TEAM: The Phospho Force

Protein Function Prediction

(Differentiating Kinases for Targeted Drug Discovery)

Dhanashree Lokesh, Yumeng Li, Lara Brindisi

THE ERDŐS INSTITUTE
Helping PhDs get jobs they love at every stage of their career.

Data science Boot Camp: Spring 2023

Why Kinases?

- Central Role in Cell Signalling
- Abundance and Diversity
- Druggability
- Potential for Combination Therapies
- Biomarkers and Pharmacodynamics
- Versatility and Adaptability

Stakeholders

- Pharmaceutical & Biotechnology Companies
- Academic and Research Institutions
- Regulatory Bodies
- Investors and Funding Agencies
- Government and Policy Makers
- Ethics Committees

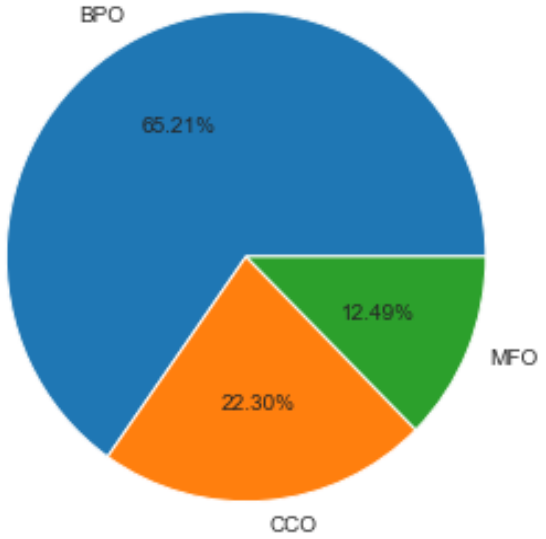
Objective

- **Predict the GO term based on the sequence**
- **A classification problem**

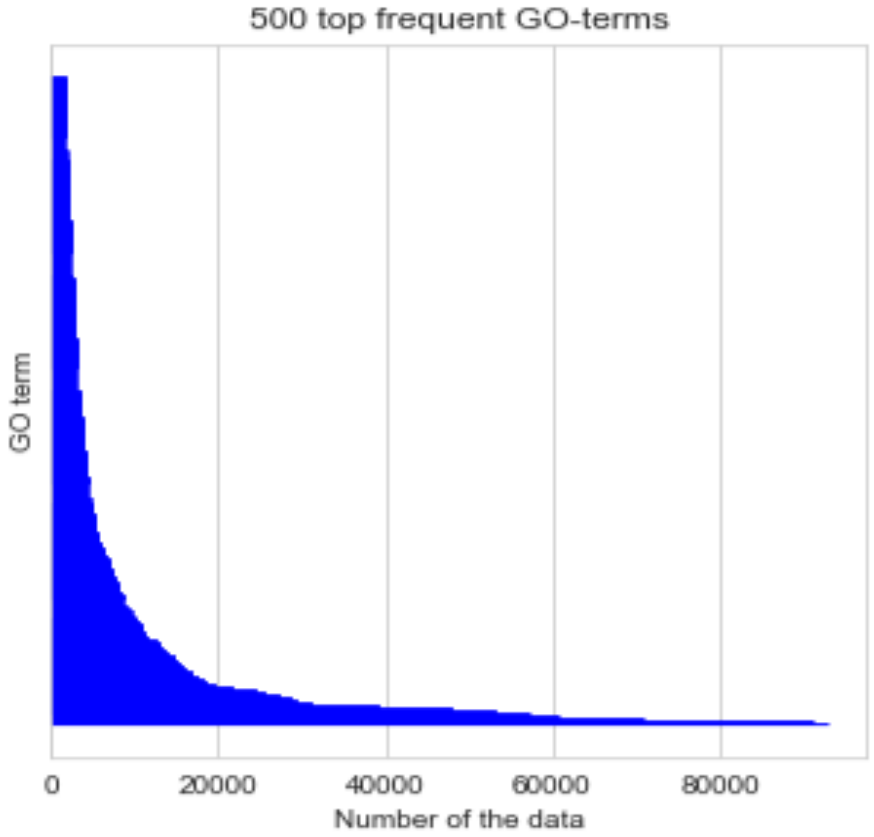
Data

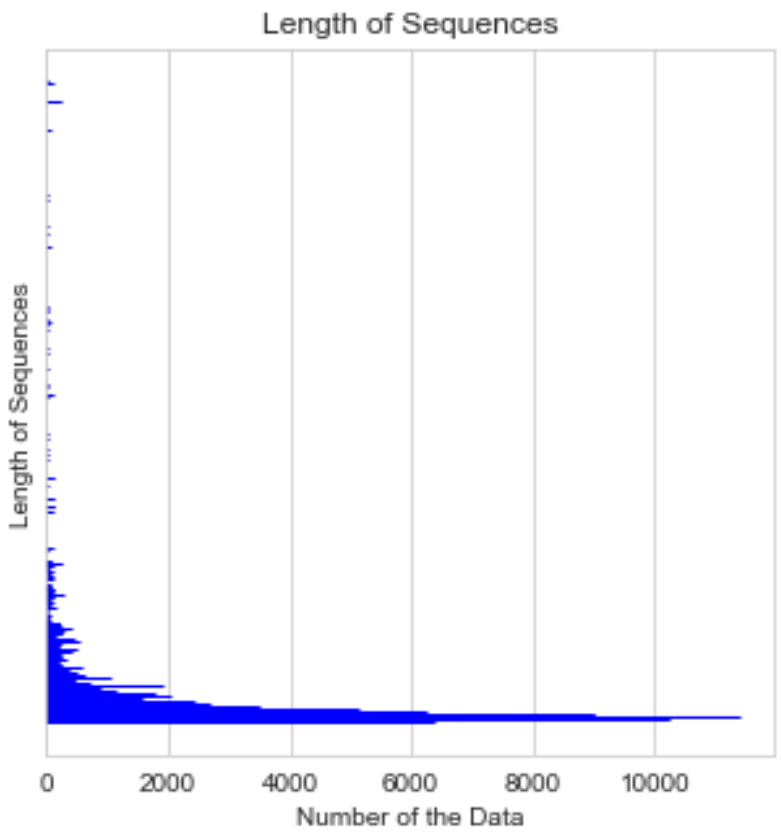
- **5,363,863 proteins (= # of rows)**
- **Length of sequence ranges from 3 to 11391 with a focus under 1500.**
- **31,466 different GO terms (classes)**

	seq_id	sequence	term	aspect
