SIMPLE DNA SEQUENCE SPLICE CLASSIFIER

- An attempt to classify the different sections of DNA chain for rapid analysis

Lakshmi Nainar

Date: 06-DEC-2021

Agenda and Expected Outcomes

- Definition of DNA
- Importance of DNA
- Relevance of ML & AI in DNA
- A simple technique attempted for classification
- Inference of the experiment
- Conclusion and future work

Why this topic?

- Study of DNA - important to understand human body and treat diseases
- Understanding biomarkers – detection and prediction of body conditions
- Precision medicine
- Computational complexity – can be solved by ML and AI models
- Rapid analysis and insight gathering for continuous prototyping
- Aid in research and development of treatment plans, drugs, and management of cell behavior

DNA – deoxyribonucleic acid

Chemical Molecule

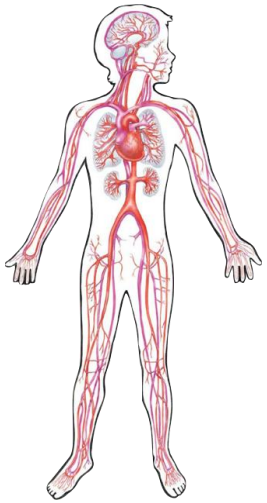
Genetic Instructions carrier(genus and species)

Hereditary Signature

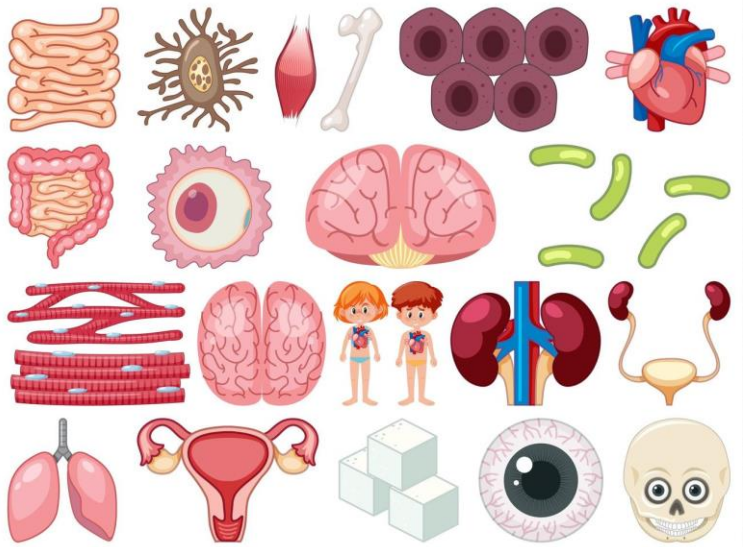
Malformation in DNA → Mutation and Diseases

Behavior of body functions → DNA script: ON and OFF, Intensity of gene expression

