CLASSIFICATION MODELS FOR GENE FAMILIES BASED ON DNA SEQUENCING

Apoorv Saraogee
Northwestern University, Practical Machine Learning
June 5, 2022

Abstract

The volume of DNA sequencing data has been growing at an unprecedented pace. Predictive algorithms can help leverage this data for useful applications such as gene family classification by using new sequencing data. This study analyzes a dataset for protein coding DNA sequences of 7 different gene families across three different species – human, chimpanzee and dog. Naïve Bayes, random forest, convolutional neural networks and recurrent neural networks were used to classify each gene family with a k-mer count encoding. The Naïve Bayes classifier performed the best on the holdout test dataset with F-1 scores of 0.982, 0.993 and 0.983 for human, chimpanzee and dog respectively. Overfitting was seen with the other models used in the study.

1. Introduction

Proteins are incredibly complex three-dimensional biological structures that are encoded in nature using a sequence of four possible nucleotides - Adenine (A), Thymine (T), Guanine (G), and Cytosine (C). There are almost 100 million unique protein sequences in the UnitProt database and this is growing at a rate faster than Moore's law (Greenhalgh, Saraogee, and Romero 2021). The volume of data together with the complexity of working with categorical sequence data makes predictive algorithms an attractive choice for analysis (Chauhan 2021).

This study's central research topic is to analyze a dataset derived from Kaggle (Chauhan 2021) of DNA sequences in the protein coding regions of three species – 4380 human, 1682 chimpanzee and 820 dog sequences by comparing four different predictive algorithms – Naïve Bayes, random forest, convolutional and recurrent neural networks. Although one-hot encoding is more widely used in genomics (Shen, Bao, and Huang 2018) and performs better than k-mer count encoding with some DNA sequencing data (Li et al. 2021), this study used k-mer count encoding to deal with the unequal lengths of sequences in the dataset. This study uses a k-mer count encoding with k=3 for a 'word' to mimic how nature uses three nucleotides to code for a single amino acid – the basic building block of proteins. Research questions include whether traditional machine learning models such as Naïve Bayes and random forests are more suitable than convolutional or recurrent neural networks (scored using the F-1 precision-recall based accuracy score). Other questions will explore the generalizability of the trained models across the three species as well as hyperparameters and structure of each model.

2. Literature Review

A variety of machine learning classification models including Naïve Bayes, random forests and neural networks have been used with other datasets of DNA sequencing data using both one-hot encoding and k-mer count encoding. Naïve Bayes has been shown to outperform other sophisticated models with DNA sequencing data, perhaps due to the large number of features (Monaco et al. 2021). Only a Naïve Bayes classification with a k-mer encoding (k=6 instead of k=3 in this study) has been employed with the same gene family dataset reporting an F-1 score of 0.984, 0.994, 0.925 for the human, chimpanzee and dog datasets respectively (Chauhan 2021). Deep learning architectures have been gaining in popularity for DNA sequence classification due to the high dimensionality of the data and increasing computational power (Min, Lee, and Yoon 2017; Li et al. 2021; Shen, Bao, and Huang 2018; lo Bosco and di Gangi 2017; Gunasekaran et al. 2021). This study uses some of the same parameter settings as other studies – LeNet architectures for convolutional neural networks (lo Bosco and di Gangi 2017) and long short-term memory (LSTM) layers and gated recurrent units (GRU) layers. However, researchers in other studies had more computational power and notably used more units in recurrent neural networks with 100 LSTM units (Gunasekaran et al. 2021) and 5 million GRU units (Shen, Bao, and Huang 2018).