0	P20536	MNSVTVSHAPYTITYHDDWEPVMSQLVEFYNEVASWLLRDETSPIP...	0008152	BPO
1	P20536	MNSVTVSHAPYTITYHDDWEPVMSQLVEFYNEVASWLLRDETSPIP...	0071897	BPO
2	P20536	MNSVTVSHAPYTITYHDDWEPVMSQLVEFYNEVASWLLRDETSPIP...	0044249	BPO
3	P20536	MNSVTVSHAPYTITYHDDWEPVMSQLVEFYNEVASWLLRDETSPIP...	0006259	BPO
4	P20536	MNSVTVSHAPYTITYHDDWEPVMSQLVEFYNEVASWLLRDETSPIP...	0009059	BPO



Exploratory Data Analysis





Data Setup

term	name	namespace	EntryID	aspect	sequence
GO:0033549	MAP kinase phosphatase activity	molecular_function	P35182	MFO	MSNHSEILERPETPYDITYRVGVAENKNSKFRRTMEDVHTYVKNFA...
GO:1990439	MAP kinase serine/threonine phosphatase activity	molecular_function	P35182	MFO	MSNHSEILERPETPYDITYRVGVAENKNSKFRRTMEDVHTYVKNFA...
GO:0004672	protein kinase activity	molecular_function	Q0KHV6	MFO	MSVRLLTVRLIKHGRYILRSYCKRDIHANILDQNQLKTRSKRGFPL...
GO:0004674	protein serine/threonine kinase activity	molecular_function	Q0KHV6	MFO	MSVRLLTVRLIKHGRYILRSYCKRDIHANILDQNQLKTRSKRGFPL...
GO:0016301	kinase activity	molecular_function	Q0KHV6	MFO	MSVRLLTVRLIKHGRYILRSYCKRDIHANILDQNQLKTRSKRGFPL...