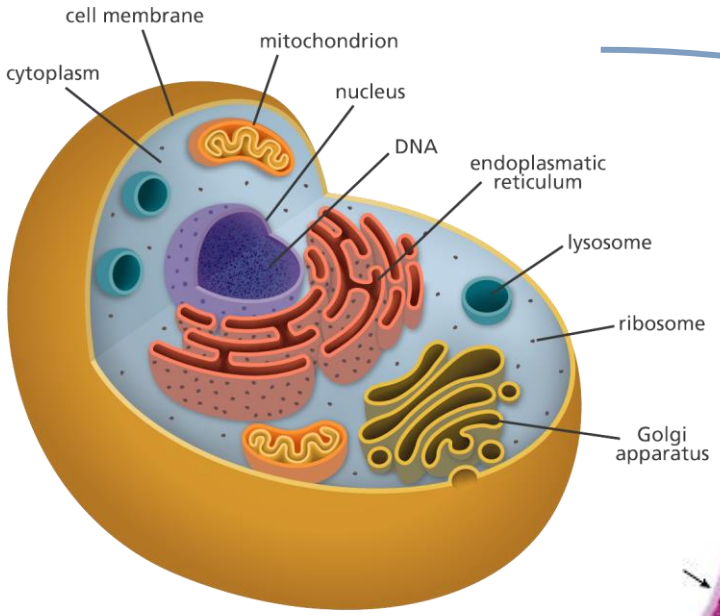
DNA Representation



Human Body

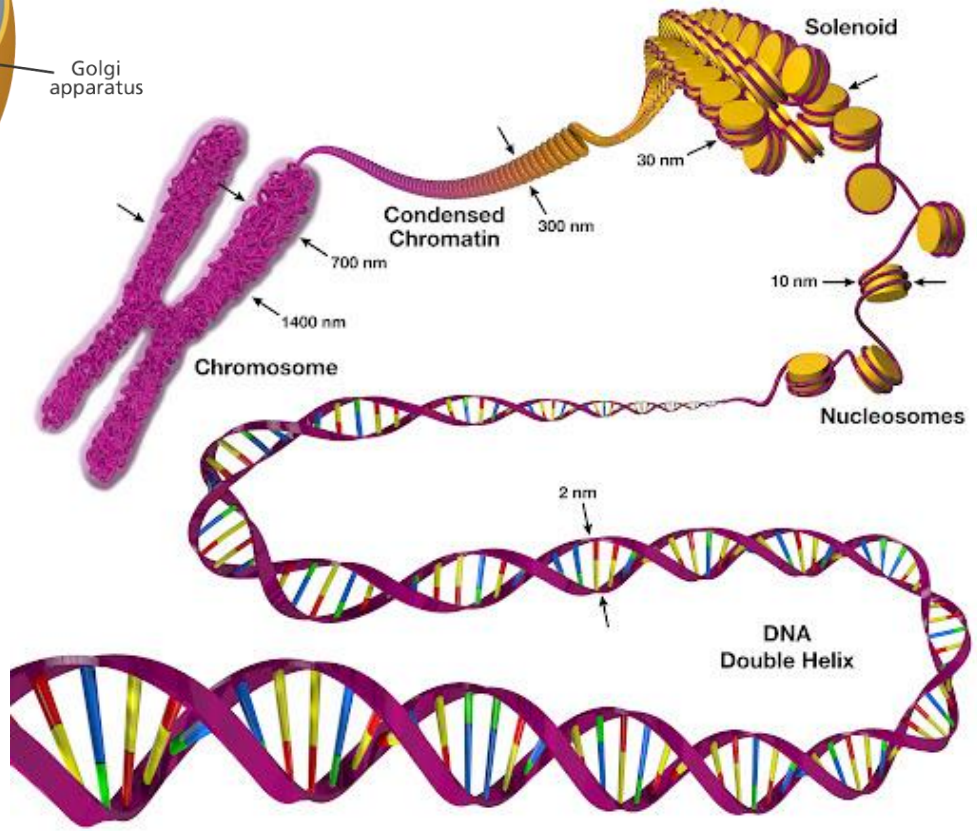


Multicellular organs

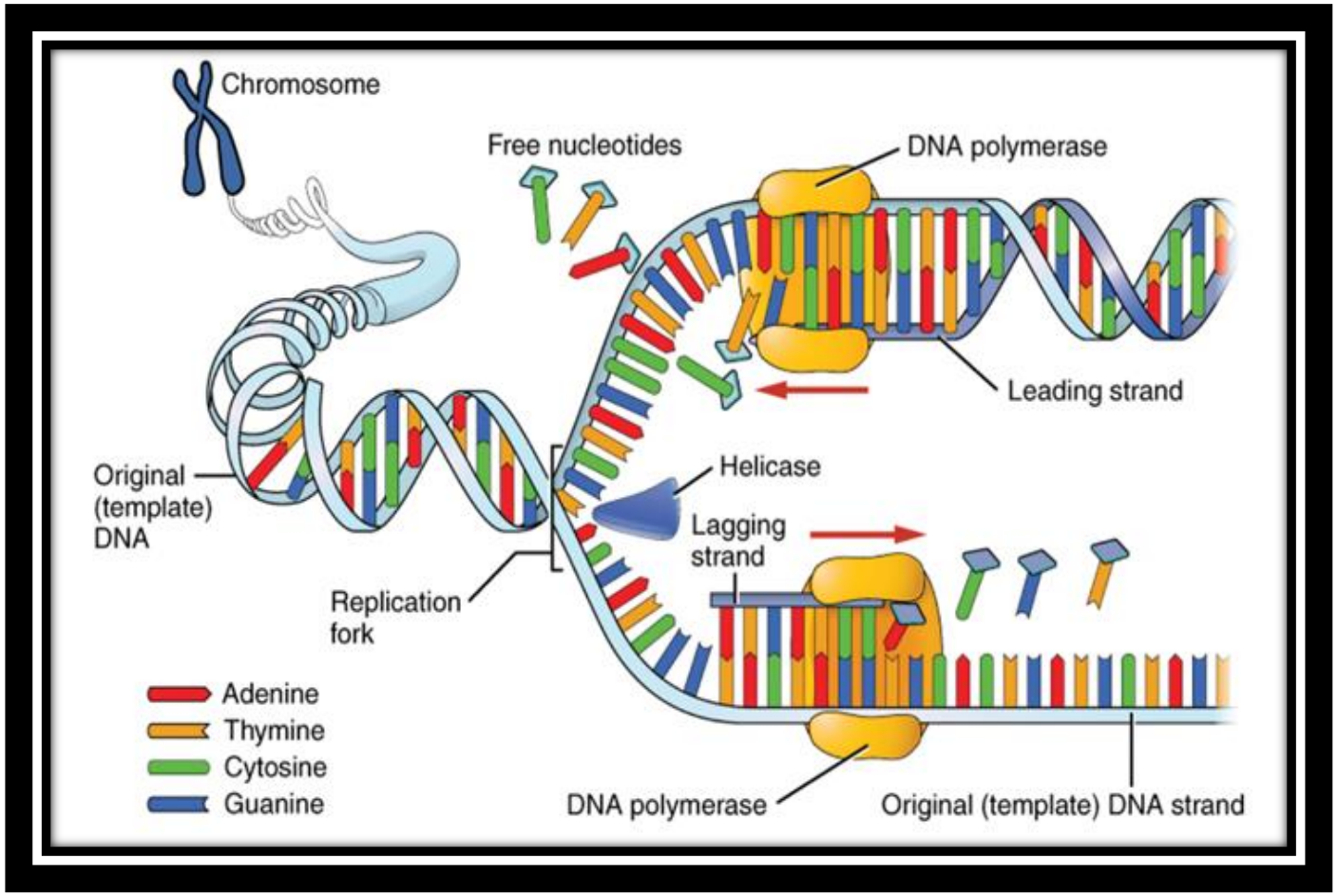


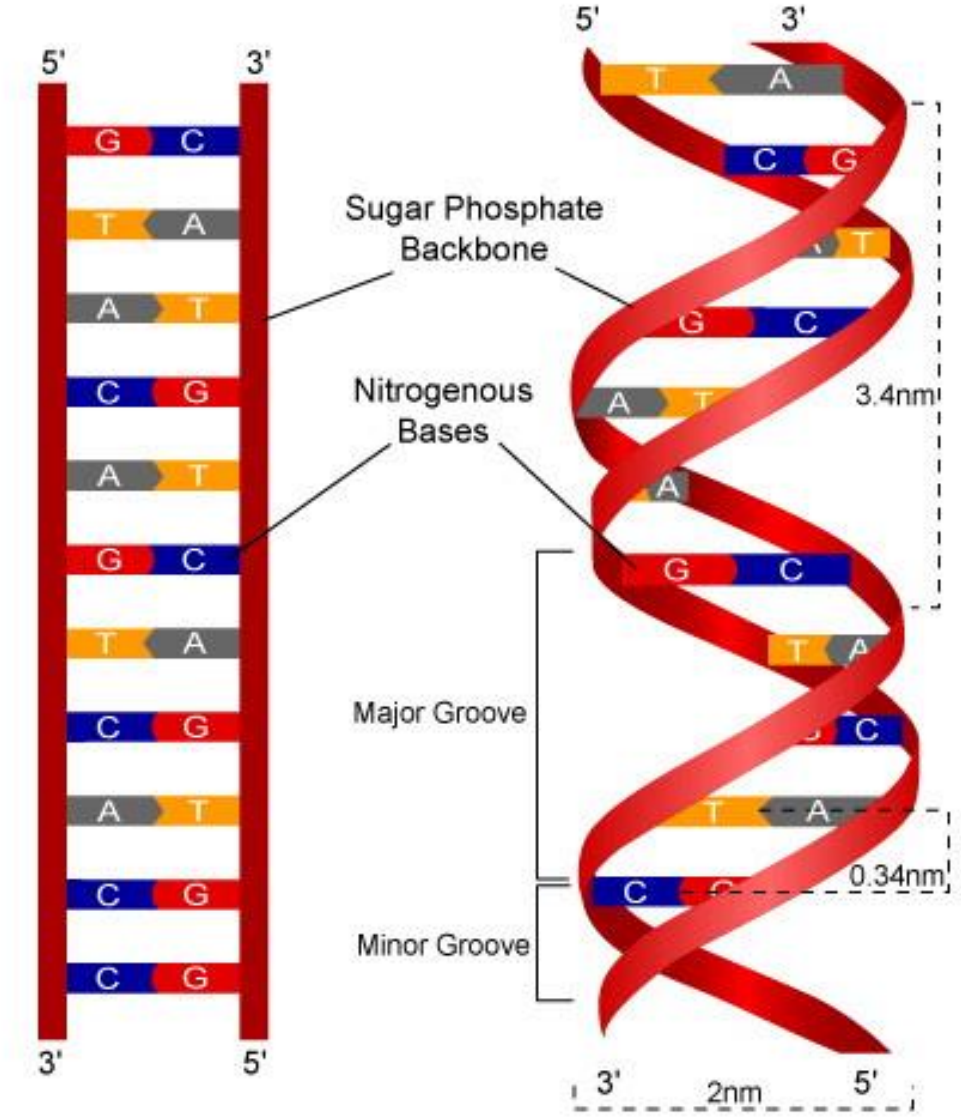
Cell

Inside the nucleus



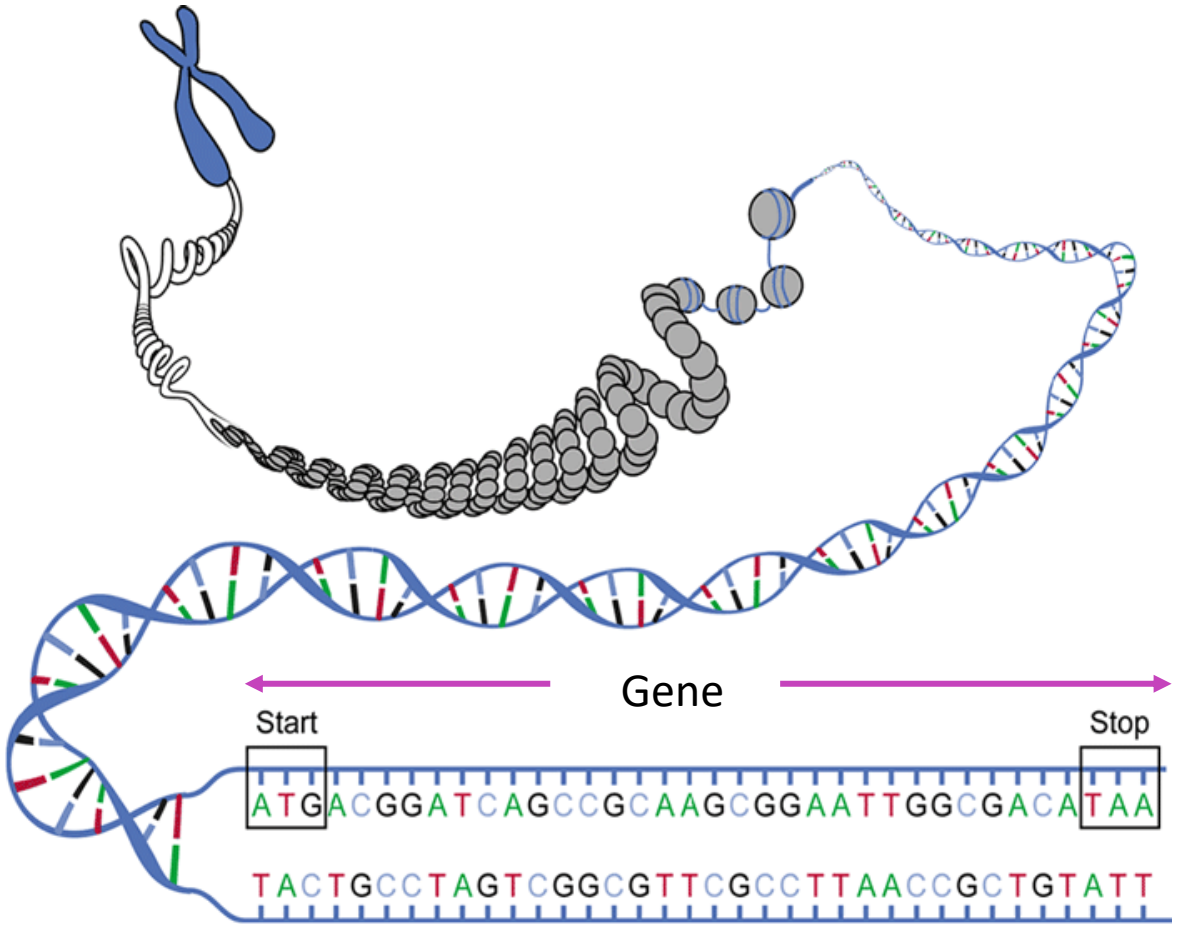
