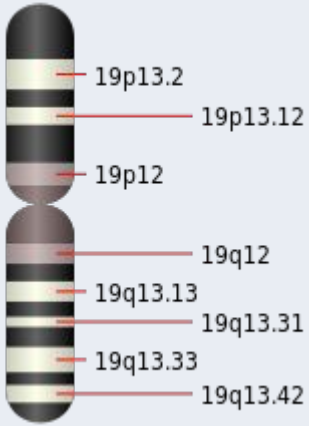
With current gene mapping techniques, can I predict the location of a gene based on its composition?

Beth Fawcett

Background

- DNA
- Chromosomes
- Genes
  - Locus
- Proteins
  - Strand +/-
  - Amino acids



Business Case

- Drug Development
- Individualized Medicine
- Advancement in understanding mechanism behind chromosomal based disease processes
- Health applications

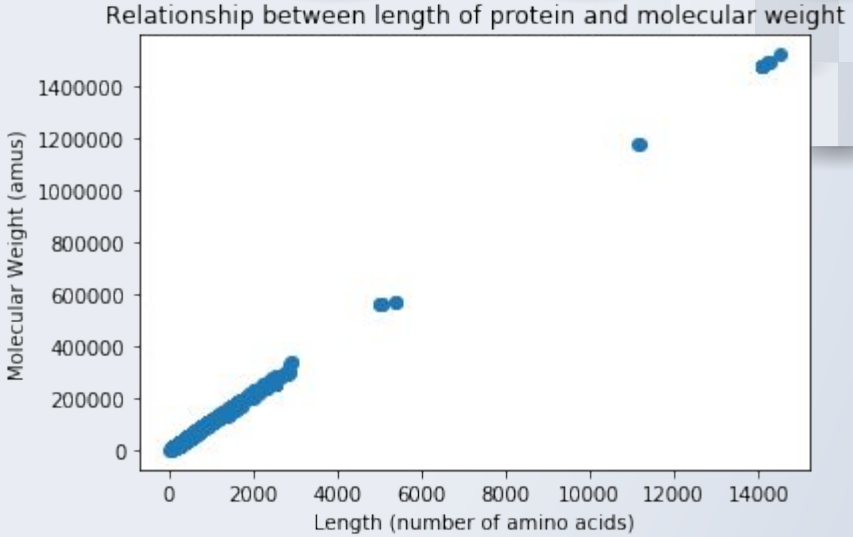
Methodology

- Supervised Learning Techniques
- Loci with > 50 occurrences
- Chromosome 19 has 48 instances of loci with > 50 occurrences
- Locus Study
- Strand Study

- Obtain Data
- Scrub Data
- Explore Data
- Modeling
- Interpret

Findings

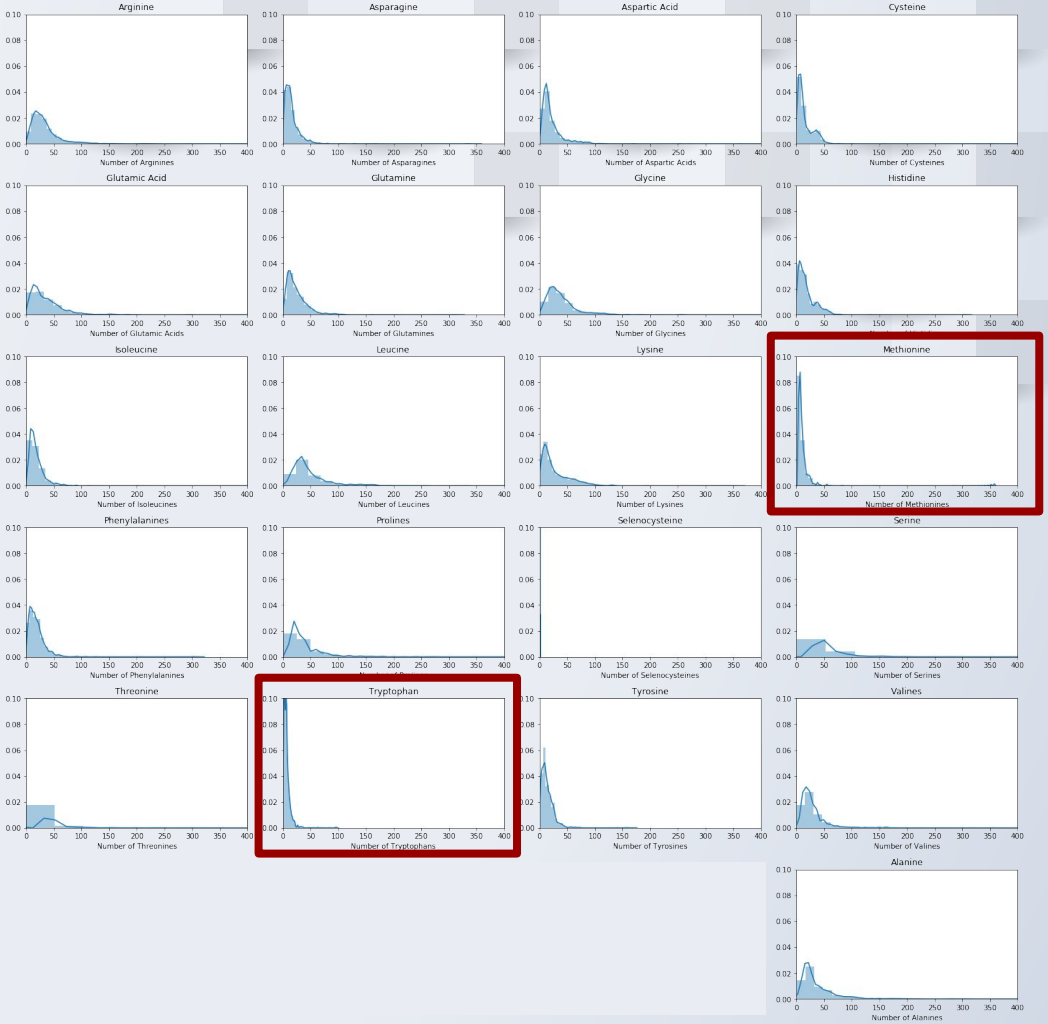
A commonly known fact is that as length of amino acids in protein increases, the molecular weight increases.