Pre-processing:

- Random subsample (n=2000, 10000, 60000)
- Select 100 most frequent GO terms
- Filter for kinases

Classification Models:

- Random Forest
- SVM
- Keras Neural Network

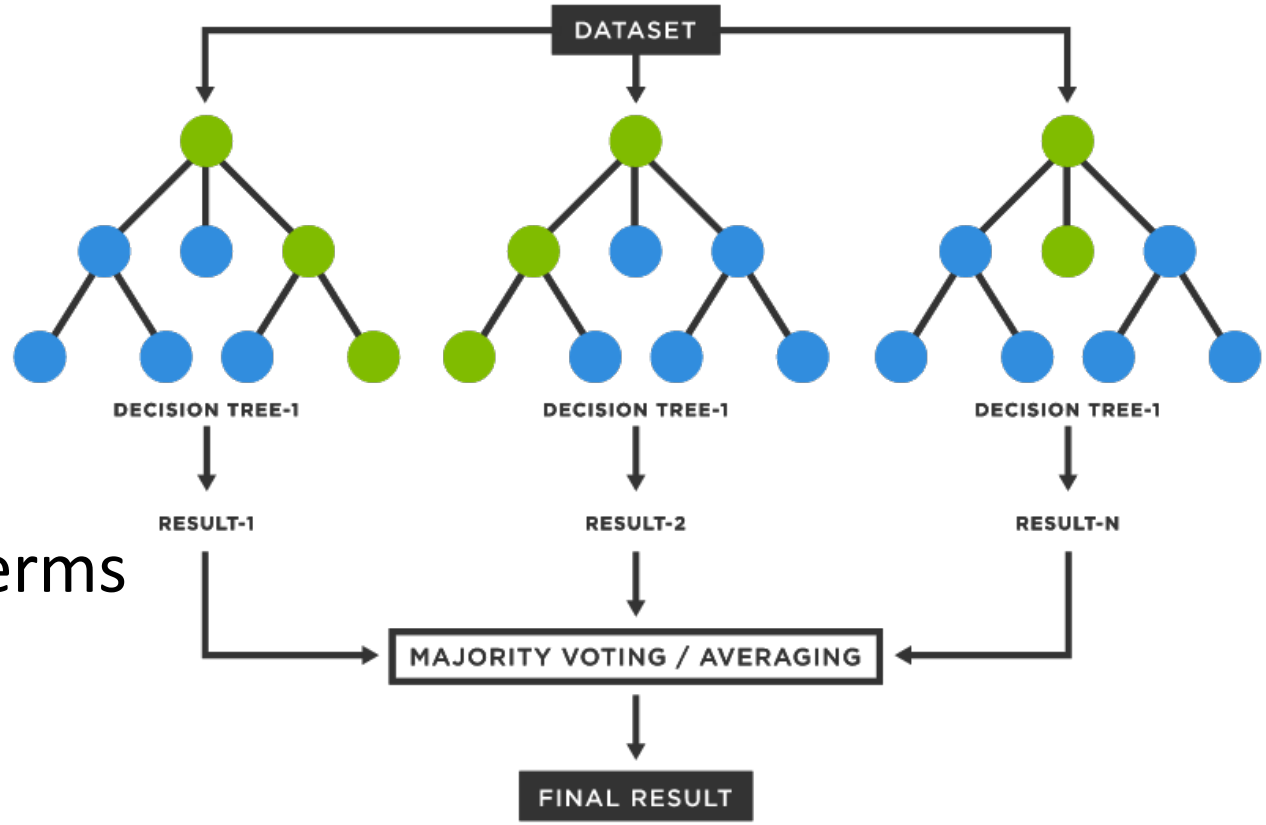
Random Forest

Limitations:

- Computational resources
- Time

Solution:

- Select 100 most frequent GO terms
- Truncate sequence
- Encode target labels with Label_Encoder()
- Accuracy = 2.1%



https://www.tibco.com/reference-center/what-is-a-random-forest

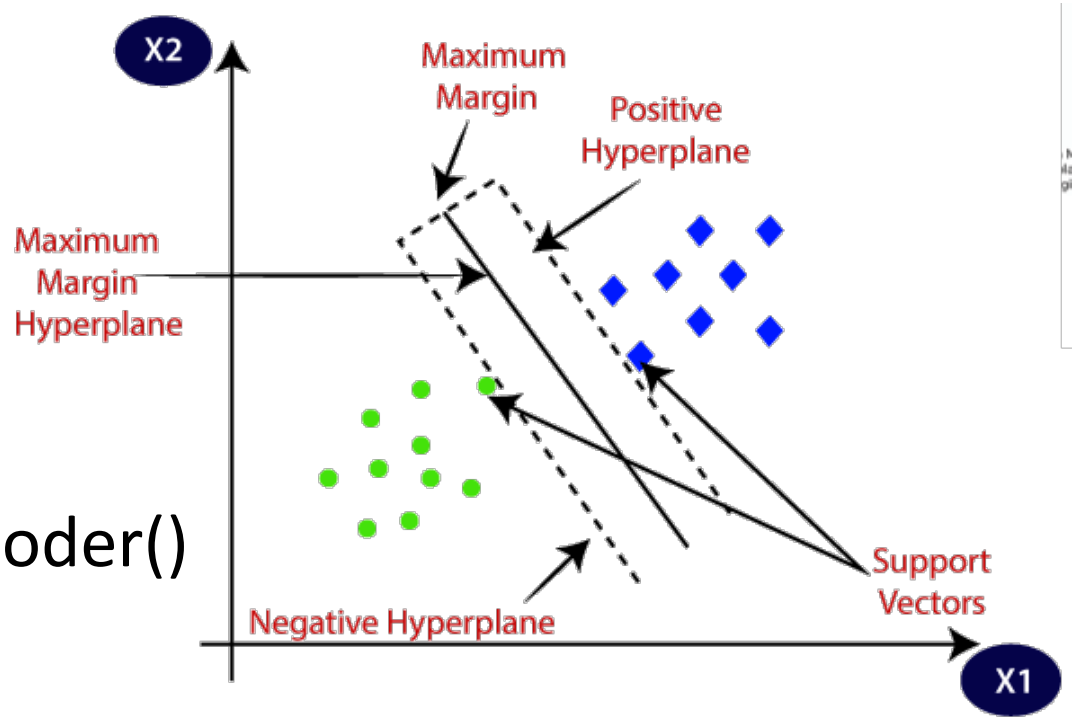
SVM

Limitations:

- Computational resources
- Time

Solution 1:

- Select 100 most frequent GO terms
- Truncate sequence
- Encode target labels with Label_Encoder()
- Accuracy = 2.1%



Solution 2:

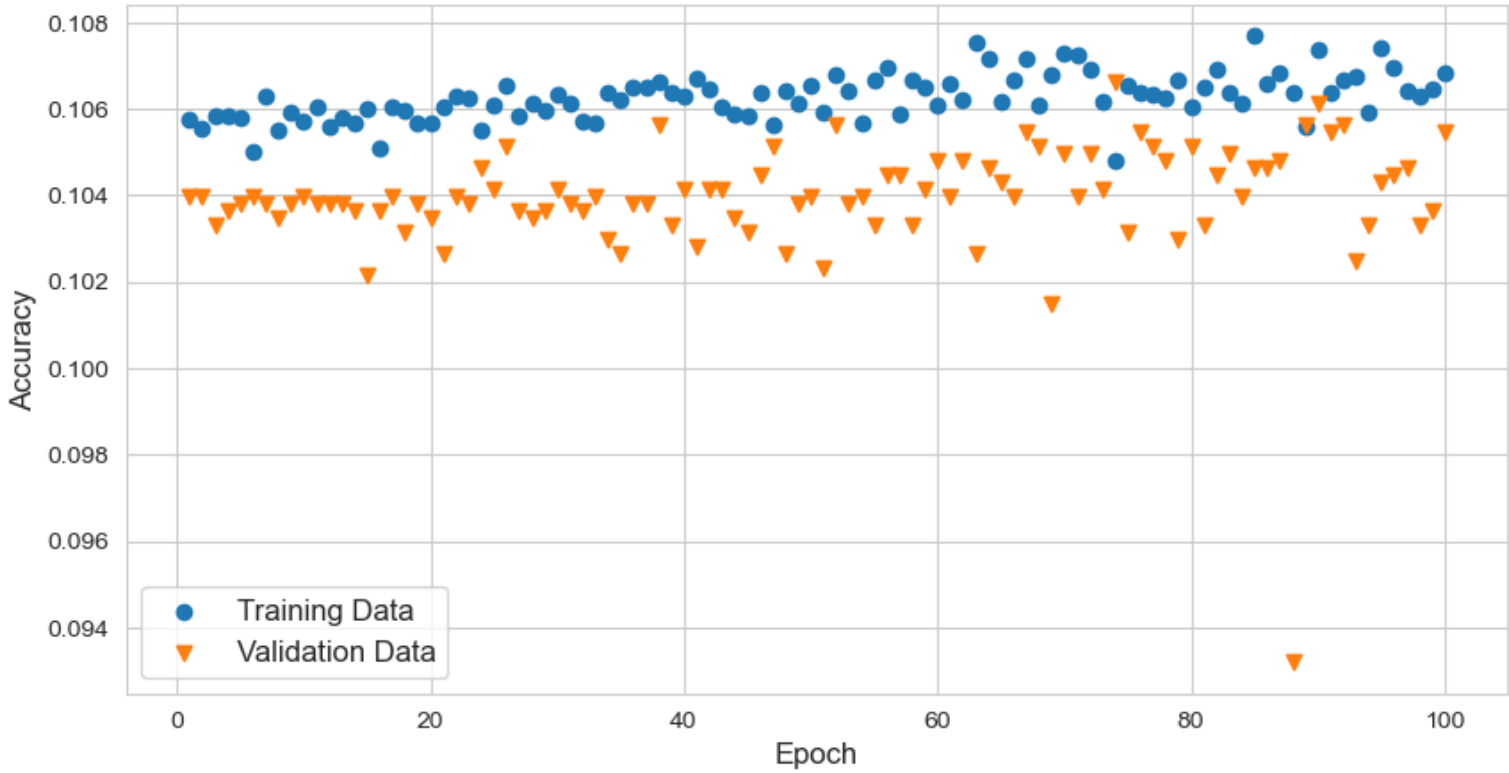
- Filter for kinases
- Vectorize with TfidfVectorizer()
- Accuracy = 8.9%

https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

Keras

- **Limitations:**
  - Computational resources
  - Time
  - For ex)
    - Epochs = 100
    - Subsample_size = 10,000
    - Batch_size = 32

    --> 6.3 years to run



- **Solution:**
  - K-mer numeric representation
  - Filter for kinases
  - Accuracy = 10-11%
  - Choose less expensive layers

    ex) GlobalAveragePooling1D over LSTM

Conclusion

- Keras neural network performed best at 10-11%
- Optimization would require greater memory capacity
- Future prospective:
  - Optimize:
    - Layer choice
    - Batch size
    - Encoding strategy
  - Identify signature sequences on predicted kinases
  - Differentiate prokaryotic and eukaryotic kinases
  - Comparative proteomics studies