DNA Representation





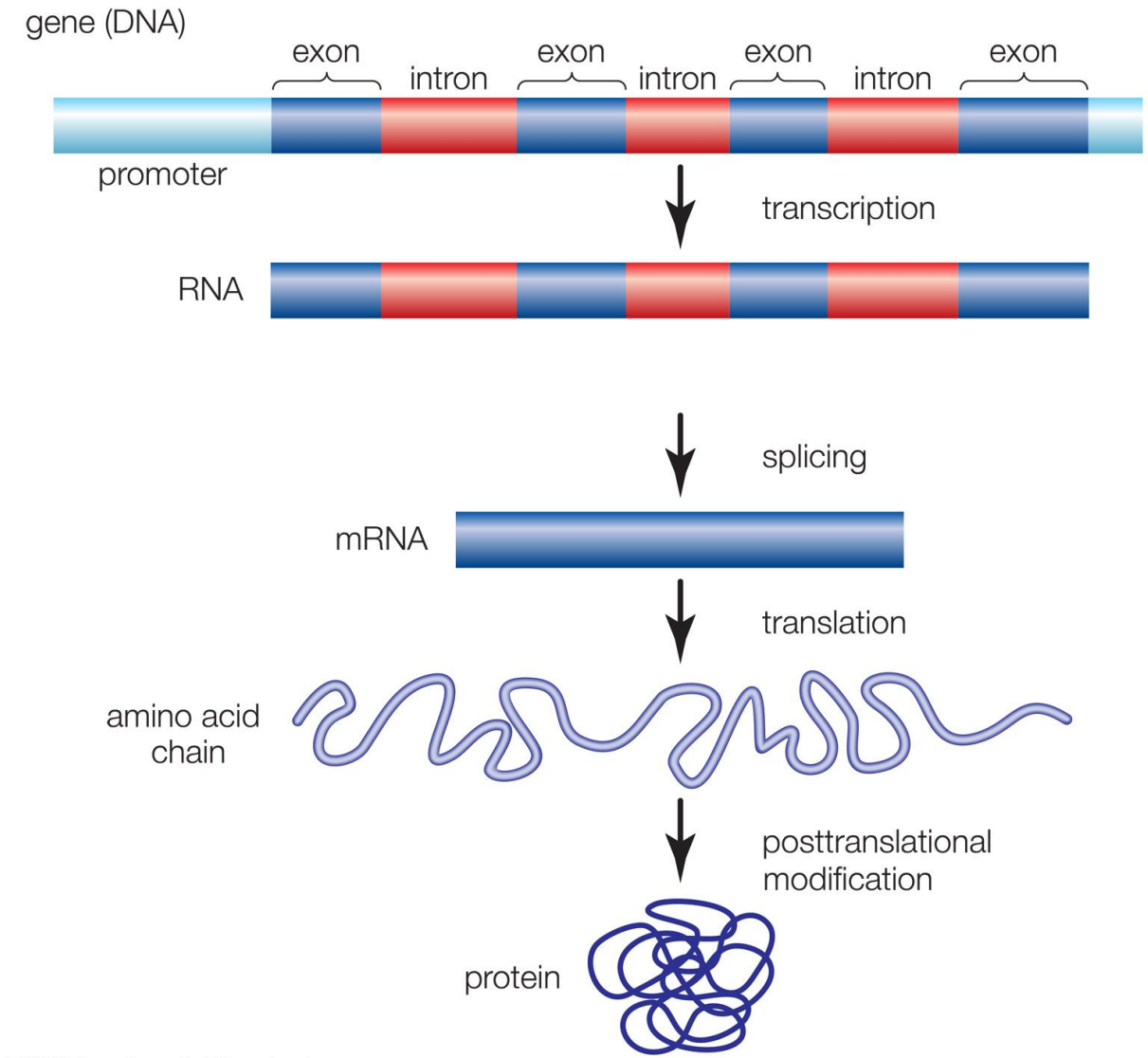
Sequence Representation in Text: GTACAGTCACC

Gene



Gene:

- Segments of DNA sequence
- Code for a particular trait
- Undergoes transcription –
  - DNA sequence → RNA sequence → mRNA → Protein



EXONS:
- Subsequence that codes a protein

INTRONS:
- Non-coding subsequences

PROMOTER:
- ON/OFF gene expression

TRANSCRIPTION:
- DNA → RNA process

RNA:
- Single stranded nucleotide sequence

SPLICING:
- Splitting of introns and exons

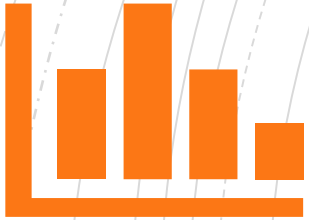
mRNA :
- Messenger RNA – protein coding instructions

TRANSLATION
- Instruction → Production

AMINO ACID CHAIN
- Unclassified composition of protein composing acids

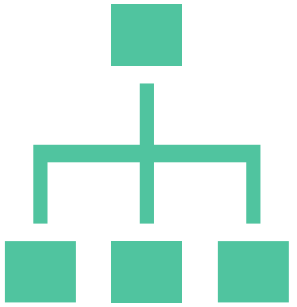
PROTEIN:
- Chemical coding for body activities/biological components

Project Description and Goal



Dataset of exons, introns, and none coding DNA sequences

Represented as a sequence of DNA nucleotides



Classification classes:

exon-exon

intron-exon

Normal sequence

Purpose:

Smooth processing of large DNA sequence data

Separation and categorization of DNA data for further analysis

Scope for research and study

Project Details:

Language: Python 3.10

Packages:

- NumPy
- Pandas
- Sci-kit learn

IDE: Jupyter Notebook

Platform: Local computer

Algorithms:

- Multinomial Naïve Bayes:
  - Probabilistic-learning based NLP algorithm for text processing
- K-Mer Counting:
  - What? – Algorithm to process data by splitting them into sequences of size k
  - Why? – DNA sequence splitting and bagging
- Count Vectorizer:
  - What? – Algorithm to vectorize text on frequency basis
  - Why? – Pattern identification and analysis of nucleotides

Dataset:

Point_of_Connectivity	Gene_ID	DNA_Sequence
EI	ATRINS-DONOR-521	CCAGCTGCATCACAGGAGGCCAGCGAGCAGGTCTGTTCCAAGGGCCTTCGAGCCAGTCTG
EI	ATRINS-DONOR-905	AGACCCGCCGGGAGGCGGAGGACCTGCAGGGTGAGCCCCACCGCCCCTCCGTGCCCCCGC
EI	BABAPOE-DONOR-30	GAGGTGAAGGACGTCCTTCCCCAGGAGCCGGTGAGAAGCGCAGTCGGGGGCACGGGGATG
EI	BABAPOE-DONOR-867	GGGCTGCGTTGCTGGTCACATTCCTGGCAGGTATGGGGCGGGGCTTGCTCGGTTTTCCCC
EI	BABAPOE-DONOR-2817	GCTCAGCCCCCAGGTCACCCAGGAACTGACGTGAGTGTCCCCATCCCGGCCCTTGACCCT
EI	CHPIGECA-DONOR-378	CAGACTGGGTGGACAACAAAACCTTCAGCGGTAAGAGAGGGCCAAGCTCAGAGACCACAG
EI	CHPIGECA-DONOR-903	CCTTTGAGGACAGCACCAAGAAGTGTGCAGGTACGTTCCCACCTGCCCTGGTGGCCGCCA
EI	CHPIGECA-DONOR-1313	CCCTCGTGCGGTCCACGACCAAGACCAGCGGTGAGCCACGGGCAGGCCGGGGTCGTGGGG
EI	GCRHBBA1-DONOR-1260	TGGCGACTACGGCGCGGAGGCCCTGGAGAGGTGAGGACCCTCCTGTCCCTGCTCCAGTCC
EI	GCRHBBA1-DONOR-1590	AAGCTGACAGTGGACCCGGTCAACTTCAAGGTGAGCCAGGAGTCGGGTGGGAGGGTGAGA
EI	GCRHBBA6-DONOR-461	TGGCGACTACGGCGCGGAGGCCCTGGAGAGGTGAGGACCCTGGTATCCCTGCTGCCAGTC
EI	GCRHBBA6-DONOR-795	AAGCTGAGAGTGGACCCTGTCAACTTCAAGGTGAGCCACCAGTCGGGTGGGGAGGGTGAG
EI	GIBHBGGL-DONOR-2278	GGAAGATGCTGGAGGAGAAACCCTGGGAAGGTAGGCTCTGGTGACCAGGACAAGGGAGGG
EI	GIBHBGGL-DONOR-2624	AAGCTGCATGTGGATCCTGAGAACTTCAGGGTGAGTACAGGAGATGTTTCAGCCCTGTTG
EI	GIBHBGGL-DONOR-7198	GGAAGATGTTGGAGGAGAAACCCTGGGAAGGTAGGCTCTGGTGACCAGGACAAGGGAGGG
EI	GIBHBGGL-DONOR-7544	AAGCTGCATGTGGATCCTGAGAACTTCAGGGTGAGTACAGGAGATGTTTCAGCCCTGTTG
EI	HUMA1ATP-DONOR-1972	GGCACCACCACTGACCTGGGACAGTGAATCGTAAGTATGCCTTTCACTGCGAGGGGTTCT

