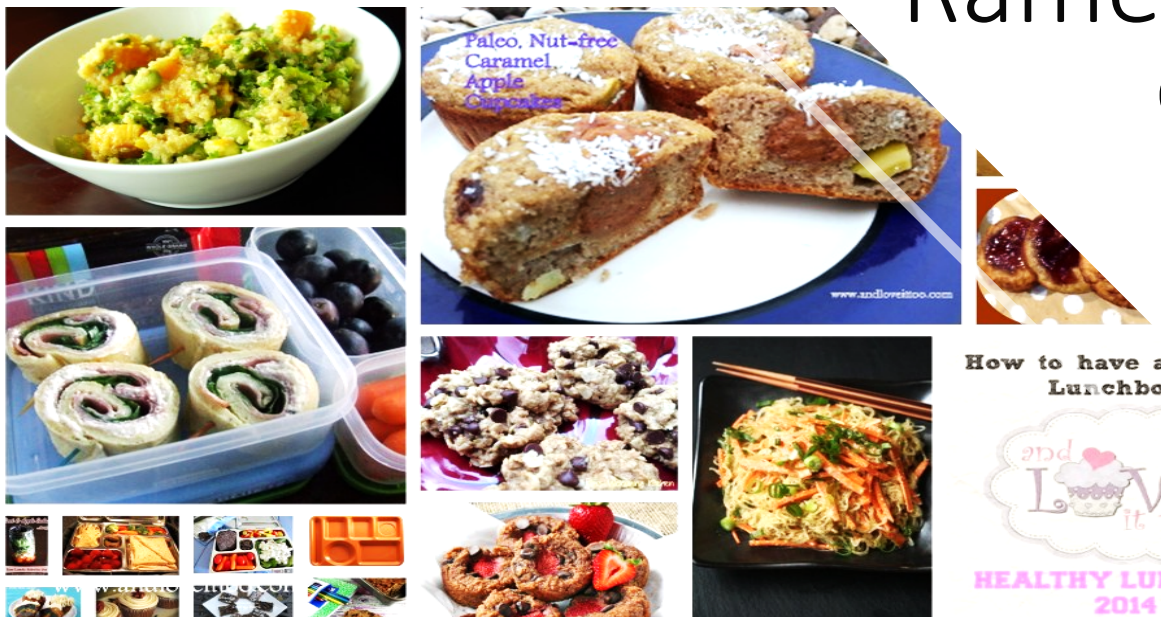
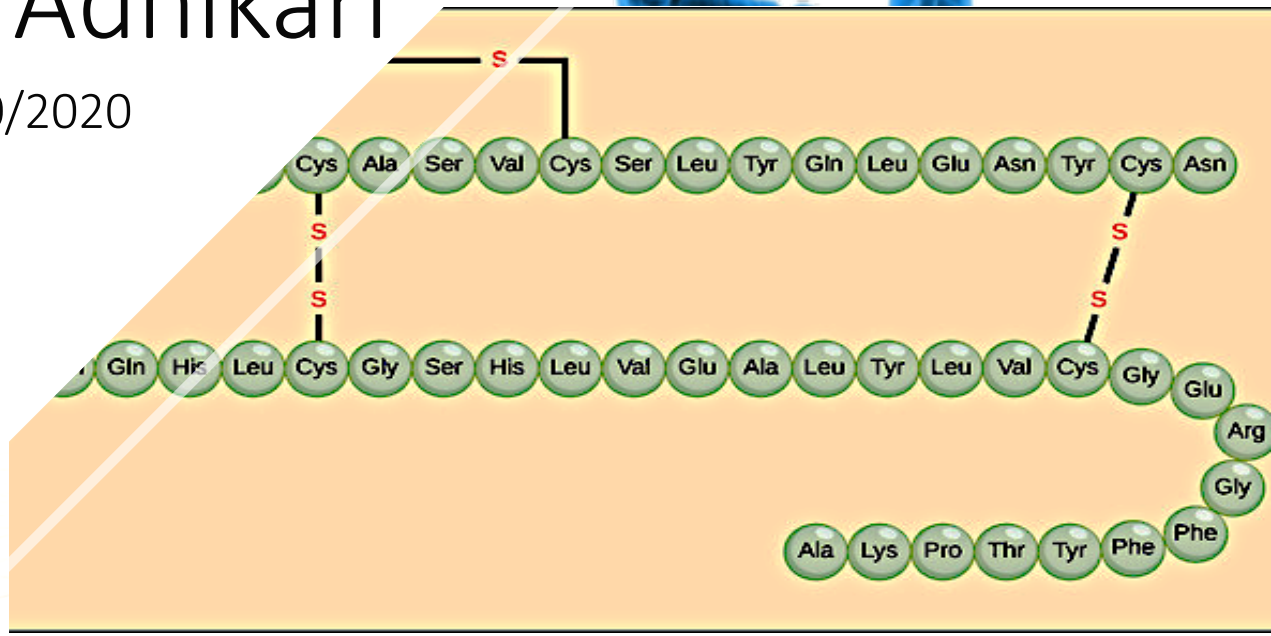
Deep Learning Models to Classify Protein Families

Ramesh Adhikari

03/10/2020





Data Story

- Over 60, 000 protein families have been defined. (We had 5,050 classes of proteins in the present datafile. Classes with less than 1200 samples were removed.)

- Classification of protein by matching physical and chemical properties is tedious.

- Machine learning and deep learning models can predict the family of any unknown protein.

