Name:	Dave Mannion
Course:	UCSD Machine Learning & Engineering
Date:	October 22, 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 NLP applications 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 classifica-

[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

tion[6], Ribosmal RNA analysis[7,8,9,10,11], 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

2 Data

2.1 Source

The data for this project came from a Kaggle notebook, DNA Sequencing with Machine Learning[18].

2.2 Labeling

2.3 Preprocessing: K-Mer vectorization

2.4 Feature Extraction/Embedding

Use CountVectorizer to convert k-merized sequences to a matrix of token counts[19].

2.5 Model Building

2.6 Model Performance

2.7 Model Deployment

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

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