In silico prediction of precursor microRNA in insects

A Thesis Submitted in Partial Fulfilment of The Requirement for The Degree Of

Doctorate of Philosophy

By

Adhiraj Nath

(166106106)

Under the Supervision of

Prof. Utpal Bora

Professor

Department of Biosciences and Biotechnology

Indian Institute of Technology Guwahati

Guwahati – 781039, Assam, India.

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

Department of Biosciences and Bioengineering

DECLARATION

This is to declare that the content embodied in this thesis entitled **"*In silico* prediction of precursor microRNA in insects"** is the result of investigations carried out by me under the supervission of **Prof. Utpal Bora**, and is submitted to Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Assam-781039, India for the award of degree of **Doctor of Philosophy in Biosciences and Bioengineering**. This work has not been submitted elsewhere for any degree or diploma of any institute or university to the best of knowledge and belief.

In keeping with the general practice of reporting scientific investigations, due acknowledgements have been made wherever the work of other investigators are referred.

Adhiraj Nath
Roll No- 166106106
Department of Biosciences and Bioengineering
Indian Institute of Technology Guwahati,
Guwahati, Assam-781039, India

Guwahati

Jan, 2023

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

Department of Biosciences and Bioengineering

CERTIFICATE

This is to certify that the work embodied in this thesis entitled "***In-silico* prediction of precursor microRNA in insects**" is the result of the investigation carried out by **Adhiraj Nath** (Roll No. 166106106) under my supervision in the Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Assam-781039, India and is submitted for the award of degree of **Doctorate of Philosophy in Biosciences and Bioengineering.** This work has not been submitted elsewhere for a degree.

Guwahati

Jan, 2023

Prof. Utpal Bora

Thesis Supervisor,

Department of Biosciences and Bioengineering

Indian Institute of Technology Guwahati, Guwahati,

Assam-781039, India

ACKNOWLEDGEMENTS

I am extremely grateful and indebted to my thesis supervisor ***Prof. Utpal Bora*** for introducing me to the exciting world of omics and data science. I thank him for providing me with the opportunity to work in one of the most emerging fields of science. I also thank him for allowing me to conduct my research independently and for coaching me to develop my scientific communication and interpersonal skills. I hope that I have been able to imbibe his enthusiasm and boldness in me.

I would like thank my doctoral committee members ***Prof. Ranjan Tamuli, Dr. Priyadarshi Satpati and Prof. Karuna Kalita*** for their valuable suggestions, motivation and scientific guidance which always helped me to make my work better.

I would also like to convey my gratitude to the ***Department of Biosciences and Bioengineering***, ***Institutional Biotech Hub at the Centre for the Environment*** and***Param-Ishan, IIT Guwahati's high-performance computing cluster*** for providing me all the necessary facilities to pursue my research.

I sincerely acknowledge the financial support from ***Ministry of Human Resource Development (MHRD)***, Government of India for providing me fellowship as well as ***Department of Biotechnology (DBT)***, Government of India and ***Central Silk Board*** for funding our laboratory.

My laboratory mates have been a wonderful lot with diverse personalities and interests and interacting with them has shaped me. A note of gratitude to my seniors ***Deepika, Hasnahana, Debajyoti, Vimal, Jonjyoti*** for their guidance and companionship; to labmates ***Dharitri***

Biju and ***Pulak*** and my friendly neighbor, ***Pankaj*** for sharing responsibilities, insights and laughter on all occasions; and to ***Pragya Ma'am*** for her affection, humor and culinary prowess.

I would like to thank my batchmates and friends ***Pratap, Suvankar, Ratan, Deepak, Angshu, Vineet, Sayan, Barnali*** and many more who have always supported me during thick and thins.

I express my deepest sense of gratitude and love to ***Maa, Deuta*** and ***Maina*** for their overwhelming love, support and patience. Without them, I would have been nowhere.

-Adhiraj

Table of Contents

Synopsis	:	i-iv
List of Abbreviations	:	v-viii
List of Figures	:	ix-xi
List of Tables	:	xii-xiii

	Content	Page Number
CHAPTER 1:	**Introduction and Literature Review**	**1 - 35**
	1.1. microRNA biogenesis	5-8
	1.2. Role of miRNA in insects	9-13
	1.3. Resources and techniques for pre-miRNA characterization and function	14-26
	1.4. Artificial Intelligence and Machine Learning in pre-miRNA prediction	27-34
	1.5. Objectives	35
CHAPTER 2:	**Training binary machine learning classifiers using insect precursor microRNA features.**	**36 - 66**
	2.1. Introduction	37-38
	2.2. Methods	39-49
	2.3. Results	50-59
	2.4. Discussion	60-62
	2.5. Conclusion	63
	2.6. Annexures	64-66
CHAPTER 3:	**Development of cloud-based web interface for the trained machine learning models to predict novel insect precursor microRNA and search target in model organism, *Drosophila melanogaster*.**	**67 - 90**
	3.1. Introduction	68-69
	3.2. Methods	70-78
	3.3. Results	79-87

	3.4. Discussion	88-89
	3.5. Conclusion	90
CHAPTER 4:	**Comparative analysis of sequential and thermodynamic features of insect precursor microRNA with other groups of organisms.**	**91 - 126**
	4.1. Introduction	92-93
	4.2. Methods	94-99
	4.3. Results	100-111
	4.4. Discussion	112-113
	4.5. Conclusion	114
	4.6. Annexures	115-126
CHAPTER 5	**Conclusion and future prospects**	127-129
CHAPTER 6	**References**	130-153

Synopsis

Synopsis of the thesis:

CHAPTER 1: Introduction and Literature Review:

This chapter contains an overview of the researches that have been carried out for prediction of precursor microRNA. It discusses the biogenesis of miRNA and its role in various biological processes in insects. Various studies reporting the differential expression of miRNA during insecticide resistance and immune response has been mentioned in this chapter. The role of miRNA during metamorphosis and reproduction has also been discussed.

State-of-the-art techniques for the discovery of novel pre-miRNA has been discussed, giving a timeline of both the experimental and computational techniques. Different experimental laboratory techniques such as microarray, nucleic acid-based amplification, nanomaterial-based techniques for identification of miRNA has been discussed here.

A detail account of various miRNA and pre-miRNA databases, prediction and target searching tools have been given in this chapter, this chapter also contains various machine learning algorithms and their applications in the prediction of novel pre-miRNA.

CHAPTER 2: Training binary machine learning classifiers using insect pre-miRNA features.

In Chapter 2, we explain the development of a machine learning model for prediction of precursor microRNA in insects. Different features of insect precursor microRNA such as length, GC content, hairpin loop thermodynamics, etc. were used to train machine learning algorithms. A total of 93 features were considered for this experiment. Machine learning algorithms such as Support Vector Machine, Random Forest, Logistic Regression and k-Nearest Neighbours were used to learn from insect precursor microRNA features and classify them from hairpin sequences which do not form precursor microRNA. The parameters were optimized with 10 Fold Cross Validation using GridSearchCV and RandomSearchCV.

The machine learning models trained on Support Vector Machine and Random Forest were selected based upon their performance evaluation with accuracy of 92.19% and 80.28% respectively which was better than the previously reported tools such as Mipred, HuntMI, miPred, microPred and Triplet-SVM. The trained classifiers were also tested on precursor microRNA of related phyla.

CHAPTER 3: Development of cloud based web interface for the trained ML models to predict novel insect pre-miRNA and search target in model organism, *Drosophila melanogaster*.

Implementation of the Support Vector Machine and Random Forest predictive model discussed in Chapter 2, as a web-based tool has been outlined in this chapter. Furthermore, experimentally validated miRNA targets of model organism *Drosophila melanogaster* were collected from various databases which was used as a database for searching target of positively identified pre-miRNA sequences by the predictive model. This chapter also highlights various aspects of API development for interaction with predictive machine learning models for web-development, along with security features and exception handling while getting requests from users. This chapter provides a roadmap for building full-stack web application using raw Linux server. The API is based on python's Flask framework which is bound to the public Linux server IP with Gunicorn. The monitoring of the running of Guniron, restarting when it crashes and bug reporting is done by Supervisor. NGINX is used as the reverse proxy to get the precursor microRNA query from the User as HTTP request. For Target searching, users need to specify the relevant information for choosing the mature miRNA sequence from the pre-miRNA sequence. RNAinsecta is freely available at : https://rnainsecta.in/ and the source code at : https://github.com/adhiraj141092/RNAinsecta

CHAPTER 4: Comparative analysis of sequential and thermodynamic features of insect pre-miRNA with other groups of organisms.

This chapter contains a discussion about the comparative statistical analysis of various features used the in development of machine learning based predictive tools. The sequence features of insect precursor microRNA were compared with precursor microRNA of other available organisms. We initially established that features such as Length, GC content, Minimum Free Energy (MFE) of folding, etc., differs in insects as compared to other organisms using chi-square test. We further trained a predictive model for classification using XGBoost between insects, human, monocots, aves, ruminants, sauria, dogs and rodents. We performed PCA and retained 20 best features for classification. Various parameters of XGBoost was tuned with 5-fold CV and the parameter values with highest CV score were considered. The accuracy of precursor microRNA prediction in insect, plants, rodents, human, ruminants, sauria, aves and dogs were found to be 0.9125, 0.9325, 0.8575, 0.86, 0.8875, 0.8225, 0.845 and 0.8675

respectively. The AUC was found to be 0.934 for insects, 0.985 for plants (rice), 0.941 for sauria, 0.859 for dog, 0.815 for ruminant, 0.743 for human, 0.75 for rodent and 0.765 for aves.

CHAPTER 5: Conclusion

This chapter contains an overview of the development of a web-based tool RNAinsecta for prediction of pre-miRNA in insects. It highlights the training of machine learning classifiers to predict insect precursor microRNA and analysis of their performance with previously developed tools. This chapter also summarises the implementation of the predictive models for a free, web-based tool that predicts novel insect precursor microRNAs and further allows users to search the genome of the model organism, *Drosophila melanogaster*, for the mature microRNA's targets. The developed tool can be run through any shell script in Linux which can be easily implemented in any pipeline.

List of Abbreviations:

Acc	:	Accuracy
ago	:	Argonaute
AI	:	Artificial Intelligence
ANN	:	Artificial Neural Network
API	:	Application Programming Interface
AUC	:	Area Under the Curve
CDS	:	Coding sequences
CRISPR	:	Clustered Regularly Interspaced Short Palindromic Repeats
cm	:	Confusion Matrix
CV	:	Cross Validation
DNA	:	Deoxyribonucleic Acid
G4	:	G-Quadruplex
k-NN	:	*k*-Nearest Neighbours
LogR	:	Logistic Regression
MCC	:	Matthews Correlation Coefficient
miRNA	:	microRNA
ML	:	Machine Learning
MFE	:	Minimum Free Energy
mre	:	miRNA Response Element
mRNA	:	Messenger RNA

NCBI	:	National Center for Biotechnology Information
NM	:	Near-Miss
PCA	:	Principal Component Analysis
PCG	:	Protein Coding Genes
dD	:	Base-pair Distance
dQ	:	Shannon Entropy
dG	:	Normalised MFE
DSN	:	Duplex Specific Nuclease
F1	:	Harmonic Mean of SN and p
FDR	:	False Discovery Rate
FP	:	False Positive
FN	:	False Negative
FNR	:	False Negative Rate
HTML	:	Hypertext Markup Language
HTTPS	:	Hypertext Transfer Protocol Secure
LAMP	:	Loop Mediated Isothermal Amplification
IP	:	Internet Protocol
p	:	Precision
PACT	:	Protein Activator of Protein Kinase R
PCR	:	Polymerase Chain Reaction
pb	:	Base Propensity

pid	:	Process ID
pre-miRNA	:	Precursor microRNA
qPCR	:	Realtime PCR
RAN	:	RAs-related Nuclear protein
RCA	:	Rolling Circle Amplification
RISC	:	RNA-Induced Silencing Complex
RF	:	Random Forest
ROC	:	Receiver Operating Characteristic
RNA	:	Ribonucleic Acid
rRNA	:	Ribosomal RNA
SDA	:	Strand-displacement Amplification
SHA	:	Secure Hashing Algorithm
SMOTE	:	Synthetic Minority Over-sampling Technique
SN	:	Sensitivity
SP	:	Specificity
SSL	:	Secure Sockets Layer
SVM	:	Support Vector Machine
TLS	:	Transport Layer Security
TRBP	:	Transactivation Response Element RNA-Binding Protein
tRNA	:	Transfer RNA

TN	:	True Negative
TP	:	True Positive
zQ	:	Normalised Shannon Entropy
zP	:	Normalised pb
zD	:	Normalised dD

List of Figures:

Figure 1.1	:	Timeline of miRNA research. The evolution of experimental and computational aspects of miRNA is shown. Red, green and blue represent miRNA biology, experimental technology and development of computational resources respectively.
Figure 1.2	:	Biogenesis of microRNA production. miRNA genes are transcribed in the nucleus as pri-miRNA which is processed by Drosha along with DGCR8 to produce pre-miRNA. pre-miRNA is then transported out of the nucleus to cytoplasm by Exportin-5 and Ran protein. pre-miRNA is then further processed by Dicer1 in presence of TRBP or PACT to cleave the loop to form mature miRNA which is then loaded onto AGO1-4 protein for its functionality. Ago1–4 regulate the processing of pre-miRNA and the assembly of the RNA-induced silencing complex (RISC).
Figure 1.3	:	Pie chart for ML based tools developed using various algorithms for miRNA research. Most number of tools have been designed using SVM since it's one of the simple and oldest algorithms, followed by Neural Network and Random forest. Naïve bayes and k-NN had are the least implemented ML algorithms for pre-miRNA detection.
Figure 2.1	:	Workflow for training ML classifier to predict insect pre-miRNA.
Figure 2.2	:	Radar plot for comparative performance analysis of RNAinsecta SVM and RF with already published tools.

Figure 2.3	:	ROC-AUC for comparative performance analysis of RNAinsecta with available tools for detection of insects' pre-miRNA. The validation dataset used for this figure contains 464 positive and 536 negative sequences. Y axis contains the True Positive Rate (TPR) and X axis contains the False Positive Rate (FPR). More the AUC (Area Under the Curve) better is the performance
Figure 2.4	:	AUPRC of the ML models
Figure 3.1	:	Chromosome wise miRNA target distribution.
Figure 3.2	:	Web-server implementation of RNAinsecta. The figure shows the full-stack implementation of the website. NGINX takes nucleotide sequence information as HTTP request from User through the Homepage. Flask runs the API for pre-miRNA and target prediction as a localhost. Gunicorn works as mediator between NGINX and Flask which allows public IP to interact with the APIs. The monitoring of Gunicorn is done by supervisor, which prepares error reports and restarts the server if it stops unexpectedly.
Figure 3.3	:	Compilation and deployment of development server created using python FLASK.
Figure 3.4	:	Supervisor running WSGI server.
Figure 3.5	:	NGINX successfully running in the backend server. The Active status shows running along with a master process and 2 worker processes.
Figure 3.6	:	The homepage of RNAinsecta. It contains search option for both the Random Forest and SVM classifier for single and batch query. Example sequences are also provided which is displayed upon clicking on 'Example".
Figure 3.7	:	The predicted results of RNAinsecta. A. shows the result for batch queries. It accepts a maximum of 200 sequences. B. shows the result

		for single sequence queries. It shows the graphs and charts corresponding to the feature of the input nucleotide sequence.
Figure 3.8	:	RNAinsecta's user interface for searching miRNA targets in Drosophila melanogaster. B. The result of miR target search containing Transcript ID and its hyperlink to FlyBase as well as miRBase ID and its hyperlink.
Figure 3.9	:	Site information provided by Google Chrome. This information is displayed when clicked on the padlock located in the top left corner.
Figure 3.10	:	SSL certificate of RNAinsecta. This certificate is generated by Let's Encrypt using certbot. This certificate renews every 3 months
Figure 3.11	:	Example of exception handling using Regex in JavaScript. As the input sequence contains characters other than the desired ATGC nucleotide bases, the request was not sent to the backend web server.
Figure 3.12	:	QR code for RNAinsecta.
Figure 4.1	:	Workflow for XGBoost training.
Figure 4.2	:	Comparison of insect pre-miRNA with other class of organisms. The comparation of various features of 500 randomly sampled pre-miRNA from insects, human, monocots, aves, ruminants, sauria, dogs and rodents are is shown in the pair-plot scatter diagrams. Features such as GC percentage (%G+C), Length (Len) and dG (MFE/Length) were considered out of 57 features mentioned below. Multivariate gaussian distribution plot is given in the diagonals.
Figure 4.3	:	Accuracy, MCC and F1 Score of classifiers trained on 7, 14,20,21,28 and 35 PCA estimation features evaluated on insect pre-miRNA. Performance with 20 features was found to be most optimum.
Figure 4.4	:	The performance comparison of XGBoost classifier.

Figure 4.5 : ROC-AUC of XGBoost classifier.

List of Tables:

Table 1.1	:	List of previously reported tools trained on various machine learning algorithms, adopted from previous review (Stegmayer et al., 2019)
Table 2.1	:	List of all the features calculated for training the binary machine learning classifiers.
Table 2.2	:	3391 pre-miRNA sequences downloaded for each species from miRBase.
Table 2.3	:	Sets of best performing parameters obtained after gridsearching through different values for each classification algorithm. Selection was based on the highest 10-fold cross-validation score. Parameters tuned for SVM were Cost function CSVM, Kernel and Gamma value. Parameters considered for RF were No. estimators, Max depth, Min sample leaf and Min sample split. For LogR the parameters considered were Cost Function CLR and solver. For KNN, No. of neighbours and Metric Distance was considered for tuning. The CV score shows the mean accuracy score of the best classifier obtained with the corresponding parameters.
Table 2.4	:	Performance measure for each classifier of SMOTE from Sl. No. 1-4, NM from Sl. No. 5-8, Imbalance from Sl. No. 9-12 and 8494 human_CDS negative datasets from Sl. No. 13-16. Accuracy, Specificity (SP), Sensitivity (SN), Matthew's correlation coefficient (MCC), Precision (p), harmonic mean of sensitivity and precision (F1) are given corresponding to each ML classifier.
Table 2.5	:	Comparative performance analysis between available tools and trained models tested upon independent insect pre-miRNA validation dataset. 1-5 shows the performance of previous tools. 6-9 shows the performance on the SMOTE classifiers, 10-13 shows the performance on Imbalance set. The parameters for evaluation are Accuracy, Specificity (SP), Sensitivity (SN), Matthew's correlation coefficient

		(MCC Precision (p), harmonic mean of sensitivity and precision (F1) are given for each corresponding ML classifier.
Table 2.6	:	Performance evaluation on imbalance class dataset (M_test) containing 116230 negative and 464 positive samples.
Table 2.7	:	Performance of RNAinsecta_RF in comparison with miPred for prediction of pre-miRNA across related phyla. pre-miRNA of different species from Nematoda, Platyhelminthes, Virus and Mollusca and their performance based on TP and SN is shown.
Table 3.1	:	No. of targets from each chromosome of Drosophila melanogaster.
Table 4.1	:	Total sequences collected for the analysis.
Table 4.2	:	Parameters calculated for feature extraction.
Table 4.3	:	Weight associated with all the features calculated from permutation feature importance package of Scikit-Learn.
Table 4.4	:	Parameter values considered for XGBoost training obtained after 5 fold CV during Grid Searching.
Table 4.5	:	Performance evaluation of the XGBoost classification.

CHAPTER 1

Introduction and Literature Review

1. Introduction and Literature Review

Pre-microRNAs (pre-miRNA) are the precursor of microRNAs (miRNA) from which one or more miRNAs are produced that regulate gene expression. They are typically ~22 bp long and bind to the 3′ untranslated region (3′ UTR) of target mRNAs to induce mRNA degradation and translational repression (Ha & Kim, 2014). However, recent studies have suggested that they also bind to the 5' UTR, coding region and gene promoters (Broughton et al., 2016). *Lin-4* was the first microRNA to be discovered by Ambros and Ruvkun groups in *Caenorhabditis elegans* in the year 1993 (R. C. Lee et al., 1993; Wightman et al., 1993). Since then it has been discovered in a large number of species across different kingdoms.

The mode of action of miRNA is similar across all the kingdoms. They bind to the miRNA transcript and thus, downregulate the gene expression. (O'Brien et al., 2018). In humans, they are found to be associated with various disease conditions such as, cancer where their distinct signatures are found in B cell chronic lymphocytic leukemias (Calin et al., 2004), mutation in miR-96 causes progressive hereditary hearing loss (Mencía et al., 2009), reactivation distinct set fetal miRNA causes heart disease and cardiac-arrest (Thum et al., 2007). In other animals, miRNA takes part in many developmental stages such as early vertebrate development, late vertebrate development, nervous system development (Wienholds & Plasterk, 2005). It is also found to take part in brain morphogenesis (Giraldez et al., 2005). It has been reported that animals that are unable to produce mature miRNAs do not survive or reproduce (Bushati & Cohen, 2007).

The role of miRNA is crucial in insects as it is reported to participate in a wide range of biological activities (Belles et al., 2012) . Changes in the miRNA profiles have been observed during metamorphosis where miR-100/let-7/miR-125 cluster has been found to participate in wing morphogenesis in both hemimetabolan and holometabolan species (E. Gomez-Orte &

Belles, 2009; Ling et al., 2014; Q. Zhang et al., 2019). In reproduction, during ovarian development miR-309 plays a critical role in female *A. aegypti* mosquitoes and during spermatogenesis miR-7911c-5p is upregulated in *B. dorsalis* (K. Tariq et al., 2016; Yang Zhang et al., 2016). Several miRNAs has been found to play important role in the regulation of immune related genes (Yin et al., 2018; X. Zhang et al., 2014) and also during insecticide resistance where the genes responsible are downregulated with the help of miRNAs, miR-2b-3p is found to be involved in regulation of metabolic resistance (K. Etebari et al., 2018; Y. Zhang et al., 2018).

In recent years numerous tools have been designed to predict pre-miRNA using machine learning approaches. Tools for species specific pre-miRNA detection like TarMirPred (Jaiswal et al., 2019) and phylum specific such as ViralMir (K.-Y. Huang et al., 2015) have also been developed. Most of the tools use the characteristics of the hairpin loop as features for the classification (Gkirtzou et al., 2010; Hertel & Stadler, 2006; T.-H. Huang et al., 2007; Jiang et al., 2007; M. E. Rahman et al., 2012; Y. Xu et al., 2008; Xue et al., 2005). Most tools consider 8,494 non-redundant human pseudo hairpins as the negative dataset (J. Chen et al., 2016; Fu et al., 2019; Jiang et al., 2007; Ng & Mishra, 2007; J.-H. Xu et al., 2008; Xue et al., 2005), however selection of negative dataset still remains a challenge and careful consideration is required to make efficient binary supervised classification models (Allmer & Yousef, 2012; Gomes et al., 2013).

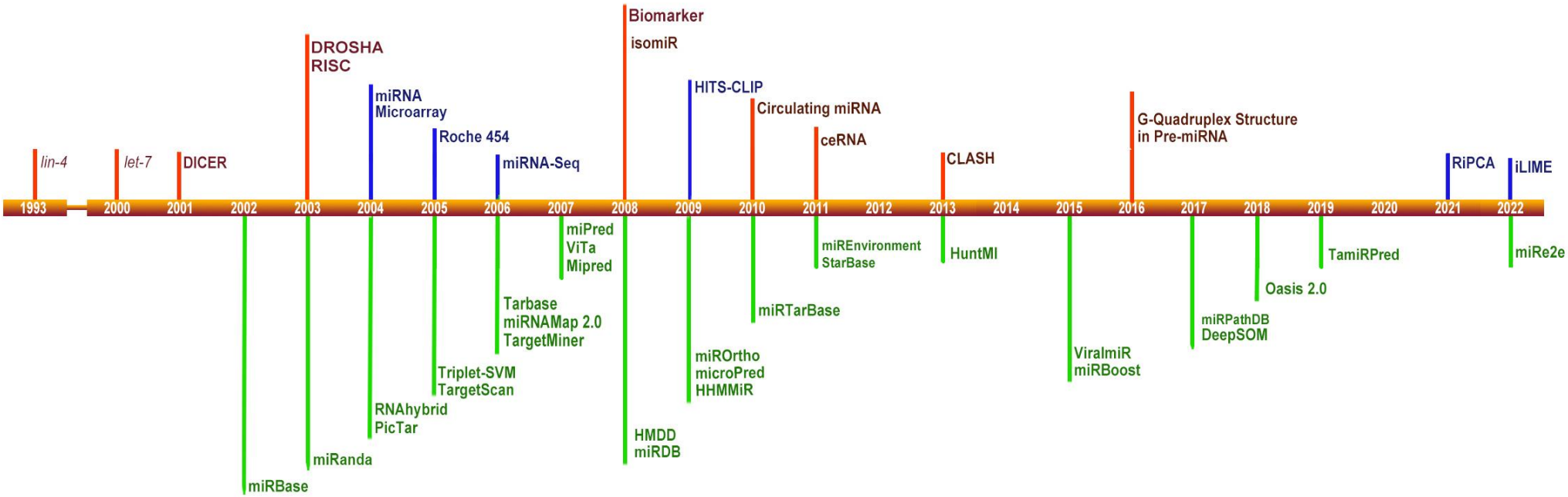

Figure 1.1: Timeline of miRNA research. The evolution of experimental and computational aspects of miRNA is shown. Red, green and blue represent miRNA biology, experimental technology and development of computational resources respectively.

1.1. microRNA biogenesis

microRNAs are produced as primary-microRNA (pri-miRNA) in the nucleus by RNA Polymerase II and III (Borchert et al., 2006; Y. Lee et al., 2004). pri-miRNA contains a hairpin stem, a terminal loop and a flanking single stranded sequence of several bases (Romero-Cordoba et al., 2014). pri-miRNAs undergo post transcriptional modification where they are capped at the 5´end and polyadenylated at the 3´end (Cai et al., 2004). They are further processed by RNAse III Drosha, cleaving them few bases from the hairpin stem (Blaszczyk et al., 2001). Drosha may form two distinct complexes during miRNA biogenesis to facilitate pri-miRNA cleavage. One is made up of RNA helicases p68 and p72, as well as heterogeneous nuclear ribonucleoproteins (hnRNPs). The other complex, known as the microprocessor, is made up of Drosha and the DiGeorge syndrome Critical Region 8 protein (DGCR8), a dsRNA-binding protein that interacts with Drosha's C-terminal domain to stabilise it (Han et al., 2004a, 2009). DGCR8, also serves as a molecular ruler, directing the cleavage of Drosha to the hairpin stem. Drosha digestion can occur co-transcriptionally or before splicing, and the outcome is an intermediary RNA molecule known as pre-miRNA, which has ~22 nucleotides in the stem and ~48 nucleotides in the terminal loop in humans (Han et al., 2006; Y. Lee et al., 2003; Morlando et al., 2008; Zeng & Cullen, 2003).

Exportin-5, a Ran-GTP-dependent dsRNA-binding protein, transports the generated pre-miRNAs to the cytoplasm in a GTP-dependent process (Yi et al., 2003). pre-miRNAs also are protected from nuclear degradation by exportin-5 (Zeng & Cullen, 2004). Dicer, another RNase III enzyme, digests the pre-miRNA in the cytoplasm to form a ~22nt mature duplex miRNA (miRNA:miRNA*, where miRNA* is the passenger strand) (Feng et al., 2012; Gregory et al., 2005). Dicer is associated with other proteins such as TAR RNA binding protein (TRBP) and kinase R-activating protein (PACT) during this process to increase its stability and processing activity (Chendrimada et al., 2005; Y. Lee et al., 2006). Dicer is an essential protein

in miRNA maturation, and its inhibition reduces mature miRNA levels. In fact, the absence of Dicer is lethal under certain conditions (Bernstein et al., 2003; Davis et al., 2008; Wienholds et al., 2003). The strands are unwound in an ATP-independent process after the miRNA duplex is formed. The miRNA-guide strand is loaded onto the RNA-induced silencing complex (RISC), which is formed by the association of Dicer, TRBP (transactivation response element RNA-binding protein), PACT (Protein Activator of Protein kinase R), and, Argonaute 1- 4 protein, most commonly 2 (Ago2) (Chendrimada et al., 2005; Maniataki & Mourelatos, 2005). (Figure 1.2).