Findings

Tryptophan, a heavy protein, was frequently found.

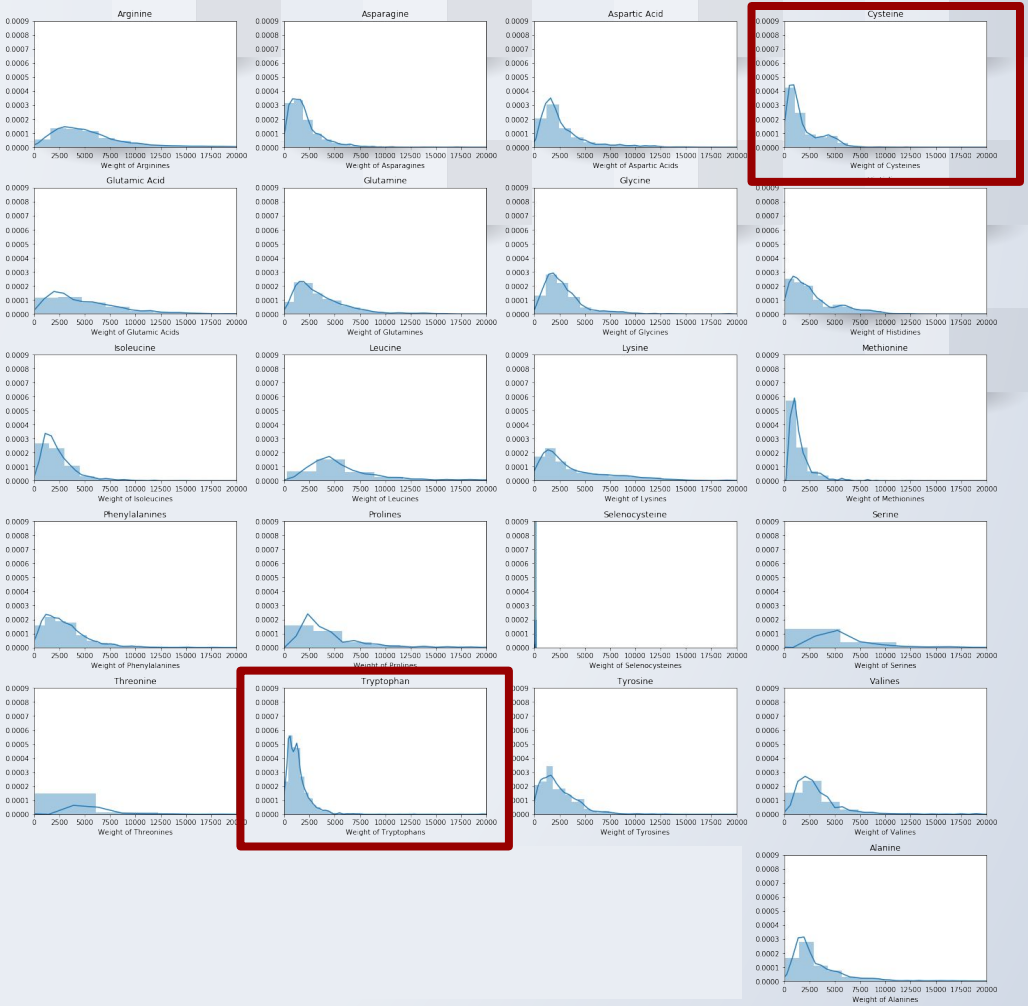
Methenamine, was also found frequently, but it is the start of all proteins.



Frequency of Amino Acids

Findings

- **Cystine and Tryptophan's weights are high for the distribution.**



Contribution of Amino Acid Weight

Findings - Random Forest

Locus Study

Test Accuracy: 99%

Strand Study

Test Accuracy: 72%

Future Work

- Grid Search with PCA
- Other types of classifiers: SVM, XGBoost
- Additional chromosomes
- SMOTE (Synthetic Minority Over-sampling Technique) to assist with expanding to all loci in chromosome
- Neural Network
- Build a chromosome from predicted values

Thank you!

For more information:

Beth Fawcett

elizabethfawcett47@gmail.com

github.com/eannefawcett