Data wrangling:

- Protein sequences along with physical and chemical features were downloaded from www.kaggle.com

- Whole dataset was divided into two parts:
  a) with sequence of amino acids (for deep learning)
  b) with given physical and chemical features (for machine learning).

- Sequence data were converted into character level features

Features and Targets:

Amino acid represented by alphabets

Amino Acid	3-Letter Code	1-Letter Code
Alanine	Ala	A
Cysteine	Cys	C
Aspartic acid or aspartate	Asp	D
Glutamic acid or glutamate	Glu	E
Phenylalanine	Phe	F
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Lysine	Lys	K
Leucine	Leu	L
Methionine	Met	M
Asparagine	Asn	N
Proline	Pro	P
Glutamine	Gln	Q
Arginine	Arg	R
Serine	Ser	S
Threonine	Thr	T
Valine	Val	V
Tryptophan	Trp	W
Tyrosine	Tyr	Y

A protein sequence under family 'hydrolase' with 286 amino acid units

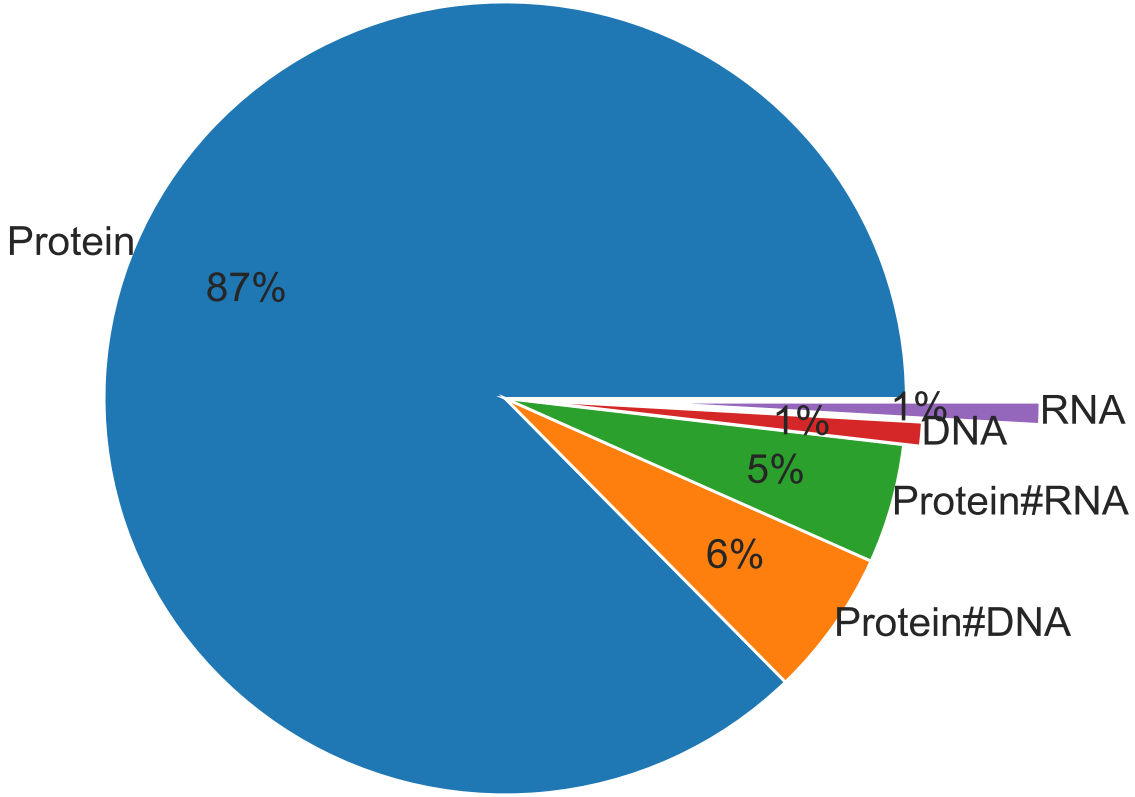
```
'TYTTRQIGAKNTLEYKV
YIEKDGKPVSAFHDIPLY
ADKENNIFNMVVEIPRWT
NAKLEITKEETLNPIIQD
TKKGKLRFVRNCFPHHGY
IHNYGAFPQTWEDPNVSH
PETKAVGDNEPIDVLEIG
ETIAYTGQVKQVKALGIM
ALLDEGETDWKVIAIDIN
DPLAPKLNDIEDVEKYFP
GLLRATNEWFRIYKIPDG
KPENQFAFSGEAKNKKYA
LDIIKETHDSWKQLIAGK
SSDSKGIDLTNVTLPDTP
TYSKAASDAIPPASLKAD
APIDKSIDKWFFISGSV'
```