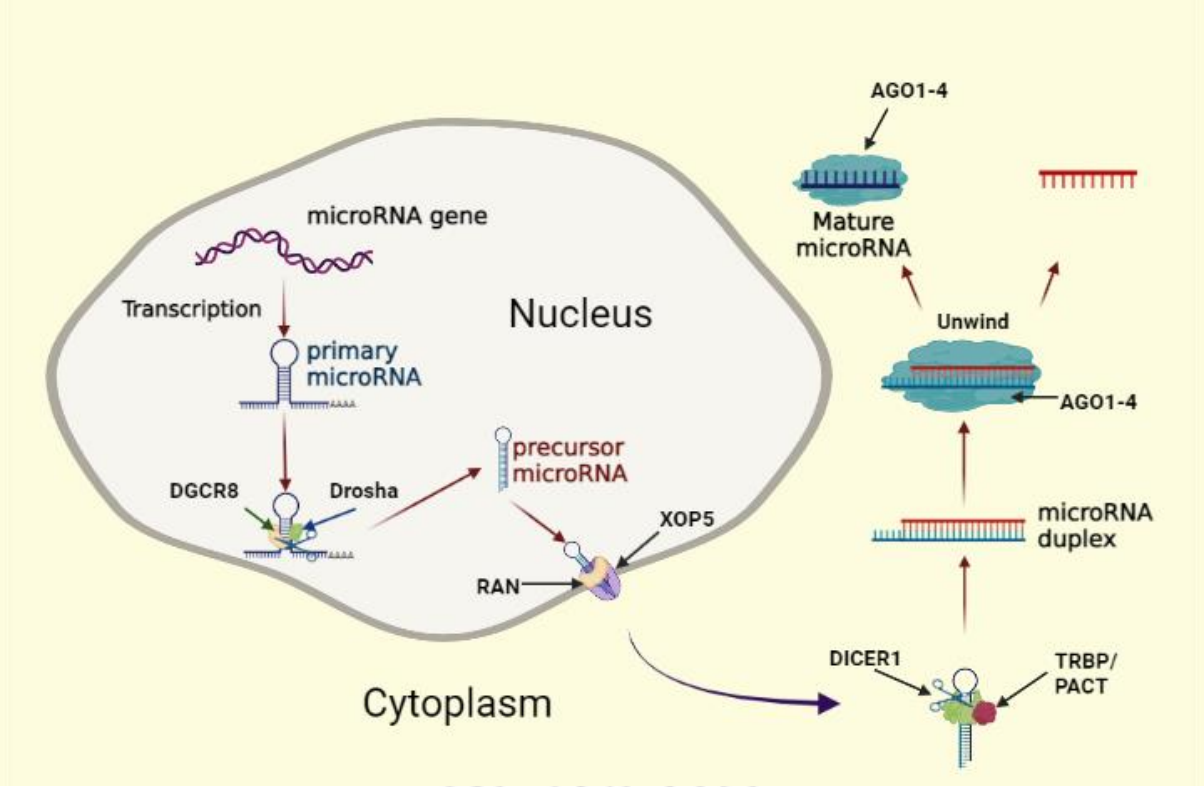

Figure: 1.2: Biogenesis of microRNA production. miRNA genes are transcribed in the nucleus as pri-miRNA which is processed by Drosha along with DGCR8 to produce pre-miRNA. pre-miRNA is then transported out of the nucleus to cytoplasm by Exportin-5 and Ran protein. pre-miRNA is then further processed by Dicer1 in presence of TRBP or PACT to cleave the loop to form mature miRNA which is then loaded onto AGO1-4 protein for its functionality. Ago1–4 regulate the processing of pre-miRNA and the assembly of the RNA-induced silencing complex (RISC).

Compared to the annotated miRNA sequences, there are also shorter miRNA sequence that

are not produced by the canonical pathway. They are primarily derived from Drosha and Dicer imprecise cleavage, 3′ addition events, RNA editing, and single nucleotide polymorphisms (L. Guo & Chen, 2014). These miRNAs are termed as isomiR since they are isoforms of miRNA produced from the same pre-miRNA with different 5' or 3' (Morin et al., 2008).

Intercellular signalling is another important function of miRNA. Although the majority of miRNAs are found within the cell, a significant proportion migrates outside of it and can be found in bodily fluids (Weber et al., 2010; Zubakov et al., 2010). Circulating miRNAs are released in blood, urine, saliva, seminal fluid, breast milk, and other fluids as a result of tissue damage, apoptosis, and necrosis (Condrat et al., 2020; Zen & Zhang, 2012). miRNAs were first identified as cancer biomarkers in 2008, when Lawrie et al. used them to examine the serum of patients with diffused large B-cell lymphoma. (Lawrie et al., 2008). Circulating miRNA has been found to act as potential biomarkers in cancer prognosis (Mitchell et al., 2008).

In addition to the traditional microRNA→RNA function, there is a reversed RNA→microRNA scenario in which coding and noncoding RNA targets can communicate by competing for microRNA binding which is mediated by microRNA-binding sites ("microRNA response elements," or "MREs") is termed as the competing endogenous RNA (ceRNA) hypothesis (Salmena et al., 2011). A conservative estimate suggests that miRNAs control more than 60% of mammalian genes, but the majority of the target sites are unknown (Friedman et al., 2009).

Multiple vertically stacked guanine tetrads compose a G-quadruplex. Four guanine bases form a guanine tetrad plane (G-quartet) by forming Hoogsteen hydrogen bonds, and then two or more G-quartet planes stack on top of each other to form a G4 structure. (Qin & Hurley, 2008). G4 structures are also observed in pre-miRNA structures for successful production

mature microRNA like miR-27a, miR-149, miR-92b, let-7e, etc. (Kwok et al., 2016; Pandey et al., 2015).

1.2. Role of miRNA in insects:

There are currently 3855 pre-miRNAs reported for insects in miRBase. Most number of reported miRNA of insects is from *Drosophila* genus. However, the highest number of microRNA from a single reported species is *Bombyx mori.* There are 469 mature miRNA reported from 258 pre-miRNA sequences for *Drosophila melanogaster*, 328 mature miRNA of *Drosophila virilis* from 180 precursors has been reported in miRBase. *Bombyx mori* has 487 pre-miRNA producing 563 mature miRNA.

1.2.1. miRNA in metamorphosis:

Metamorphosis is the process of transformation of insects from nymph to adult in hemimetabolous insects and larva-pupa-adult in holometabolous insects. Hemimetabolous insects gradually grow from nymph to adult, whereas in holometabolous insects the larval and pupal stages appear contrastingly different from the adult stage. Once adult, they are sexually active miRNA plays a crucial role in metamorphosis in both hemimetabolous and holometabolous insects. (Eva Gomez-Orte & Belles, 2009; Song & Zhou, 2019)

let-7 which was originally discovered in *C. elegans* as a part of a pathway of heterochronic genes promoting stage-specific cell fate decisions. It was found to be essential component of metamorphosis in both hemimetabolous and holometabolous insects (Caygill & Johnston, 2008; Eva Gomez-Orte & Belles, 2009; S. Liu et al., 2007). Interestingly the abnormal expression of *let-7* has been found in cancer prognosis (Büssing et al., 2008).

In holometabolous insect *Drosophila m. let-7* contains in a single copy at chromosome 2L in 18472034 to 18472111 position. It is found that *let-7* and *mir-125* plays a vital role in molting to adult stages. Defects in these miRNAs is found to cause delays in two metamorphic processes, i.e. terminal cell-cycle exit in the wing and maturation of neuromuscular junctions (Caygill & Johnston, 2008). In *Bombyx mori* the expression of *let-7* begins in the early first

moult. It remains low until the early third instar, and then rapidly grows to a peak in the early third moult, reaching a maximum at pupa and imago. (S. Liu et al., 2007). It is reported to be present in NW_004582019.1 (NCBI nucleotide accession) scaffold in miRbase from 3070428 to 3070450 nucleotide position.

In hemimetabolous insects, absence of *let-7* causes wing deformation. In *Blattella germanica,* cockroach which is a hemimetabolous insect, loss of Dicer-1 led to reduced levels of mature Let-7. This results in nymphoid formation instead of nymph where the wings are twisted and shortened, forms black abdominal sternites and has deformities in genital region. They die within 9-14 days (E. Gomez-Orte & Belles, 2009).

1.2.2. miRNA in reproduction:

miRNA plays a crucial role in insect reproduction. In *Aedes aegypti* (yellow fever mosquito), miR-309 participate in ovarian development while mir-277 regulates insulin-like peptides 7 and 8 to control lipid metabolism. CRISPR/Cas9 mediated knockout of both the miRNAs led to breakdown of lipid storage and ovarian development, and caused failure in primary follicle formation. Further, miR-1890 has been found to downregulate serine protease JHA15 which affects blood digestion, ovarian development, and egg deposition in *A. aegypti*. (Ling et al., 2017; Lucas et al., 2015; Q. Zhang et al., 2021; Yang Zhang et al., 2016)

miR-8-2p has been found to regulate spermatogenesis in *Bactrocera dorsalis* (Kaleem Tariq et al., 2016b). It targets mitoferrin which is involved in the transportation of iron inside the mitochondria (Brazzolotto et al., 2014; Froschauer et al., 2009). Feeding miR-8-2p mimic resulted in lower spermatozoa count as well as viability (Kaleem Tariq et al., 2016b). Additionally, miRNA-1-3p act as a sex determining factor in *B. dorsalis.* Its mimic injected in early embryogenesis resulted in most phenotypic males while the knockdown by CRISPR/Cas9 resulted in more phenotypic females (Peng et al., 2020).

Similar miRNA functions in reproduction has been reported in various other insects as well. In *Helicoverpa armigera,* feeding miR-2002b mimic resulted in significant decrease in reproduction rate due to the inhibition of *trypsin-like serine protease* (Jayachandran et al., 2013). In *Locusta migratoria,* miR-278 is reported to regulate oogenesis along with let-7 by targeting the transcription factor Krüppel-homolog 1 (Song et al., 2018). Additionally, miR-2/13/71 cluster was found to regulate oogenesis by targeting Notch transmembrane protein (Song et al., 2019).

1.2.3. miRNA in immune response:

miRNAs play a vital role in regulating immune related genes in pest insects. In *Laodelphax striatellus*, miR-315-5p enhances rice black-streaked dwarf virus infection by targeting the melatonin receptor. Upregulation of miR-375 and miR-927 in *Aedes aegypti* promotes Dengue virus serotype 2 infection, while inhibition of miR-2b promotes Chikungunya virus replication. miR-8 is downregulated in *Plutella xylostella* (diamondback moth) after being affected by the parasite *Diadegma semiclausum*, resulting in enhanced production of anti-microbial peptides. The JNK pathway is activated in *A. pisum* (pea aphid) in response to bacterial infection, and agomir-184 (mimic) injected insects resulted in higher fatality rate than control aphids. (Avila-Bonilla et al., 2020; Dubey et al., 2017; Kayvan Etebari & Asgari, 2013; Hussain et al., 2013; Ma et al., 2020; J. Zhang et al., 2021; Q. Zhang et al., 2021)

Furthermore, several researches have shown that miRNAs can play a cross-species regulatory role in pest parasitism. Parasite miRNAs can directly modulate host immunity to favour and increase parasite replication. Two *Snellenius manilae bracovirus* (SmBV) miRNAs, miR-199b-5p and miR-2989, repress host genes domeless and toll-7, respectively, to modulate host innate immune responses upon parasitism in infected *S. litura*. In turn, host miRNA can influence parasite reproduction. Rice stripe virus (RSV) replication is inhibited in *L. striatellus* by host

miR-263a targeting the viral RNA1 region. The findings revealed that miRNAs in hosts or parasites play essential regulatory functions during parasitism, laying the groundwork for pest or insect-borne virus management. (Tang et al., 2021; Q. Zhang et al., 2021; Zhao et al., 2021)

1.2.4. miRNA in Insecticide resistance

Pests' detoxification genes play essential roles in pesticide resistance and plant toxin tolerance at the molecular level in insects. miRNAs have been found to be involved in insecticide resistance by targeting detoxification genes. (X. Li et al., 2020; B. Liu et al., 2016)

Cytochrome P450 (CYP) enzyme genes are commonly described as targets and are regulated by distinct miRNAs among the detoxification genes. miRNAs (miR-13664, miR-2/miR-13, miR-71 and miR-278-3p) have been found to regulate deltamethrin resistance in the deltamethrin resistant strain of *Culex pipiens pallens.* Injecting mimics of these miRNA led to the downregulation of CYP314A1, CYP9J35, CYP325BG3, and CYP6AG11 genes thereby increasing the resistance to deltamethrin. (Q. Guo et al., 2017; Lei et al., 2015; Sun et al., 2019) In *Aphis gossypii,* during *de-novo* fatty acid synthesis, acetyl-CoA carboxylase (ACC) catalyses the carboxylation of acetyl-CoA to produce malonyl-CoA. miR-276 and miR-3016 together regulate the expression of ACC transcript to control spirotetramat resistance. (Wei et al., 2016). The diamide insecticides work by activating the ryanodine receptors (RyR) in muscle fibres, causing feeding cessation, muscle paralysis, and eventually death (Troczka et al., 2012; Xingliang Wang & Wu, 2012; Zalucki et al., 2012). In *Plutella xylostella,* diamide insecticide resistance has been linked to amino acid mutations in the RyR and increased activity of detoxification enzymes which is controlled by two miRNAs, miR-7a and miR-8519 (X. Li et al., 2015).

miRNAs have recently been reported to influence *Bacillus thuringiensis* (Bt) resistance in

herbivorous pests. miR-998-3p controlled Cry1Ac resistance in three lepidopteran pests (*H. armigera, Spodoptera exigua, and P. xylostella*) by targeting ATP-binding cassette subfamily C member 2. (ABCC2). miR-998-3p mimic injection substantially improved resistance to Cry1Ac toxin in *H. armigera, S. exigua, and P. xylostella* (Cry-sensitive) larvae, while suppression of miR-998-3p greatly raised ABCC2 expression and decreased survival rates in the *P. xylostella* Cry1Ac-resistant population. (Q. Zhang et al., 2021; Zhu et al., 2020)

1.3. Resources and techniques for pre-miRNA characterization and function

Numerous techniques have been implemented to understand the biology of miRNA including both experimental and computational aspects.

1.3.1. Experimental Techniques:

Significant efforts have been made in recent decades to develop new detection methods for miRNA discovery and their role in regulation of gene expression (Cheng et al., 2018). These methods can be classified into two categories: traditional methods and new technology methods (Ye et al., 2019).

1.3.1.1. Traditional methods:

These are the early methods implemented for the detection of miRNA and are still being widely used.

- **Northern Blotting**

Among the early methods adopted for the discovery of miRNA, Northern Blotting is considered one of the standard ways for detection of miRNA and is still being widely used. It can be used to detect not only mature miRNAs, but also their precursors. It does not require the use of specialised equipment. The fundamental principle is: The RNA sample is digested with a restriction endonuclease, separated by agarose gel electrophoresis, denatured, and transferred to a nitrocellulose film based on its position in the gel, then fixed before reacting with isotope or other marker labeled probes. miRNAs can be detected using autoradiography or other suitable techniques after washing the free probe (Pall et al., 2007; Torres et al., 2011). Additionally, it is semi-quantitative, has a poor throughput, is difficult and time-consuming, and RNA is easily degraded. As a result, it demands a severe experimental condition. Northern

blot has low sensitivity, preventing the detection of RNAs with low molecular weight. Specific probes labelled with radioisotopes will increase the sensitivity, but also the risk of a reaction (Cheng et al., 2018; Ye et al., 2019).

- **Real-time qPCR:**

It is the current gold standard for detection of miRNA and is being routinely used because of its large dynamic range, high sensitivity, and high sequence specificity (Gan et al., 2011). The target miRNA is initially converted to cDNA through reverse transcription. PCR is then conducted to achieve real-time fluorescence detection. cDNA synthesis of miRNA by reverse transcription is mostly done using stem-loop primer (Czimmerer et al., 2013; Mohammadi-Yeganeh et al., 2013). It is also reported to be synthesized with a gene-specific primer (GSP) with a tail sequence (Raymond et al., 2005) or using a poly-(T) adapter (Niu et al., 2015; R. Shi & Chiang, 2005). Two fluorescent methods are utilised for monitoring miRNA qPCR: the TaqMan probe method and the SYBR Green fluorescent dye method (Varkonyi-Gasic et al., 2007).

The primary advantages of real-time qPCR over northern blotting are increased sensitivity, specificity, and quantification range. Its major disadvantages are variability in RNA template, improper experimental design, inconsistent data processing, and inappropriate data normalisation (Ye et al., 2019).

- **Microarray technology:**

Microarrays are the most common method for rapid and high-throughput miRNA detection (W. Li & Ruan, 2009). The sample RNA is reverse-transcribed using a labelled probe, and the fluorophores or biotin-labelled cDNAs are detected using solid-phase oligonucleotides with the same sequence as the target miRNA. The labelled cDNA sample is loaded into each well, followed by a series of washing steps designed to eliminate free DNAs. If the hybridised cDNA

is biotinylated, the streptavidin-labelled fluorophore can be labelled, and the fluorescence intensity of each well can be measured directly if the cDNA has been labelled with the fluorophore. The fluorescence intensity of each well can be used to determine the expression level of miRNAs.

Although microarray is one of the most sophisticated tool, yet suffers from numerous drawbacks such as miRNAs that are too short or have a low copy number cannot be detected, and the specificity of analysing miRNAs with similar sequences is poor (Ye et al., 2019). Microarray experiments are also very expensive.

1.3.1.2. New technology methods:

Research on miRNA has accelerated with the cutting-edge state-of-the-art techniques. Methods such as Nano-material based and Nucleic acid amplification methods are used to detect miRNA.

- **Nanomaterial based techniques:**

Various nanomaterial based techniques have been applied to detect miRNA such as gold nanoparticles (R. D. Li et al., 2016; Persano et al., 2016), silver nanoparticles (R. Liu et al., 2017; Pan et al., 2018; Salahandish et al., 2018), magnetic nanoparticle (Hosseinzadeh et al., 2018; Oishi et al., 2016) and quantum dots (Foda et al., 2014). Due to their enormous surface area, high electrochemical conductivity, and remarkable chemical stability, nanoparticles are potent instruments for enhancing the effectiveness of conventional detection techniques. In addition, when it comes to in vivo imaging, their potent variety of cellular transfection, high photostability, and minimal immunogenicity should be considered. However, their inherent cytotoxicity and self-aggregation within living cells continue to be problems for sustained use. The application of nanobiosensors being investigated further in an effort to address these shortcomings (Ye et al., 2019).

- **Nucleic acid amplification techniques:**

Usually, nucleic acid amplification techniques are applied to improve the sensitivity of miRNA detection. Due to their greater applicability to point-of-care testing devices than standard PCR-based assays, various innovative technologies based on isothermal amplification have emerged. Numerous nucleic acid amplification techniques, including rolling circle amplification (RCA), duplex-specific nuclease (DSN)-based amplification, loop-mediated isothermal amplification (LAMP), strand-displacement amplification (SDA), and some enzyme-free amplifications, have been utilised. With SYBR Green as fluorescent DNA-intercalating dyes, they are easily applicable to the real-time based assay. However, SYBR Green dyes can reduce the effectiveness of amplification in a dose-dependent manner and are susceptible to nonspecific amplification (Gudnason et al., 2007).

i. **RCA** (rolling circle amplification): Due to its ease of use, specificity, and high sensitivity, RCA has gained popularity in the detection of miRNA. In the majority of instances, miRNA functions as a ligation template, and the padlock probe will hybridise with the target miRNA, which will be ligated by T4 RNA ligase or SplintR enzyme, forming a circular ssDNA, followed by extension around the circle with an external primer or miRNA itself as a primer, displacing the conjoined miRNA and continuing to produce long cascaded nucleic acid products (Tian et al., 2019; H. Xu et al., 2018, 2019). Yang et al. described a multicomponent nucleic acid enzyme-mediated rolling circle amplification on a gold electrode for the ultrasensitive and precise detection of microRNA (J. Yang et al., 2016). Hong et al. presented a straightforward and practical method for the quantitative detection of miRNA using RCA, graphene oxide (GO), and fluorescently tagged peptide nucleic acid (F-PNA) (Hong et al., 2016). Xu et al. devised a sensitive and specific fluorescent technique based on the combination of

RCA and SDA for the detection of let-7a miRNA, which utilised a multifunctional molecular beacon (MMB) to perform SDA without the use of nucleic acid (H. Xu et al., 2018).

ii. **DSN** (duplex-specific nuclease)**:** DSN can hydrolyze DNA in DNA/RNA or double-stranded DNA (dsDNA) of a certain length, irrespective of the nucleotide sequence, and does not cleave single-stranded DNA (ssDNA) or RNA (Qiu et al., 2015). Based on this unique property of DSN, miRNAs can be recycled in the process, and thermal amplification was accomplished. Le et al. established a miRNA-21 detection method with DSN-based amplification that exhibited high sensitivity and specificity (one base mismatch discrimination). Amplifications based on DSN were compatible with many platforms, including colorimetric, fluorescence, and electrochemical assays (Le et al., 2018).

iii. **LAMP** (loop-mediated isothermal amplification): LAMP is a widely used isothermal reaction for the amplification of DNAs and RNAs which shows high sensitivity due to its exponential amplification property. It uses 4–6 distinct primers to simultaneously identify 6–8 different target sequences, thereby significantly enhancing the selectivity (Ye et al., 2019). In the majority of LAMP-based miRNA quantification techniques, miRNAs serve as reaction initiators. Primers can perform extension with the help of DNA polymerase and strand displacement DNA synthesis only in the presence of the miRNA target. The LAMP template has 4–6 pre-defined sequences for stem-loop creation, which also diminish the sensitivity of the method due to the synergistic hybridization and extension of the numerous primers along the lengthy template (C. Li et al., 2011).

iv. **SDA** (strand-displacement amplification): The mechanisms of SDA are nicking, polymerase extension, and strand displacement. Since miRNAs (in their overall role

as a template) are recycled in the process based on polymerase extension-driven strand displacement, linear amplification is usually the result. Within two cycles of nicking, polymerization, and displacement processes caused by target miRNA, Shi et al. reported an exponential SDA method. Notably, this one-pot assay could detect as little as 16 zmol of the target miRNA within 90 minutes (C. Shi et al., 2014). Enzyme-mediated reactions are limited by a variety of parameters, such as temperature and ionic composition, which makes the usual use of enzymes an impediment to the widespread application of SDA.

1.3.2. Computational Resources:

Various databases, pre-miRNA prediction tools, target searching tools have been developed in recent years that has helped researchers and scientists working in the experimental fields to understand the biology of miRNA. NGS based bioinformatics studies have also been carried out to understand the influence of miRNA abundance and differential gene expression.

There are various tools for prediction of pre-miRNA developed using different approaches. The first comprehensive collection of miRNA, 19iRbase (Kozomara et al., 2019), was released in 2002. Most of the tools were developed after that. The early tools were based on homology search algorithm such as miRscan, miRSeeker, etc. (Lai et al., 2003; Lim et al., 2003)

1.3.2.1. Databases:

After the discovery of the first miRNA, subsequently several new miRNAs were discovered in various organisms. The need to store this information in a well annotated database was crucial, hence, the first database dedicated for miRNA and pre-miRNA annotation information called miRbase was developed (Kozomara et al., 2019). Following this, a number of databases

on miRNAs were created such as miRNAMap (S. Da Hsu et al., 2008), miRTarBase (Chou et al., 2016; H.-Y. Huang et al., 2019), miRWalk (Dweep & Gretz, 2015), etc. Tools4mirs is an online database containing information about various computational resources available for miRNA (Lukasik et al., 2016). There are a total of 68 databases currently listed in Tools4mir at the time of writing this dissertation.

miRBase: The miRBase database, which was created in 2002 and is continuously updated, is a searchable collection of annotated and published miRNA sequences. The database contains data on the location and sequence of each item, which represents a predicted hairpin section of a mature miRNA sequence (also known as mir) (termed miR). At http://www.mirbase.org, both hairpin and mature sequences can be searched using a web browser. Access to the entries is possible via name, keyword, references, and annotation (Kozomara et al., 2019).

miRTarBase: Data on experimentally validated miRNA-target interactions can be found in miRTarBase (MTIs). Utilizing 138 crosslinking and immunoprecipitation sequencing (CLIP-seq) data sets supplied by 21 different studies, the most recent version of miRTarBase included the ability to systematically find Argonaute-miRNA-RNA connections. The database includes 3,48,007 MTIs from CLIP-seq, 7,439 MTIs with high validation (using reporter assays or western blots), and 4,966 articles. The database was improved after the most recent update in 2022. You can access the new web server at https://mirtarbase.cuhk.edu.cn (H.-Y. Huang et al., 2019; H. Y. Huang et al., 2022; Lukasik et al., 2016).

miRDB: The miRDB's recently updated features include 2.1 million projected gene targets controlled by 6709 miRNAs. It was first released in 2008 (Xiaowei Wang, 2008). Since then, regular maintenance has been performed, most recently in 2020. The web server interface is a new feature that allows user-supplied sequences to be uploaded for miRNA target prediction in addition to showing precompiled prediction results. Users are so allowed to

research any customised miRNAs or target genes of interest (Y. Chen & Wang, 2020).

miREnvironment: MicroRNA (miRNA), a class of recently discovered genetic variables, as well as interactions between environmental factors and their related phenotypes are all compiled in the database miREnvironment. It also contains a bioinformatics tool that forecasts the effectiveness of cancer treatment and links EFs to human disease (Q. Yang et al., 2011).

TarBase: TarBase provides for the first time thousands of hand curated, empirically proven, high-quality miRNA:gene interactions with extensive meta-data. It was first released in 2006 (Sethupathy et al., 2006). The analysis of favourable and unfavourable outcomes has been made simpler with DIANA-TarBase v7.0. Customization of the experimental settings, including cell/tissue type and treatment, as well as the experimental methodologies used is possible through the interface. The new interface provides sophisticated information ranging from the location of the binding site that has been physically and computationally discovered to the primer sequences utilised in cloning research (Lukasik et al., 2016; Vlachos et al., 2015).

miRNAMap 2.0: miRNAMap was initially released as a target searching tool in 2006. (Paul W.C. Hsu et al., 2006). However, miRNAMap 2.0 was developed as a database by assembling experimentally validated microRNAs and microRNA target genes from the human, mouse, rat, and other metazoan genomes. Three computational methods, miRanda, RNAhybrid, and TargetScan, were used to find miRNA targets in the 3'-UTR of genes, in addition to the known miRNA targets. (S.-D. Hsu et al., 2007) In addition, miRNA expression profiles can provide useful information regarding the properties of miRNAs, such as tissue specificity and differential expression in cancer/normal cells. (S. Da Hsu et al., 2008; Lukasik et al., 2016).

HMDD: It is a manually curated database of miRNA-related illnesses derived from research findings. The human miRNA-disease association data are more thoroughly and precisely annotated in HMDD v2.0. This comprises information about miRNA and disease that was

gleaned from genetic, epigenetic, circulating, and miRNA-target interaction findings. 2014's. (Y. Li et al., 2014).

StarBase: It is designed for decoding Pan-Cancer and Interaction Networks of lncRNAs, miRNAs, competing endogenous RNAs(ceRNAs), It was first released in 2011 and since then has been regularly maintained (J. H. Yang et al., 2011). RNA-binding proteins (RBPs), and mRNAs from massive CLIP-Seq data and tumour samples. starBase is also designed to decode Protein-RNA and miRNA-target interactions, including protein-lncRNA, protein-sncRNA, protein-mRNA, protein-pseudogene, miRNA-lncRNA, miRNA-mRNA, miRNA-circRNA, miRNA-pseudogene, miRNA-sncRNA relationships, and ceRNA networks from CLIP-Seq datasets. (J. H. Li et al., 2014)

miRPathDB: miRPathDB aims to complement existing target pathway web-servers by facilitating researchers' access to information regarding which pathways are regulated by a miRNA, which miRNAs target a pathway, and how specific the regulation is. The database includes a significant number of microRNAs, distinct microRNA target groups, and a wide range of functional biochemical categories. M. musculus data are also kept and can be compared to human target pathway information. (Backes et al., 2017) The version 2 of miRPathDB was released in 2020. (Kehl et al., 2020)

miROrtho: Combining orthology and a Support Vector Machine, miROrtho includes predictions of precursor miRNA genes across many animal genomes. We give extended homology alignments of previously identified miRBase families and potential miRNA families uniquely predicted by our SVM and orthology process (Gerlach et al., 2009).

