometry rather than analysis of DNA. This alternative sequencing technique really intrigued me, as did the biological importance of knowing the structure and sequence of antibiotics. Using mass spectrometry to sequence antibiotics was a chapter in our textbook, which also drew me to this project as the textbook and its companion website provided explanations, datasets, and pseudocode that helped further this project.

Data Formatting

The datasets that I obtained came either from the textbook, *Bioinformatics Algorithms,* or from its companion website Rosalind. Most of the larger datasets came from Rosalind, while smaller and toy examples came from the textbook. The only manipulation of datasets was turning sets into lists so that they fit well into my program.

High-Level Steps

My program can: generate both theoretical linear & cyclic spectrums of a given peptide and provide a possible mass sequence for a cyclic peptide, given ideal experimental data.

I. Theoretical Spectra

a. Linear Spectrum: takes peptide string and outputs a list of all the possible masses of the different slices of the peptide.

b. Cyclospectrum: same thing as linear except the mass slices corresponding to the connected end and start are also added to the output list.

II. Sequencing a Cyclopeptide

a. Given an ideal experimental mass spectrum of a cyclic peptide, output its possible peptide sequence (converted also into a sequence of the component amino acids).

i. While no peptide yet matches the ideal cyclospectrum:

1. Expand peptides: add each of the 20 amino acids to candidate peptides.
2. If expanded peptides' mass is in the experimental spectrum, keep them. If not, don't include them.
3. Once a peptide's theoretical cyclospectrum matches the experimental, output that peptide.

Results, Discussion, and Conclusions

I discovered that the runtime of my final Cyclopeptide Sequencing function limits the length of the peptide it can sequence in a decent amount of time to around 8-10 amino acids. However, it can reliably sequence a peptide and keep the lists from building up unhelpful or already tested peptides. There were many assumptions that this program made that allowed the results it did: it assumed that there were no missing masses, no incorrect masses, and that only 20 amino acids are present in the peptide. This general method that incorporated elements of the branch-and-

bound algorithm structure, could be used to sequence other peptides and sequence-able elements of an organism that are not found in the genome.