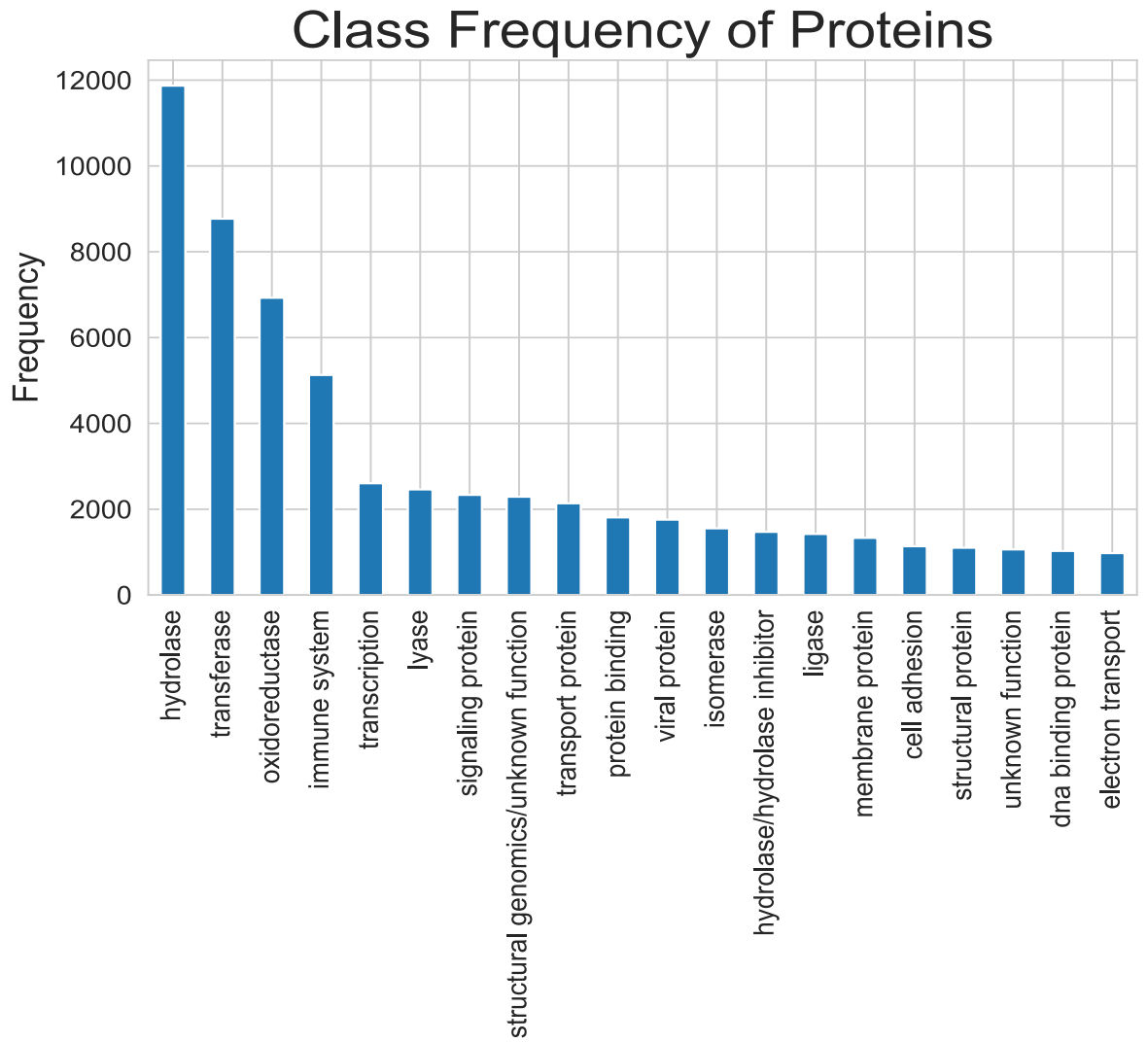
Labeling the target variable:

Protein families	Numerical Labels:
'hydrolase'	: 1
'hydrolase/hydrolase inhibitor'	: 13
'immune system'	: 4
'isomerase'	: 12
'ligase'	: 14
'lyase'	: 6
'membrane protein'	: 15
'oxidoreductase'	: 3
'protein binding'	: 10
'signaling protein'	: 7
'structural genomics/unknown function'	: 8
'transcription'	: 5
'transferase'	: 2
'transport protein'	: 9
'viral protein'	: 11

Exploratory Data Analysis:

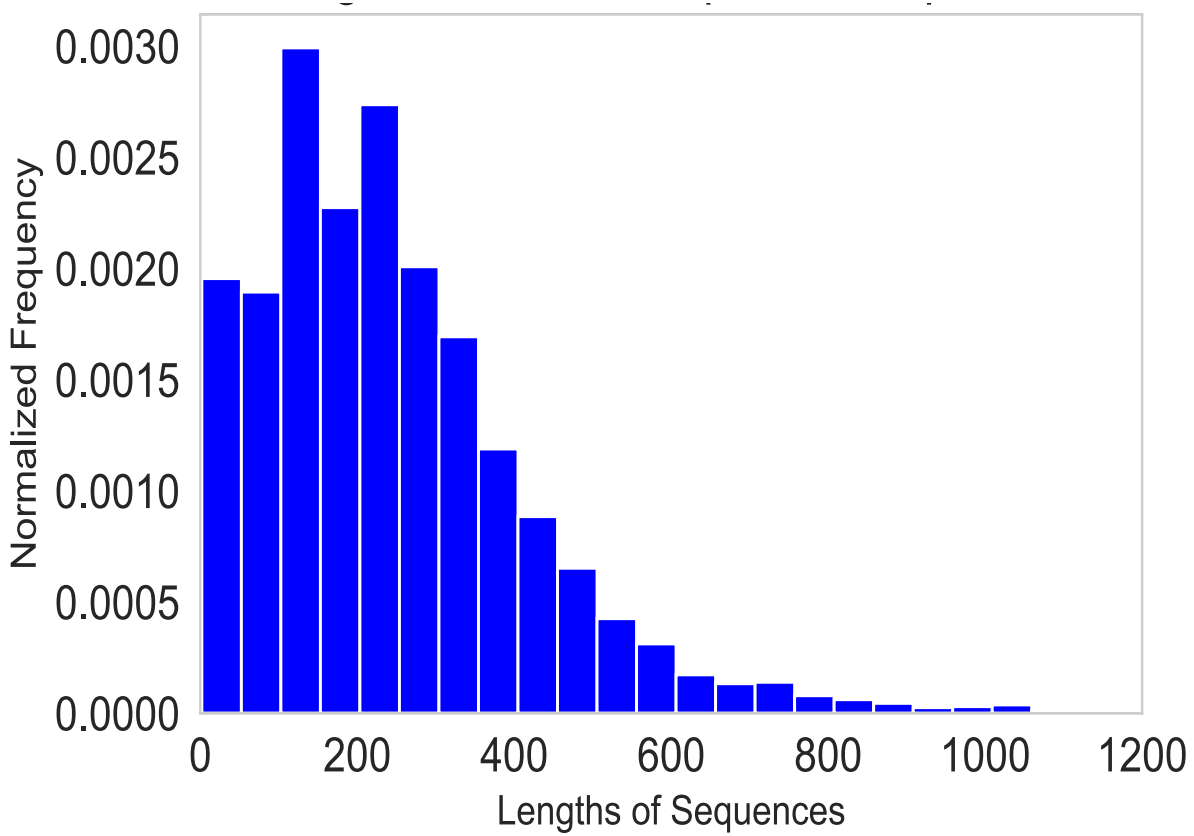
87% of the macromolecules in the data set were protein. We omitted the samples other than proteins only