ViTa: It is a database that collects virus data from miRBase and ICTV, VirGne, VBRC, etc., as well as known miRNAs on viruses and projected host miRNA targets from miRanda and TargetScan. ViTa also provides useful annotations, such as human miRNA expression, virus-

infected tissues, virus annotation, and comparisons (Paul Wei Che Hsu et al., 2007).

1.3.2.2. Prediction tools

There are 23 pre-miRNA prediction tools listed in Tools4miR. Many tools were designed to detect the hairpin-loop present in pre-miRNA. Various features of the pre-miRNA sequence such as Length, MFE (Minimum Free Energy), GC%, nucleotide counts, stem loop length, etc. are used to train machine learning classifiers. Triplet-SVM and MiPred were among the first tools developed for detection of pre-miRNA which used machine learning algorithms (Jiang et al., 2007; Xue et al., 2005). DeepSOM and HuntMi uses class imbalance algorithm for training the classifiers (Gudyś et al., 2013; Stegmayer et al., 2017).

Triplet-SVM: It was designed to predict pre-miRNA using triplet element scores to train machine learning classifiers. The triplicate features are calculated from the secondary structure of the pre-miRNA, predicted with RNAfold package from ViennaRNA software. Pseudo pre-miRNA were generated from the human genome consisting 8494 sequences which were labelled as the negative dataset whereas true human pre-miRNA were taken as true dataset upon which machine learning algorithm called SVM was trained (Xue et al., 2005).

miPred: Developed by Stanley Ng long, It is one of the oldest tools developed for detection of pre-miRNA hairpin loops based on parameters such as Shannon entropy, MFE, base pairing distance, dinucleotide shuffling, etc. The thermodynamic features are calculated using RNAfold and perl scripts. It is trained upon SVM machine learning algorithm (Ng & Mishra, 2007).

microPred: The microPred classifier system distinguishes genuine human pre-miRNA hairpins from pseudo hairpins and other noncoding RNAs. In both comparative and non-comparative settings, the microPred classifier could be utilised to predict unique human pre-miRNAs (Batuwita & Palade, 2009).

HHMMIR: It uses Hierarchical Hidden Markov Model (HHMM) to predict pre-miRNA. First, a template for the structure of a typical miRNA hairpin was constructed by compiling data from public databases. This template was then utilised to construct the HHMM topology (Kadri et al., 2009).

MiPred: It uses a hybrid feature consisting of local contiguous structure-sequence composition, minimum of free energy (MFE) of the secondary structure, and P-value of randomization test to identify actual pre-miRNAs from other hairpin sequences with similar stem-loops (pseudo pre-miRNAs). MiPred is trained upon Random Forest algorithm. (Jiang et al., 2007)

HuntMi: It is a machine learning miRNA classification tool that addresses the class imbalance issue known as ROC-select, which is based on the thresholding score function generated by conventional classifiers (Gudyś et al., 2013).

miRBoost: A classification method for microRNA precursors that employs a boosting methodology with support vector machine components to handle imbalanced training data. Following a feature selection of 187 fresh and existing features, classification is done. miRBoost delivers an optimal balance between prediction accuracy and execution speed (Tran et al., 2015).

DeepSOM: It is a machine learning-based tool for predicting pre-miRNA in genome-wide data. Clustering unlabeled sequences of a genome with well-known miRNA precursors for the organism under study enables the rapid identification of the best candidates for miRNA, as the unlabeled sequences cluster with the known precursors. A deepSOM model is utilised to solve the issue of having few positive class labels. (Stegmayer et al., 2017).

ViralmiR: It is a pre-miRNA prediction tool developed specifically for virus. The tool uses SVM for classification of viral pre-miRNA data and pseudo viral pre-miRNA data. It was

developed in 2015 and has a prediction accuracy of above 80%. (K.-Y. Huang et al., 2015)

TamiRPred: It is a pre-miRNA prediction tool for rice. It also searches for the target which is predicted by miRanda. The tool uses SVM for the classification. The predicted miRNA targets mapped from the rice genome using sliding window protocol has more than 4464 putative miRNA genes. (Jaiswal et al., 2019)

miRe2e: It is the first end-to-end Deep Learning model for the prediction of pre-miRNA. The model is based on Transformers which is a neural architecture that uses attention mechanisms to infer global relationships between inputs and outputs. It is able to accept genome-wide raw data as input, without pre-processing or feature engineering. After training using known pre-miRNAs, hairpin and non-hairpin sequences, it is capable of identifying all pre-miRNA sequences in a genome. Several experimental settings utilising the human genome were used to validate the model (Raad et al., 2022).

1.3.2.3. Target searching tools:

Various tools have been designed to search the target of miRNA in any given mRNA transcript such as miRScan, miRanda, etc. There are currently 60 tools listed in tools4miRs for target prediction.

miRanda: It is an algorithm used for searching miRNA targets in a given reference sequence. Initially a dynamic programming local alignment is carried out. This algorithm scoring is based on sequence complementarity and not sequence identity. The algorithm also allows the G:U wobble pair. (John et al., 2004)

PicTar: PicTar is a computational technique for locating typical microRNA targets. PicTar has an excellent success rate in predicting targets for both individual microRNAs and for combinations of microRNAs, according to statistical tests using genome-wide alignments of

eight vertebrate genomes, its capacity to specifically recover published microRNA targets, and experimental validation of seven predicted targets. (Krek et al., 2005)

Oasis 2.0: Oasis is a web application that makes it simple and quick to analyse small-RNA-seq (sRNA-seq) data online. It offers robust biomarker discovery through classification, multivariate sample analysis, and a sophisticated computational interface for batch task submission. Both modules perform functional analyses, including GO and pathway enrichment, on novel miRNAs, miRNA targets, and other data. (R.-U. Rahman et al., 2018)

RNAhybrid: RNAhybrid is a programme that calculates the least free energy of a long and a short RNA hybridization. The hybridization is done in a domain mode, which means that the short sequence is hybridised to the most suitable section of the long one. The tool's primary purpose is to forecast microRNA targets. (Krüger & Rehmsmeier, 2006)

TargetMiner: A programme called RNAhybrid determines which hybridization of a long and short RNA requires the least amount of free energy. The short sequence gets hybridised to the best portion of the lengthy sequence since the hybridization is carried out in a domain mode. The main objective of the technology is to predict microRNA targets. (Bandyopadhyay & Mitra, 2009)

1.4. Artificial Intelligence and Machine Learning in pre-miRNA prediction:

Artificial intelligence is founded on the idea that human intellect may be characterised in such a manner that a computer can easily imitate it and complete tasks ranging from the most basic to the most complicated. Learning, thinking, and perception are all aims of artificial intelligence. John McCarthy coined the term "artificial intelligence" (AI) in 1955, defining it as "the science and engineering of making intelligent machines". He was very influential in the early development of AI. With his colleagues he founded the field of AI in 1956 at a Dartmouth College conference on artificial intelligence.(Dick, 2019)

Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed. —Arthur Samuel, 1959

Machine learning (ML) is a branch of research concerned with understanding and developing techniques that 'learn,' that is, methods that use data to improve performance on a set of tasks (T Mitchell et al., 2003).

1.4.1. ML algorithms:

ML algorithms are divided into supervised, unsupervised and semi-supervised based on the availability of labels. When labels are provided for each class, it is known as supervised learning. Some of the supervised learning algorithms include Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), etc. Unsupervised learning is applied when the labels are not provided and the classification is done based on clustering. Some of the unsupervised learning algorithms include k-Nearest Neighbour (kNN), kMeans, DBScan, etc. (Ray, 2019)

Supervised learning algorithms: The goal of supervised algorithm is to create a mapping function *f(x)* which gives an output *y* for each input *x*. There are many different mapping

functions such as Support Vector Machine (SVM), Naïve Bayes, Random Forest, Decision Trees, etc. (Jordan & Mitchell, 2015)

- **SVM:** In this method, a hyperplane is calculated that serves as a decision boundary for the given set of data. This is calculated from the data points serving as supporting vectors. Based on the complexity and separability of the data, different kernels can be implemented such as linear kernel: $K(x_i, x_j) = x_i^T x_j$, polynomial kernel: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d$ and radial basis function (RBF) kernel: $K(x_i, x_j) = \exp\{-\gamma||x_i - x_j||^2\}$, where $x \in \mathrm{R}$, (i=1,2,3..,N) are inputs and $\gamma, r, d > 0$ are kernel parameters.

- **Naive Bayes:** It is the use of conditional probability to classify data with the help of Bayesian inference. The assumption taken is that occurrence of an even is independent of the other and hence, Naïve term is used. Bayes' theorem states the following relationship, given class variable y and dependent feature vector x_1 through x_n :

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n \mid y)}{P(x_1, \dots, x_n)}$$

Using the naive conditional independence assumption that

$$P(x_i|y, x_1, \dots, x_i - 1, x_i + 1, \dots, x_n) = P(x_i|y)$$

for all i, this relationship is simplified to

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)\prod_n^{i=1} P(x_i \mid y)}{P(x_1, \dots, x_n)}$$

Since $P(x1, \dots, xn)$ is constant given the input, we can use the following classification rule:

$$P(y \mid x_1, \dots, x_n) \propto P(y)\prod_n^{i=1} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \underset{y}{\arg max}\, P(y)\prod_n^{i=1} P(x_i \mid y)$$

we can use Maximum A Posteriori (MAP) estimation to estimate $P(y)$ and $P(x_i \mid y)$; the former is then the relative frequency of class y in the training set.

- **Decision Trees:** This method is mostly used for classification problem. There are two nodes in a decision tree. One is the Decision node and the other is Leaf node. The decision node makes the decision while the leaf node are the output of those decision. Decision nodes contain branches while Leaf nodes do not. A decision node or root node is divided into sub-nodes by Splitting. Pruning is the process of removing unwanted branches from the node. Entropy is the amount of information needed to accurately describe data. So, if data is homogenous that is all elements are similar then entropy is 0 (that is pure), else if elements are equally divided then entropy move towards 1 (that is impure). Mathematically it is written as:

$$Entropy = \sum_{i=1}^{n} p_i \log(p_i)$$

Giny impurity is the measure of amount of impurity in the node which is given by;

$$Gini\ index = 1 - \sum_{i=1}^{n} {p_i}^2$$

- **Logistic Regression:** It is another supervised learning algorithm which is used to classify data based on labels. In this method, instead of giving an output of a class such as 0 or 1, it gives probabilistic value based on the sigmoid function.

Mathematically it can be written as:

$$\sigma(t) = \frac{1}{1 + e^t}$$

where σ is the sigmoid function for the given input t. Assuming t to be a linear function of a single variable x:

$$t = \beta_0 + \beta_1 x$$

where β_0 is the y-intercept and β_1 is the slope.

- **Random Forest:** This approach creates a decision forest or random forest by combining a large number of decision trees. Averaging is used to increase predictive accuracy and control over-fitting. It works on:

$$RFfi_i = \frac{\sum_{j \in N} normfi_{ij}}{T},$$

where $RFfi_i$ is the importance of feature *i* (fi_i) calculated from N which denotes all trees in the Random Forest model, $normfi_{ij}$ is the normalized feature importance for i in tree j, i.e. $normfi_i = \frac{fi_i}{\sum_{j \in F} fi_j}$ and T is total number of trees and F denotes all features.

***k*-NN:** It is a non-parametric supervised learning algorithm which implies it does not take any assumption into account. It is based on the distance between *k* nearest data point of a category which is used for training, so when a new datapoint is given as input, it should predict the category where it falls into. There are several distance matrices which is used for calculating the distance between the neighbours such as

$$euclidean = \{\sum_{i=1}^{k} (x_{1i} - x_{2i})^2\}^{\frac{1}{2}},$$

$$manhattan = \sum_{i=1}^{k} |x_{1i} - x_{2i}|,$$

$$minkowski = \{\sum_{i=1}^{k} |x_{1i} - x_{2i}|^p\}^{\frac{1}{p}},$$

where k is the number of neighbours to be considered for calculating the distance and $x \in \mathrm{R}$, (i=1,2,3..,N) are inputs.

- **ANN:** It aims to imitate the network of neurons that makes up the human brain so that computers can understand and make judgements in a human-like manner. Computers are programmed to act like interlinked brain cells in the artificial neural network. A typical ANN architecture contains an input layer that takes in the input, a hidden layer which does all the calculation for feature extraction and prediction, and finally the output layer which gives the output of the calculations. The artificial neural network receives input and computes the weighted total of the inputs, as well as a bias. A transfer function is used to represent this calculation.

$$y = \sum_{i=0}^{n} w_i x_i + b$$

There are various activation functions used to set the threshold if a particular neuron should fire or not, such as Sigmoidal, ReLu, TanH, etc.

1.4.2. ML based pre-miRNA prediction tools:

Various tools have been developed using supervised ml algorithms for prediction of pre-miRNA. Supervised learning approach involves a true miRNA dataset and a pseudo miRNA dataset and the goal is to approximate a target function that maps these two datasets corresponding to their class based on sequential, thermodynamic, etc features. Thus, in supervised learning, the labels of the two classes must be all known beforehand. Let m be training samples as n-dimensional vectors $x_i = [x_{i1}, \ldots, x_{in}]^T$ such that $L = \{(x_i, y_i)\};\ i = 1, \ldots, m$, where $x_i \in R^n$ and $y_i \in \{-1, +1\}$ are the response variables.

SVM is the most popular algorithm for pre-miRNA prediction as it is based on binary class prediction. Although ANN based approaches are relatively lower than SVM, it is rapidly increasing in the recent years after advancements in Deep Learning. Random Forest which is

an ensemble based technique is also widely used for pre-miRNA prediction. (Stegmayer et al., 2019). Figure 1.3 contains the number of tools developed for each algorithm which has been adopted from previously reported review given in Table 1.1 (Stegmayer et al., 2019)

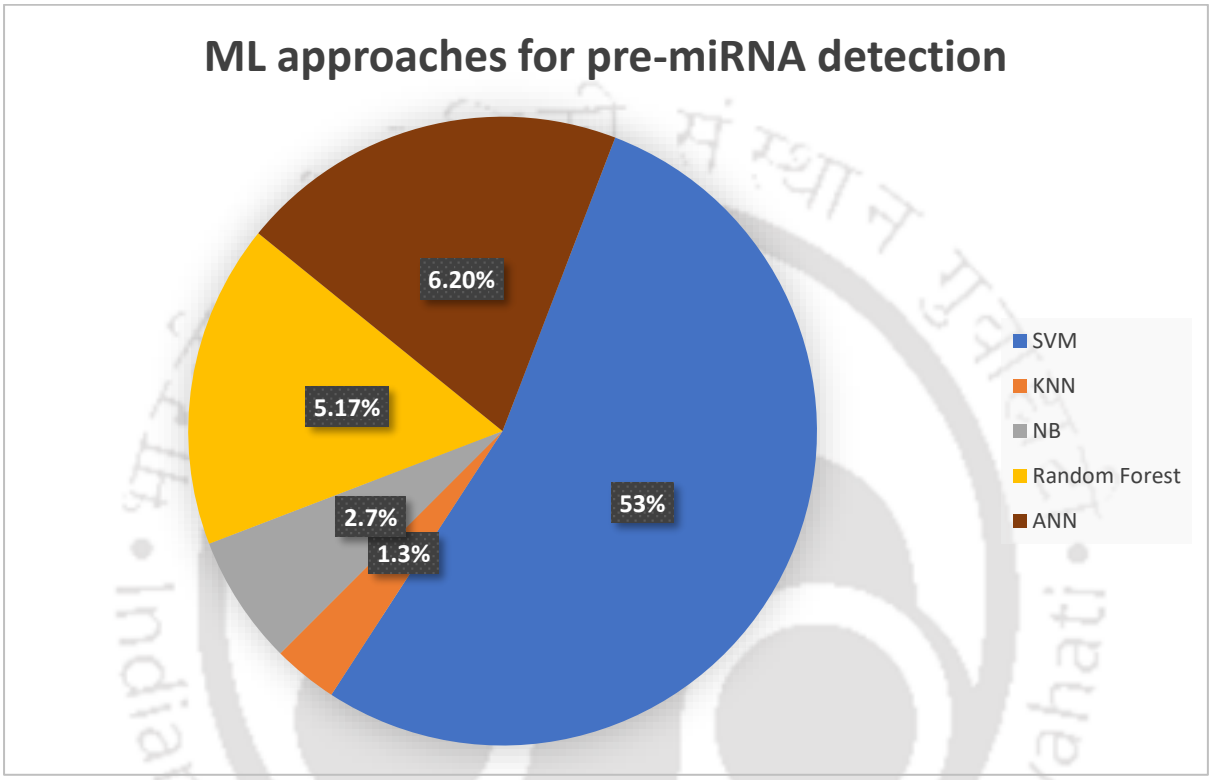

Figure 1.3: Pie chart for ML based tools developed using various algorithms for miRNA research. Most number of tools have been designed using SVM since it's one of the simple and oldest algorithms, followed by Neural Network and Random forest. Naïve bayes and k-NN had are the least implemented ML algorithms for pre-miRNA detection.

ML method	Name	Positive class	Negative class
SVM	Triplet-SVM	*Homo sapiens*	Random pseudo hairpins from *H. sapiens*
	MirAbela	*H. sapiens*, *Mus musculus*, *Rattus norvegicus*	tRNA, rRNA and mRNA from *H. sapiens*
	RNAmicro	Animals (nematodes, insects and vertebrates)	Random shuffling of animal miRNA features and tRNAs
	Micro-processorSVM	*H. sapiens*	ncRNA from *H. sapiens*
	MiRFinder	Animals (human, mouse, pig, cattle, dog and sheep)	Random sequences from human and mouse
	MIRenSVM	*H. sapiens*, *Anopheles gambiae*	Pseudo hairpins
	mirCos	*H. sapiens*, *M. musculus*	Random sampling of training genomes
	microPred	*H. sapiens*	Pseudo hairpins and other ncRNAs from *H. sapiens*
	PlantMiRNAPred	All miRNA plants in miRBase	Pseudo hairpins from the protein coding sequences of *A. thaliana* and *G. max*
	miRPara	Animals, plants and virus in miRBase	Sequences with pri-miRNAs identical to the positive class
	SMIRP	Species-specific positive sets from miRBase	ncRNA
	iMiRNA-SSF	*H. sapiens*	Pseudo pre-miRNAs from *H. sapiens*
	ViralmiR	Virus	Random virus sequences, human pre-miRNAs and pseudo hairpins from *H. sapiens*
	YamiPred	*H. sapiens*, animals	Random pseudo hairpins from *H. sapiens* and ncRNAs
	iMcRNA-PseSSC	*H. sapiens*	Random pseudo hairpins from *H. sapiens*
	MiRNA-dis	*H. sapiens*	Random pseudo hairpins from *H. sapiens*
KNN	MinDist	*Drosophila melanogaster*, *A. gambiae*	Random sequences
NB	BayesMirnaFind	*H. sapiens*, *M. musculus*	Potential negative stem-loops
	MatureBayes	*H. sapiens*, *M. musculus*	Random sequences from *H. sapiens*, *M. musculus*
Random Forest	miPred	*H. sapiens*	Pseudo pre-miRNAs from *H. sapiens*
	HuntMi	*H. sapiens*, *A. thaliana*, animals, plants	*Homo sapiens*, *A. thaliana*, animals, plants
	pMIRNA	*H. sapiens*, *O. sativa* and *A. thaliana*	Pseudo hairpins and ncRNAs
	PlantMirP-Rice		

	miR-BAG	Animals (human, mouse, rat, dog, nematode and fruit fly)	Pseudo hairpins of tRNA, rRNA, sRNA, mRNA
ANN	DP-miRNA	*H. sapiens*, animals	Random pseudo hairpins from *H. sapiens* and ncRNAs
	miRe2e		
	hybrid CNN-LSTM		
	Pre-miRNA sequence prediction using convolutional neural network		
	PlantMirP2		
	deepSOM	*H. sapiens*, *A. thaliana*, animals, plants, *Caenorhabditis elegans*	

Table 1.1. List of previously reported tools trained on various machine learning algorithms, adopted from previous review (Stegmayer et al., 2019)

1.5. Objectives:

Although there are many available tools for the detection of pre-miRNA, however, there is no existing tool that is exclusively trained on insect pre-miRNA. Based on the literature review, this thesis has the following objectives:

Objective 1: Training binary machine learning classifiers using insect pre-miRNA features.

Objective 2: The development of a cloud-based web interface for the trained ML models to predict novel insect pre-miRNA and search targets in the model organism, Drosophila melanogaster.

Objective 3: Comparative analysis of sequential and thermodynamic features of insect pre-miRNA with other organisms.

CHAPTER 2

Training binary machine learning classifiers using insect pre-miRNA features.

CHAPTER 2: Training binary machine learning classifiers using insect pre-miRNA features.

Executive Summary:

The development of a machine-learning model for the prediction of precursor microRNA in insects is explained in this chapter. Different features of insect precursor microRNA such as sequence composition, hairpin loop thermodynamics, length, etc. were used to train machine learning algorithms. For this experiment, 93 features in all were considered for the machine-learning classification. To distinguish insect precursor microRNA characteristics from hairpin sequences that do not generate precursor microRNA, machine learning methods including Support Vector Machine, Random Forest, Logistic Regression, and k-Nearest Neighbors were used. GridSearchCV and RandomSearchCV were used in 10 Fold Cross Validation to optimise the parameters.

The machine learning models trained on Support Vector Machine and Random Forest were selected based upon their performance evaluation with an accuracy of 92.19% and 80.28% respectively which was better than the previously reported tools such as Mipred, HuntMI, miPred, microPred, and Triplet-SVM. The trained classifiers were also tested on precursor microRNA of related phyla.

2.1. Introduction:

The role of microRNA (miRNA) is crucial in insects which have been reported to participate in a wide range of biological activities (Belles et al., 2012; Kayvan Etebari & Asgari, 2013; Song et al., 2018; Kaleem Tariq et al., 2016a; Q. Zhang et al., 2019, 2021) . Changes in the miRNA profiles have been observed during metamorphosis where miR-100/let-7/miR-125 cluster has been found to participate in wing morphogenesis in both hemimetabolan and holometabolan species (E. Gomez-Orte & Belles, 2009; Ling et al., 2014; Q. Zhang et al., 2019). In reproduction, during ovarian development miR-309 plays a critical role in female *A. aegypti* mosquitoes and during spermatogenesis miR-7911c-5p is upregulated in *B. dorsalis* (K. Tariq et al., 2016; Yang Zhang et al., 2016). Several miRNAs have been found to play important role in the regulation of immune-related genes (Yin et al., 2018; X. Zhang et al., 2014) and also during insecticide resistance where the genes responsible are downregulated with the help of miRNAs, miR-2b-3p is found to be involved in the regulation of metabolic resistance(K. Etebari et al., 2018; Y. Zhang et al., 2018).

As discussed in Chapter 1.3.2., many tools have been designed to predict novel precursor microRNA (pre-miRNA) using ML (machine-learning) approaches by training data to classify pre-miRNA hairpins from pseudo-pre-miRNA hairpins. Tools for species-specific novel pre-miRNA detection like TarMirPred (Jaiswal et al., 2019) and phylum specific such as ViralMir(K.-Y. Huang et al., 2015) have also been developed. Most of the tools use the characteristics of the hairpin loop as features for the classification (Gkirtzou et al., 2010; Hertel & Stadler, 2006; T.-H. Huang et al., 2007; Jiang et al., 2007; M. E. Rahman et al., 2012; Y. Xu et al., 2008; Xue et al., 2005). Most tools consider 8,494 non-redundant human pseudo hairpins as the negative dataset (J. Chen et al., 2016; Fu et al., 2019; Jiang et al., 2007; Ng & Mishra, 2007; J.-H. Xu et al., 2008; Xue et al., 2005), however selection of negative dataset remains a challenge and careful consideration is required to make efficient binary supervised classification models (Allmer & Yousef, 2012; Gomes et al., 2013).

Genomic hairpin sequences which are not pre-miRNA viz., mRNA, tRNA and rRNA are also used as the negative set (Mendes et al., 2009). However, the inclusion of a such collection of pseudo-hairpins gives rise to the class-imbalance problem. This issue is addressed in tools like HuntMi, where thresholding classifier score function is combined with receiver operating characteristics (ROC) (Gudyś et al., 2013), microPred where the concept of undersampling majority class and oversampling minority class was used (Batuwita & Palade, 2009) and DeepSOM addresses this issue by creating self-organizing maps (Stegmayer et al., 2017).

Tools have also been developed to search for potential miRNA target sites in a genomic sequence such as miRanda, PicTar, mirmap (John et al., 2005; Krek et al., 2005; Vejnar & Zdobnov, 2012) etc. These tools search for potential target sites for a given sequence in a gene by calculating likelihood, allowing wobble base-pairing and reward and complementarity at 5' end. Recently tool for genome-wide pre-miRNA detection, MiRe2e was also developed using a deep learning model (Raad et al., 2022).

The pre-miRNA sequences of insects differ from humans, plants, and mice in length, MFE, GC%, etc. upon which most of the available tools are trained. As miRNA plays a major role in insects and yet a tool that is exclusively dedicated to its detection is not available, this chapter explains the development of an ML based pre-miRNA prediction model for insect pre-miRNA. This chapter thus addresses Objective 1 the of thesis: "training binary machine learning classifiers using insect pre-miRNA features".

2.2. Methods:

2.2.1. Data Collection and Preprocessing for binary ML classification:

- **Data Collection and Pre-processing:**

Data was collected from miRBase (Griffiths-Jones, Sam, et al. 2007) which is a regularly updated database containing precursor and mature miRNA sequence and labelled as true miRNA. A total of 3391 sequences were collected and labelled as positive set for the ML classification (available at: https://github.com/adhiraj141092/RNAinsecta/blob/master/dataset/insect_miRNA.fasta).

The negative dataset, i.e. pseudo miRNA sequences chosen for the classification was obtained from 8494 non-redundant human pseudo hairpin sequences which have been previously used in triple-SVM, MiPred and microPred methods. The negative dataset was not sufficient to predict the model as: a. It is pseudo miRNA which is artificially generated; b. It is based on the human genome.

Hence, along with artificially generated sequences, Protein Coding Genes of different species from the insect phyla was used. Hairpins of insect mRNAs with length below 250 bp since the longest pre-microRNA reported for insects is 222 bp long produced by Plutella xylostella (miRBase ID: MI0027331). Secondary structure and minimum free energy (MFE) was calculated using RNAfold of ViennaRNA package. We then filtered the sequences based on MFE from -5 to -180, since dvi-mir-315b in Drosophila virilis is found to have the highest MFE of -5.4 (miRBase Accession: MI0009499) and the same pre-miRNA from Plutella xylostella (miRBase Accession: MI0027331) was found to have lowest MFE of -174.9. GC content (%G+C) was calculated by in-house python script and we chose the sequences with GC between 10–85%, since the lowest GC content was found to be 12.28% in pxy-mir-8547 (miRBase Accession: MI0027419) and the highest in pxy-mir-8517a (miRBase Accession: MI0027332). After filtering, a total of 23,252 negative dataset sequences from insect PCGs and previous pseudo pre-miRNA was prepared. This dataset can be found in RNAinsecta GitHub repository: (https://github.com/adhiraj141092/RNAinsecta/blob/master/dataset/pseudo_insect_pre-mIR.csv).

2.2.2. Feature Calculation: A total of 93 thermodynamic and sequence features based on miPred and triplet-SVM (Loong, et al., 2007; Xue, et al., 2005) was used as features to train the models.

- **Triplet Element scores:**

We used TripletSVM's method for calculating the triplet element scores where, given any three adjacent nucleotides, there are eight (2^3) possible structure compositions: '(((', '((.', '(..', '…', '.((', '..(', '.(.' and '(.(', taking '(' for both instances of paired nucleotide. Considering the middle nucleotide, there are 32 (4 × 8) possible structure-sequence combinations, which are denoted as 'C(((', 'G((.', etc. We used a perl script for the triplet calculation (Xue et al., 2005).

