Monday, October 22, 2018

Protein Secondary Structure Prediction

Broad Goal

- Is deep mutational scan data better for predicting protein secondary structure than evolutionary alignments?

The above was my question but I don't think that this has to be their question. They could just download a faster Amino acid sequence alignment for a protein of choice (viruses are great for obvious evolutionary reasons) and see if they can use the amino acid sequence to predict secondary structure.

Learning Goals

1. What are proteins, what is evolution
2. What is secondary structure, what role does it play in evolution
3. How do amino acid sequence patterns influence secondary structure
    1. https://www.cs.princeton.edu/~mona/Chapter29.pdf
4. What secondary structure predictors currently exist: how do they compare?
    1. http://www.compbio.dundee.ac.uk/jpred/ (Jpred)
    2. http://bioinf.cs.ucl.ac.uk/psipred/

Steps (simple)

The below steps are a simple approach to see if creating a "consensus?" sequence for a protein by downloading an evolutionary alignment and finding the most common amino acid in each position will generate an accurate secondary structure using psi red or Jpred.

1. Download an amino acid alignment for protein of choice
2. Read the FASTA file into python and make a dataframe/matrix

3. Quantify the amino acids in each position (Go crazy and make a frequency plot for each AA at each position?)
4. Generate a “consensus” sequence using the most common AA in each position
5. Paste the sequence into Psipred and Jared to get a secondary structure.
6. Download Pymol
7. Get the solved structure of protein of choice from PDB
8. Compare the structure outputted by Jared and Psi prod to that of the Solved structure by overlaying

Steps (if the above is quick)

Once they’ve done the above if you have a published DMS dataset for their protein of choice they could do they same thing but on that dataset. The hypothesis is that a DMS dataset is better because it looks at all the possible AA in each position. We may not have had the chance to sequence every protein with every possible AA through evolution yet.

In reality I don’t think I saw better results with DMS data.

1. Once they are comfortable with the evolutionary data looking at DMS data is basically the same process. Obviously, they will have to learn what DMS is.
2. They can try to predict secondary structure themselves and compare it to Psipred and J-pred by learning what patterns of amino acids make what secondary structure
3. If number 2 works then they can try to add the Chau-Faussman correction to see if that makes their prediction better.