Manual for Transposon-derived piRNA Prediction

1. Introduction

The program is developed for piRNA prediction. Primary sequences were transformed into feature vectors using statistical and position-relative moment-based features. The file name is `Feature_Extraction_Code.ipynb`.

2. Datasets

There are three datasets available, covering three species: Human, Mouse, and Drosophila. Each dataset includes the same number of positive instances and negative instances. The datasets are included in the `/Datasets/` directory in the `.fasta` format.

3. Codes

All the codes for piRNA prediction are included in the `/Codes/` directory. Before running the prediction, please copy the datasets of the relevant species to the `/Codes/` directory and rename the files as `posi_samples.fasta` and `nega_samples.fasta`. All predicted results are saved in `.csv` files.

4. Feature Extraction

After installing the necessary dependencies, such as Biopython, CSV, NumPy, etc., perform feature extraction using the provided code. Make sure to import the required libraries and execute the feature extraction code.

```
!pip install BioPython
import csv
import math
import numpy as np
from Bio import SeqIO
```

5. Ensemble Learning

Once feature extraction is completed, ensemble learning algorithms can be applied to enhance prediction accuracy. Use the appropriate ensemble learning techniques, such as bagging, boosting, or stacking, to improve the performance of the piRNA prediction model. All these codes also in it.