- **Base Composition:**

The nucleotide and its percentage:

$\%X = \frac{|X|}{(L)} * 100$, where $X \in \{A, C, G, U\}$ and $\mathrm{L} = \mathrm{Length}$

dinucleotide counts and their percentage:

$\%XY = \frac{|XY|}{(L-1)} * 100$, where $X, Y \in \{A, C, G, U\}$ and $\mathrm{L} = \mathrm{Length}$

base pair composition:

$\%(X + Y) = \frac{|X|+|Y|}{L} * 100$, where $\mathrm{L} = \mathrm{Length}$ and $X, Y = \begin{cases} X = C \ and \ Y = G \\ or \\ X = A \ and \ Y = U \end{cases}$

- **Structural and thermodynamic features:**

Number of stems, loops, loop length and number of basepairs were calculated from the secondary structure using regular expression and were used as features. A motif containing more than three contiguous base pairs in the secondary structure is termed as stem. The features dG, dP, dD, dQ, normalized Shannon entropy, MFE1 and MFE2 were adapted from miPred perl script (Ng & Mishra, 2007). dG is calculated by taking the ratio of MFE to the Length i.e. $dG = \frac{MFE}{L}$. Normalized base-pairing propensity, $dP = \frac{tot_{bp}}{L}$, where tot_{bp} is the total

basepairs and *L* is the Length. MFE1 is the ratio between dG and GC content, i.e. $MFE1 = \frac{dG}{(\%G+C)}$ and MFE2 is the ratio between dG and number of stems, i.e. $MFE2 = \frac{dG}{n\_stems}$, where *n_stems* is a structural motif containing more than three contiguous base pairs. All these features were calculated using in-house python script.

MFE3 and MFE4 features were implemented from microPred (Batuwita & Palade, 2009). MFE3 is the ratio between dG and number of loops, i.e. $MFE3 = \frac{dG}{n\_loops}$, where *n _loops* is the number of loops in the secondary structure. MFE4 is the ratio between dG and the total number of bases i.e. $MFE3 = \frac{dG}{tot\_bases}$ where *tot_bases* is the total number of base pairs in the secondary structure. dD is the adjusted basepair distance and zD is normalized dD. Normalized Shannon entropy is given by $dQ = -\sum_{i=1}^{j} \frac{(p_{ij}).\log_2 p_{ij}}{L}$, where the probability that base *i* pair with base *j* is then given by p_{ij} and L is the Length (Freyhult et al., 2005). Average basepair was calculated by taking the ratio of total bases and *n_stems*, i.e. $avg\_bp = \frac{tot\_bases}{n\_stems}$.

Table 2.1. contains all the features that were used

Sl. No.	Feature	Description
1	A...	For the predicted secondary structure, an opening parenthesis "(" or a closing parenthesis ")" and a dot "." were used to denote paired and unpaired nucleotides. Generally, the opening parenthesis "("represents a paired nucleotide located near the 5'-end that can be paired with another nucleotide at the 3'-end, which is denoted by the closing parenthesis ")". This study used "(" for both situations. For any three adjacent nucleotides, there are eight (2^3) possible triple-structure compositions: "(((", "((.", "(.(", ".((", "(..", ".(.", "..(", and "...". Considering the middle nucleotide, there are 32 (4 × 8) possible structure-sequence combinations, which are denoted as "C(((", "A.((", etc. (Xue et al., 2005)
2	A..(	
3	A.(.	
4	A.((	
5	A(..	
6	A(.(	
7	A((.	
8	A(((	
9	G...	
10	G..(	
11	G.(.	
12	G.((	
13	G(..	
14	G(.(	
15	G((.	
16	G(((	
17	C...	
18	C..(	
19	C.(.	

20	C.((	
21	C(..	
22	C(.(	
23	C((.	
24	C(((	
25	U...	
26	U..(	
27	U.(.	
28	U.((	
29	U(..	
30	U(.(	
31	U((.	
32	U(((	
33	Len	Length of the sequence
34	A	Single nucleotide count
35	C	
36	G	
37	U	
38	G+C	basepair counts
39	A+U	
40	AA	Dinucleotide counts
41	AC	
42	AG	
43	AU	
44	CA	
45	CC	
46	CG	
47	CU	
48	GA	
49	GC	
50	GG	
51	GU	
52	UA	
53	UC	
54	UG	
55	UU	
56	%A	Nucleotide percentage counts
57	%C	
58	%G	
59	%U	
60	%G+C	Basepair percentage counts
61	%A+U	
62	%AA	Dinucleotide percentage counts

63	%AC	
64	%AG	
65	%AU	
66	%CA	
67	%CC	
68	%CG	
69	%CU	
70	%GA	
71	%GC	
72	%GG	
73	%GU	
74	%UA	
75	%UC	
76	%UG	
77	%UU	
78	pb	Base propensity (total basepair/length)
79	zP	Normalised pb
80	mfe	Minimum free energy of folding
81	dG	Normalised mfe
82	dQ	Shannon Entropy
83	zQ	Normalised Shannon Entropy
84	dD	Basepair distance The base pair distance between S_α and S_β is equal to $d_{BP}(S_\alpha, S_\beta) = \sum_{i<j}(\delta_{ij}^{\alpha} + \delta_{ij}^{\beta} - 2\delta_{ij}^{\alpha}\delta_{ij}^{\beta})$, where δ_{ij}^{α} is 1 if bases *i* and *j* is a base pair in S_α, and 0 otherwise (Freyhult et al., 2005).
85	zD	Normalised basepair distance
86	avg_bp	Average base pair
87	nstem	Number of stems,
88	MFE1	$\frac{dG}{(\%G + C)}$
89	MFE2	$\frac{dG}{n\_stems}$
90	MFE3	$\frac{dG}{n\_loops}$
91	tot_bases	Total bases
92	MFE4	$\frac{dG}{tot\_bases}$
93	nstems	3 or more of '('

Table 2.1. List of all the features calculated for training the binary machine learning classifiers.

2.2.3. Imbalance class handling

As majority class (negative dataset) has 23,252 records and the minority class has 3,391 records. This leads to class imbalance problem. Imbalance class ML training leads to overfitting as the model learns heavily from one class. The issue is addressed by 2 methods:

- **SMOTE:** Synthetic Minority Oversampling Technique (SMOTE) is a widely used process to upsample minority class. SMOTE module available in the "imbalance" package of python was used to create a balanced class dataset.

- **Near-Miss:** Near Miss (NM) refers to a collection of undersampling methods that select examples based on the distance of majority class examples to minority class examples.

2.2.4. Hyperparameter tuning:

Different parameters of the ML algorithms were applied to the SMOTE , NM and unbalanced datasets to classify the positive and negative miRNA. For SVM, we used linear kernel: $K(x_i, x_j) = x_i^T x_j$, polynomial kernel: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d$ and radial basis function (RBF) kernel: $K(x_i, x_j) = \exp\{-\gamma||x_i - x_j||^2\}$, where $x \in \mathrm{R}$, (i=1,2,3..,N) are inputs and $\gamma, r, d > 0$ are kernel parameters. Different values of the Cost function (C_{SVM} value) and gamma were adjusted in the SVM algorithm optimization.

In the case of RF the model works on: $RFfi_i = \frac{\sum_{j \in N} normfi_{ij}}{T}$, where $RFfi_i$ is the importance of feature *i* (fi_i) calculated from N which denotes all trees in the Random Forest model, $normfi_{ij}$ is the normalized feature importance for i in tree j, i.e. $normfi_i = \frac{fi_i}{\sum_{j \in F} fi_j}$ and T is total number of trees and F denotes all features. We henceforth, chose different values for the number of trees, learning rate, maximum depth, minimum number of sample split and sample leaf were used.

For kNN, number of neighbours and different distance matrices were used such as $euclidean = \{\sum_{i=1}^{k} (x_{1i} - x_{2i})^2\}^{\frac{1}{2}}$, $manhattan = \sum_{i=1}^{k} |x_{1i} - x_{2i}|$, $minkowski =$

$\{\sum_{i=1}^{k} |x_{1i} - x_{2i}|^p\}^{\frac{1}{p}}$, where k is the number of neighbours to be considered for calculating the distance and $x \in \mathrm{R}$, (i=1,2,3..,N) are inputs.

The logistic regression algorithm works on: $LR = \frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$ where $\beta_0 + \beta_1 x$ is the equation of straight line with $\beta_1 x$ as slope and β_0 as y-intercept which is converted to natural log. Regularization strength (C_{LR} value) similar to cost function of SVM provides a penalty score and different solvers were used to optimize the LR algorithm.

We used python's scikit-learn package to choose the hyperparameters for training each algorithm (Pedregosa et al., 2011). Initially, we chose a wide range of hyperparameters for each of the above-mentioned parameters and classified using a model selection package called RandomizedsearchCV which randomly chooses different parameters to train the ML algorithm with 10-fold cross validation. We then fine-tuned the parameters using GridsearchCV, where the training was performed using each of the possible combinations of the provided parameters along with 10-fold Cross Validation (CV).

Python codes used for parameter tuning using GridSearchCV:

```
#Defining parameters:

param_svm = {
            'C': [1,5,6,7,10,20,40,60],
            'kernel': ['rbf','linear','polynomial],
            'gamma' : ['auto',0.1, 0.01, 0.001, 1, 10]
}

param_rf = {'n_estimators': [10,30,100,250,800,1200,1400,2000,2500,5000],
               'max_depth': [7,20,30,40,50,60,150,300,500],
               'min_samples_split': [0.5,1,2,5,8,10],
               'min_samples_leaf': [0.5,1,2,3,4,8,10],
               }
param_knn = {
            'n_neighbors': [1,2,3,5,7,10,12,20,50,70,100],
            'metric': ['euclidean','manhattan']
        }
param_log = {
            'C': [3,5,40,50,70,80,100],
            'solver' : ['liblinear','sag','saga','newton-cg']
        }

#Assigning variable name to classifiers

svc_rbf = SVC(shrinking=False)
```

```
rf = RandomForestClassifier()

log = LogisticRegression(multi_class='auto')

knn = KNeighborsClassifier()


#Binding parameters to classifiers

search_SVM = GridSearchCV(svc_rbf,
                      param_svm,
                      cv = 10,
                      return_train_score=True,
                      n_jobs=len(c))

search_RF = GridSearchCV(rf,
                      param_rf,
                      return_train_score=True,
                      cv = 10,
                      n_jobs=len(c))
search_LR = GridSearchCV(log,
                      param_log,
                      cv = 10,
                      return_train_score=True,
                      n_jobs=len(c))
searchKNN = GridSearchCV(knn,
                      param_knn,
                      return_train_score=True,
                      cv = 10,
                      n_jobs=len(c))

#The training process using parallel computing:
#X is the input features and y contains the corresponding labels.

with parallel_backend('ipyparallel'):
    search_SVM.fit(X, y)
    search_RF.fit(X, y)
    search_LR.fit(X, y)
    search_KNN.fit(X, y)

#Getting the scores for best performing parameters:

scores.append({
        'model': "SVM",
        'best_score': search_SVM.best_score_,
        'best_params': search_SVM.best_params_
    })

scores.append({
        'model': "RF",
        'best_score': search_RF.best_score_,
        'best_params': search_RF.best_params_
```

```
    })

scores.append({
        'model': "log",
        'best_score': search_LR.best_score_,
        'best_params': search3.best_params_
    })

scores.append({
        'model': "knn",
        'best_score': searchKNN.best_score_,
        'best_params': search4.best_params_
    })
```

10-Fold CV essentially splits the dataset into 10 parts and trains 9 parts with a given parameter and use 1 to test the data. This repeats for all 10 parts and the mean accuracy score is provided as CV score. For example, in case of RF, we used No. estimators: 10 to 5000, Max depth: 5 to 100, Bootstrap: True and False, Min sample leaf: 1 to 10, Min sample split: 1 to 10. For SVM we used, Cost Function (C_{svm}): 0.01 to 100, Kernel: Linear, Polynomial and RBF, Gamma: 0 to 10. For KNN, No. of neighbours between 1 and 50, distance metrics: Euclidean, manhattan and minkowski were used. In case of LR, we used cost function (C_{LR}): 10 to 100. RandomSearchCV initially selects random parameter values from the range and performs a 10-fold training resulting in mean accuracy score CV for a given algorithm. Then using GridSearchCV, exact parameter value was provided from the range of obtained values from RandomSearchCV. In GridSearchCV, one optimum parameter value once reached is fixed and the remaining parameter is optimized one at a time, resulting in the fine-tuning of the parameters for the ML classification model.

2.2.5. Performance Evaluation:

- **Test Set:**

Initially X_test was used to evaluate performance for all the classifiers. Further, to generalize the model, an independent test dataset was created from the pre-miRNA sequences of the insects *Spodoptera frugiperda* (Kakumani et al., 2015) and *Tribolium castaneum* (Marco et al., 2010; Singh & Nagaraju, 2008) which were not used in the initial data collection and hence remained entirely unseen to the project, were considered as positive dataset. Insect CDS of 250bp length fetched from GenBank using the same steps as mentioned above in "Data Collection" were considered as negative dataset. A total of 999 sequences were considered as the validation dataset (V_test) of which 464 were positive and 535 were negative as given

in the GitHub link:
(https://github.com/adhiraj141092/RNAinsecta/tree/master/dataset/pos.fold and https://github.com/adhiraj141092/RNAinsecta/tree/master/dataset/neg.fold respectively).

- **Testing Parameters:**

Performance was calculated based on the following classical classification measures: sensitivity (*SN*): $SN = \frac{TP}{TP+FN}$, specificity (*SP*): $SP = \frac{TN}{TN+FP}$, Accuracy (*Acc*): $Acc = \frac{TN+TP}{TN+FP+TP+FN}$, precision (*p*): $p = \frac{TP}{TP+FP}$, harmonic mean of sensitivity and precision (F_1): $F_1 = 2\frac{SN \,.\, p}{SN + p}$ and Matthew's correlation coefficient (*MCC*): $MCC = \frac{(TP \,.\, TN) + (FP \,.\, FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$, where *TP*, *TN*, *FP* and *FN* are the number of true-positive, true-negative, false-positive and false-negative classifications, respectively. For given false positive rate (α) and true positive rate ($1 - \beta$) at different threshold values, the AUC-ROC was computed as: $AUC = \sum_{n=1}^{i} \left\{(1 - \beta_i \Delta\alpha) + \frac{1}{2}[\Delta(1 - \beta)\Delta\alpha]\right\}$, where $\Delta(1-\beta)=(1-\beta_i)-(1-\beta_{i-1})$ and $\Delta\alpha=\alpha_i-\alpha_{i-1}$ and i = 1, 2, …, m (number of test data points) (Jaiswal et al., 2019). In imbalance class testing data, accuracy, sensitivity, specificity, precision and F_1 are not the best measures to analyse performance of models as they are not based on the entire confusion matrix. MCC is a better estimator of performance in such cases as it produces a high score only if good results are obtained in all of the four confusion matrix categories. (Chicco & Jurman, 2020).

test.py (https://github.com/adhiraj141092/RNAinsecta/blob/master/test.py) can be used to replicate the results.

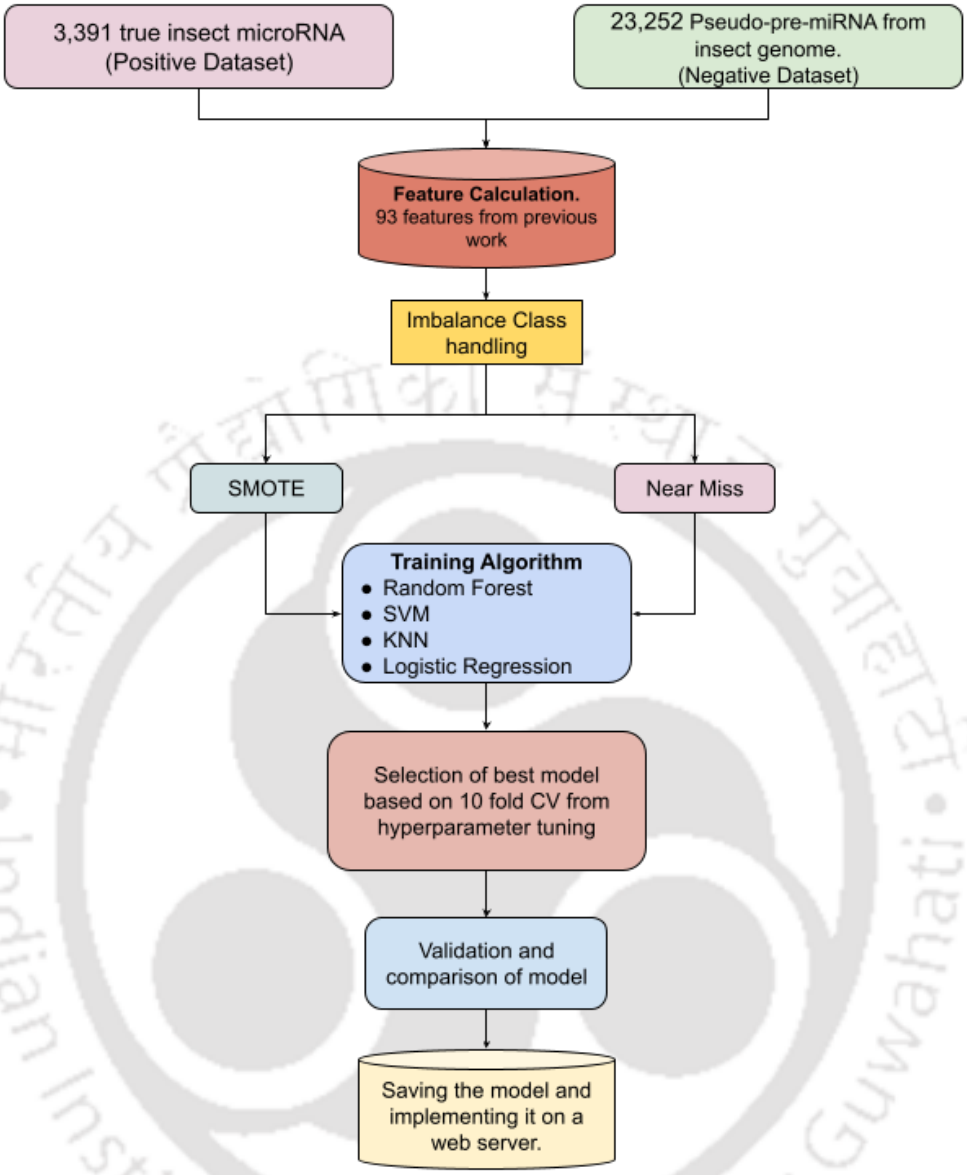

Figure 2.1. Workflow for training ML classifier to predict insect pre-miRNA.

2.3. Results:

2.3.1. Data Preprocessing for binary ML classification:

Dataset:

A total of 3391 sequences were collected and labelled as positive set for the ML classification as shown in Table 2.2 (available at: https://github.com/adhiraj141092/RNAinsecta/blob/master/dataset/insect_miRNA.fasta). A total of 23,252 negative dataset sequences from insect PCGs and previous pseudo pre-miRNA was prepared. This dataset can be found in RNAinsecta GitHub repository (https://github.com/adhiraj141092/RNAinsecta/blob/master/dataset/pseudo_insect_pre-mIR.csv).

Organism	No. of precursors
Bombyx mori	487
Drosophila melanogaster	258
Apis mellifera	254
Drosophila pseudoobscura	210
Drosophila virilis	180
Aedes aegypti	155
Drosophila simulans	148
Plutella xylostella	133
Anopheles gambiae	130
Acyrthosiphon pisum	123
Drosophila sechellia	103
Dinoponera quadriceps	102
Drosophila erecta	101
Manduca sexta	98
Heliconius Melpomene	92
Drosophila yakuba	89
Drosophila grimshawi	82
Bactrocera dorsalis	80
Drosophila willistoni	77
Drosophila ananassae	76
Drosophila persimilis	75
Culex quinquefasciatus	74
Polistes canadensis	73
Drosophila mojavensis	71
Nasonia vitripennis	53
Nasonia giraulti	32
Nasonia longicornis	28

Locusta migratoria	7

Table 2.2 3391 pre-miRNA sequences downloaded for each species from miRBase.

The calculated features are given in Table 2.1. There are a total of 93 features which was calculated for the binary machine learning classification.

2.3.2 Hyperparameter Tuning:

The initial parameters were selected based on the best performing models for SMOTE, NM and imbalance dataset. The overall CV score of NM was found to be lower than the SMOTE dataset. The best parameters for all the classifiers are given in Table 2.3 which was used in final model preparation and evaluation.

Classifier	SMOTE Best CV Score	SMOTE Best Parameters	NM Best CV Score	NM Best Parameters	Imbalance Best CV Score	Imbalance Best Parameters
SVM	0.99459	C_{SVM} = 10, Kernel = RBF, Gamma= 0.01	0.952789	C_{SVM} = 1, Kernel = RBF, Gamma= 0.001	0.96871	C_{SVM} = 6, Kernel = RBF, Gamma= 0.001
RF	0.98609	No. estimators= 1600 Max depth = 30, Bootstrap = 'False', Min sample leaf = 1, Min sample split = 2,	0.944617	No. estimators= 1400 Max depth = 30, Bootstrap = 'False', Min sample leaf = 1, Min sample split = 2,	0.96281	No. estimators= 30 Max depth = 10, Bootstrap = 'False', Min sample leaf = 2, Min sample split = 2,
LogR	0.9627	C_{LR} = 70, Solver = Newton-cg	0.946952	C_{LR} = 100, Solver = Newton-cg	0.96703	C_{LR} = 30, Solver = Newton-cg
KNN	0.98199	Metric Distance: Manhattan, No. of neighbours = 2	0.92166	Metric Distance = Euclidean, No. of neighbours = 7	0.96032	Metric Distance = Euclidean, No. of neighbours = 7

Table 2.3: Sets of best performing parameters obtained after gridsearching through different values for each classification algorithm. Selection was based on the highest 10-fold cross-validation score. Parameters tuned for SVM were Cost function C_{SVM}, Kernel and Gamma value. Parameters considered for RF were No. estimators, Max depth, Min sample leaf and Min sample split. For LogR the parameters considered were Cost Function C_{LR} and solver. For KNN, No. of neighbours and Metric Distance was considered for tuning. The CV score shows the mean accuracy score of the best classifier obtained with the corresponding parameters.

2.3.3. Performance Evaluation:

The X_test consisted of 853 positive and 5818 negative entries. Each model obtained from SMOTE, NM and imbalance dataset was tested on the same dataset so that there is no sampling bias in their comparison. The selection of such large datasets for testing gives a better understanding of their performance. Table 2.4. contains the performance measures for all the selected classifiers.

Sl. No.	Classifier	Acc	SP	SN	MCC	P	F_1
1	SVM_SMOTE	0.97451	0.975685	0.967611	0.90525	0.871468	0.917026
2	RF_SMOTE	0.983498	0.992757	0.92915	0.934044	0.95625	0.942505
3	LogR_SMOTE	0.969501	0.971029	0.960526	0.8882	0.849597	0.901663
4	KNN_SMOTE	0.970679	0.983101	0.897773	0.885446	0.900508	0.899138
5	SVM_NM	0.421836	0.331954	0.949393	0.270949	0.194929	0.323448
6	RF_NM	0.436865	0.35075	0.942308	0.281079	0.198254	0.327586
7	LogR_NM	0.446589	0.361097	0.948381	0.284791	0.201853	0.33286
8	KNN_NM	0.431266	0.349026	0.913968	0.285773	0.193031	0.318743
9	SVM_imbalanced	0.983351	0.993848	0.917647	0.928914	0.959732	0.938218
10	RF_imbalance	0.978341	0.994874	0.874866	0.906552	0.964623	0.917555
11	LogR _imbalance	0.975394	0.991285	0.875936	0.894082	0.941379	0.907479
12	KNN_imbalance	0.97672	0.991114	0.886631	0.900102	0.940976	0.912996
13	SVM_human_CDS	0.438338	0.36056	0.925134	0.210874	0.187758	0.312162
14	RF_human_CDS	0.137763	0	1	0	0.137763	0.242165

15	LogR_human_CDS	0.433623	0.357143	0.912299	0.199071	0.184832	0.307387
16	KNN _human_CDS	0.264329	0.146787	1	0.152158	0.157726	0.272476

Table 2.4: Performance measure for each classifier of SMOTE from Sl. No. 1-4, NM from Sl. No. 5-8, Imbalance from Sl. No. 9-12 and 8494 human_CDS negative datasets from Sl. No. 13-16. Accuracy, Specificity (SP), Sensitivity (SN), Matthew's correlation coefficient (MCC), Precision (p), harmonic mean of sensitivity and precision (F_1) are given corresponding to each ML classifier.

2.3.4. Comparision with previous tools using V_test:

The SMOTE and imbalance models were further analysed for their performance in comparison to the already developed tools, viz., miPred, microPred, Triplet-SVM, HuntMi and MiPred on the same validation dataset, V_test, comprising of 464 positive and 536 negative sequences. Table 2.5 consists of the performance measures for each of the tools along with the imbalance set.

From Sl. Nos. 10 to 13, the performance for the Imbalance dataset is provided. SVM_imbalance, RF_imbalance, LogR_imbalance, and KNN_imbalance all have accuracy values of 0.5355, 0.7297, 0.5756 and 0.5265, respectively. For SVM_imbalance, RF_imbalance, LogR_imbalance, and KNN_imbalance, the Specificity and Sensitivity were 0 and 1, 0.8579 and 0.5819, 0.9103 and 0.1897, and 0.9832 and 0, respectively. MCC for SVM_imbalance was 0, RF_imbalance was 0.4611, LogR_imbalance was 0.1453, and KNN_imbalance was also 0.

Sl. No.	**Tool**	**Acc**	**SP**	**SN**	**MCC**	**P**	**F_1**
1	Triplet-SVM	0.754755	0.798131	0.704741	0.625745	0.751724	0.727475
2	MiPred	0.48048	0.128972	0.885776	0.325557	0.468643	0.612975
3	miPred	0.778779	0.785047	0.771552	0.654075	0.756871	0.764141
4	microPred	0.713714	0.570093	0.87931	0.574294	0.639498	0.740472
5	HuntMI	0.618619	0.306542	0.978448	0.414076	0.550303	0.704422
6	SVM_SMOTE	0.921922	0.945794	0.894397	0.854779	0.934685	0.914097
7	RF_SMOTE	0.802803	0.695327	0.926724	0.676982	0.725126	0.813623

8	LogR_SMOTE	0.677678	0.450467	0.939655	0.513201	0.59726	0.730318
9	KNN_SMOTE	0.761762	0.583178	0.967672	0.614122	0.668155	0.790493
10	SVM_imbalance	0.535536	1	0	0	0	0
11	RF_imbalance	0.72973	0.857944	0.581897	0.461038	0.780347	0.666667
12	LogR_imbalance	0.575576	0.91028	0.189655	0.145339	0.647059	0.293333
13	KNN_imbalance	0.526527	0.983178	0	0	0	0

Table 2.5: Comparative performance analysis between available tools and trained models tested upon independent insect pre-miRNA validation dataset. 1-5 shows the performance of previous tools. 6-9 shows the performance on the SMOTE classifiers, 10-13 shows the performance on Imbalance set. The parameters for evaluation are Accuracy, Specificity (SP), Sensitivity (SN), Matthew's correlation coefficient (MCC Precision (p), harmonic mean of sensitivity and precision (F_1) are given for each corresponding ML classifier.