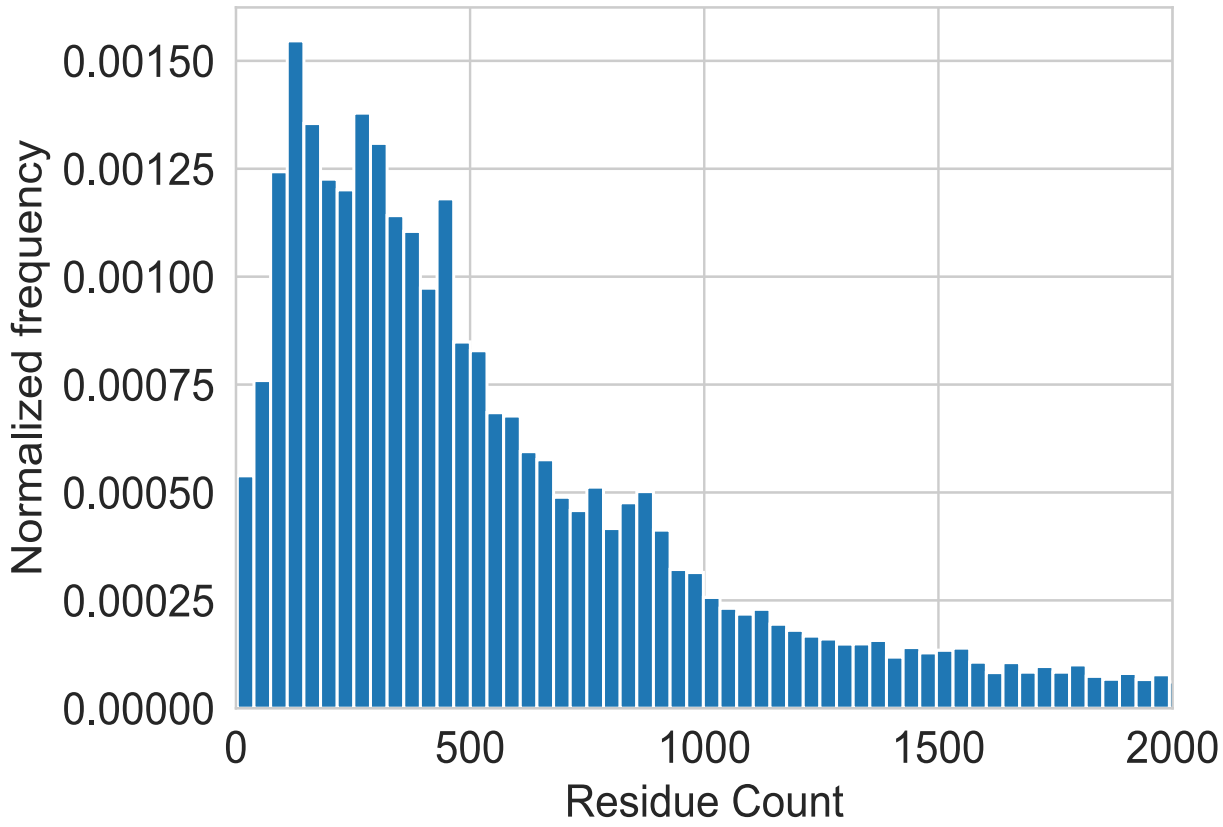


Exploratory Data Analysis:

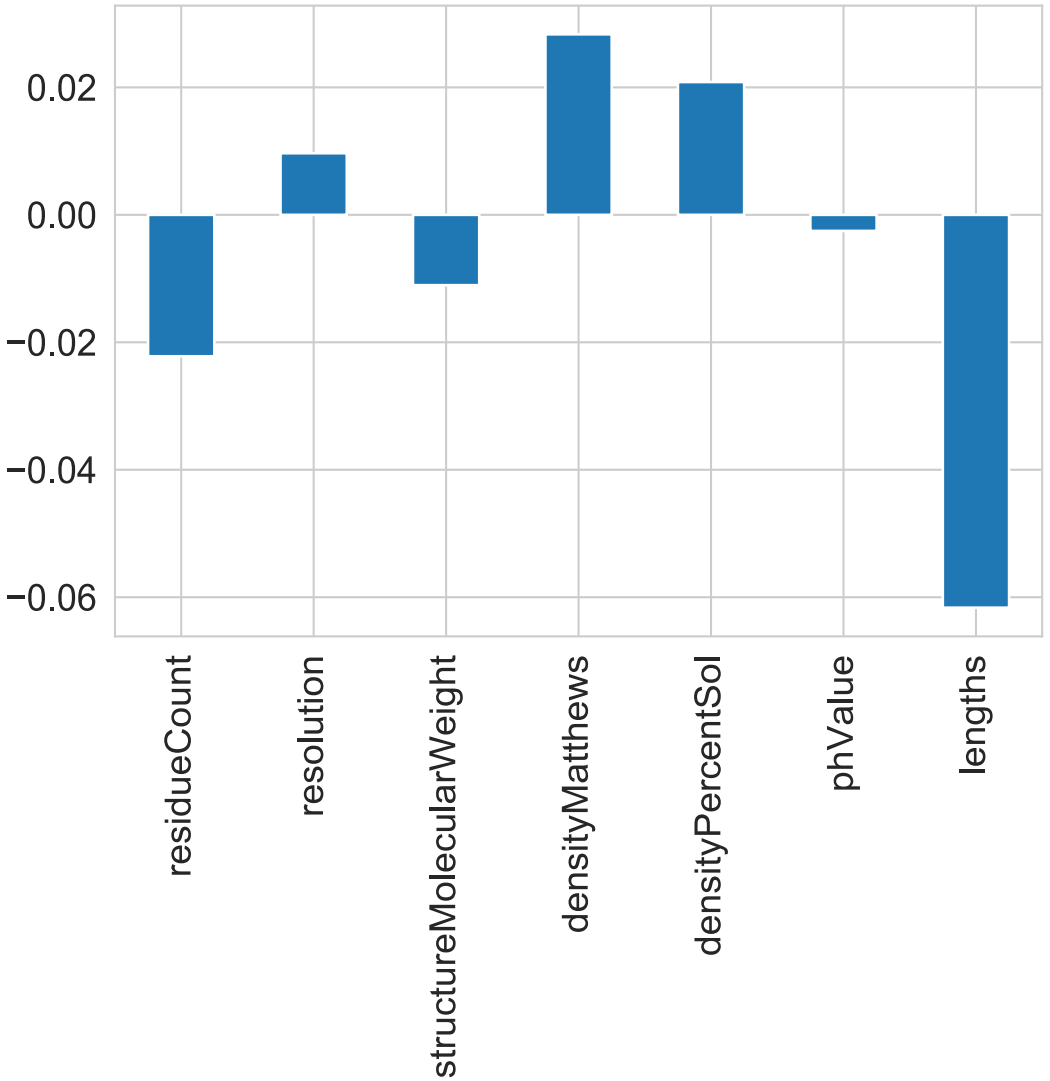
Majority of the sequences are not longer than 600 units.

