Name:	Dave Mannion
Course:	UCSD Machine Learning & Engineering
Date:	October 23, 2022

Bootcamp Capstone Report

Gene Family Prediction & ML

Machine Learning and DNA Sequence Classification

Dave Mannion

Contents

Gene Family	**Count**	**Class Label**
G protein coupled receptors	531	1
Tyrosine kinase	534	2
Tyrosine phospotase	349	3
Synthetase	632	4
Synthase	711	5
Ion channel	240	6
Transcription factor	1341	7

Table 1: Gene Families represented by Human DNA coding sequences.

1 Introduction

Here we examine NLP algorithms to develop a Machine Learning model for making predictions on sequences of Nucleic Acid text. Specifically, we train a Multinomial Naïve Bayes classifier on Human DNA Coding Sequences to predict Gene Families. In addition to Human sequences, predictive testing was also performed in Chimpanzee and Dog DNA coding sequences. Table 1 above shows Gene Families in the dataset.

1.1 Motivation

This Capstone project was selected to provide an opportunity to explore the application of NLP methods to biological text-based sequence data.

1.2 Exploratory Work

Prior to this Capstone project, recent work in ML applied to biological sequences was surveyed in order to identify the feasability and scope of a project appropriate for this course. Method feasability focused on LDA, NB and LSTM models. Some of this work is linked to Jupyter notebooks or original papers and includes protein classification[1], Sapien/Neandertal sequence classification[2], Neandertal DNA introgression[3], CRISPR gRNA design tools[4], Bacterial Classification with RNNs[5], Bacteria taxonomic classifi-

[1] https://github.com/ai-dave/ml-gene-pub/blob/main/protein-classification.ipynb

[2] https://github.com/ai-dave/ml-gene-pub/blob/main/Deep_Learning_on_Ancient_DNA.ipynb

[3] https://github.com/ai-dave/ml-gene-pub/blob/main/Deep_Learning_Neanderthal_Introgression.ipynb

[4] https://pubmed.ncbi.nlm.nih.gov/31533522/

[5] https://github.com/lelugom/wgs_classifier

cation[6], Ribosmal RNA analysis[7,8,9,10,11], Viral and Single-celled Eukaryote Taxonomy classification[12,13], Wolf/Dhole classification[14,15,16].

1.3 Capstone Project Work

Jupyter notebook for the Capstone Project is is available from a github repo[17].

[6] https://www.researchgate.net/publication/348432006_Bacteria_taxonomic_classification_using_Machine_learning_models

[7] https://github.com/ai-dave/ml-gene-pub/blob/main/gnb-rdp.ipynb

[8] https://github.com/ai-dave/ml-gene-pub/blob/main/lda-compare-silva.ipynb

[9] https://github.com/ai-dave/ml-gene-pub/blob/main/lda-rdp.ipynb

[10] https://github.com/ai-dave/ml-gene-pub/blob/main/lstm-random-noise.ipynb

[11] https://github.com/ai-dave/ml-gene-pub/blob/main/lstm-rdp.ipynb

[12] https://github.com/ai-dave/ml-gene-pub/blob/main/ncbi_lstm_poc.ipynb

[13] https://github.com/ai-dave/ml-gene-pub/blob/main/Eukaryota-Amoebozoa-Discosea.ipynb

[14] https://github.com/ai-dave/ml-gene-pub/blob/main/Dhole-Wolfe-LDA.ipynb

[15] https://github.com/ai-dave/ml-gene-pub/blob/main/Dhole-Wolfe-LSTM.ipynb

[16] https://github.com/ai-dave/ml-gene-pub/blob/main/Dhole-Wolfe.ipynb

[17] https://github.com/ai-dave/ml-gene-pub/blob/main/gene-family-human-chimp-dog.ipynb

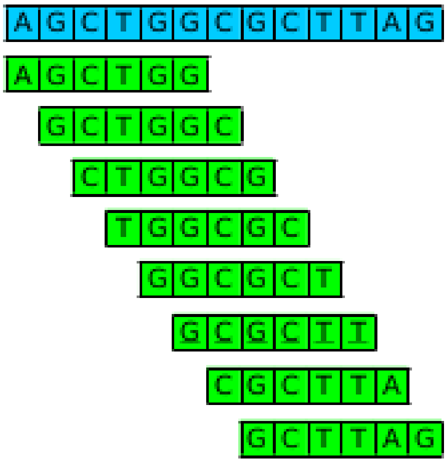

Figure 1: 6-mer "*word*" production from original sequence.

2 Data

2.1 Source

The data for this project came from a Kaggle notebook, *DNA Sequencing with Machine Learning*[18]. DNA Coding Sequences are mostly comprised of exons and, in this corpus, range in length from ~1000 to ~3,500 bases.

2.2 Labeling

Each sequence in the corpus of sequence data was annotated with one of seven Gene Families. The Gene Family annotation was translated into an integer class label c in $\{1, 2, 3, 4, 5, 6, 7\}$

2.3 Preprocessing: K-mer vectorization

The nucleotide bases comprise an *alphabet* of the four letters b in $\{A, G, C, T\}$. A *K-mer* is a k-letter word comprised from *letters* of our *alphabet*. Here we let k=6. The *dictionary* contains up to 4^6 *words*, each 6 letters in length.[19]

Figure 1 above depicts 6bp "*words*" derived from a sequence of the original corpus. From this point on, all vectorization is based on a corpus of *6-mer* "*words*" rather than the original sequences.

[18] https://www.kaggle.com/code/nageshsingh/demystify-dna-sequencing-with-machine-learning/data

[19] Dictionary size calculated as $alphabetSize^{wordLength} \Rightarrow 4^6 \Rightarrow$ 4,096 words.

3 Feature Engineering

3.1 Feature Extraction/Embedding

Use CountVectorizer to convert k-merized sequences to a matrix of token counts[20].

BOW embedding

[20] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

4 Modeling

4.1 Training

4.2 Performance

4.3 Deployment