The visualisation of Table 2.5. is given in Figure 2.2. The performance of each tool was plotted in a Radar chart corresponding each parameter for evaluation. The figure shows maximum area to be covered by SVM classifier followed by RF. Therefore, we calculated the "ROC-Area Under the Curve" for further analysis

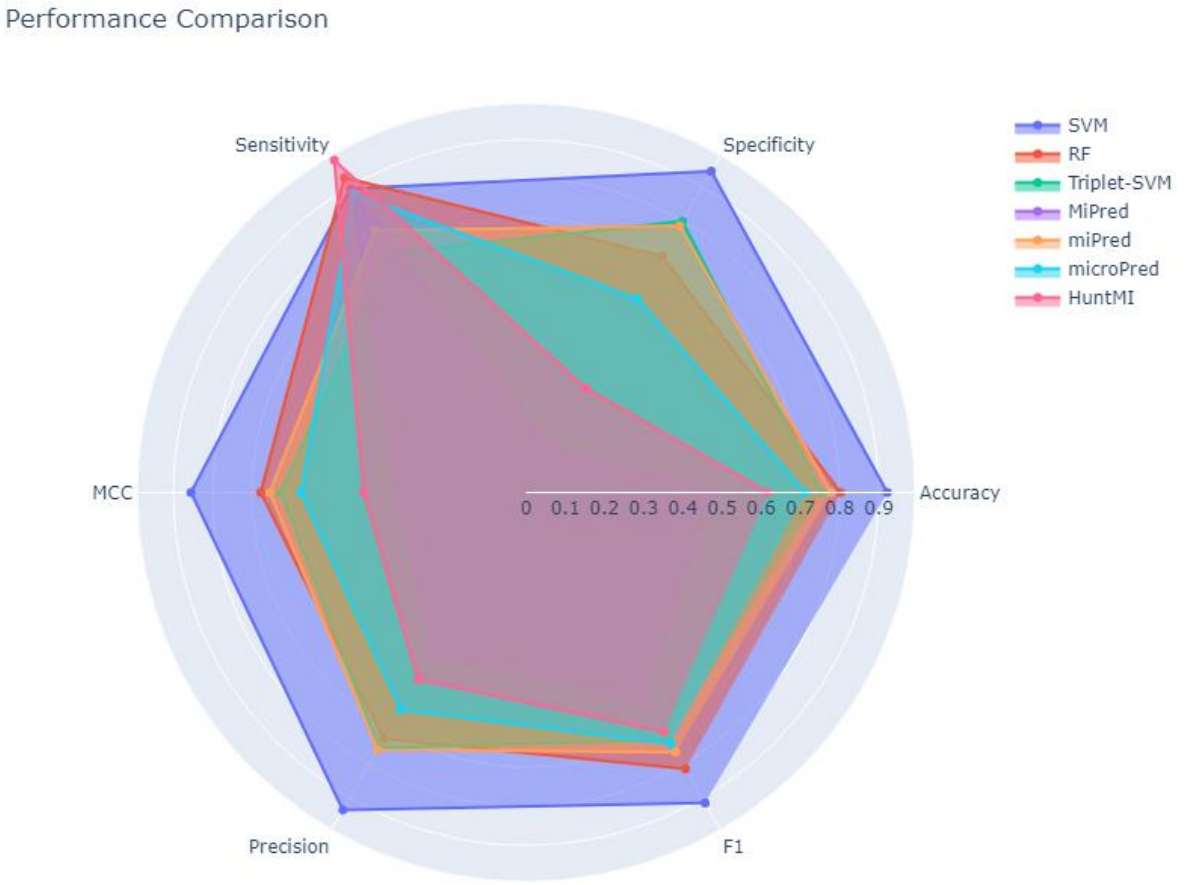

Figure 2.2: Radar plot for comparative performance analysis of RNAinsecta SVM and RF with already published tools.

ROC: The ROC curve (Receiver Operating Characteristic curve) is used to measure the performance of a model at different classification thresholds. It is a plot between True Positive Rate (Sensitivity) and False Positive Rate (1 - Specificity). Higher the AUC (Area Under the Curve) of ROC, the better is the model at classifying, i.e. higher degree of separability. The ROC-AUC of the models is given in Figure 2.3.

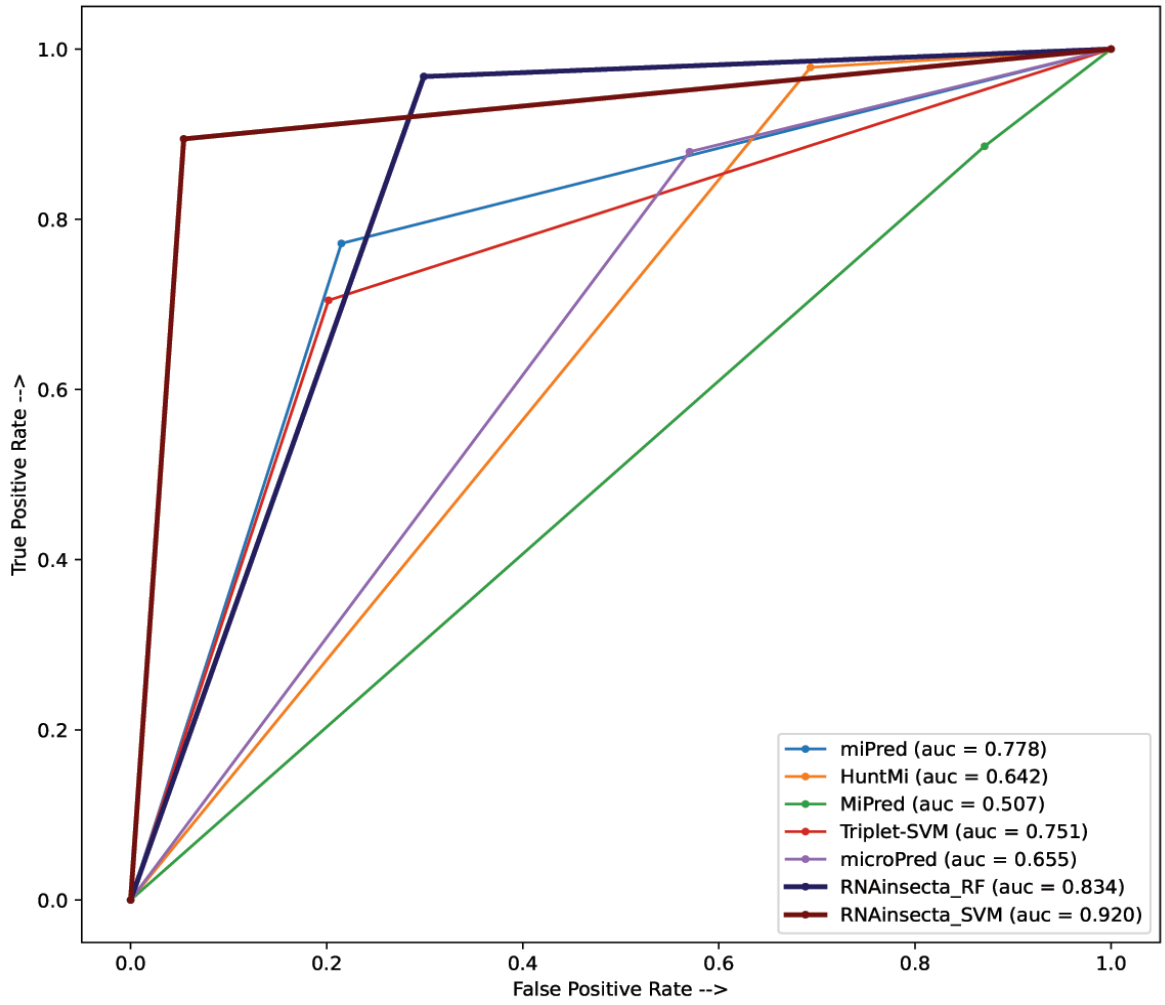

Figure 2.3: ROC-AUC for comparative performance analysis of RNAinsecta with available tools for detection of insects' pre-miRNA. The validation dataset used for this figure contains 464 positive and 536 negative sequences. Y axis contains the True Positive Rate (TPR) and X axis contains the False Positive Rate (FPR). More the AUC (Area Under the Curve) better is the performance.

2.3.5. Imbalance set performance: We further tested our RNAinsectaRF and RNAinsectaSVM with imbalance dataset (M_test) to measure their performance on large imbalanced data.The negative class of the dataset was constructed with ncRNA since mRNAs are significantly longer which would overfit the data with a large number of TN. The result is given in Table 2.6. Out of 116,230 negative sequences 94,136 and 111,147 were correctly classified by RNAinsectaRF and RNAinsectaSVM respectively.

The accuracy, sensitivity, and specificity of RNAinsectaRF were 81%, 96.77%, and 81%, respectively. The accuracy, sensitivity, and specificity of RNAinsectaSVM were 95.6%, 89.43%, and 95.62%, respectively. However, the MCC, precision and recall for RNAinsectaSVM were 0.2526, 0.075482 and 0.139215 respectively. The MCC, precision and recall for RNAinsectaRF

were 0.12395, 0.019917 and 0.039032 respectively. AUPRC (Area Under Precision Recall Curve) is considered optimal for evaluating binomial classification with imbalance class dataset (Ozenne et al., 2015). The AUPRC which is bound to be less as there is a huge difference in the testing set for both the classes. The AUPRC plot is given in Figure 2.4. The AUPRC of SVM and RF model were found to be 0.62 and 0.08 respectively. This is primarily due to poor F1 and precision value as the negative sample is huge. This suggest it is not suitable for RNA-Seq pipeline yet.

Classifier	TN	FP	FN	TP	acc	SP	SN	MCC	p	F1
RNAinsecta_RF	94136	22094	15	449	0.810539	0.809911	0.967672	0.12395	0.019917	0.039032
RNAinsecta_SVM	111147	5083	49	415	0.956022	0.956268	0.894397	0.252656	0.075482	0.139215

Table 2.6: Performance evaluation on imbalance class dataset (M_test) containing 116230 negative and 464 positive samples.

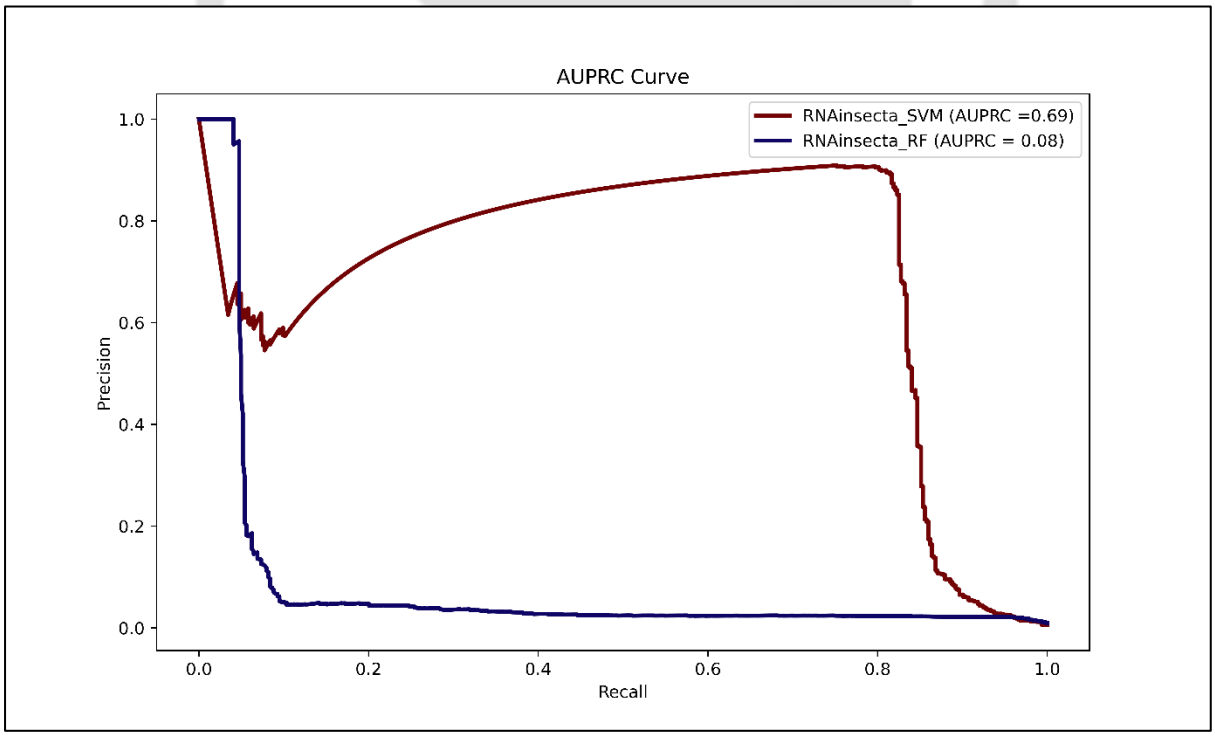

Figure 2.4: AUPRC of the ML models

2.3.6. Performance on related phyla:

The performance of RNAinsecta was measured across species from other phyla and compared with miPred which has so far been better than other tools in our analysis. Their comparative performance is given in Table 2.7. pre-miRNA of various species from Nematoda, Platyhelminthes, Virus and Mollusca were taken.

In Platyhelminthes for instance, out of 148 pre-miRNAs from *Schmidtea mediterranea* RNAinsecta_RF correctly predicted 126 with a sensitivity 0.8514 while miPred correctly predicted 114 with a sensitivity of 0.7703. In *Gyrodactylus salaris*, out of 60 pre-miRNAs RNAinsecta_RF correctly predicted 52 with a sensitivity of 0.8667 while miPred identified 43 with a sensitivity of 0.7166.

In Nematoda, RNAinsecta_RF correctly predicted 208 of the 214 pre-miRNA from *Caenorhabditis brenneri* with a sensitivity of 0.9719 while miPred correctly predicted 194 with sensitivity of 0.9065. In *Brugia malayi,* out of 157, RNAinsecta_RF correctly predicted 122 with sensitivity of 0.7770 while miPred correctly predicted 119 with sensitivity of 0.7579.

In Virus such as *Duck enteritis*, out of 24 sequences RNAinsecta_RF correctly predicted 19 while miPred predicted 13. In Mollusca such as *Melibe leonina* out of 90 RNAinsecta_RF correctly predicted 83 sequences while miPred identified 85 sequences correctly.

Phylum	Species	Total	RNAinsecta_RF		miPred	
			TP	SN	TP	SN
Nematoda	*Brugia malayi*	157	122	0.7770	119	0.7579
	Caenorhabditis brenneri	214	208	0.9719	194	0.9065
	Caenorhabditis elegans	253	207	0.8181	209	0.8260
	Ascaris suum	97	82	0.8453	86	0.8865
	Pristionchus pacificus	353	302	0.8555	307	0.8696

Platyhelminthes	*Fasciola hepatica*	38	27	0.7105	25	0.6578
	Gyrodactylus salaris	60	52	0.8667	43	0.7166
	Schistosoma mansoni	115	86	0.7478	53	0.4609
	Echinococcus granulosus	111	72	0.6486	81	0.7298
	Schmidtea mediterranea	148	126	0.8514	114	0.7703
Virus	*Duck enteritis*	24	19	0.7917	13	0.5417
	Epstein barr	25	21	0.84	23	0.92
	Human cytomegalovirus	15	10	0.6667	9	0.6
	Mouse cytomegalovirus	18	12	0.6667	12	0.6667
Mollusca	*Lottia gigantea*	59	46	0.7797	55	0.9322
	Melibe leonina	90	83	0.9222	85	0.9444

Table 2.7: Performance of RNAinsecta_RF in comparison with miPred for prediction of pre-miRNA across related phyla. pre-miRNA of different species from Nematoda, Platyhelminthes, Virus and Mollusca and their performance based on TP and SN is shown.

2.4. Discussion:

2.4.1. Dataset:

A total of 28 organisms were considered for the study out of which *B. mori* had 427 pre-miRNA which was the highest among all. A large number of pre-miRNAs belonged to the *Drosophila* genus. The consideration of large negative datasets made it a typical imbalanced dataset classification problem since the positive to negative class ratio was approximately 1:6. The problem with such classification is that the majority of data considered for the classification belongs to a single class and hence the results are misleading.

2.4.2. Performance Evaluation:

CV score helps in the initial choice of the hyperparameters, however, to regularize the classifiers which performed well, they were tested on X_test that was initially kept separate. The NM models did not perform well on unseen data as was expected from their CV scores. The accuracy of the NM classifiers dropped quite below their CV Scores. As the training data of NM had lesser negative class, the models could not learn to classify non-miRNA which closely resemble true miRNAs and hence suffered from Type I error. The SN of these models were quite high suggesting they learned fairly well to classify positive miRNAs but produced a lot of FP as their SP was low. Hence, these models had poor precision and accuracy for which they were discarded from further analysis. Also, the 8494 human_CDS performed extremely poor with X_test and hence, it was discarded from 10-Fold CV parameter optimization.

The SMOTE models performed well on the test data as the models learned to correctly classify non-miRNAs which included insect CDS hairpins that closely resembled true miRNAs. The accuracy of Logistic Regression was the lowest among all the SMOTE models but still had higher MCC and F_1 than the KNN model. The MCC of SVM and RF were the highest among all the models.

The performance of SMOTE was better than NM suggesting that with increase in the amount of training data, the performance of these classifiers improves. Hence, for the validation test the SMOTE models were selected for their performance analysis and comparison with already existing tools.

Although the imbalance dataset models performed extremely well with X_test, their performance drastically dropped with V_test. The models failed to produce any proper prediction, showing it to be a typical case of overfitting which was expected due to class imbalance. The models over learned from the majority class and classified every sequence as negative pre-miRNA. Hence, it indicated that our initial assumption to balance the dataset was necessary.

2.4.3. Comparison with previous tools:

Tools such as microPred, miPred, Triplet-SVM and MiPred are trained on the 8494 human pseudo pre-miRNA sequence as the negative dataset. HuntMi on the other hand uses many classes of CDS such as plant, virus, human, arabidopsis and along with 8494 human pseudo pre-miRNA as their negative dataset and have different classifiers for them. However, it does not contain any insect specific classifier. In our study we used the negative dataset that closely resembled with true insect pre-miRNA. Hence, most of the tools classified them as true miRNA making the Type I error. HuntMi and MiPred had the least Specificity with 0.31 and 0.13 respectively. HuntMI had an F_1 score of 70.44% but precision was 55.03%. microPred although had 71.37% accuracy, the specificity was 57%. Triplet-SVM and miPred performed well with MCC of 62.57% and 65.4% respectively, which was the highest among the previously developed tools considered for this experiment.

Triplet-SVM is trained solely on SVM classifier, with the 32 triplicate features. There is no mention of CV optimization of their hyperparameters. (Xue et al., 2005) MiPred is exclusively trained on Random Forest with the 32 triplet-SVM features along with dinucleotide shuffling and p-value of randomization.(Jiang et al., 2007) miPred uses SVM RBF kernel with nucleotide thermodynamics features.(Ng & Mishra, 2007) microPred uses 29 features from miPred and along with 12 modified features and is trained only on SVM classifier.(Batuwita & Palade, 2009) HuntMi uses 21 feaatures from microPred and uses 7 additional features such as loop length, orf, etc. (Gudyś et al., 2013) These tools are based on command-line interface without UI/UX support. The provision for target prediction is not available is these tools. In our approach, we trained 4 datasets, on 93 features with 4 different ML algorithms and have also provided provisions for further analysis of the miRNA targets.

The SMOTE trained models of RF and SVM in our experiment had fairly good sensitivity but the Logistic Regression and KNN model suffered from the same Type I error. The RF model had accuracy and precision of 80.28% and 81.36% which was higher than all the previous tools tested on V_test. However, the best performance was given by the SVM model with specificity of 94.58% which was the highest among all models used in the experiment indicating it had the least Type I error. The accuracy, precision and F_1 score of the SVM model was also highest with 92.19%, 93.47% and 91.41% respectively. However, to achieve such low FP the model was allowed to make few Type II errors for which sensitivity of the model was lower than RF but yet was more than Triplet-SVM and miPred. The MCC score of SVM was 85.48% which was found to be the highest. As RF and SVM models performed better than all the models, both were considered for implementation in a web server called RNAinsecta and the choice of model to select will depend on the user's requirement of specificity in their experiment.

The outcome of M_test suggest that the model has good specificity, sensitivity and accuracy but since the ratio of positive to negative class was huge, hence the MCC recall and precision dropped significantly. However, the usability of the tool is not RNA-Seq data analysis but rather PCR products or small-scale synthesis of pre-miRNA in which respect the works fairly well. We believe the RNAinsectaSVM model is better suited for the prediction. While RNAinsectaSVM is available on the web, we have not removed RNAinsectaRF based on the precision and recall of M_test since it is still a better estimator than the other available tool which share the same PCR based methodology for detection as discussed

2.4.4. ROC:

Triplet-SVM and miPred have lesser FPR than RNAinsecta_RF model but more than RNAinsecta_SVM. Tools like microPred and HuntMi although have high TPR also have high FPR for which their AUC is less. RNAinsecta SVM and RF had the highest AUC with 0.92 and 0.83 respectively followed by miPred and Triplet-SVM with 0.78 and 0.75 respectively.

2.4.5. Performance on other Phyla:

RNAinsecta_RF performed well on Nematoda with highest prediction specificity. The performance on Platyhelminthes was better as compared to miPred. The performance on Virus was almost same as miPred whereas in case of Mollusc, miPred performed better.

2.5. Conclusion

In this chapter we demonstrated the development of a machine learning based predictive model that was trained upon insect pre-miRNA. We used 93 features sequence and thermodynamic characteristics of pre-miRNA. These features were trained on various ML algorithms such as SVM, Random Forest, Logistic Regression and KNN for binary classification of true and pseudo pre-miRNA. SMOTE and Near-Miss were used to handle the imbalance in the class, along with 10-fold cross-validation. Two models viz., SVM and RF were selected upon their performance evaluation with accuracy of 92.19% and 80.28% respectively, tested on independent validation dataset along with other previous tools.

2.6. Annexures

Annexure 2.1. F-Score of the features between true and pseudo insect pre-miRNA datasets.

FEATURE	F SCORE	P-VALUE
A...	653.9571378	1.50E-142
A..(	475.1035278	1.96E-104
A.(.	8.265539286	0.004043606
A.((	344.2484575	2.25E-76
A(..	481.1705809	9.86E-106
A(.(	50.83676427	1.03E-12
A((.	10.84192727	0.000993542
A(((	2342.157864	0
G...	366.1019035	4.52E-81
G..(	388.0696827	8.67E-86
G.(.	55.29922195	1.07E-13
G.((	738.835983	1.52E-160
G(..	289.3480646	1.49E-64
G(.(	2.200490664	0.137978425
G((.	213.2725659	4.04E-48
G(((	19.5908167	9.63E-06
C...	660.7704762	5.38E-144
C..(	65.09593727	7.43E-16
C.(.	59.39020982	1.34E-14
C.((	200.6641009	2.17E-45
C(..	179.8239881	7.15E-41
C(.(	436.3610977	3.83E-96
C((.	794.5526722	2.51E-172
C(((	4.8203234	0.0281345
U...	261.8002174	1.31E-58
U..(	14.27417843	0.000158355
U.(.	12.38372726	0.0004338
U.((	0.017679988	0.894221162
U(..	38.34739128	6.01E-10
U(.(	601.3862739	2.25E-131
U((.	182.7937852	1.62E-41
U(((	2352.841866	0
LEN	5209.808399	0
A	3015.687698	0
C	6493.233838	0
G	5817.344837	0
U	1578.61082	0
G+C	7012.206004	0
A+U	2680.701972	0
AA	1370.750259	9.13E-293

AC	2581.792227	0
AG	4418.395234	0
AU	1689.460568	0
CA	3959.247479	0
CC	4471.273683	0
CG	2107.28148	0
CU	4097.016618	0
GA	3912.739365	0
GC	4304.143049	0
GG	4809.252333	0
GU	1028.640962	1.46E-221
UA	579.2715806	1.15E-126
UC	3361.542805	0
UG	2763.723186	0
UU	0.05361183	0.816895284
%A	14.17730868	0.000166715
%C	1833.873645	0
%G	1026.022195	5.17E-221
%U	4883.55807	0
%G+C	2276.787186	0
%A+U	2276.787186	0
%AA	1.156919887	0.282114805
%AC	18.03621415	2.17E-05
%AG	357.3016573	3.52E-79
%AU	284.5372643	1.62E-63
%CA	243.957714	9.34E-55
%CC	2260.626986	0
%CG	49.33894959	2.20E-12
%CU	59.89530782	1.04E-14
%GA	441.5570639	2.95E-97
%GC	400.129732	2.24E-88
%GG	2422.584515	0
%GU	2649.694379	0
%UA	1110.24128	1.26E-238
%UC	75.58491378	3.69E-18
%UG	533.9564054	5.25E-117
%UU	3984.526652	0
PB	3027.63623	0
NPB	8749.405663	0
MFE	1823.575316	0
DG	3418.150407	0
Q	6189.715378	0
NQ	7792.065012	0
D	6377.943609	0
ND	7359.388605	0
NSTEM	4559.792343	0

MFE1	752.8934757	1.61E-163
MFE2	5209.182167	0
MFE3	8608.090703	0
MFE4	3976.186129	0
TOT_BASE	2959.690429	0
N_STEMS	4257.420691	0
AVG_BP	2331.344471	0

CHAPTER 3

Development of a cloud-based web interface for the trained ML models to predict novel insect precursor microRNA and search mature microRNA targets in the model organism, *Drosophila melanogaster*.

CHAPTER 3: Development of a cloud-based web interface for the trained ML models to predict novel insect precursor microRNA and search mature microRNA targets in the model organism, *Drosophila melanogaster*.

Executive Summary:

As discussed in Chapter 2, the Support Vector Machine and Random Forest predictive models were trained for the prediction of precursor microRNA. This chapter contains a detailed account of the implementation of the trained machine learning models as a web-based tool. Furthermore, experimentally validated miRNA targets of model organism *Drosophila melanogaster* were collected from various databases which were used in a local database for searching mature microRNA targets of positively identified precursor microRNA sequences by the predictive model. This chapter also highlights various aspects of API development for interaction with predictive machine learning models for web-development, along with security features and exception handling while getting requests from users. This chapter provides a roadmap for building a full-stack web application using a raw Linux server. The API is based on python's Flask framework which is bound to the public Linux server IP with Gunicorn. The monitoring of the running of Guniron, restarting when it crashes and bug reporting are done by Supervisor. NGINX is used as the reverse proxy to get the precursor microRNA query from the User as an HTTP request. For Target searching, users need to specify the relevant information for choosing the mature miRNA sequence from the pre-miRNA sequence. RNAinsecta is freely available at: https://rnainsecta.in/ and the source code at: https://github.com/adhiraj141092/RNAinsecta

3.1. Introduction:

Precursor microRNAs (pre-miRNAs) are processed by DICER and RNA Binding Protein (TRBP) and thereafter by RNA-induced silencing complex (RISC) which produces mature miRNA that targets gene regulation by base-pairing with the target mRNA (Ha & Kim, 2014; Han et al., 2004a, 2009). However, it is difficult to predict miRNA targets as multiple mRNAs are targeted by a single miRNA and vice versa (Riffo-Campos et al., 2016).

Hence, various tools have been developed to search for potential miRNA target sites in a genomic sequence such as miRanda, Pictar, RNAhybrid, etc (John et al., 2005; Krek et al., 2005; Krüger & Rehmsmeier, 2006; Vejnar & Zdobnov, 2012). These tools search for potential target sites for a given sequence in a gene by calculating likelihood, allowing wobble base-

pairing and reward and complementarity at the 5' end. Recently tool for genome-wide pre-miRNA detection, MiRe2e was also developed using a deep learning model (Raad et al., 2022).

Databases have also been designed for experimentally validated miRNA targets (H.-Y. Huang et al., 2019; Q. Yang et al., 2011). These databases help to annotate genes regulated by a particular miRNA and vice versa, thus, enabling to make a network pathway of gene expression. Genome coordinates of the genes regulated by miRNA in the model organism *Drosophila melanogaster* is available in miRbase (Kozomara et al., 2019). The parent ID of these genes belong to FlyBase (Larkin et al., 2021).

As we have already discussed about the development of a predictive model in Chapter 2, henceforth, implementation of the model as a web-based tool has been outlined in this chapter. Furthermore, experimentally validated miRNA targets of model organism *Drosophila melanogaster* were collected from various databases which were used as a local database for searching target of positively identified pre-miRNA sequences by the predictive model.

This chapter also highlights various aspects of API development for interaction with predictive machine learning models for web-development, along with security features and exception handling while getting requests from users. This chapter provides a roadmap for building full-stack web application using raw Linux server. This chapter thus addresses Objective 2 of this thesis: "The development of a cloud-based web interface for the trained ML models to predict novel insect pre-miRNA and search targets in the model organism, *Drosophila melanogaster*".

