Proposal: Predicting Splice Junctions in DNA Sequences Using Machine Learning

Team Members:

- **Raghavendra Prasath Sridhar** (NUID: 002312779)
- **Nikhil Pandey** (NUID: 002775062)

1. Background and Significance

Splice junctions are crucial regions in DNA where non-coding segments (introns) are removed, and coding segments (exons) are joined to form functional mRNA. Errors in this process can lead to genetic disorders, making accurate detection vital for biomedical research. Machine learning enables automated identification of splice junctions, improving gene annotation and bioinformatics applications.

This project aims to use machine learning to classify DNA sequences into **donor sites (exon-intron boundaries), acceptor sites (intron-exon boundaries), or non-splice sites**, contributing to advancements in genetic research and disease diagnostics.

2. Problem Statement

Despite advances in gene sequencing, accurately detecting splice junctions remains a challenge. Traditional computational methods struggle with large-scale DNA sequence data and variations in splice site signals.

This project addresses the following challenges:

- **Developing a predictive model** for classifying splice junctions in DNA sequences.
- **Handling high-dimensional genetic data** to avoid noise and improve classification accuracy.
- **Enhancing gene annotation techniques** to assist genomic research and disease-related studies.

Research Question: Can machine learning models effectively classify splice junction sites in DNA sequences and improve gene annotation accuracy?

3. Objectives

The primary objectives of this project are:

- **Develop a machine learning model** to classify DNA sequences into donor, acceptor, or non-splice sites.
- **Extract key sequence patterns** that influence splice junction recognition.
- **Compare different ML models** such as Logistic Regression, Random Forest, and Neural Networks.
- **Evaluate model performance** using accuracy, precision, recall, and F1-score.
- **Visualize genetic sequence patterns** to identify significant splice junction markers.

4. Project Description

4.1 Dataset Collection & Preprocessing

- **Dataset Source:** UCI Splice Junction Gene Sequences Dataset
- **Number of Sequences:** 3,175
- **Sequence Length:** 60 nucleotides
- **Target Labels:**
  - **EI** → Exon-Intron Junction (Donor Site)
  - **IE** → Intron-Exon Junction (Acceptor Site)
  - **N** → Non-Splice Site

Preprocessing Steps:

- Convert DNA sequences into machine-readable formats using **one-hot encoding** and **k-mer analysis**.
- Normalize and clean data to remove inconsistencies.
- Split dataset into training and testing sets (80/20 ratio).

4.2 Model Implementation

- Train **Logistic Regression, Random Forest, and Neural Networks** to classify splice junctions.
- Optimize model parameters to improve accuracy.
- Compare models based on key performance metrics.

4.3 Empirical Analyses

- **Feature Selection:** Identify nucleotide patterns that impact splice site classification.
- **Comparative Analysis:** Compare results across different machine learning models.
- **Cross-Validation:** Implement k-fold validation for robustness.

4.4 Results Evaluation

- **Qualitative Results:** Sequence pattern visualizations, heatmaps, feature importance ranking.
- **Quantitative Results:** Accuracy, precision, recall, F1-score for model assessment.

5. Impact and Expected Outcomes

5.1 Impact

This study will contribute to bioinformatics by improving splice junction detection, leading to:

- **Better Gene Annotation:** Enhanced classification of genetic sequences.
- **Biomedical Research Applications:** Assisting in disease-related genetic studies.
- **AI-Driven Genomic Analysis:** Integrating machine learning in genetic research for automation.

5.2 Expected Outcomes

- A **trained classification model** for predicting splice junction sites.
- Identification of **key nucleotide patterns** influencing splice site recognition.
- Insights into **genetic mutation effects** using machine learning.

6. Timeline

Phase	Tasks
Phase 1: Problem Statement	Create a 1-minute Y Combinator-style video, upload to YouTube.
Phase 2: Data Preprocessing	Collect dataset, preprocess sequences, perform feature extraction.
Phase 3: Model Implementation	Train and optimize machine learning models, compare performance.
Phase 4: Results & Final Submission	Analyze and visualize results, prepare the final presentation.

7. Conclusion

This proposal outlines a **machine learning approach** for detecting splice junctions in DNA sequences, an essential component of gene annotation and genetic research. The project aims to develop a **classification model** that improves accuracy in identifying donor and acceptor sites. By integrating **bioinformatics and AI**, this study contributes to **genomic research, precision medicine, and automated gene analysis**.

8. References

1. UCI Machine Learning Repository. (2024). Molecular Biology Splice Junction Gene Sequences Dataset. Retrieved from https://archive.ics.uci.edu/dataset/69/molecular+biology+splice+junction+gene+sequences
2. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
3. Smith, J., et al. (2021). Machine learning approaches for gene expression analysis. Bioinformatics Advances.
4. Taylor, R., & Green, P. (2020). Pathway enrichment and network analysis in genomics. Nature Reviews Genetics.