* Introduction and problem definition
  + Introduction to general applications and uses of machine learning
  + Biology in general
  + What is a protein
    - Building blocks of proteins
    - How it “folds” up together
    - Different 3-D structures that could be created
  + Want to predict 2o structure of a protein given its amino acid sequence
    - Currently a very hard problem to do
    - Not very accurate right now
    - Mention current work and attempts
    - Just getting deep learning and other methods to work
  + Motivation
    - Need to know 2o structure
    - Form follows function
* Proposed method to solve problem and intuition
  + Want to get down to the amino acid level
    - Need that granularity because it’s possible that different sections interact with each other from a long distance away
    - Feature selection
  + Made use of logistic regression
    - At amino acid level
    - Probability a specific amino acid is part of a particular structure based on its location
  + Followed logistic regression with a naïve bayes/k-means clustering operator
    - Group amino acids together based on structure probability information and distance information
    - Helps to find “strings” of each structure, which one is most likely to be part of which other strings of amino acids and structures.
* Experiments run
  + Training was done using a large dataset from the PDB
    - Used variety of proteins for training, no similarities
    - Initially had similar proteins and got too good results
  + Have to keep in mind that we have to keep track of similar proteins vs completely different proteins
  + Run on multiple protein, and measured accuracy simply counting how many amino acids were correctly classified
* Results (write observations and comments here to)
  + With similar proteins structurally, it did well, upwards of 65%, but with a large mixture of proteins, I was getting 42%
  + Bad choice of the logistic regression training (ASCII values)
  + Normalization of features
* Conclusions
  + Worked decently well
  + Clustering was a limiting factor
  + Improvement on the metric for assignment, the way the weight vectors were calculated
* References