**PlantLncBoost Data Repository**

This repository contains all the datasets (training, test, and experimentally validated) used in the training and testing of PlantLncBoost, including experimentally validated lncRNA sequences.

**Directory Structure**

1. **Training Data**

Located in<https://github.com/xuechantian/PlantLncBoost/tree/master/data/>

* **lncRNA\_training.fasta**: Positive training set containing lncRNA sequences from various plant species
* **mRNA\_training.fasta**: Negative training set containing protein-coding RNA sequences from various plant species

1. **Test Data**

Located in <https://github.com/xuechantian/PlantLncBoost/tree/master/data/testdata/>

This directory contains test datasets from 20 diverse plant species across a broad range of plant lineages:

* **Spermatophytes**

*Ananas comosus*

*Amborella trichopoda*

*Arabidopsis thaliana*

*Brachypodium distachyon*

*Cucumis sativus*

*Glycine max*

*Manihot esculenta*

*Medicago truncatula*

*Musa acuminata*

*Oryza sativa*

*Populus trichocarpa*

*Solanum lycopersicum*

*Sorghum bicolor*

*Vitis vinifera*

*Zea mays*

* **Bryophyte**

*Physcomitrella patens*

* **Archaeplastida**

*Chlamydomonas reinhardtii*

*Coccomyxa subellipsoidea*

*Micromonas pusilla*

*Volvox carteri*

1. **Experimentally validated lncRNA**

Located in [https://github.com/xuechantian/PlantLncBoost/tree/master/data/](https://github.com/xuechantian/PlantLncBoost/tree/master/data/testdata/)

* **Validated\_lncRNA.fasta:** Collection of experimentally validated lncRNA sequences used for model validation and evaluation

**\*Notes**

* Sequences were preprocessed to remove redundancy and ensure quality
* Each test dataset contains a balanced number of positive and negative samples