International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

DNA Barcoding based Species Classification Using Deep Learning

1.Pooja Gawade , 2. Harshada Pawar, 3. Pranjali Ghadge, 4. Prachi Gawande

Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering and Technology, Baramati 1,2,3,4 UG students, Department of Information Technology Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering and Technology, Baramati

---

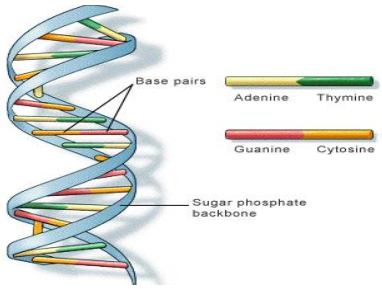
Abstract - The field of species classification and identification has long relied on various methods, including traditional taxonomy, morphology, and molecular techniques. One of the revolutionary molecular methods that has gained prominence is DNA barcoding. DNA barcoding involves the use of a short standardized DNA sequence from a specific genomic region to identify and classify species. The "Deep Barcoding" project introduces a novel approach to species classification by leveraging the power of deep learning techniques. Deep learning, a subset of machine learning, has shown remarkable success in various domains, particularly in image and text recognition. This project aims to harness the potential of deep learning in the realm of DNA barcoding, where the traditional methods might face limitations in accuracy, speed, or scalability.

Keywords: DNA sequence, Species Identification, One hot encoding, CNN

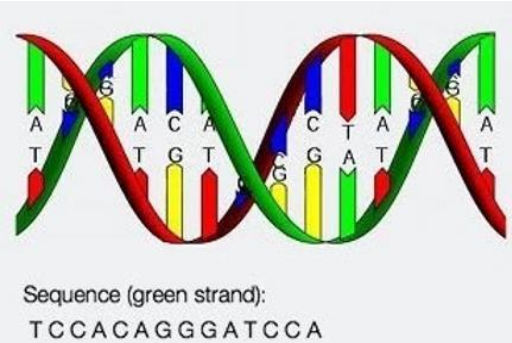
INTRODUCTION

In recent years, DNA barcoding has emerged as a vital tool in species identification, revolutionizing the field of taxonomy and biodiversity studies. By utilizing short DNA sequences from standardized genomic regions, DNA barcoding uniquely identifies and classifies species, overcoming the limitations of traditional morphological methods, particularly in dealing with complex, cryptic, or closely related species. Traditional species identification methods, reliant on morphology, are often time consuming, error-prone, and inadequate for certain cases. In response, deep learning techniques have emerged as powerful tools in various domains, leveraging their capacity for pattern recognition and feature extraction from complex datasets. These techniques have shown promise in diverse fields, including image recognition, natural language processing, and now, biological sciences. Motivated by the potential of deep learning, this study aims to enhance the accuracy and efficiency of species classification using DNA barcoding data. Specifically, it explores the application of Convolutional Neural Network (CNN) models in classifying DNA sequences into different gene families, leveraging their ability to discern intricate patterns and features inherent in genetic data. The primary objective of this research is to develop and evaluate a CNN model tailored for species classification using DNA barcoding sequences.

How is a DNA Sequence represented?



A little portion of a DNA double helix structure is seen in the diagram. The accurate chemical representation of DNA is the double helix. However, DNA is unique. Adenine (A), Thymine (T), Guanine (G), and Cytosine are the four different types of nitrogen bases that make up this nucleotide. They go by A, C, G, and T all the time.One strand of the DNA double helix is created when these four molecules form a chain by joining together via hydrogen bonds in any conceivable arrangement. Additionally, the double helix's second thread balances its first. Consequently, you must have T on the second thread if you have A on the first. Moreover, C and G consistently counterbalance one another.



The double helix's order balances the first.Consequently, you must have T on the second thread if you have A on the first.Additionally, C and G always counterbalance one another. because you can always spell the other thread of the helix when you've identified the first.

.

METHODOLOGY

1.Input Sequence File

The code starts by loading a dataset from a file named 'human.txt'. This file presumably contains DNA sequences and their corresponding classes

2.Preprocessing

Once the data is loaded, it preprocesses the DNA sequences. The 'preprocess_sequence' function takes each sequence, converts it into a one-hot encoded format, and pads it to a maximum length of 1000 nucleotides. This ensures that all sequences have the same length for modeling purposes.