3. Methods

This research will be conducted on the Kaggle dataset (Chauhan 2021) of DNA sequences using a k-mer encoding (k=3) and n-gram size of 8 to create 602,855 features for each sequence. Analysis is done in a Jupyter notebook using kernels for Python3 run locally with models in the keras and sklearn packages for modeling and the biopython package for preprocessing of sequences. Real DNA sequencing data is used in the classification models with 20% of the data used for testing using accuracy scores. The remaining 80% of the data is used to train the Naïve Bayes algorithm and a five-fold cross-validated random forest model. The training set is further split up 80/20 into a subtraining set and validation dataset for neural networks. For 1-D convolutional neural networks, a modified LeNet architecture (Géron 2017) is used with fewer Dense layers, 16 to 24 filters for the two layers in the architecture as well as varying kernel size (3 or 5) and constant stride of 3. Various recurrent neural network architectures were tested with two or three 1-D convolutional and pooling layers with (filter size of 8, kernel size of 5, constant stride of 3) and two recurrent layers with 2-10 units (SimpleRNN, LSTM or GRU). The key objectives include comparing different neural network architectures (recurrent and convolutional) with Naïve Bayes and random forest models using the precision-recall based accuracy metric F-1 score.

3. Results

The F-1 accuracy scores for each of the tested models across the three species is included in Table 1 with Naïve Bayes performing the best across all test datasets with an average score of

0.986 compared to 0.920, 0.935 and 0.147 for the random forest, convolutional and recurrent neural networks respectively. The random forest and convolutional neural network models both indicate overfitting as both had high accuracy scores in the training sets with scores of 1.00 and 0.99 respectively, but poor scores on data the model had not seen. The convolutional neural networks saw the model perform slightly better with the higher kernel size of 5, but lowering the filter size to 16 greatly increased the performance indicating low interaction effects between features. All of the recurrent neural network architectures trained performed poorly and classified all the sequences as transcription factors (class label 6). Interestingly, GRU performed the best with an accuracy score of 0.30 compared to 0.17 and 0.13 for the SimpleRNN and LSTM models respectively. These results suggest that overfitting is a common issue when using machine learning models such as random forests and convolutional neural networks for analyzing DNA sequencing data as also reported in the literature (Greenhalgh, Saraogee, and Romero 2021) and the unsuitability of using recurrent neural networks in the classification of gene families.

4. Conclusions

Within this study, the Naïve Bayes model is the best because it had the highest accuracy score of 0.986 on the test sets. This is quite good as it generalizes well across the three different species compared to the literature value of 0.967 on the same dataset (Chauhan 2021). This suggests that using a k-mer encoding with k=3 works better than k=6. Both random forest and convolutional neural networks were found to be overfitting the data with high training accuracy scores but lower test accuracy scores. Recurrent neural networks were found to be unsuitable for the classification task but will likely perform better with more units or fewer convolutional layers as reported in the literature (Shen, Bao, and Huang 2018).