3.2. Methods:

3.2.1 Data Collection and annotation: We initially downloaded the genome coordinates of *Drosophila melanogaster* pre-miRNA from miRBase (https://www.mirbase.org/ftp/CURRENT/genomes/dme.gff3). Experimentally verified miRNA and their target gene IDs of *Drosophila melanogaster* were fetched from MirTarBase (H.-Y. Huang et al., 2019) which was used to extract the relevant IDs from the genome coordinates obtained from miRBase. Parent IDs were annotated from the target gene ID list using e-utilities (Kans, 2020), with which the CDS of the genes were downloaded from Flybase (Larkin et al., 2021). We used regular expression in e-utilities to match the patterns and retrieve the IDs since single miRNA regulates multiple transcripts.

3.2.2. Target Searching:

MiRanda (John et al., 2005), which is a popular miRNA target searching tool, was implemented to enable users to search for potential miRNA targets for their pre-miRNA in the CDS of reported genes regulated by miRNA.

3.2.3. Web server implementation:

3.2.3.1. API development in localhost:

The selected trained models were implemented in a backend server using python's Flask API (Application Programming Interface) on a cloud platform. Flask use routing with decorators with which each API can be programmed to interact with the backend processor, for example:

```
from Flask import Flask
app = Flask(__name__)

@app.route('/')
def hello_world():
   return 'Hello World'

if __name__ == '__main__':
   app.run()
```

This will return `Hello World` on a web-browser with Flask running at: `localhost/` or `http://127.0.0.1:5000/` since the decorated route is at '`/`'. The `route()` function of the

Flask class is a decorator, which tells the application which URL should call the associated function. Similarly, routes were decorated for `/` (root), `/predict` and `/mirtar` for the homepage, predicted result and target prediction respectively. For prediction, the ML (machine-learning) models trained on Support Vector Machine (SVM) and Random Forest (RF) as discussed in Chapter 2 were used.

For example, the pseudo code for root route acting as a homepage is:

```
@app.route('/')
def home():
    return render_template('home.html')
```

Here we rendered HTML templates to our API route for displaying the content of our homepage. Similarly, we used respectively HTML templates for rendering in `predict` and `mirtar` routes.

For batch and single sequence submission for prediction, two routes were employed, namely `predict` and `predict2`. Each of these routes acted as endpoint of the API that sent the response from the server to the front-end web-server through Jinja2 templates. An easy calling convention called the Web Server Gateway Interface (WSGI) allows web servers to send requests to web apps or frameworks created in the Python programming language (Xavier Leitão et al., 2020).

3.2.3.2. Backend server setup:

The backend cloud server was based on Linux OS with Ubuntu 18 LTS with 2 CPUs at 2200.00 MHz clock speed running on Intel Core Processor (Broadwell, IBRS), with 20GB cloud space and 2GB RAM. For security measures, the app is executed from guest user account so that the root user stays safe. To log in the cloud server `ssh` command was used and to transfer files `scp` command was used.

The goal was to run the same FLASK web app that we discussed in the previous section but with a public IP, i.e. instead of localhost, it should be accessible remotely from any location over the web. However, to reflect an app over a public IP requires lot of configuration to handle user data, management for flow of data, assigning load to each worker node that will be taking the request and processing, getting request from backend server. Along with keeping

track of logging information, bug reporting and restarting the app in case it stops running unexpectedly or exits.

The first step to serve the FLASK web app over the cloud server was to bind the app with the IP address of the cloud server. This was done using Gunicorn (Green Unicorn) which takes the FLASK app request and works as a WSGI server. It is invoked using:

```
gunicorn app:app -b localhost:8000 &
```

This command makes the web server listen to localhost at port 8000. This process only works until the gunicorn app is running in the terminal. Hence, to improve it Supervisor which is a monitoring system designed for WSGI like Gunicorn that enables the execution of such web apps in background even when the webmaster is not logged in through `ssh` tunnel. The supervisor script has to be run from the location `/etc/supervisor/` as a file called `conf.d` . This script contained the following codes:

```
[program:mirna]

directory=/home/XXX/rnainsecta/source/RNAinsecta/web
command = /usr/bin/gunicorn3 -w 3 app:app -b localhost:8000
user=XXX
autostart = true
autorestart = true
stopasgroup = true
killasgroup = true
stderr_logfile = /home/XXX/err/mirna.err.log
stdout_logfile = /home/XXX/err/mirna.out.log
```

This script auto restarts if the program crashes and sends log and error information to the specified path. The domain name can be added along with the IP so that user can access the app through HTTP request which NGINX was used (Nedelcu, 2013). To configure NGINX web engine, a configuration file was added to `site-enabled` directory in the location `/etc/nginx/site-enabled`. The file contains the following:

```
server {
    listen 80;
    server_name rnainsecta.in;
    return 301 https://rnainsecta.in$request_uri;
    location /static {
        alias /home/XXX/rnainsecta/web/static/;
    }
```

```
    location / {
        proxy_pass http://localhost:8000;
        include /etc/nginx/proxy_params;
        proxy_redirect off;
    }
}
```

Here, the domain name `rnainsecta.in` is added to the server name which return the http request. The files that will be accessed during the execution of the program will be stored in `static` directory.

3.2.3.3. Front End:

The request from FLASK API was served on the web-browsers of the users with which they interact with the tool RNAinsecta. The webpage is the front end of the tool which is designed using HTML, CSS and JavaScript. FLASK uses Jinja2 templates through which any value can be passed from the API to the front-end using a function for example:

```
@app.route('/predict',methods = ['POST', 'GET'])
def predict():
   if request.method == 'POST':
      return render_template("predict.html",display = "This is a pre-miRNA")
```

Here the function `display` carries the string "`This is a pre-miRNA`" which is served from back-end server. To display this string in the webpage, a tag "`{{display}}`" has to be used in the rendered template HTML file.

The chart elements displayed during prediction is done using chart.js package. The element were dynamically created using javascript using `getElementById` method.

Different nucleotide counts on clicks was achieved using JQuery where each graph was created on click and the remaining were made to hide. The sample pseudo code to achieve this is:

```
var ctx = document.getElementById('myChart').getContext('2d');
    var chart = new Chart(ctx, {
// The type of chart we want to create

        type: 'bar',

        // The data for our dataset
        data: {
            labels: features,
            datasets: [{
```

```
                label: 'Nucleotide Percentage',
                backgroundColor: 'rgb(0, 99, 132)',
                borderColor: 'rgb(0, 99, 132)',
                data: score
            }]
        },
        // Configuration options go here
        options: {
            scales: {
            yAxes: [{
                ticks: {
                    beginAtZero: true
                }
            }],
            xAxes: [{
                // Change here
                  barPercentage: 0.4
            }]
            }
      }
    });

$('#0').on('click', function (e) {
        e.preventDefault;
        chart.config.data = {
            labels: di_features,
            datasets: [{
                label: "Dinucleotide Percentage",
                backgroundColor: 'rgb(19, 74, 1)',
                borderColor: 'rgb(19, 74, 1)',
                data: di_score,

            }],
        }
        chart.update();
    });
$('#1').on('click', function (e) {
        e.preventDefault;
        chart.config.data = {
            labels: features,
            datasets: [{
                label: 'Nucleotide Percentage',
                backgroundColor: 'rgb(0, 99, 132)',
                borderColor: 'rgb(0, 99, 132)',
                data: score

            }],
        }
        chart.update();
    });
$('#2').on('click', function (e) {
        e.preventDefault;
        chart.config.data = {
```

```
            labels: a_fo_features,
            datasets: [{
                label: 'Adenine Folding',
                backgroundColor: 'rgb(1, 11, 37)',
                borderColor: 'rgb(2, 28, 103)',
                data: a_fo_score

            }],
        }
        chart.update();
    });
$('#3').on('click', function (e) {
        e.preventDefault;
        chart.config.data = {
            labels: g_fo_features,
            datasets: [{
                label: 'Guanine Folding',
                backgroundColor: 'rgb(1, 11, 37)',
                borderColor: 'rgb(2, 28, 103)',
                data: g_fo_score

            }],
        }
        chart.update();
    });
$('#4').on('click', function (e) {
        e.preventDefault;
        chart.config.data = {
            labels: c_fo_features,
            datasets: [{
                label: 'Cytosine Folding',
                backgroundColor: 'rgb(1, 11, 37)',
                borderColor: 'rgb(2, 28, 103)',
                data: c_fo_score

            }],
        }
        chart.update();
    });
$('#5').on('click', function (e) {
        e.preventDefault;
        chart.config.data = {
            labels: u_fo_features,

            datasets: [{
                label: 'Uracil Folding',
                backgroundColor: 'rgb(1, 11, 37)',
                borderColor: 'rgb(2, 28, 103)',
                data: u_fo_score

            }],
        }
        chart.update();
```

```
    });
});
```

In the target prediction, JSON elements were parsed from the backend FLASK server as

```
result = df.to_json(orient="records")
            parsed = json.loads(result)
            return render_template('mirtar.html',
targets=json.dumps(parsed))
```

Here, the HTML template used is called mirtar.html, with which JSON data is exported from the FLASK server which is parsed by the front end JavaScript. This is then converted to dynamic table with the table elements `<td> <tr>` for columns and rows respectively. The pseudo code used to achieve this is:

```
var t =
JSON.parse(JSON.parse(document.getElementById("targetid").dataset.targets))
;

function mir(mi) {
    return `

     <tr>
        <td style="width:220px"><a href="${mi.urlF}"
target='_blank'><b>Target Transcript:</b> ${mi.name}</a> </td>
        <td style="width:200px"><b>Range transcript:
</b>${mi.Ref_range}  </td>
        <td
style="width:100px"><b>Score:  </b>${mi.score}; </td>
        <td style="width:180px"><b>Alignment Length: </b>${mi.aln_len}
 </td>
        <td style="width:130px"><b>Energy:</b> ${mi.energy}</td>
        <td><a href="${mi.urlm}"
target='_blank'><b>MicroRNA:  </b> ${mi.mirna}</a> </td>
    </tr>
    <tr><td colspan="6"> <br></td>
    <td colspan="6" align="center">
    <pre><b>${mi.algn}</b></pre>
    </td>
    </tr>
     <tr>
```

```
    <td colspan="6">
    <br>
    </td>
    </tr>
      `
}

document.getElementById("mir").innerHTML = `
<table style="float:left">
${t.map(mir).join("")}
</table>
;
```

3.2.3.4. Security and exception handling:

The website loads securely with SSL (Secure Sockets Layer) certificate generated by Let's Encrypt which was implemented using certbot. This certificate ensures no malicious activity is being done (Tiefenau et al., 2019). Certbot was executed using the following command:

```
sudo certbot --nginx -d rnainsecta.in -d www.rnainsecta.in
```

This takes few minutes to check the cache and cookies collected from the webpage and finally returns a certificate which enables the use of secured http requests (`https://`).
The exceptions were handled using Regular Expression (Regex) so the working server does not require to process any error prone data. Regex looks for:

- Any missing data
- Any character apart from the four nucleotides (A,T/U,G and C)
- Line Break in the nucleotide sequence

The codes written in JavaScript for handling these exceptions are as follows:

```
function check_Beta(letters){
    var regex = /^>.*\n(?![ATGUCatucg]*$).*/gim;
    var reg1 = /(>.*)(\n)[ATGCuUatcg]+(\n)+[ATGCuUatgc]/;
    var reg2 = /(>.*)(\n)/;


    if(reg2.test(letters.rna1.value) == false){
   alert("Enter sequence with fasta header.");
   letters.rna1.focus();
   return false;
    }

    else if(reg1.test(letters.rna1.value) == true){
```

```
 alert("Line break detected! \nThe sequence must be in the same line.");
 letters.rna1.focus();
 return false;
  }
  else if(regex.test(letters.rna1.value) == true){
 alert("Sequence must contain A,T/U,G & C only");
 letters.rna1.focus();
 return false;
  }

  return true;
}
```

3.3. Results:

3.3.1. miRNA Target Collection:

We have listed the chromosome-wise miRNA target distribution of *Drosophila melanogaster* in Table 3.1. There are total 176 targets of which chromosome 2 Left and Right (2L and 2R) has 34 and 7 respectively. There are 46 targets for left and 55 for right of Chromosome 3 (3L and 3R) respectively. Chromosome 4 has 9 targets whereas sex chromosome X has 23 targets.

Chromosome	No. of Sequence
2L	34
2R	7
3L	46
3R	55
X	23
4	9

Table 3.1: No. of targets from each chromosome of Drosophila melanogaster.

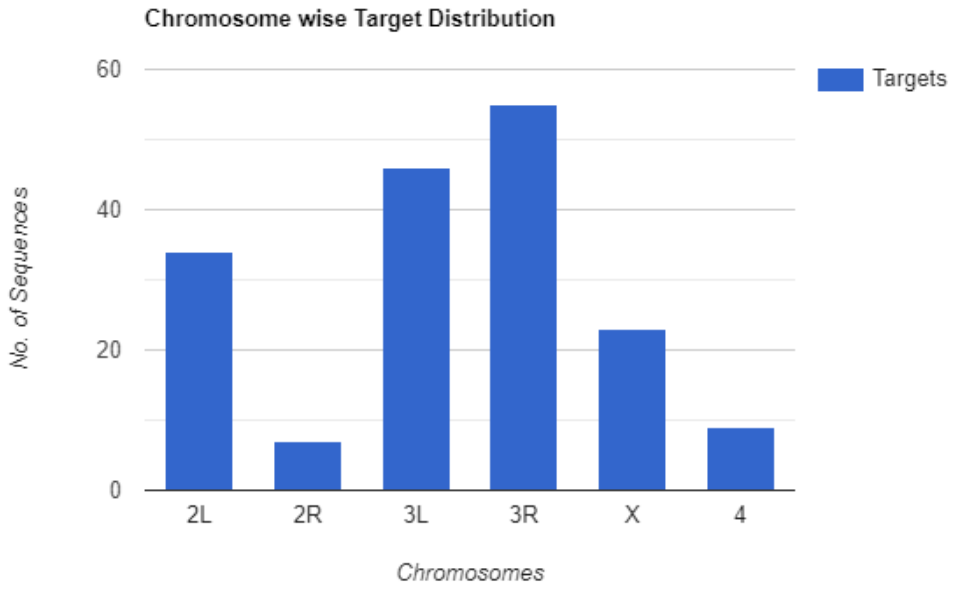

Figure 3.1. Chromosome wise miRNA target distribution.

3.3.2. Web Server implementation:

The web-architecture designed using FLASK, NGINX, Gunicorn and supervisor for building the web tool RNAinsecta is given in Figure 3.2.

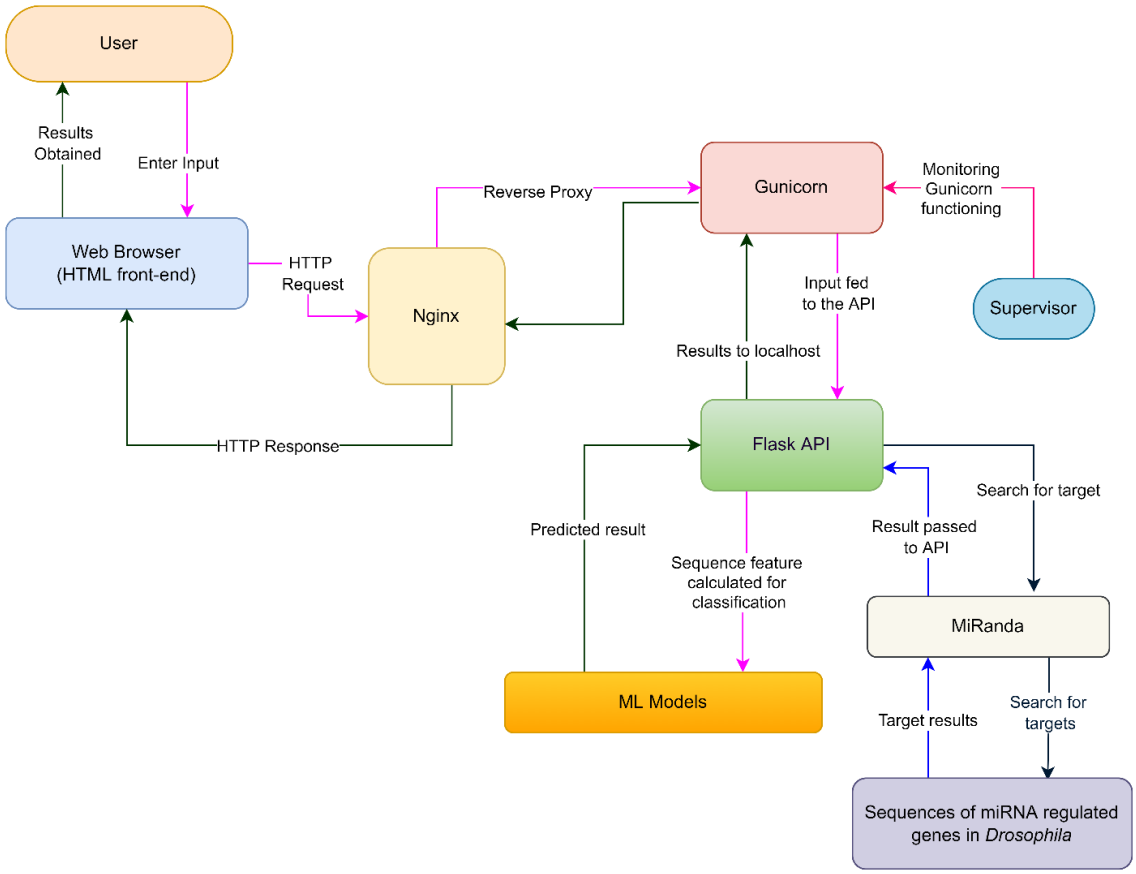

Figure 3.2: Web-server implementation of RNAinsecta. The figure shows the full-stack implementation of the website. NGINX takes nucleotide sequence information as HTTP request from User through the Homepage. Flask runs the API for pre-miRNA and target prediction as a localhost. Gunicorn works as mediator between NGINX and Flask which allows public IP to interact with the APIs. The monitoring of Gunicorn is done by supervisor, which prepares error reports and restarts the server if it stops unexpectedly.

3.3.2.1. API development in localhost:

The FLASK local server was compiled with the result given in Figure 3.3. The web-app was running successfully on the development server. The web-app was available at http://127.0.0.1:5000/ which corresponds to the localhost.

```
* Serving Flask app "app" (lazy loading)
* Environment: production
  WARNING: This is a development server. Do not use it in a production deployment.
  Use a production WSGI server instead.
* Debug mode: on
* Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
* Restarting with stat
* Debugger is active!
* Debugger PIN: 829-988-276
```

Figure 3.3. Compilation and deployment of development server created using python FLASK.

3.3.2.2. Backend server setup:

Upon checking supervisor with:

```
sudo supervisorctl status
```

The backend Gunicorn WSGI server running supervisor gave the results shown in Figure 3.4.

```
adhiraj@localhost:~$ sudo supervisorctl status
mirna                            RUNNING   pid XXXXXX, uptime 0:03:22
```

Figure 3.4. Supervisor successfully running WSGI server. For security purposes, the `pid` has not been disclosed.

This ensures the FLASK server has been bound to the WSGI. To check the result of NGINX server the following command is used:

```
sudo systemctl status nginx
```

Upon execution of the above command, the result displayed is shown in Figure 3.5. The NGINX server uses one Master process and 2 Worker process for balancing the loads. With these worker nodes, it can parallelly execute multiple requests. This shows our web-application RNAinsecta was successfully deployed in the production server.

```
adhiraj@localhost:~$ sudo systemctl status nginx
● nginx.service - A high performance web server and a reverse proxy server
   Loaded: loaded (/lib/systemd/system/nginx.service; enabled; vendor preset: enabled)
   Active: active (running) (Result: exit-code) since Sun 2022-09-11 17:56:12 UTC; 3 months 26 days ago
     Docs: man:nginx(8)
  Process: XXXXX ExecReload=/usr/sbin/nginx -g daemon on; master_process on; -s reload (code=exited, status=0/SUCCESS)
 Main PID: XXXXX (nginx)
    Tasks: 3 (limit: 2342)
   CGroup: /system.slice/nginx.service
           ├─10374 nginx: master process /usr/sbin/nginx -g daemon on; master_process on;
           ├─15080 nginx: worker process
           └─15081 nginx: worker process
```

Figure 3.5. NGINX successfully running in the backend server. The Active status shows running along with a master process and 2 worker processes. For security purposes Main PID and Process ID has not been disclosed.

3.3.2.3. Front End:

The Front end for RNAinsecta is available at https://rnainsecta.in/. The web interface of RNAinsecta is given in Figure 3.6. It contains both RNAinsecta_RF and RNAinsecta_SVM classifiers with batch and single sequence query.

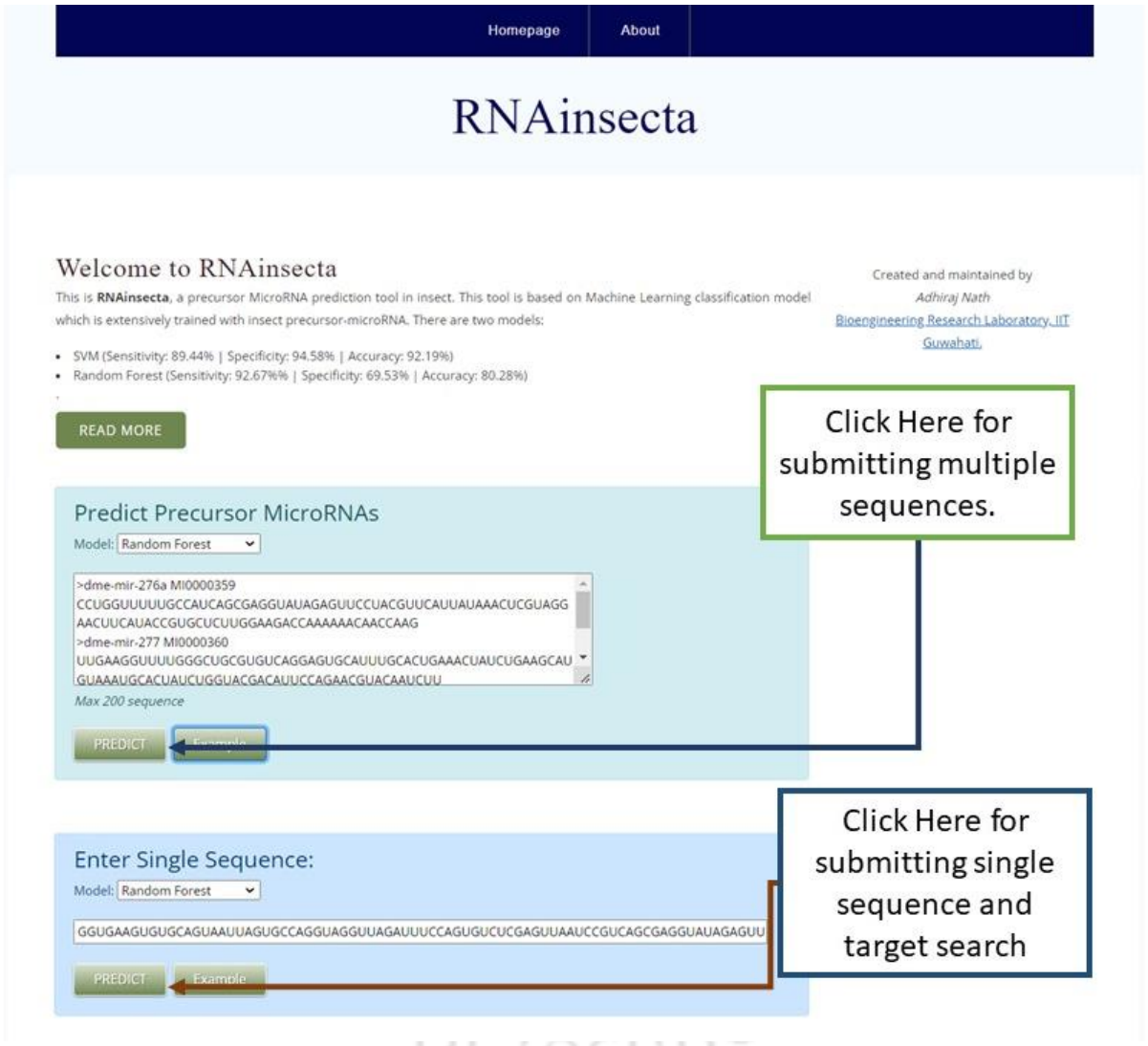

Figure 3.6. The homepage of RNAinsecta. It contains search option for both the Random Forest and SVM classifier for single and batch query. Example sequences are also provided which is displayed upon clicking on 'Example".

The result displays the prediction result and probability score for both batch and single sequence query which is given on Figure 3.7. On the left A. the results for batch input sequence is given where the HTML elements are dynamically created depending on the number of submitted query which is achieved using JQuery.

On the right B. shows the results for single sequence queries where users can download the secondary structure predicted for their pre-miRNA by RNAfold. Also, users can visualize various parameters of their sequence such as nucleotide and dinucleotide counts, and Triplet folding pattern which is achieved using chart.js JavaScript library.

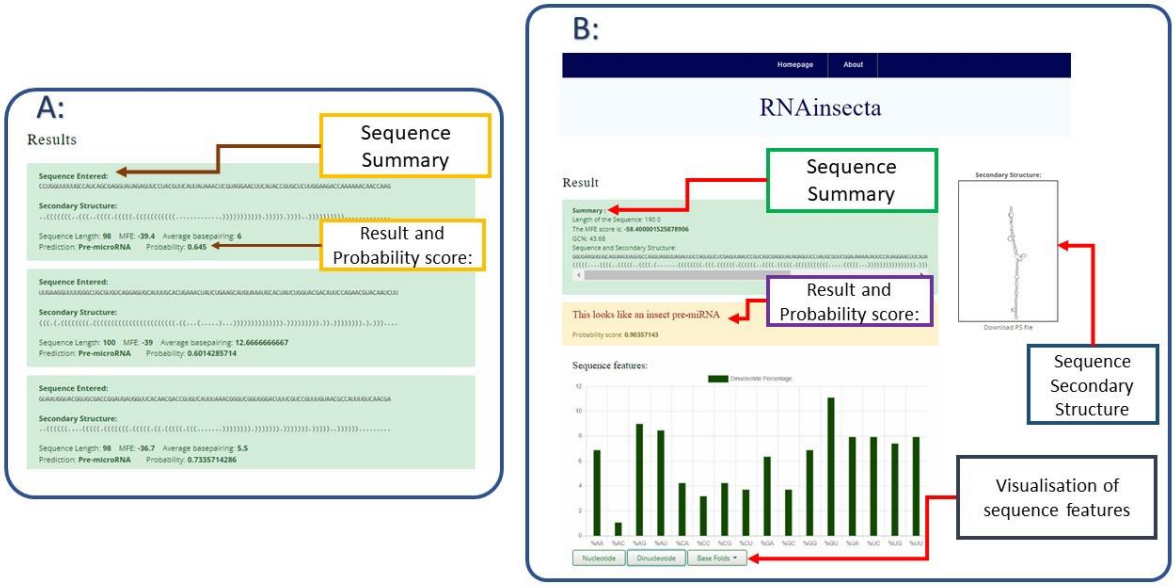

Figure 3.7. The predicted results of RNAinsecta. A. shows the result for batch queries. It accepts a maximum of 200 sequences. B. shows the result for single sequence queries. It shows the graphs and charts corresponding to the feature of the input nucleotide sequence.

For Target searching, users need to specify the relevant information for choosing the mature miRNA sequence from the pre-miRNA sequence. The interface for target prediction is given in Figure 3.8. For a particular single sequence query predicted to be true pre-miRNA, users can search for the mature miRNA targets in each chromosome given in Table 3.1. miRanda, miEMap and RNAhybrid program runs in the back-end which searches the targets and displays the top 100 results. The selection of cleavage site of pre-miRNA to give mature miRNA is dynamically created depending upon the occurrence and position of the hairpin loops predicted in the secondary structure on the selected 3' or 5' directionality. User defined miRNA sequence can also be provided as query in case user already has mature miRNA sequence from the pre-miRNA given in Figure 3.8.A. The result page contains a hyperlink to the corresponding FlyBase Transcript ID. It displays the gene with a probable target match. It also includes a hyperlink to the miRBase ID of the miRNA that has been experimentally found to control the particular FlyBase Transcript shown in Figure 3.8.B.