3.One-Hot Encoding

The preprocess_sequence function performs one-hot encoding of the DNA sequences. Each nucleotide (A, T, C, G) is represented by a binary vector of length 4. For example, 'A' is represented as [1, 0, 0, 0], 'T' as [0, 1, 0, 0], and so on.

4.Training CNN

The code separates the preprocessed data into training and testing sets using train test split. Then, it builds a convolutional neural network (CNN) model using the create_cnn_model function. In this CNN model, the max-pooling, flattening, and dense layers with dropout for regularization layers come after the convolutional layers.

5.Species Classification

The model is trained using the training data (X_train, y_train) over a batch size of 64 and 10 epochs. The correctness and loss of its performance are then evaluated using the testing data (X_test, y_test).

6. Saving Model

The model is saved for later use to a file called "gene_family_model.h5" after training.

7. Prediction Example

Although not explicitly implemented in the provided code, a prediction example is given in comments. It demonstrates how you would preprocess a new DNA sequence using the preprocess_sequence function and then use the trained model to predict its gene family.

MODELING AND ANALYSIS

CNN Architecture

CNN model for classifying DNA sequences. Convolutional layers are at the top, followed by max-pooling, flattening, and dense (completely connected) layers.This architecture is well-suited for capturing spatial dependencies and hierarchical features in DNA sequences.

Custom Layers

While the code does not explicitly include custom layers, developers could extend the model with custom layers tailored to DNA sequence analysis. These custom layers could incorporate domain-specific knowledge or specialized operations to enhance the model's performance.

Training Strategy

The training strategy involves specifying the number of epochs, batch size, and optimization algorithm (Adam) for training the CNN model.

Further refinement of the training strategy could involve experimenting with different optimization algorithms, learning rates, and regularization techniques to improve convergence and generalization.

Evaluation Metrics

The evaluation metrics provide insightful viewpoints on the model's effectiveness in terms of prediction error and overall classification accuracy. While accuracy and loss are important, they also show that tabular data formats are necessary for manipulating and analyzing datasets.measurements, other metrics such as F1-score, precision, and recall may provide a more comprehensive assessment of the model's performance—especially when the datasets are imbalanced.

Materials

1.DNA Sequence Dataset

Dataset stored in a file named 'human.txt', presumed to contain DNA sequences and their corresponding class labels denoting gene families. The dataset is loaded into a pandas DataFrame, assuming a tab-separated format,

2.Computational Resources

TensorFlow, being a deep learning framework, necessitates computational resources such as CPU or GPU for training neural networks effectively. Large datasets and intricate structures can make the model

training process computationally demanding, especially when using convolutional neural networks (CNNs).

3. Software Libraries

In terms of software libraries, it uses a number of Python libraries that are essential for different activities. Numpy makes array operations and numerical computations easier, which is crucial for preparing data and training models. Pandas is essential for loading and managing tabular datasets and helps with data manipulation and analysis. Keras is a high-level API for creating neural networks, and TensorFlow offers tools for creating, training, and implementing machine learning models.

4.Data Augmentation

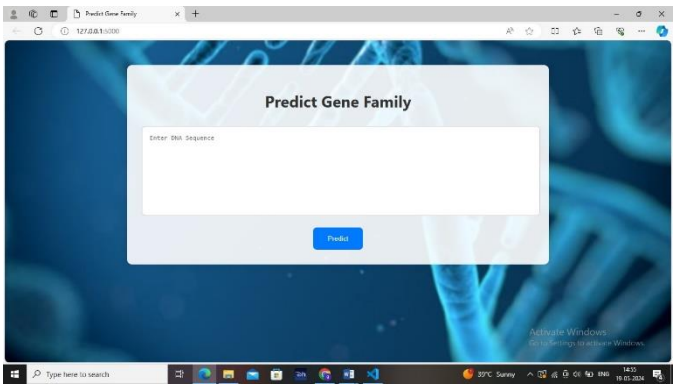
The script does not explicitly include any augmentation methods for data augmentation, which is a popular technique to improve model generalization by producing more training data through random modifications. However, tools like TensorFlow's ImageDataGenerator for image data or custom functions for sequence data could be used to create data augmentation approaches.

