Title: Multilabel Classification for Gene Ontology Prediction Using Yeast Sequences

Group Members:

Sher ALI- P21-8024

Objective:

This project aims to develop a machine learning model capable of predicting gene ontology (GO) terms for yeast sequences through multilabel classification. By assigning GO terms, this project seeks to enhance the functional annotation of yeast gene sequences.

Problem Description:

In bioinformatics, multilabel classification is a complex task where each sequence may relate to multiple GO terms. The provided dataset includes:

1. Yeast Sequence File – Contains the sequences in FASTA format.
2. Gene Ontology File – Maps each sequence to its associated GO terms.

Our goal is to preprocess these data files, extract meaningful features from the sequences, and then train a classification model to accurately predict the GO terms for each input sequence.

Solution Approach:

1. Data Exploration:

Overview of Files: We will examine both files to understand data characteristics. For the FASTA file, we will assess sequence length and variety, while in the GO file, we will study the distribution and frequency of each GO term across sequences.

Objective Mapping: Assess the complexity of assigning GO terms due to the multilabel nature of the problem.

2. Data Preprocessing

Sequence Preprocessing: We will parse the FASTA file, standardizing sequences for feature extraction. This may involve transforming sequences into a consistent format for downstream analysis.

Label Transformation: Convert GO terms into a binary label format where each GO term corresponds to a column. Each sequence will have a binary vector indicating the presence or absence of each GO term.

3. Feature Extraction

Sequence Encoding: Convert sequences into numeric formats suitable for machine learning, with encoding methods such as:

One-hot Encoding for shorter sequences.

K-mer Frequency Analysis to capture repeating sequence patterns associated with biological functions.

Dimensionality Reduction (Optional): If feature space becomes high-dimensional, apply techniques like PCA to improve computational efficiency.

4. Model Selection

Algorithm Choice: Since this is a multilabel classification task, we will consider:

Traditional Machine Learning Models: Random Forest or Support Vector Machines (SVM) adapted for multilabel data.

Deep Learning Models: If feasible with the data, we will explore Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) for sequence-based modeling.

Libraries: Use machine learning libraries such as scikit-learn for traditional models or TensorFlow/Keras for deep learning models.

5. Model Training and Cross-Validation

Data Splitting: Divide the dataset into training, validation, and test sets, ensuring stratified distribution of GO terms across these sets.

Cross-Validation: Implement K-fold cross-validation to evaluate model stability and mitigate overfitting, while maintaining a balanced GO term representation in each fold.

6. Hyperparameter Tuning

Optimization Technique: Use grid search or random search to tune hyperparameters for optimal model performance.

Regularization and Dropout (for Neural Networks): Regularization methods will help prevent overfitting, especially with deep learning models.

7. Model Evaluation

Evaluation Metrics: We will use multilabel metrics to assess the model's performance:

Hamming Loss – Measures the fraction of labels predicted incorrectly.

Subset Accuracy – Considers the exact match across all GO terms.

F1 Score (Macro and Micro) – Evaluates precision and recall across labels.

8. Generalization and Testing

Testing with Unseen Sequences: Evaluate the model’s generalization by testing it on new sequences not present in the training set, assessing its ability to correctly assign GO terms based on sequence patterns.

9. Documentation

Project Documentation: Throughout the process, maintain detailed documentation of preprocessing steps, model choices, training, and evaluation for a clear and replicable workflow.

Final Report: Submit a report describing each step of the approach and detailing the results, evaluation, and possible improvements for future work.

Expected Outcome:

> This project is expected to deliver a trained model capable of predicting multiple GO terms for each yeast sequence, providing valuable insights into the biological functions associated with each gene. The model's success will be evaluated based on its accuracy in assigning GO terms to sequences, as well as its robustness to new data.