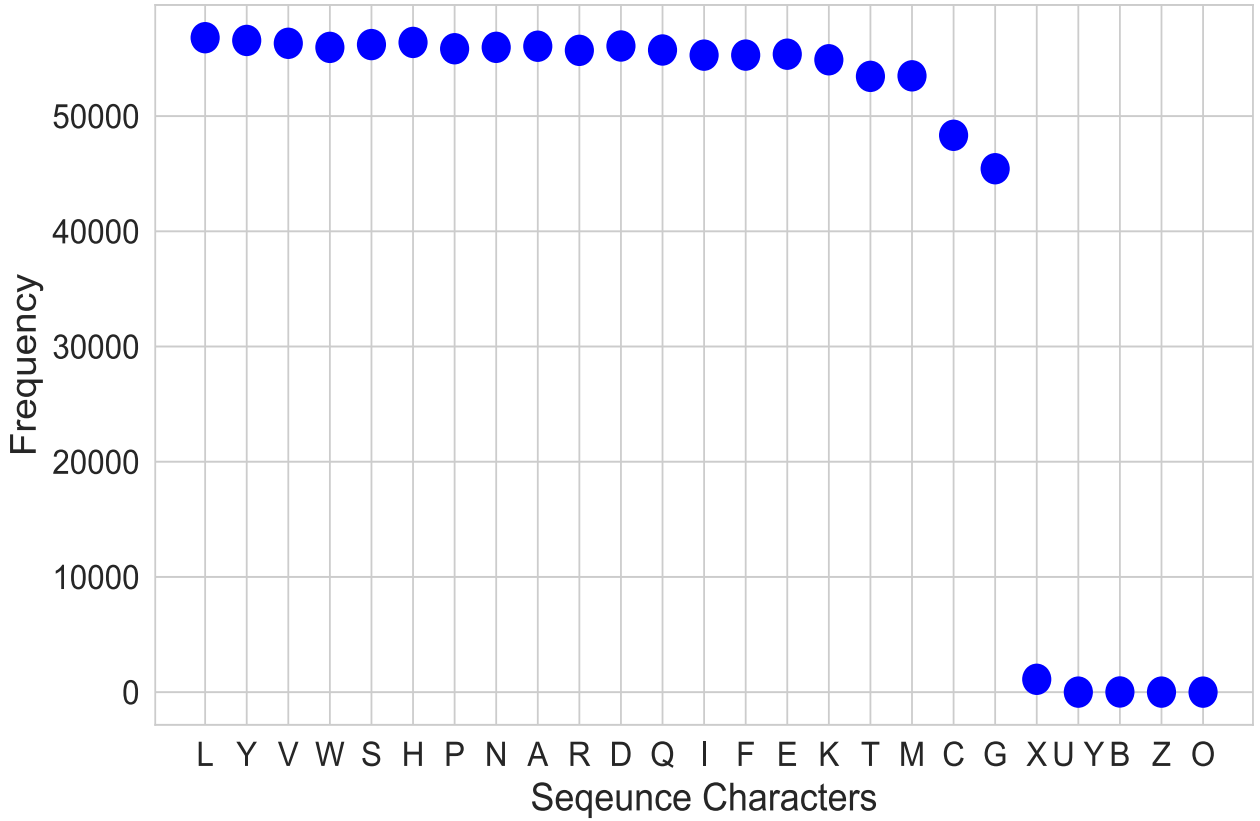


Are Residue counts same as the sequence lengths?



Exploratory Data Analysis:

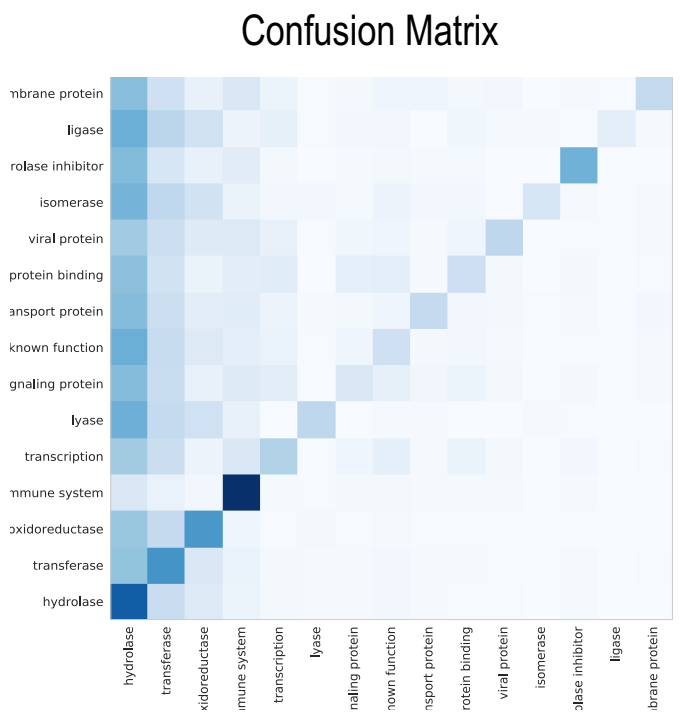


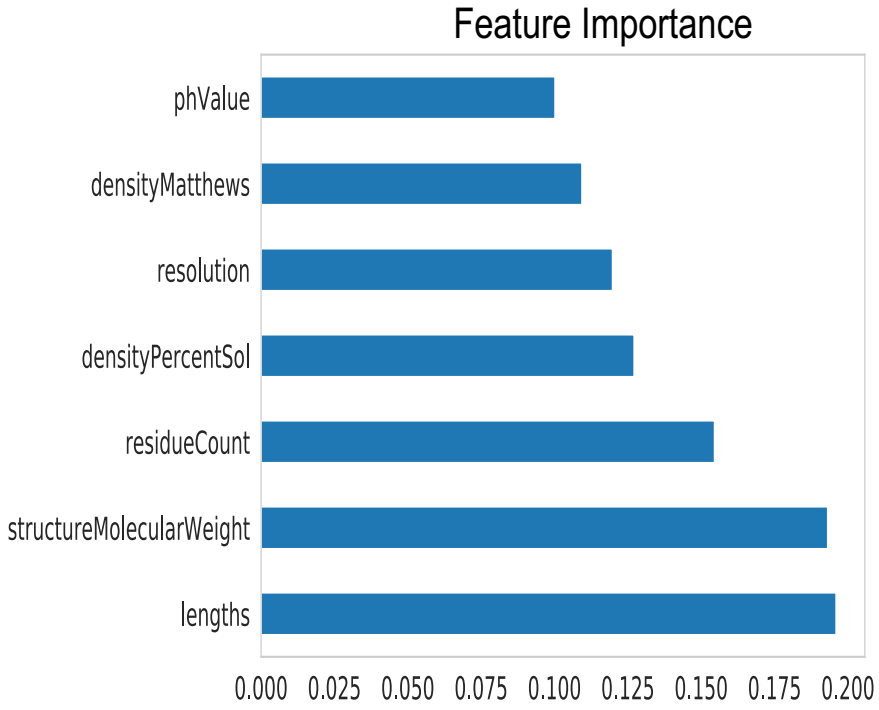


Machine Learning Models:

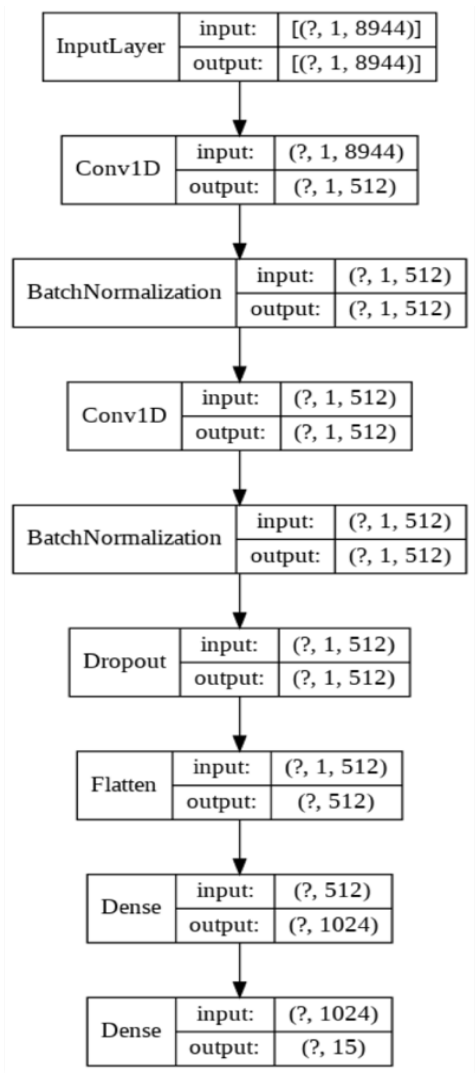
Models:	Test Accuracy
Random Forest	0.40
Decision Tree	0.31
Gradient Boosting	0.21
Gaussian Naïve Bayes	0.12
Multinomial NB	0.09
Support Vector Machine	0.11
Logistic Regression classifier (OneVsRest)	0.158
K-nearest neighbor	0.22

Comparing to the listed machine learning models, Random Forest model performed better. However, the RF model accuracy is less than 50%

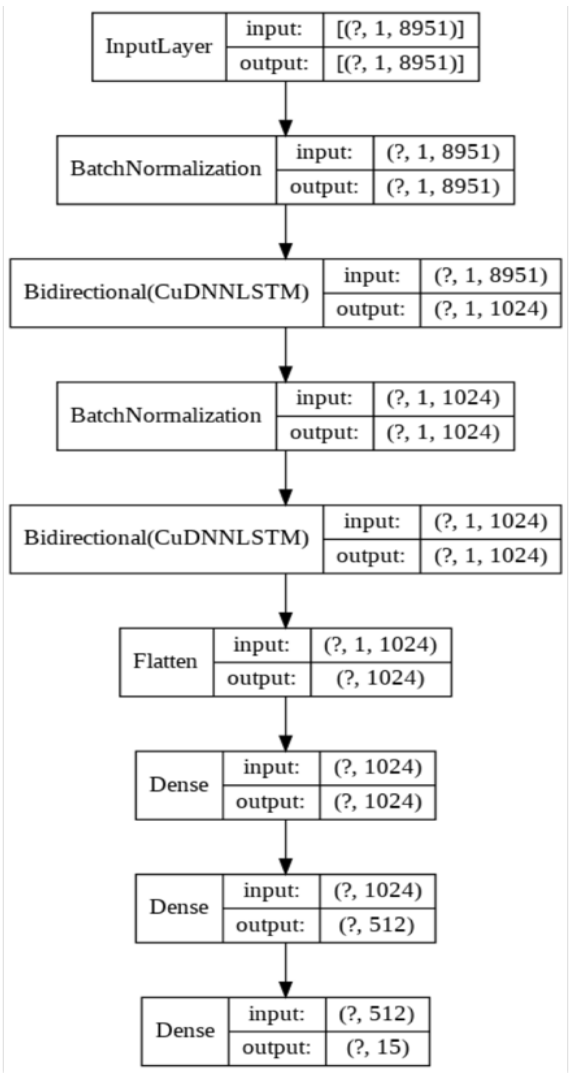




Deep Learning Models: Architectures

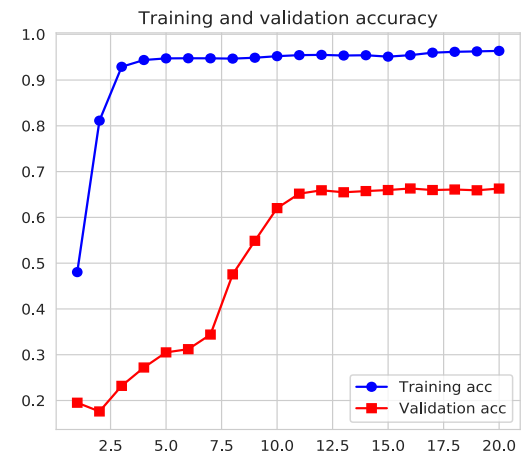


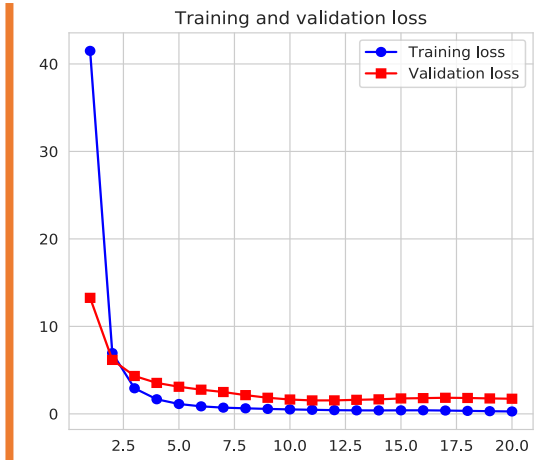
CNN (Conv1D)