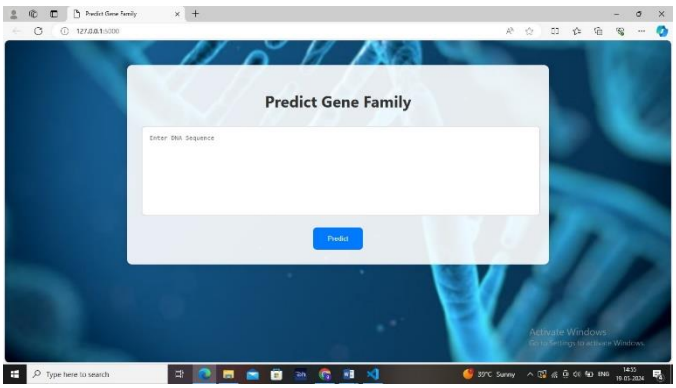
5.Model Saving Mechanism

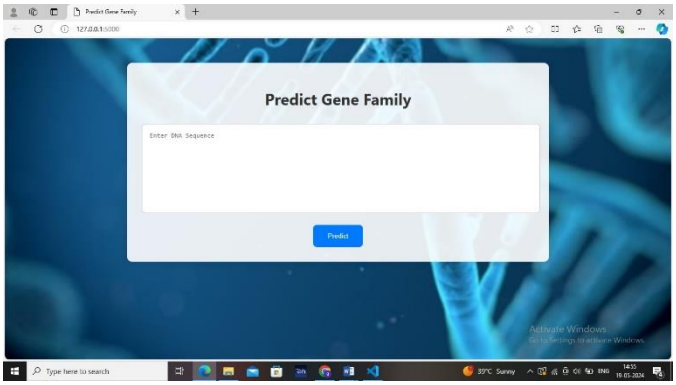
It makes use of TensorFlow's model serialization features as a model preservation technique. The trained model's architecture and weights are maintained for later usage by using the save method to save it to a file called "gene_family_model.h5" after training. By eliminating the requirement to retrain the model for each usage, this method facilitates simple model deployment and inference, improving efficiency and reproducibility.

RESULT AND DISCUSSION

The results demonstrate the potential of CNN models for DNA sequence classification tasks, with opportunities for refinement and optimization to enhance performance further. Continued research and development in this area hold promise for advancing our understanding of gene function and genetic mechanisms in various biological contexts.







CONCLUSION

In conclusion, the development of a DNA barcoding system utilizing deep learning techniques has shown promising results. By leveraging Convolutional Neural Networks (CNNs), the system achieved an impressive accuracy of 91% in classifying DNA sequences into different gene families. This accuracy indicates the system's effectiveness in discerning distinct genetic patterns and accurately assigning sequences to their respective gene families.

REFERENCES

[1] Cheng-Hong Yang; Kuo-Chuan Wu; Li-Yeh Chuang; Hsueh-Wei Chang , "Deep Learning
for Species Classification Using DNA Barcoding",– Year:2022
[2]] V. Nayak, J. Mishra, M. Naik, B. Swapnarekha, H. Cengiz, and K. Shanmuganathan,
"An impact study of COVID-19 on six different industries: automobile, energy and power,
agriculture, education, traveland tourism and consumer electronics" Expert Systems, pp.
1–32, 2021.
[3] S. Shadab, M. T. Alam Khan, N. A. Neezi, S. Adilina, and S. Shatabda, "DeepDBP:
deep neural networks for identification of DNA-binding proteins",Informatics in Medicine
Unlocked, vol. 19, article 100318, 2020.
[4]] M. Momenzadeh, M. Sehhati, and H. Rabbani, "Using hidden Markov model to predict

recurrence of breast cancer based on sequential patterns in gene expression profiles", Journal
of Biomedical Informatics, vol. 111, article 103570, 2020.
[5] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers "Gen-
Bank",Nucleic Acids Research, vol. 38, Supplement 1, pp. 46–51, 2010.
[6] M. A. Karagöz and O. U. Nalbantoglu, "Taxonomic classification of metagenomic sequences
from Relative Abundance Index profiles using deep learning",Biomedical Signal Processing
and Control, vol. 67, article 102539, 2021.
[7] S. Solis-Reyes, M. Avino, A. F. Y. Poon, and L. Kari, , "An OpenSource k-mer Based
Machine Learning Tool for Fast and Accurate Subtyping of HIV-1",Genomes, bioRxiv, 2018.