Deep Learning-Based Classification of DNA Sequences into Functional Gene Families

Ghada Abu Shaqra[1] **and Sara Alsanajleh**[1]

[1]Department of Computer Science, Jordan University of Science and Technology, Jordan
Emails: gmabushaqra22@cit.just.edu.jo, sjalsanajleh22@cit.just.edu.jo

Abstract

The classification of DNA [1] sequences into functional gene families is a crucial task in computational biology. In this study, we employ deep learning models to classify human DNA sequences into predefined categories based on their functional characteristics. The dataset, obtained from Kaggle, includes DNA sequences from humans, chimpanzees, and dogs. Initially, we attempted species classification but achieved suboptimal accuracy. Consequently, we focused solely on human DNA sequence classification. We performed data preprocessing, including outlier removal and missing value handling, followed by exploratory data analysis to understand sequence distributions. We experimented with multiple deep learning architectures, including LSTM and a Transformer-based classifier. The results demonstrate varying performance across models, highlighting the effectiveness of transformer-based approaches in DNA sequence classification.

Introduction

DNA sequence classification is a fundamental problem in bioinformatics, aiding in the understanding of genetic functions, evolutionary relationships, and disease mechanisms. Traditional sequence classification methods rely on alignment-based approaches, which can be computationally expensive and less effective for large-scale datasets. Recent advancements in deep learning have demonstrated promising results, particularly for tasks in natural language processing (NLP).

In this study, we aim to classify human DNA sequences based on their functional characteristics. The dataset includes sequences from human, chimpanzee, and dog. Initially, we explored species classification, but the results were suboptimal, with accuracy scores in the 60% range. Thus, we shifted our focus to classifying human DNA sequences into predefined gene families.

Our approach involves preprocessing the dataset to remove outliers, handle missing values, and analyze sequence distributions. We experimented with different deep learning architectures, including LSTM and a Transformer-based classifier, to assess their effectiveness in accurately classifying the sequences.

Methods

A. Data Preprocessing. The dataset was obtained from Kaggle and includes DNA sequences from human, chimpanzee, and dog. We focused on the human DNA sequences due to their larger subset. The preprocessing steps included:

- **Outlier Removal**: Outliers were identified (see Fig. 1) and handled using appropriate statistical methods to mitigate their impact on analysis [3].

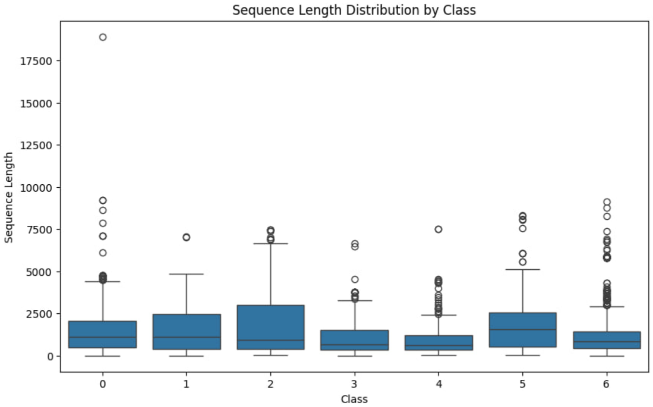

Fig. 1. Visualization of outliers identified in the dataset, highlighting values that deviate significantly from the rest of the data.

- **Missing Value Handling**: The dataset contained one missing record, which was removed during preprocessing.

- **Exploratory Data Analysis (EDA)**: We analyzed class distribution and DNA sequence lengths. Visualization was performed using count plots, along with functions like `data.describe()` and `data.nunique()`.

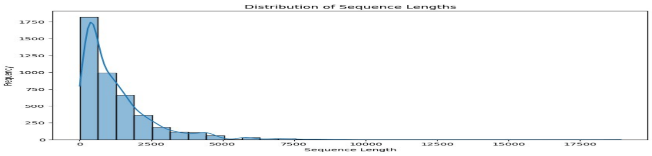

Fig. 2. Histogram illustrating the distribution of DNA sequence lengths in the dataset.

B. Model Selection. For the classification task, we used the LSTM [2] (Long Short-Term Memory) model, a type of recurrent neural network (RNN) well-suited for sequence prediction tasks due to its ability to capture long-range dependencies.

We built a Bidirectional LSTM architecture consisting of an embedding layer, followed by two Bidirectional LSTM layers with dropout and recurrent dropout for regularization. The final layers included BatchNormalization, a dense layer, and a softmax output layer. We used the Adam optimizer (learning rate = 0.001) and sparse categorical cross-entropy loss.

Experiments

C. Data Splitting. We divided the dataset using an 80-20 split: 80% for training and 20% for testing.

D. Classification Approach. Initially, our goal was to classify DNA sequences based on species (human, chimpanzee, dog), but this approach yielded low accuracy (around 60%). We then focused on classifying human DNA sequences into six predefined gene families:

- **Class 0**: G-Protein Coupled Receptors (GPCRs)
- **Class 1**: Tyrosine Kinase
- **Class 2**: Tyrosine Phosphatase
- **Class 3**: Synthase
- **Class 4**: Ion Channel
- **Class 5**: Transcription Factor

E. Model Comparison. We trained both the LSTM model and a Transformer-based classifier using the same dataset split. Their performance was compared based on accuracy and loss.

Results

Model performance on the test set is summarized in the table below:

Model	**Accuracy (%)**	**Loss**
LSTM	73.06%	0.0868
Transformer	74.77%	1.7148

Table 1. Performance of the models on the test set.

Discussion

This study explored the effectiveness of deep learning models—specifically LSTM and Transformer architectures—in classifying human DNA sequences into functional gene families. Both models showed promising performance, but with notable differences in learning behavior.

The LSTM model achieved 73.06% accuracy with a low loss of 0.0868, indicating its ability to capture long-range sequential dependencies effectively.

The Transformer model achieved slightly higher accuracy (74.77%) but suffered from a much higher loss (1.7148), suggesting overconfidence in wrong predictions or overfitting. This emphasizes the importance of evaluating models using multiple metrics. Future work could explore regularization techniques, label smoothing, or loss functions like focal loss to improve Transformer calibration.

Both models demonstrate potential for biological sequence classification. Incorporating gene annotations and other metadata may further enhance model accuracy and interpretability.

Conclusion

We investigated the use of deep learning models to classify human DNA sequences into gene families. The study compared LSTM and Transformer architectures, supported by proper data preprocessing and systematic experimentation.

The LSTM model effectively captured sequence dependencies, while the Transformer model showed competitive accuracy, with room for improvement in generalization. These results validate the usefulness of deep learning in genomics and open future research opportunities in optimizing models and understanding biological signals.

References

- https://en.wikipedia.org/wiki/DNA
- Scribbr. *Interquartile Range*. Available at: https://www.scribbr.com/statistics/interquartile-range/. Accessed: 2025-03-28.
- PyTorch. *torch.nn.LSTM*. Available at: https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html. Accessed: 2025-03-28.

et al. | DNA Classification Using Deep Learning