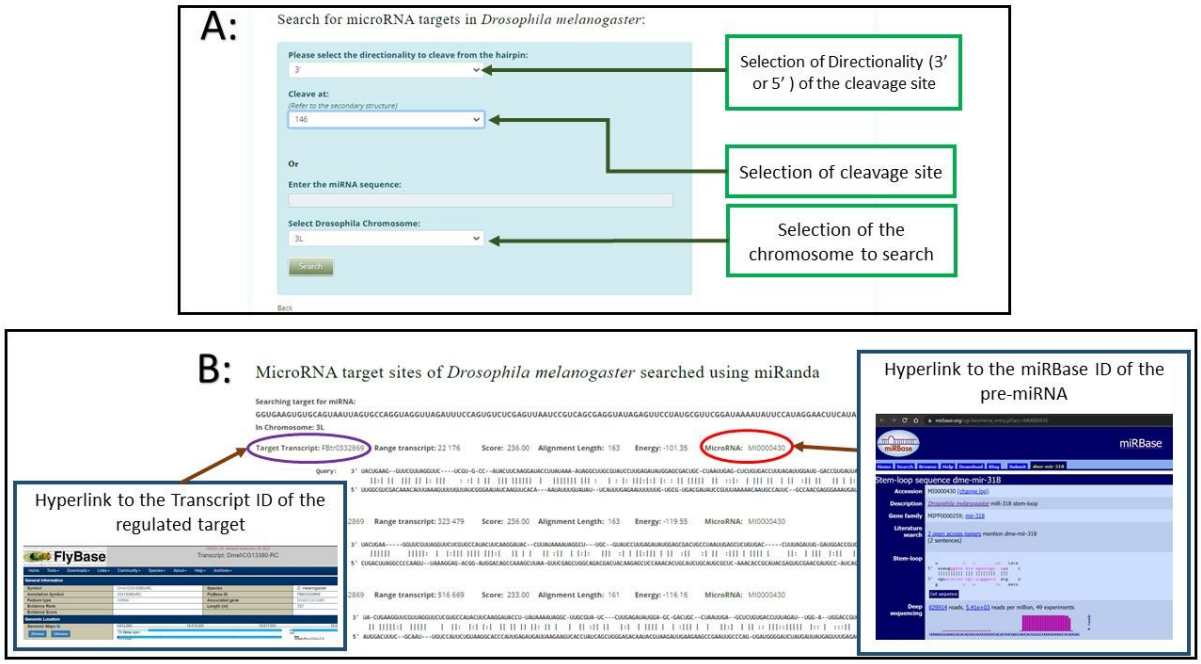

Figure 3.8. A. RNAinsecta's user interface for searching miRNA targets in Drosophila melanogaster. B. The result of miR target search containing Transcript ID and its hyperlink to FlyBase as well as miRBase ID and its hyperlink.

3.3.2.4. Security and exception handling:

The site information when accessed through Google Chrome is given in Figure 3.9. It shows a locked padlock on the top left corner suggesting the website is secure. This ensures that the website does not store any user information or performs malicious activity.

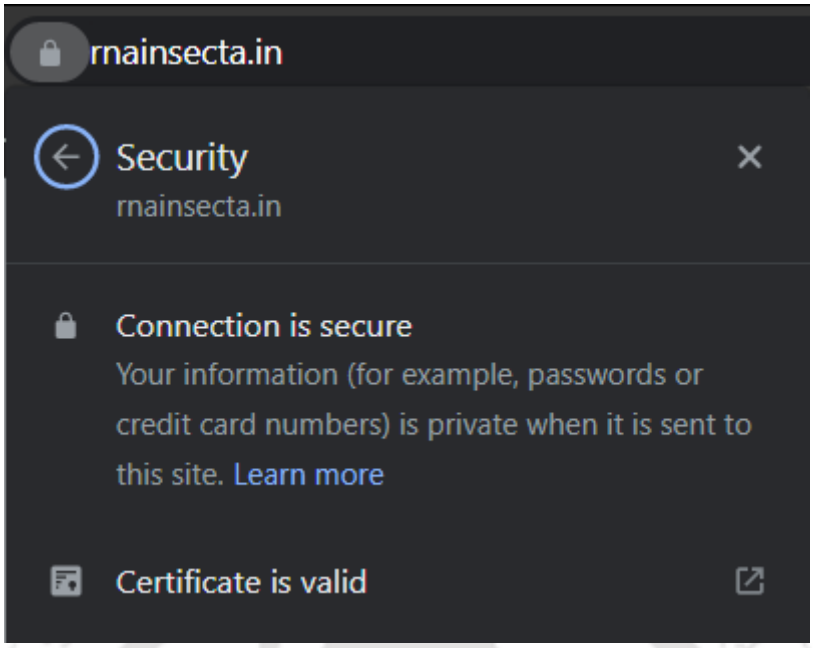

Figure 3.9. Site information provided by Google Chrome. This information is displayed when clicked on the padlock located in the top left corner.

SSL certificate issued by "Let's Encrypt" makes the connection secure. The SSL certificate accessed by Google Chrome is given in Figure 3.10. The Fingerprints issued as SHA (secure hashing algorithm) is given at the bottom of the figure. The certificate renews every 3 months.

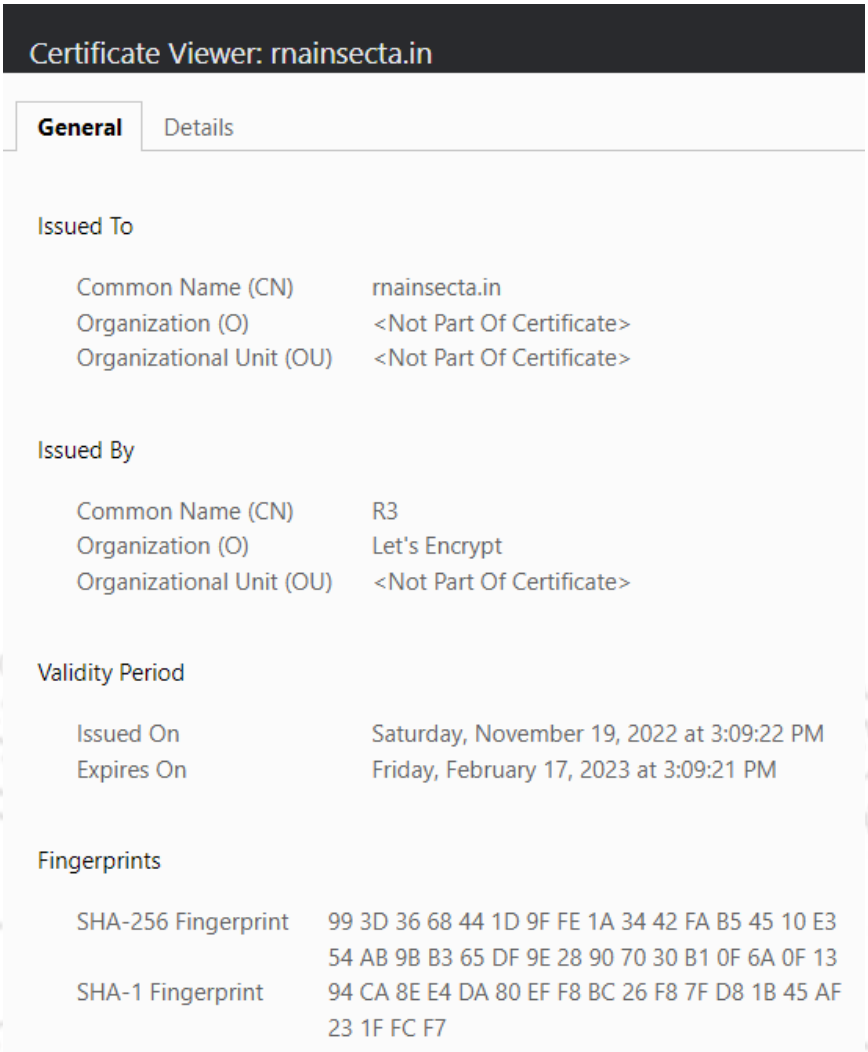

Figure 3.10. SSL certificate of RNAinsecta. This certificate is generated by Let's Encrypt using certbot. This certificate renews every 3 months

.

An example of exception handling is given in 3.11. were the nucleotide sequence contained characters other than ATGC and hence the request was not sent to the web server.

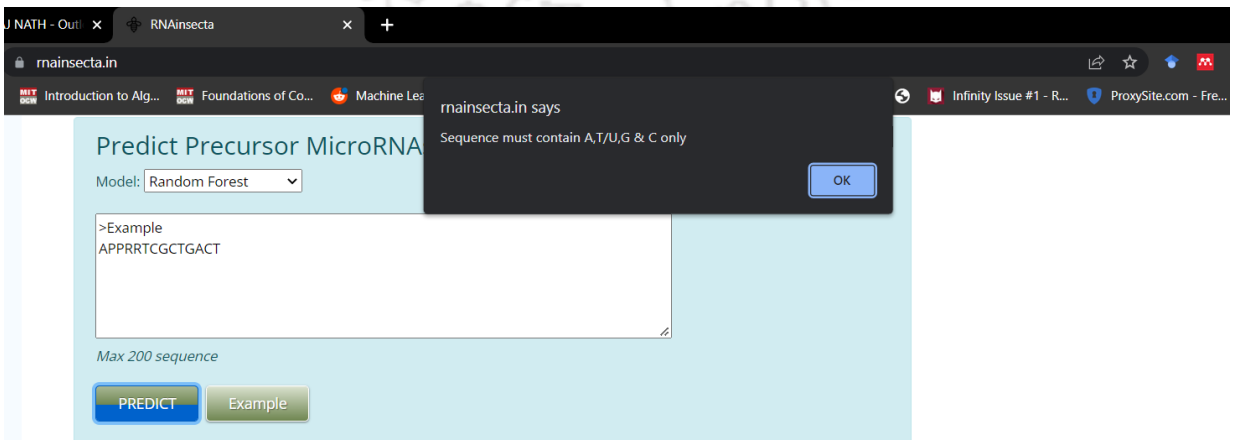

Figure 3.11. Example of exception handling using Regex in JavaScript. As the input sequence contains characters other than the desired ATGC nucleotide bases, the request was not sent to the backend web server.

Figure 3.12 is the QR code containing the URL for the tool. This image can be scan by any smart phone with QR code scanner app installed which will redirect to RNAinsecta.

Figure 3.12. QR code for RNAinsecta.

3.4. Discussion:

3.4.1. **Data collection and annotation:**

The genomic pre-miRNA location map of *Drosophila melanogaster* from the GFF file obtained from miRbase contained genes with parent ID and miRNA name. As single miRNA may target multiple mRNA, hence we enriched all the targets that is regulated by a given miRNA. The targets were fetched as gene transcripts from FlyBase. However, FlyBase ID is not readily available and had to be fetched from NCBI GenBank using e-utilities. The transcripts were stored in the backend server for Target searching using miRanda. For the ease of access, we distributed the targets chromosome wise, so that user has the freedom to choose the desired location. The highest number of target transcripts were collected from 3R chromosome and the lowest were collected from 2R chromosome.

3.4.2. **Web Server implementation:**

The implementation of Jinja2 templates in the Flask enabled content of the python backend to be easily communicated to the frontend webpage. The GET and POST method used in the Flask decorator enabled the server to request data from the user and send the results back to them. The functions containing numerical values or strings could be easily returned to the user using the Jinja2 templates.

The compilation of the backend Flask server was done in the production server. However, it is not advisable to run web application on production server. Hence, it needs IP binding. Upon successful IP binding, the app was served to NGINX which acted as a reverse proxy. NGINX supports multiple web applications to run on the same production server, hence it can manage high load application and data flow.

RNAinsecta has a user-friendly UI/UX design that enables users to submit and retrieve information easily from the web application. The use of JavaScript together with JQuery has enabled many vital features such as display of figures and chart.

Each user request is kept separate from one another by using session method in Flask. Hence, each request has its own session which does not interfere with the prediction result or the target searching.

As the exceptions are handled in the front end itself, it keeps the backend safe from unnecessary crashing due to incorrect data or making unnecessary calculations to incorrect input sequence.

The ease of switching from one classifier to another is enabled in the front end which enables researchers to select a predictive model on a click of a button. The resultant page contains the probability score for the sequence to be in positive class of insect pre-miRNA. Users can also download the image file for secondary structure predicted by RNAfold for a single sequence.

The pre-processing of pre-miRNA to mature miRNA has been made easy for the user by automated regex pattern match to find the hairpin or loop in the 3' or 5' end of the sequence and prompt the user to select the one which the believe is the optimum based on the secondary structure received. The option to select the directionality is given since miRNA has been reported to be produced from both 5' and 3' of pre-miRNA.

The target searching result contains hyperlinks to transcript of parent Flybase ID which contains the information about the gene, function and its location in the chromosome map of the model organism *Drosophila melanogaster*. It also contains the miRBase ID of the miRNA which is experimentally found to be taking part in regulation of that gene. (H. Y. Huang et al., 2022)

3.5. Conclusion:

In this chapter we developed a website for the ML predictive models discussed in the previous chapter. The website is hosted on a public IP domain through which anyone on the internet can access the web tool RNAinsecta. It contains both the Random Forest and SVM trained model for prediction depending upon the experimental design of user. It accepts both batch and single sequence queries. The tool not only provides prediction for pre-miRNA in insects but also provides a comprehensive approach for the researchers to extract the mature miRNA as well as search for its target in the genome of the model organism, *Drosophila melanogaster*. The tools RNAinsecta not only is easy to use but is also safe to use, which is certified by the SSL certificate. In future this web tool can be configured for NGS analysis where pre-miRNA collection can be estimated from the RNA-Seq data.

CHAPTER 4

Comparative analysis of sequential and thermodynamic features of insect precursor microRNA with other organisms.

CHAPTER 4: Comparative analysis of sequential and thermodynamic features of insect pre-miRNA with other organisms.

Executive Summary:

This chapter contains a discussion about the comparative statistical analysis of various features used the in development of machine learning based predictive tools. The sequence features of insect precursor microRNA were compared with precursor microRNA of other available organisms. We initially established that features such as Length, GC content, Minimum Free Energy (MFE) of folding, etc., differs in insects as compared to other organisms using chi-square test. We further trained a predictive model for classification using XGBoost between insects, human, monocots, aves, ruminants, sauria, dogs and rodents. We performed PCA and retained 20 best features for classification. Various parameters of XGBoost was tuned with 5-fold CV and the parameter values with highest CV score were considered. The accuracy of precursor microRNA prediction in insect, plants, rodents, human, ruminants, sauria, aves and dogs were found to be 0.9125, 0.9325, 0.8575, 0.86, 0.8875, 0.8225, 0.845 and 0.8675 respectively. The AUC was found to be 0.934 for insects, 0.985 for plants (rice), 0.941 for sauria, 0.859 for dog, 0.815 for ruminant, 0.743 for human, 0.75 for rodent and 0.765 for aves.

4.1. Background:

Precursor microRNA (pre-miRNA) are the non-coding RNA hairpin loops which is cleaved by Drosha to produce microRNA (miRNA) (Han et al., 2004b; O'Brien et al., 2018). Multiple miRNAs can be produced from a single pre-miRNA for which characterization and identification of pre-miRNA has been of great importance. miRNA has been found to regulate gene expression of various biological processes such as development, cell proliferation, cell differentiation, apoptosis, transposon silencing, and antiviral defense (Ambros, 2004; Cullen, 2009; Ruvkun, n.d.; Ventura & Jacks, 2009). In insects, changes in miRNA expression profile have been observed in various biological processes such as metamorphosis, reproduction, immune response, etc. (Kayvan Etebari & Asgari, 2013; Gulhane et al., 2022; Song et al., 2018; Sun et al., 2019; Kaleem Tariq et al., 2016b; Q. Zhang et al., 2019; Yang Zhang et al., 2016). miRNAs are believed to be conserved although they target diverse genes. They are believed to be similar across all the species. (Friedman et al., 2009; C. T. Lee et al., 2007; Willmann & Poethig, 2007)

Various tools are designed to predict pre-miRNAs as they give rise to mature miRNA. These data are downloaded from miRBase which contains collection of pre-miRNAs and their corresponding miRNAs of various organisms (Kozomara et al., 2019). It currently holds 38,589 miRNAs from 271 organisms. Features such as nucleotide bases, length of the sequence, GC content of pre-miRNAs are used to train machine learning classifiers to predict a true pre-miRNA (Batuwita & Palade, 2009; Gudyś et al., 2013; Ng & Mishra, 2007; Stegmayer et al., 2017; Tran et al., 2015; Xue et al., 2005). There are tools which are specifically trained on a particular organism or phylum which are capable of predicting pre-miRNA exclusively to that organism or phylum (K.-Y. Huang et al., 2015; Jaiswal et al., 2019).

In this work, we analyzed the pre-miRNA sequences of insect pre-miRNA from miRbase and performed comparative statistical analysis with other available organisms. We initially established that features such as Length, GC content, MFE, etc., differs in insects as compared to other organisms. Thus, this chapter addresses Objective 3 of this thesis: "Comparative analysis of sequential and thermodynamic features of insect pre-miRNA with other organisms". We further trained a predictive model for classification using XGBoost between insects, human, monocots, aves, ruminants, sauria, dogs and rodents.

4.2. Methods:

4.2.1. Data Collection and pre-processing:

We collected pre-miRNA sequences of insects, human, monocots, aves, ruminants, sauria, dogs and rodents from miRBase (Kozomara et al., 2019) and labelled them for comparison. The secondary structure was calculated using RNAfold software from ViennaRNA package. The fasta header, nucleotide sequence, MFE score and secondary structure for each pre-miRNA sequence was converted into tabular format using in-house python script. We calculated the parameters using the protocols mentioned in Chapter 2, section 2.2.2. The dataset is available in the GitHub repository:

https://github.com/adhiraj141092/RNAinsecta/blob/master/dataset/comp_data_labels.csv.

4.2.2. Hypothesis testing:

We resampled 500 datasets for each labelled category and performed hypothesis test.
The null hypothesis was:

$$H_0 = All\ the\ precursor\ miRNA\ features\ are\ similar\ among\ all\ organisms$$

Our alternate hypothesis states that insect pre-miRNAs are different in many aspects which are routinely used in ML (machine learning) tools, i.e.

$$H_A: Features\ of\ precursor\ miRNA\ differ\ significantly\ between\ insects\ and\ other\ organisms$$

We performed chi-square test between insect and other organism which is given by:

$$\chi^2 = \sum_{i=0} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency of a pre-miRNA feature i and E_i is the expected frequency of the pre-miRNA feature i.

4.2.3. Feature engineering:

We performed dimensional reduction techniques to extract the most feasible features for making multi-class training model. We calculated the Pearson's correlation and chose the parameters which had positive correlation. Along with these features we used Scikit-Learn's (Pedregosa et al., 2011) model inspection technique called "permutation feature importance"

to find the weight of each feature and keep the important ones, i.e. $w_i > 0$, where w is the weight of each feature i. We then performed Principal Component Analysis (PCA) with 5, 10, 15, 20, 25, 30 and 35 features.

Python codes used for feature engineering:

```
import eli5
from eli5.sklearn import PermutationImportance
import xgboost as xgb
params = {'lambda': 0.01151671654705981, 'alpha': 0.05676732265357062,
'colsample_bytree': 0.8, 'subsample': 1.0, 'learning_rate': 0.012,
'n_estimators': 160, 'max_depth': 9, 'random_state': 42,
'min_child_weight': 3}

rf= xgb.XGBClassifier(**params)
rf.fit(X_train,y_train)
perm = PermutationImportance(rf, random_state=0).fit(X_test, y_test)
eli5.show_weights(perm, feature_names = X.columns.tolist())
#X are the feature input and y are the corresponding labels to the class
#Splitting the dataset for training and testing
from sklearn.model_selection import train_test_split

X_train,X_test,y_train,y_test = train_test_split(X,y,random_state =
42,test_size=0.25)


#PCA

from sklearn.decomposition import PCA

pca = PCA(n_components=20,random_state=42)
pca_data = pca.fit_transform(X_train)
pca_test = pca.transform(X_test)
```

4.2.4. Multiclass training with XGBoost:

We kept 80% of the data for training (X_train) and used 20% of the data for testing (X_test). We used XGBoost classifier to train the multiclass data for classification. The algorithm of XGBoost constructs multiple CART models in parallel which effectively improves the computation speed. Second-order Taylor formula is used to optimize by model by calculating the error value between the predicted and true value (Ahmad et al., 2020). It can further handle missing feature values and hence does not require feature standardization (H. Yang et al., 2020). It has hence been used in the estimation and classification biological data (Bi et al., 2020; H. Li et al., 2022). It is based on minimising the loss function and regularization, $L^{(t)}$ which mathematically it can be written as:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

Where l measures the difference between prediction $\hat{y}_i$ and target y_i in the ith instance at iteration t. f_t is an independent tree for given input x_i. $\Omega(f_t)$ works has a penalty function. We used 5 fold CV while optimising the following parameters of the XGBoost package in Scikit-learn:

- **lambda**: L2 regularization range from 1e-3 to 10
- **alpha**: L1 regularization range from 1e-3 to 10.
- **colsample_bytree**: Subsample ratio of columns during construction of each tree, ranges from 0.3 to 1.0
- **subsample**: Ratio of training instances, ranges from 0.4 to 1
- **learning_rate**: Step size at each iteration while moving towards minimum of loss function, ranges from 0.001 to 0.2
- **n_estimators**: Number of trees, ranges from 50 to 400
- **max_depth**: Max depth of a tree, ranges from 5 to 17
- **min_child_weight**: Minimum instances needed to be in each node, ranges from 1 to 300

Python codes used:

```
#XGBoost parameter optimization

import optuna
from sklearn.metrics import accuracy_score
import xgboost as xgb
def objective(trial,data=X,target=y):

    # train_x, test_x, train_y, test_y = train_test_split(data, target,
test_size=0.1,random_state=42)
    param = {
        'tree_method':'gpu_hist',
        'lambda': trial.suggest_loguniform('lambda', 1e-3, 10.0),
        'alpha': trial.suggest_loguniform('alpha', 1e-3, 10.0),
        'colsample_bytree': trial.suggest_categorical('colsample_bytree',
[0.3,0.4,0.5,0.6,0.7,0.8,0.9, 1.0]),
        'subsample': trial.suggest_categorical('subsample',
[0.4,0.5,0.6,0.7,0.8,1.0]),
        'learning_rate': trial.suggest_categorical('learning_rate',
[0.008,0.01,0.012,0.014,0.016,0.018, 0.02]),
        'n_estimators': trial.suggest_int('n_estimators', 50, 400),
```

```
        'max_depth': trial.suggest_categorical('max_depth',
[5,7,9,11,13,15,17]),
        'random_state': trial.suggest_categorical('random_state', [42]),
        'min_child_weight': trial.suggest_int('min_child_weight', 1, 300),
        'objective':'multi:softmax',
        'num_class':8
    }
    model = xgb.XGBClassifier(**param)

model.fit(new_X_train,y_train,eval_set=[(new_X_test,y_test)],early_stopping
_rounds=100,verbose=False)
    preds = model.predict(new_X_test)
    #
model.fit(train_x,train_y,eval_set=[(test_x,test_y)],early_stopping_rounds=
100,verbose=False)
    # preds = model.predict(test_x)

    accuracy = accuracy_score(y_test, preds)

    return accuracy

study = optuna.create_study(direction='maximize')
study.optimize(objective, n_trials=30)
print('Number of finished trials:', len(study.trials))
print('Best trial:', study.best_trial.params)
```

4.2.5. Performance Analysis:

The best parameters were chosen and the models were evaluated on X_test dataset to check their efficiency. During performance evaluation, we considered True dataset to be the group that is being evaluated and False dataset to be the collection of all other groups in each case. The performance was calculated based on the following classical classification measures: sensitivity (*SN*): $SN = \frac{TP}{TP+FN}$, specificity (*SP*): $SP = \frac{TN}{TN+FP}$, Accuracy (*Acc*): $Acc = \frac{TN+TP}{TN+FP+TP+FN}$, precision (*p*): $p = \frac{TP}{TP+FP}$, harmonic mean of sensitivity and precision (F_1): $F_1 = 2\frac{SN \,.\, p}{SN + p}$ and Matthew's correlation coefficient (*MCC*): $MCC = \frac{(TP \,.\, TN) + (FP \,.\, FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$, where *TP*, *TN*, *FP* and *FN* are the number of true-positive, false-positive and false-negative classifications, respectively. For given false positive rate (α) and true positive rate ($1 - \beta$) at different threshold values, the AUC-ROC was computed as: $AUC = \sum_{n=1}^{i} \left\{(1 - \beta_i \Delta\alpha) + \frac{1}{2}[\Delta(1 - \beta)\Delta\alpha]\right\}$, where $\Delta(1-\beta)=(1-\beta_i)-(1-\beta_{i-1})$ and $\Delta\alpha=\alpha_i-\alpha_{i-1}$ and i = 1, 2, …, m (number of test data points) (Jaiswal et al., 2019). In imbalance class testing data, accuracy, sensitivity, specificity, precision and F_1 are not the best measures

to analyse performance of models as they are not based on the entire confusion matrix. MCC is a better estimator of performance in such cases as it produces a high score only if good results are obtained in all of the four confusion matrix categories. (Chicco & Jurman, 2020)

Python codes used for evaluating performance analysis:

```
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report

#y_test contains the true labels and y_pred contains the predicted labels
cm = confusion_matrix(y_test, y_pred)

#Calculating FP,TP,FN,TN
FP = cm.sum(axis=0) - np.diag(cm)
FN = cm.sum(axis=1) - np.diag(cm)
TP = np.diag(cm)
TN = cm.sum() - (FP + FN + TP)
FP = FP.astype(float)
FN = FN.astype(float)
TP = TP.astype(float)
TN = TN.astype(float)

# Sensitivity, hit rate, recall, or true positive rate
TPR = TP/(TP+FN)
# Specificity or true negative rate
TNR = TN/(TN+FP)
# Precision or positive predictive value
PPV = TP/(TP+FP)
# Negative predictive value
NPV = TN/(TN+FN)
# Fall out or false positive rate
FPR = FP/(FP+TN)
# False negative rate
FNR = FN/(TP+FN)
# False discovery rate
FDR = FP/(TP+FP)
# Overall accuracy for each class
ACC = (TP+TN)/(TP+FP+FN+TN)

#Calculating performance measures
```

```
sensitivity =TP/(TP+FN)
specificity = TN/(TN+FP)
precision = TP/(TP+FP)
recall = TP/(TP+FN)
f1_score = 2*precision*recall/(precision + recall)
MCC = ((TP*TN) + (FP*FN))/np.sqrt((TP + FP)*(TP + FN)*(TN + FP)*(TN + FN))
```

The workflow of the methods is given in Figure 4.1.

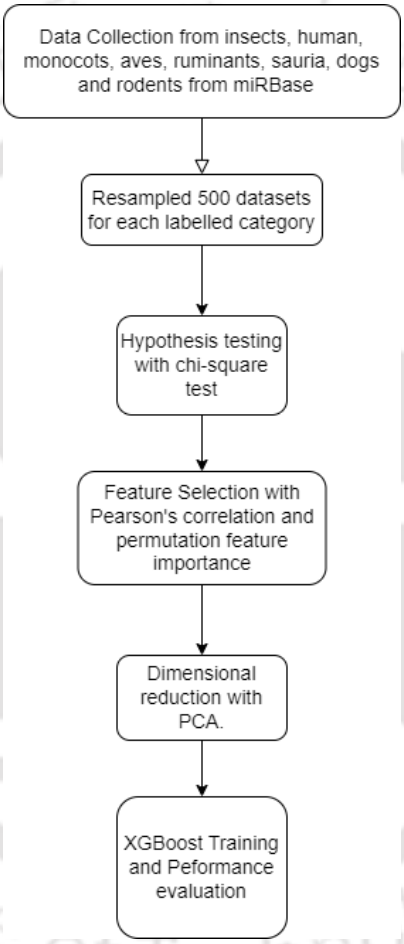

Figure 4.1. Workflow for XGBoost training.

4.3. Results:

4.3.1. Data pre-processing:

A total of 5541 sequences were collected for the analysis as given in Table 4.1.

