Abrar Abidi
Fall 2015

Computational Biology Independent Project

Data:

One of the reasons I chose this project was that I would be able to explore unfamiliar territory in my own knowledge computational biology, while largely being able to check the validity and accuracy of my code every step of the way with the help of Rosalind datasets. Thus, though I only leave you with a single data file from Rosalind in the form of "data.txt," I in fact inputted and tested every previous dataset supplied by Rosalind before the spectral convolution part of the problem. I only moved on from one part of the problem to another when I saw that my code returned the expected outputs for given inputs from Rosalind.

I designed my code to be able to read the data in the format given by Rosalind directly, so that I wouldn't have to tamper with it manually. There is also a data file called "masses.txt," which simply supplies my master code with all the masses corresponding to each amino acid. This file is again read directly, and unlike "data.txt" should be left as is, given that it contains standard values.

While initially, the motivation for my project was data-driven, i.e. to reconstruct the peptide sequence for Tyrocidine C, I soon shifted my focus more to the algorithm itself rather than the data, since it was sufficiently difficult to write it in the first place. That is why eventually I decided to simply stick with datasets provided by Rosalind. Anna also told me this would be fine.

Program outline:

1) Find the theoretical spectrum of a cyclic peptide.

2) Find the cyclic peptide corresponding to a theoretical spectrum that matches an ideal experimental spectrum.

3) Calculate the score associated with a cyclic peptide against a spectrum.

4) Find the cyclic peptide corresponding to noisier (i.e. error-containing) theoretical and experimental spectra, via leaderboard cyclopeptide sequencing.

5) Computer the spectral convolution of a given spectrum.

I could have continued with the chapter beyond here, but I decided to stop at this point, mainly because though my computer consistently returned the output I expected, it began to take much longer to compute those results. In particular, the very large "additional dataset" supplied by Rosalind for problem 4G had to be

left running overnight. It did finally return the correct peptide sequence, at which point I felt satisfied enough with the work I had done to call it quits and begin the write-up.

Discussion:

The fact that I was able to consistently get the desired output for the inputs I fed into my code meant that I had succeeded in what I set out to do. My goal here was to see if I could input a flawed experimental spectrum (as any realistic experimental spectrum would likely be) and to accurately sequence the original peptide. I thought this was a fascinating way to correct for the inevitable errors involved in taking mass spectrograms of fragments of an original amino acid.