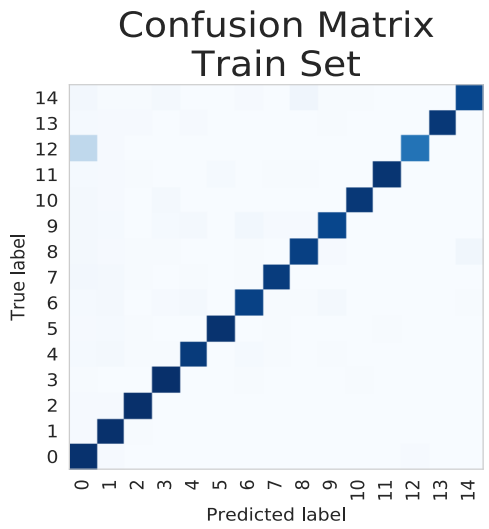


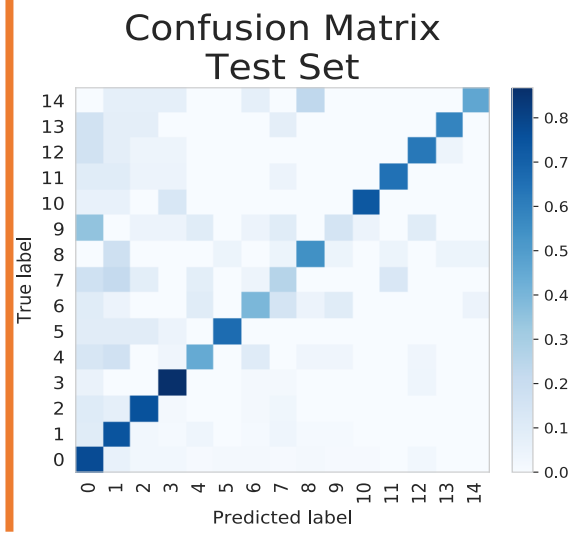
RNN (BiLSTM)

Deep Learning Models: Model Evaluation

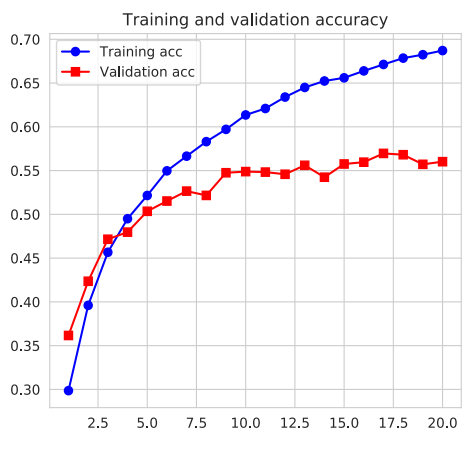


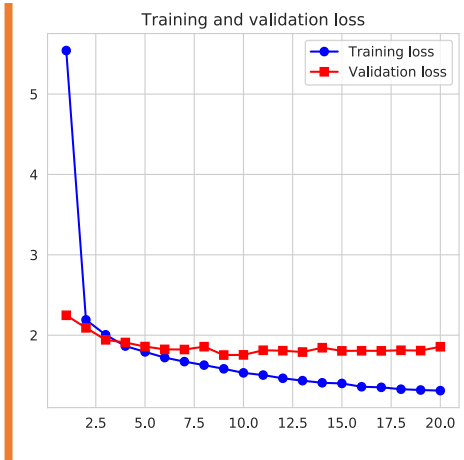


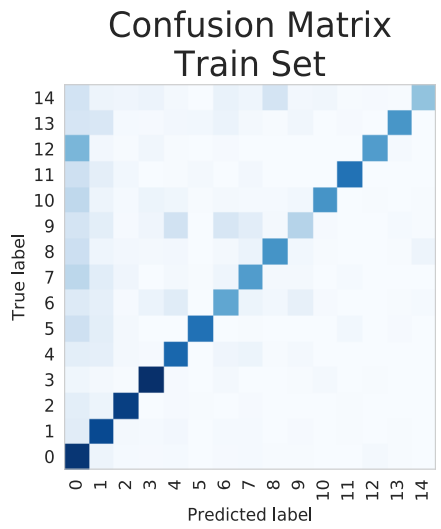


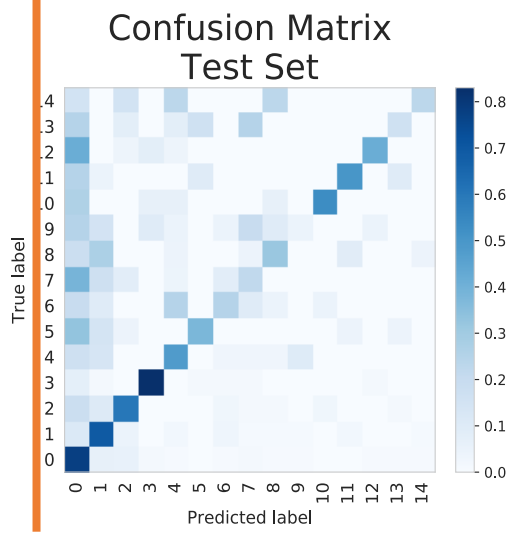


CNN (Conv1D)









RNN (BiLSTM)

Summary:

Machine learning model used to predict the protein family observing the physical and chemical features, but available features were not enough to correctly predict the protein family.

Features extracted from amino acid sequences were implemented to deep learning models.

One dimensional convolutional neural network (Conv1D) performed better than the complex bidirectional LSTM neural network (BiLSTM).