K-Mer Cluster Process

	Point_of_Connectivity	Gene_ID	DNA_Sequence	subsequences
0	EI	ATRINS-DONOR-521	CCAGCTGCATCACAGGAGGCCAGCGAGCAGGTCTGTTCCAAGGGCC...	[ccagct, cagctg, agctgc, gctgca, ctgcat, tgcat...
1	EI	ATRINS-DONOR-905	AGACCCGCCGGGAGGCGGAGGACCTGCAGGGTGAGCCCCACCGCCC...	[agaccc, gacccg, acccgc, cccgcc, ccgccg, cgccg...
2	EI	BABAPOE-DONOR-30	GAGGTGAAGGACGTCCTTCCCCAGGAGCCGGTGAGAAGCGCAGTCG...	[gaggtg, aggtga, ggtgaa, gtgaag, tgaagg, gaagg...
3	EI	BABAPOE-DONOR-867	GGGCTGCGTTGCTGGTCACATTCCTGGCAGGTATGGGGCGGGGCTT...	[gggctg, ggctgc, gctgcg, ctgcgt, tgcgtt, gcgtt...
4	EI	BABAPOE-DONOR-2817	GCTCAGCCCCCAGGTCACCCAGGAACTGACGTGAGTGTCCCCATCC...	[gctcag, ctcagc, tcagcc, cagccc, agcccc, gcccc...
...	...	...	...	...
3185	N	ORAHBPSBD-NEG-2881	TCTCTTCCCTTCCCCTCTCTCTTTCTTTCTTTTCTCTCCTCTTCTC...	[tctctt, ctcttc, tcttcc, cttccc, ttccct, tccct...
3186	N	ORAINVOL-NEG-2161	GAGCTCCCAGAGCAGCAAGAGGGCCAGCTGAAGCACCTGGAGAAGC...	[gagctc, agctcc, gctccc, ctccca, tcccag, cccag...
3187	N	ORARGIT-NEG-241	TCTCGGGGGCGGCCGGCGCGGCGGGGAGCGGTCCCCGGCCGCGGCC...	[tctcgg, ctcggg, tcgggg, cggggg, gggggc, ggggc...

Project inference and result

- Accuracy: Maximum Accuracy of 70.4%
- K = 6 for K-Mer and alpha=0.5 for MNB
- Non-linear decrease in the precision and accuracy on varying K and alpha magnitudes
- Error rate for other algorithms after adopted a ten-fold validation methodology

Algorithm	Neither	EI	IE
KBANN	4.62	7.56	8.47
BACKPROP	5.29	5.74	10.75
PEBLS	6.86	8.18	7.55
PERCEPTRON	3.99	16.32	17.41
ID3	8.84	10.58	13.99
COBWEB	11.8	15.04	9.46
NEAR NEIGHBOR	31.11	11.65	9.09

LIMITATIONS AND COMPLEXITIES

- Repositories handled by government and federal organizations
- Permission based proprietary databases
- Humungous amount of data
- Limited computational ability and resources
- Manual integration of available data not possible within the semester duration

Future work/Intended Direction

- Field under research
- CRISPR, DeepMind's Alphafold
- Algorithm/Methodology:
  - Neural Networks and Deep Learning based Models
    - Convolutional Neural Network
    - Recurrent Neural Network
- DeepDBP-ANN and DeepDBP –CNN(proposed in scientific paper(https://doi.org/10.1016/j.imu.2020.100318)

Future work/Intended Direction

- Data after classification can be used to synthesize protein, protein type and utilize in specific research purposes
- Research on pattern modification(DNA methylation) to analyze their behavior and develop reparative treatment plans or induction methodologies(example, CRISPR)
- Discover better methods of research – for example, the kind of test subjects to use for each use case

THANK YOU

REFERENCES:

- https://www.youtube.com/watch?v=uxl3_8yvbxi
- https://www.ncbi.nlm.nih.gov/
- https://archive.ics.uci.edu/ml/datasets/Molecular+Biology+(Splice-junction+Gene+Sequences)
- https://www.sciencedirect.com/science/article/pii/S2352914819304307