INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS

Name of the Student	:	Adhiraj Nath
Roll Number	:	166106106
Programme of Study	:	Ph.D.

Thesis Title: ***In silico* prediction of precursor microRNA in insects**

Name of Thesis Supervisor(s)	:	Prof. Utpal Bora
Thesis Submitted to the Department/ Center	:	BSBE
Date of completion of Thesis Viva-Voce Exam	:	28-08-2023
Key words for description of Thesis Work	:	Machine Learning, Web Application, Entomology, microRNA

SHORT ABSTRACT

Introduction and Background: Pre-MicroRNAs are the hairpin loops from which microRNAs are produced that have been found to negatively regulate gene expression in several organisms. In insects, microRNAs participate in several biological processes including metamorphosis, reproduction, immune response, etc. Numerous tools have been designed in recent years to predict novel pre-microRNA using binary machine learning classifiers where prediction models are trained with true and pseudo pre-microRNA hairpin loops. Currently, there are no existing tool that is exclusively designed for insect pre-microRNA detection.

Aim: Application of machine learning algorithms to develop an open source tool for prediction of novel precursor microRNA in insects and search for their miRNA targets in the model insect organism, *Drosophila melanogaster*.

Methods: Machine learning algorithms such as Random Forest, Support Vector Machine, Logistic Regression and K-Nearest Neighbours were used to train insect true and false pre-microRNA features with 10-fold Cross Validation on SMOTE and Near-Miss datasets. miRNA targets IDs were collected from miRTarbase and their corresponding transcripts were collected from FlyBase. We used miRanda algorithm for the target searching.

Results: In our experiment, SMOTE performed significantly better than Near-Miss for which it was used for modelling. We kept the best performing parameters after obtaining initial mean accuracy scores >90% of Cross Validation. The trained models on Support Vector Machine achieved accuracy of 92.19% while the Random Forest attained an accuracy of 80.28% on our validation dataset. These models are hosted online as web application called RNAinsecta. Further, searching target for the predicted pre-microRNA in *Drosophila melanogaster* has been provided in RNAinsecta.

Availability: RNAinsecta is freely available at https://rnainsecta.in. Source can be found at GitHub: https://github.com/adhiraj141092/RNAinsecta