5. Appendices

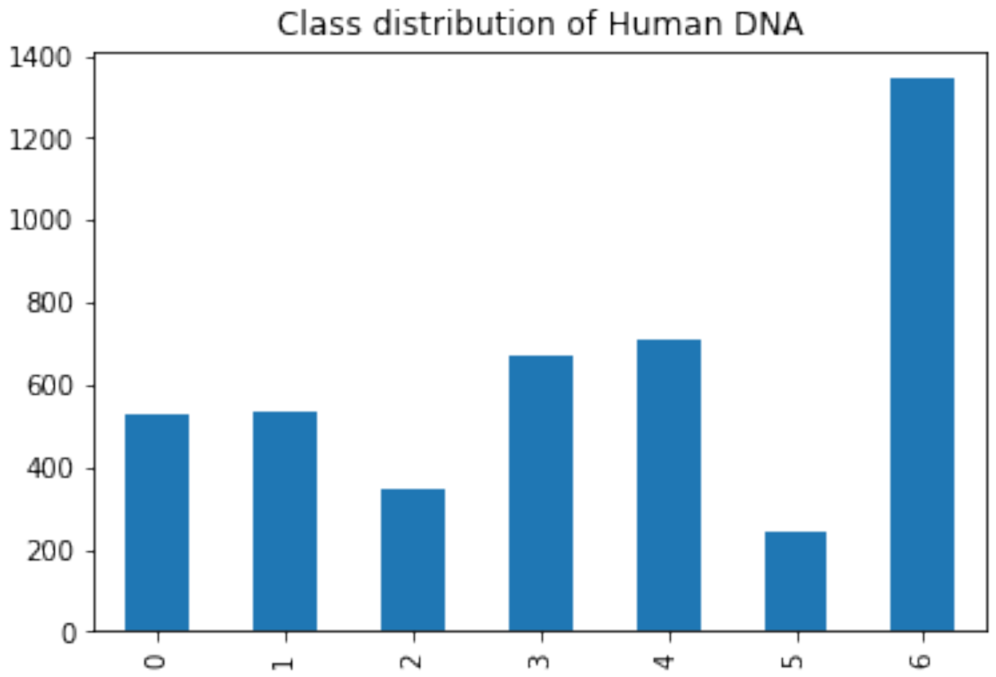

Figure 1. Class distribution of human DNA with each class 0,1,2,3,4,5 and 6 corresponding to G protein coupled receptors, tyrosine kinase, tyrosine phosphatase, synthetase, synthase, ion channel, and transcription factors respectively. This dataset (Chauhan 2021)was used for training all models.

Table 1: Precision-recall based F-1 scores for each different model across holdout test datasets

	Human	Chimpanzee	Dog	Average
Naïve Bayes	0.982	0.993	0.983	0.986 ± 0.006
Random Forest	0.925	0.988	0.846	0.920 ± 0.071
Convolutional Neural Networks	0.927	0.975	0.904	0.935 ± 0.036
Recurrent Neural Networks	0.141	0.147	0.153	0.147 ± 0.006
Naïve Bayes (Chauhan 2021)	0.984	0.993	0.925	0.967 ± 0.037

Supporting Files

- Assignment6.html
- Assignment6.py
- Assignment6.ipynb
- human.txt
- chimpanzee.txt
- dog.txt

References

Bosco, Giosué lo, and Mattia Antonino di Gangi. 2017. "Deep Learning Architectures for DNA Sequence Classification." *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10147 LNAI (February): 162–71. https://doi.org/10.1007/978-3-319-52962-2_14/TABLES/2.

Chauhan, Nagesh Singh. 2021. "Demystify DNA Sequencing with Machine Learning | Kaggle." Kaggle.Com. 2021. https://www.kaggle.com/code/nageshsingh/demystify-dna-sequencing-with-machine-learning/notebook.

Géron, Aurélien. 2017. "Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow (2019, O'reilly)." *Hands-On Machine Learning with R*, 510.

Greenhalgh, Jonathan, Apoorv Saraogee, and Philip A. Romero. 2021. "Data-Driven Protein Engineering." *Protein Engineering*, September, 133–51. https://doi.org/10.1002/9783527815128.CH6.

Gunasekaran, Hemalatha, K. Ramalakshmi, A. Rex Macedo Arokiaraj, S. Deepa Kanmani, Chandran Venkatesan, and C. Suresh Gnana Dhas. 2021. "Analysis of DNA Sequence Classification Using CNN and Hybrid Models." *Computational and Mathematical Methods in Medicine* 2021. https://doi.org/10.1155/2021/1835056.

Li, Guobin, Xiuquan Du, Xinlu Li, Le Zou, Guanhong Zhang, and Zhize Wu. 2021. "Prediction of DNA Binding Proteins Using Local Features and Long-Term Dependencies with Primary Sequences Based on Deep Learning." *PeerJ* 9 (May). https://doi.org/10.7717/PEERJ.11262.

Min, Seonwoo, Byunghan Lee, and Sungroh Yoon. 2017. "Deep Learning in Bioinformatics." *Briefings in Bioinformatics* 18 (5): 851–69. https://doi.org/10.1093/BIB/BBW068.

Monaco, Alfonso, Ester Pantaleo, Nicola Amoroso, Antonio Lacalamita, Claudio lo Giudice, Adriano Fonzino, Bruno Fosso, et al. 2021. "A Primer on Machine Learning Techniques for Genomic Applications." *Computational and Structural Biotechnology Journal* 19 (January): 4345–59. https://doi.org/10.1016/J.CSBJ.2021.07.021.

Shen, Zhen, Wenzheng Bao, and De Shuang Huang. 2018. "Recurrent Neural Network for Predicting Transcription Factor Binding Sites." *Scientific Reports* 8 (1): 15270. https://doi.org/10.1038/S41598-018-33321-1.