Group Name	Name of organism	No. of Sequences	Total No. of Sequences
Aves	Anas platyrhynchos	4	1381
	Columba livia	248	
	Gallus gallus	882	
	Taeniopygja guttata	247	
Monocotyledons	Asparagus officinalis	101	910
	Oryza sativa	604	
	Sorghum bicolor	205	
Rodentia	Cavia porcellus	346	1580
	Mus musculus	1234	
Ruminantia	Bos taurus	1064	1064
Sauria	Alligator mississippiensis	303	606
	Anolis carolinensis	303	
Humans	Homo sapiens	1916	1916
Dogs	Canis familiaris	502	502
Insects	*Bombyx mori*	487	3391
	Drosophila melanogaster	258	
	Apis mellifera	254	
	Drosophila pseudoobscura	210	
	Drosophila virilis	180	
	Aedes aegypti	155	
	Drosophila simulans	148	
	Plutella xylostella	133	
	Anopheles gambiae	130	
	Acyrthosiphon pisum	123	
	Drosophila sechellia	103	

	Dinoponera quadriceps	102
	Drosophila erecta	101
	Manduca sexta	98
	Heliconius melpomene	92
	Drosophila yakuba	89
	Drosophila grimshawi	82
	Bactrocera dorsalis	80
	Drosophila willistoni	77
	Drosophila ananassae	76
	Drosophila persimilis	75
	Culex quinquefasciatus	74
	Polistes canadensis	73
	Drosophila mojavensis	71
	Nasonia vitripennis	53
	Nasonia giraulti	32
	Nasonia longicornis	28
	Locusta migratoria	7

Table 4.1: Total sequences collected for the analysis.

4.3.2. Parameter Calculation:

The parameters used for the classification is given in Table 4.2. A total of 53 parameters were calculated out of which 16 were dinucleotide counts and dinucleotide percentage counts, 4 nucleotide counts and their percentage counts, 2 base pair counts, base propensity, Shannon entropy, etc.

Sl.No.	Feature	Description
1	Len	Length of the sequence
2	A	Single nucleotide count

3	C	a change in the genome that affects just one base in the DNA.
4	G	
5	U	
6	G+C	Base pair counts a basic building block of double-stranded nucleic acids made up of two nucleobases joined by hydrogen bonds.
7	A+U	
8	AA	Dinucleotide counts consist of two nucleotides that are linked by a 3'-5' phosphodiester bond via the 3' hydroxyl group of one nucleotide and the 5' hydroxyl group of the other one, just a as bonding occurs in native RNA.
9	AC	
10	AG	
11	AU	
12	CA	
13	CC	
14	CG	
15	CU	
16	GA	
17	GC	
18	GG	
19	GU	
20	UA	
21	UC	
22	UG	
23	UU	
24	%A	Nucleotide percentage counts
25	%C	
26	%G	
27	%U	
28	%G+C	Base Pair percentage counts

29	%A+U	
30	%AA	Dinucleotide percentage counts
31	%AC	
32	%AG	
33	%AU	
34	%CA	
35	%CC	
36	%CG	
37	%CU	
38	%GA	
39	%GC	
40	%GG	
41	%GU	
42	%UA	
43	%UC	
44	%UG	
45	%UU	

46	dP	Base propensity (total base pair/length)
47	zP	Normalized pb
48	MFE	Minimum free energy of folding
49	dG	Normalized MFE
50	dQ	Shannon Entropy
51	zQ	Normalized Shannon Entropy
52	dD	Basepair distance The base pair distance between S_α and S_β is equal to $d_{BP}(S_\alpha, S_\mu) = \sum_{i<j}(\delta_{ij}^{\alpha} + \delta_{ij}^{\beta} - 2\delta_{ij}^{\alpha}\delta_{ij}^{\beta})$, where δ_{ij}^{α} is 1 if bases *i* and *j* is a base pair in S_α, and 0 otherwise (Freyhult et al., 2005).
53	zD	Normalized base pair distance
54	MFE1	$\frac{dG}{(\%G + C)}$
55	MFE2	$\frac{dG}{n\_stems}$
56	MFE3	$\frac{dG}{n\_loops}$
57	MFE4	$\frac{dG}{tot\_bases}$

Table 4.2.: Parameters calculated for feature extraction.

4.3.3. Hypothesis Testing

The chi-square value of these parameters is given in Annexure 4.1. Out of all the Length and GC content varied the most. Figure 4.2. shows the shows the comparison between few parameters such as Length, %G+C and dG that differ in insects from other organisms.

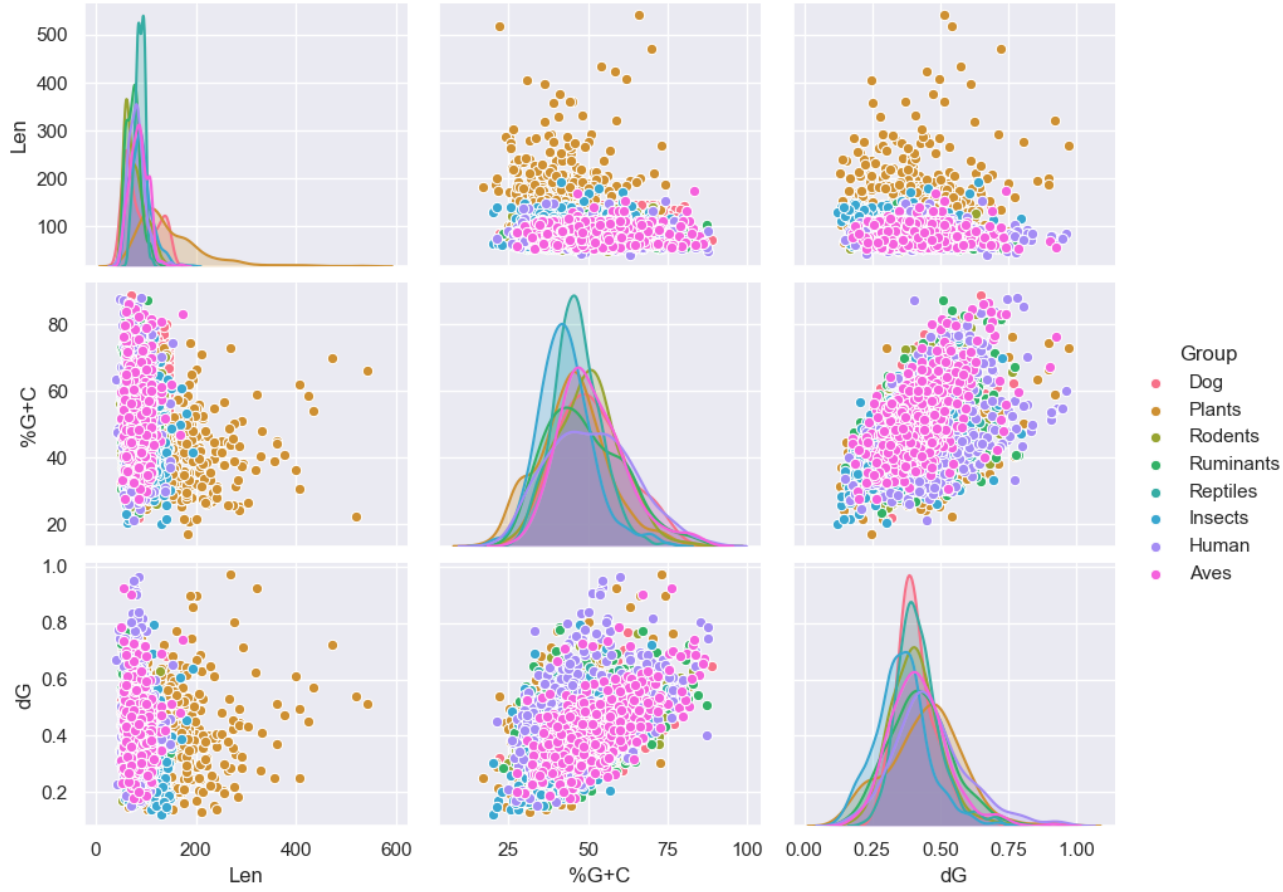

Figure 4.2: Comparison of insect pre-miRNA with other class of organisms. The comparation of various features of 500 randomly sampled pre-miRNA from insects, human, monocots, aves, ruminants, sauria, dogs and rodents are is shown in the pair-plot scatter diagrams. Features such as GC percentage (%G+C), Length (Len) and dG (MFE/Length) were considered out of 57 features mentioned below. Multivariate gaussian distribution plot is given in the diagonals.

4.3.4. Feature Engineering

The features were initially extracted based on the correlation matrix given in Annexure 4.2. The weight associated with each feature calculated from permutation feature importance is given in Table 4.3. Based on it, we only selected the features which had positive weight, i.e., from Sl No. 1 to 35. Further, we performed PCA 20%, 40%, 60%, 80% and 100% of the 35 features. We found that the performance was optimum with 20 features selected by PCA.

Sl. No.	Feature	Weight	Remark
1	Len	0.053	Retained
2	%CG	0.0355	Retained
3	%CC	0.021	Retained
4	%GC	0.0155	Retained
5	CC	0.013	Retained

6	pb	0.0125	Retained
7	%UU	0.01	Retained
8	MFE4	0.009	Retained
9	%U	0.0085	Retained
10	A+U	0.008	Retained
11	nstem	0.007	Retained
12	G	0.0065	Retained
13	%GG	0.004	Retained
14	%UC	0.004	Retained
15	A	0.0035	Retained
16	MFE	0.0035	Retained
17	%CU	0.003	Retained
18	G+C	0.003	Retained
19	%C	0.003	Retained
20	%GA	0.0025	Retained
21	UG	0.0025	Retained
22	%A	0.002	Retained
23	GU	0.0015	Retained
24	CU	0.0015	Retained
25	UU	0.0015	Retained
26	UC	0.001	Retained
27	%UG	0.001	Retained
28	total_base	0.0005	Retained
29	AC	0.0005	Retained
30	MFE3	0.0005	Retained
31	AG	0.0005	Retained
32	AA	0	Retained
33	n_stems	0	Retained
34	avg_bp	0	Retained
35	GA	0	Retained
36	AU	-0.0005	Discarded
37	U	-0.001	Discarded
38	Npb	-0.001	Discarded
39	GG	-0.0015	Discarded
40	GC	-0.002	Discarded
41	Q	-0.002	Discarded
42	C	-0.002	Discarded

43	%AA	-0.002	Discarded
44	MFE1	-0.002	Discarded
45	dG	-0.002	Discarded
46	%AC	-0.0025	Discarded
47	UA	-0.003	Discarded
48	NQ	-0.0035	Discarded
49	CA	-0.0035	Discarded
50	%G	-0.004	Discarded
51	ND	-0.004	Discarded
52	CG	-0.005	Discarded
53	%G+C	-0.005	Discarded
54	%CA	-0.005	Discarded
55	%A+U	-0.0055	Discarded
56	%AG	-0.0055	Discarded
57	%UA	-0.006	Discarded
58	%AU	-0.0065	Discarded
59	MFE2	-0.0095	Discarded
60	%GU	-0.0155	Discarded

Table 4.3: Weight associated with all the features calculated from permutation feature importance package of Scikit-Learn.

The performance of PCA for the selection of 5, 10, 15, 20, 25, 30 and 35 of 35 features is given in Figure 4.3 estimated for insects. The accuracy and sensitivity score of 5 features were found to be 0.905 and 0.629 respectively; of 10 features found to be 0.913 and 0.677; of 20 features found to be 0.912 and 0.763; of 25 features found to be 0.886 and 0.661; of 30 features found to be 0.905, and 0.645; of 35 features found to be 0.872, and 0.564

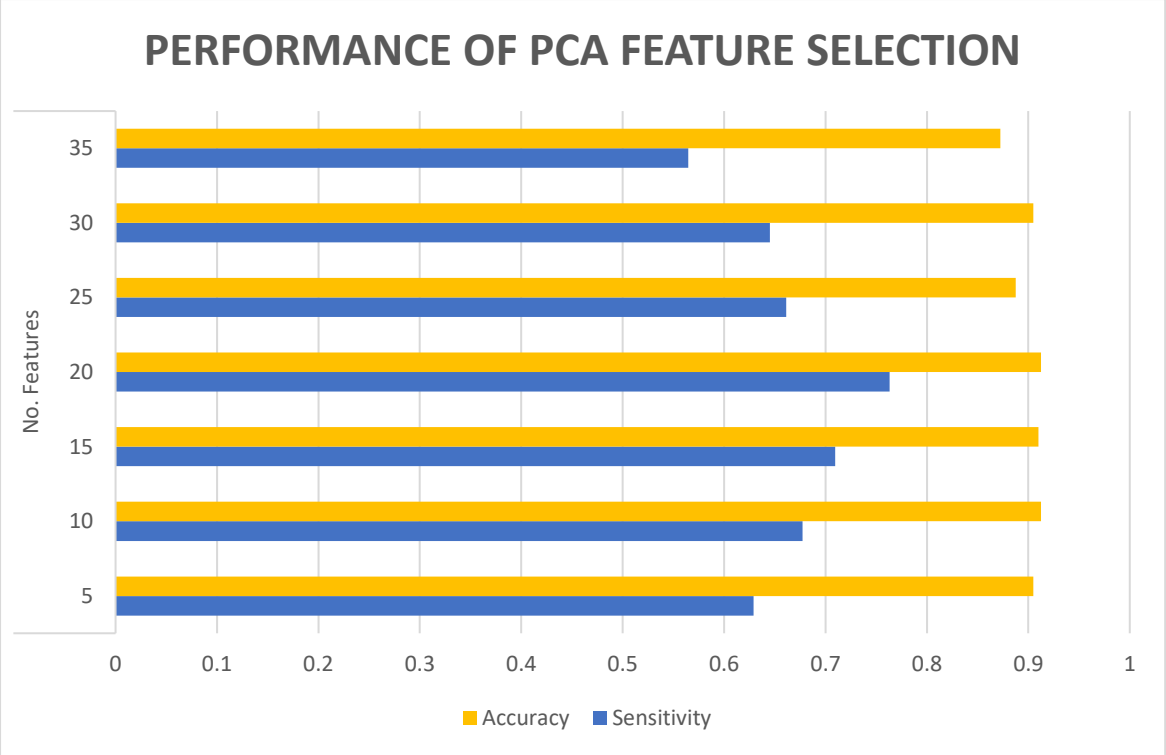

Figure 4.3: Accuracy, and Sensitivity Score of classifiers trained on 5, 10, 15, 20, 25, 30 and 35 PCA estimation features evaluated on insect pre-miRNA. Performance with 20 features was found to be most optimum.

4.3.5. Multiclass training with XGBoost

Various parameters of XGBoost was tuned with 5-fold CV and the parameter values with highest CV score were considered as given in Table 4.4. For reproducibility, we kept a random state of 42.

Parameter	Value
lambda	0.011517
alpha	0.056767
colsample_bytree	0.8
subsample	1
learning_rate	0.012
n_estimators	160
max_depth	9
min_child_weight	3

Table 4.4: Parameter values considered for XGBoost training obtained after 5 fold CV during Grid Searching.

4.3.6. Performance Evaluation:

Various performance measures for each group of organisms is given in Table 4.5. The accuracy of insect, plants, rodents, human, ruminants, sauria, aves and dogs was found to be 0.9125, 0.9325, 0.8575, 0.86, 0.8875, 0.8225, 0.845 and 0.8675 respectively. Specificity was found to be 0.928177 for Insects, 0.966197 for Plants, 0.948424 for Rodents, 0.929178 for Human, 0.948864 for Ruminants, 0.833819 for Sauria, 0.925287 for Aves and 0.93787 for Dogs. The F1 score of Insect, Plants, Rodents, Human, Ruminants, Sauria, Aves and Dogs was found to be 0.623656, 0.689655, 0.296296, 0.363636, 0.482759, 0.547771, 0.340426 and 0.530973 respectively. Sensitivity was found to be 0.763158 for Insects, 0.666667 for Plants, 0.235294 for Rodents, 0.340426 for Human, 0.4375 for Ruminants, 0.754386 for Sauria, 0.307692 for Aves and 0.483871 for Dogs. The MCC score of Insect, Plants, Rodents, Human, Ruminants, Sauria, Aves and Dogs was found to be 0.617599, 0.675554, 0.332529, 0.385423, 0.486277, 0.540746, 0.369077 and 0.527207 respectively. This has also been graphically shown in Figure 4.3.

	Insect	**Plants**	**Rodent**	**Human**	**Ruminants**	**Sauria**	**Aves**	**Dogs**
Sensitivity	0.763158	0.666667	0.235294	0.340426	0.4375	0.754386	0.307692	0.483871
Specificity	0.928177	0.966197	0.948424	0.929178	0.948864	0.833819	0.925287	0.93787
Precision	0.527273	0.714286	0.4	0.390244	0.538462	0.43	0.380952	0.588235
F1_Score	0.623656	0.689655	0.296296	0.363636	0.482759	0.547771	0.340426	0.530973
MCC	0.617599	0.675554	0.332529	0.385423	0.486277	0.540746	0.369077	0.527207
Accuracy	0.9125	0.9325	0.8575	0.86	0.8875	0.8225	0.845	0.8675

Table 4.5. Performance evaluation of the XGBoost classification.

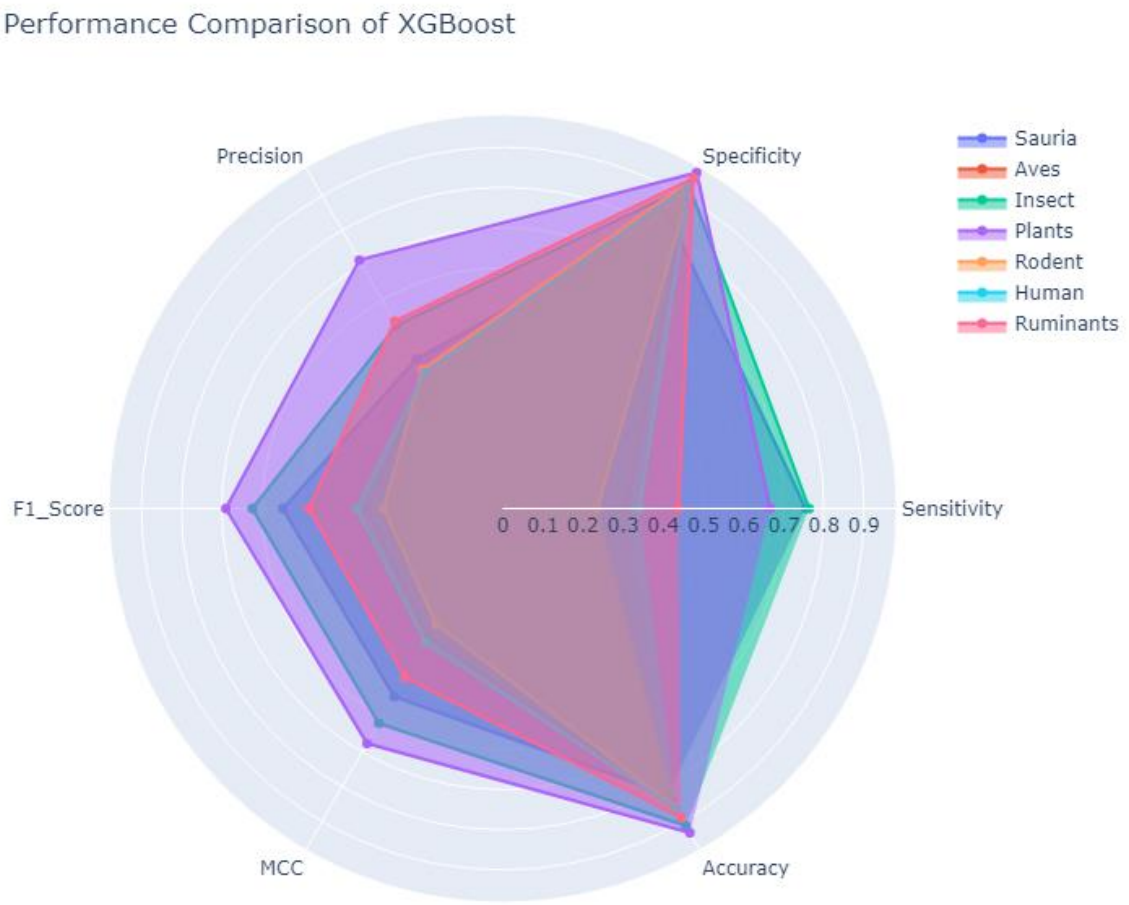

Figure 4.3. The performance comparison of XGBoost classifier.

The ROC-AUC is given in Figure 4.4. The AUC was found to be 0.934 for Insects, 0.985 for Plants (Rice), 0.941 for Sauria, 0.859 for Dog, 0.815 for Ruminant, 0.743 for Human, 0.75 for Rodent and 0.765 for Aves.

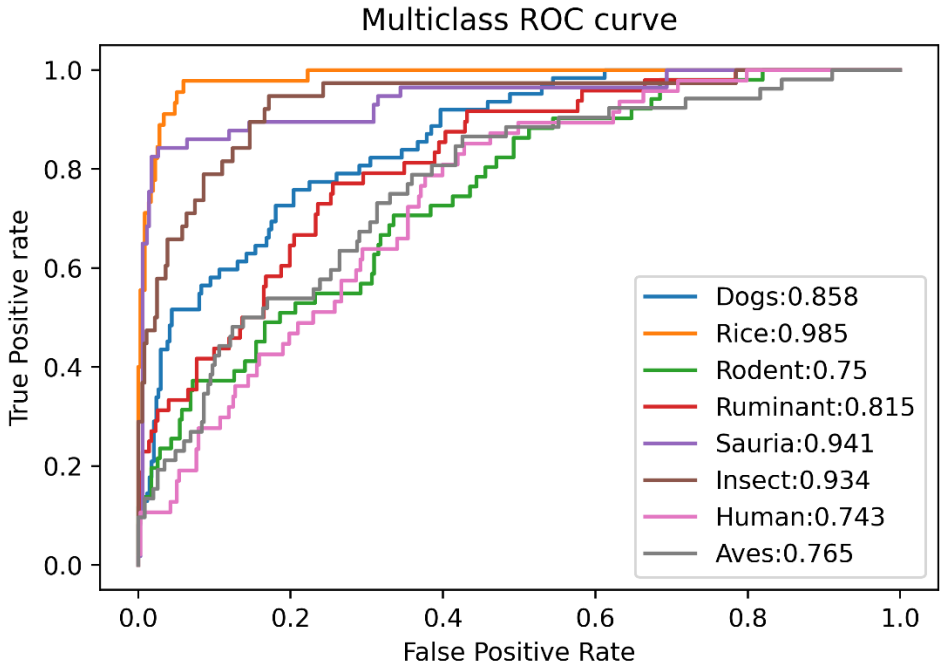

Figure 4.4: ROC-AUC of XGBoost classifier.

4.4. Discussion

4.4.1. Data collection and pre-processing:

We collected all available pre-miRNA sequences of insects as our initial focus was to distinguish them from other organisms. We also collected data from rodents, rice, aves, cattles and sauria which are reptiles. Highest number of pre-miRNA sequence from a single species was collected from humans which was followed by mouse and cattle. All the sequences formed characteristic hairpin loops which was inferred from the secondary structure calculated by RNAfold.

4.4.2. Hypothesis Testing:

Our null hypothesis was that all the pre-miRNAs are physically and compositionally similar and hence performed the comparative chi-square test of various parameters with insects. These parameters are used in various machine learning based pre-miRNA prediction tools (Batuwita & Palade, 2009; Gudyś et al., 2013; Jiang et al., 2007; Kadri et al., 2009). As the p-value of parameters such as Length, GC content, MFE1, etc were < 0.05, hence we rejected the null hypothesis and accepted the alternate hypothesis that the pre-miRNA sequences from these groups vary from insects. This indicated a possibility of multiclass supervised training for classification of pre-miRNA based on their ancestral origin. Hence, we moved forward with state-of-the-art XGBoost algorithm which can efficiently learn to classify multi group data based on given labels.

4.4.3. Feature engineering and model training:

The estimation of features which has the maximum contribution in building the model is essential (Biesiada & Duch, 2007; Saidi et al., 2019). Our approach initially gave us 35 features from correlation matrix and permutation feature importance. Length and GC content were the features with highest weight. PCA is another widely used dimension reduction technique (Kambhatla & Leen, 1997; T. Zhang & Yang, 2016). Using PCA we extracted 20 most important features for the classification model. The accuracy of 10 and 20 classifiers was highest among all and were almost the same, however, 20 classifiers had more sensitivity than any other classifier for which it was used for the classification.

To reduce bias, we sampled 500 datasets from each group before training the models. The model showed best performance with number of estimators 150-200 hence, we adjusted it to 160. Learning rate is a crucial parameter in ensemble learning which we adjusted to 0.012.

4.4.4. Performance evaluation:

The performance of XGBoost varied among the groups. Each group had fairly good accuracy however accuracy is often misleading and can not be considered as best indicator of performance (Chicco & Jurman, 2020; Sokolova et al., 2006). The predictive model had good specificity for each group. However, only insects, plants, and saurias had sensitivity more than 60%, with insects having the highest at 76.32%. MCC score which estimates all the parameter is a crucial indicator (Chicco & Jurman, 2020). The MCC and F1 score of only Plants and Insects was above 60%.

The AUC values of 0.934, 0.985 and 0.941 of insects, plants (rice), and reptiles (sauria) respectively, suggest that the parameters we used in XGBoost is capable of classifying these organisms.

4.5. Conclusion:

In this chapter we showed the peculiarity of insect pre-miRNA by comparing it with other organisms and established that their pre-miRNA are significantly different from other organisms such as plants, humans, rodents, ruminants, sauria, dogs and birds. We further developed a predictive model that was trained using XGBoost classifier that learned to differentiate the pre-miRNA between these classes of organism. In future, this model can be implemented in a web-server so that it can be used as an online tool.

4.6. Annexures

Annexure 4.1: Chi square test result and p-value of insects with other organisms

Table Title	Legend
1: Insect and Human pre-miRNA	Chi-square test score and associated p-values of various parameters between insect and human.
2: Insect and Rice pre-miRNA	Chi-square test score and associated p-values of various parameters between insect and rice.
3: Insect and Rodent pre-miRNA	Chi-square test score and associated p-values of various parameters between insect and rodent.
4: Insect and Cattle pre-miRNA	Chi-square test score and associated p-values of various parameters between insect and cattle.
5: Insect and Aves pre-miRNA	Chi-square test score and associated p-values of various parameters between insect and aves.

Chi-square value of various parameters between insect and other organisms.

Table 1-5 contains the chi-square test score and p-value for 61 features of pre-miRNA from insect and the other selected organisms. The *p-value* obtained from chi-square test suggested that insect pre-miRNA differs from Aves, Plants, Rodentia, Cattle and Humans. The %G+C of insect pre-miRNA is lowest among all the organisms that was compared. The length of pre-miRNA varied among different groups. The pre-miRNA of plants was found to be significantly longer (p-value < 0.01) than the other groups. Insects had shorter pre-miRNA than plants but longer than humans. The normalized MFE and Shannon entropy of insect pre-miRNA also varied from other groups of organisms.

1. Insect and Human pre-miRNA:

Sl. No.	Parameters	Chi-Square_Value	p-value	Sl. No.	Parameters	Chi-Square_Value	p-value
1	Len	918.3214141	1.02E-201	32	%AG	48.93155	2.65E-12
2	A	651.7312622	9.39E-144	33	%AU	562.6417	2.24E-124
3	C	9.838195879	0.001709251	34	%CA	3.254882	0.071211072
4	G	2.664868515	0.102586302	35	%CC	456.6692	2.55E-101
5	U	832.2972211	5.13E-183	36	%CG	139.0578	4.28E-32
6	G+C	11.12231632	0.000852952	37	%CU	63.34503	1.73E-15
7	A+U	1483.717795	0	38	%GA	2.564301	0.109300872
8	AA	170.4881999	5.79E-39	39	%GC	49.33994	2.15E-12
9	AC	67.15952315	2.50E-16	40	%GG	560.143	7.83E-124
10	AG	0.586444529	0.4437973	41	%GU	16.41303	5.09E-05
11	AU	827.9536001	4.51E-182	42	%UA	592.9923	5.60E-131
12	CA	26.75923317	2.30E-07	43	%UC	14.85345	0.000116195
13	CC	163.3947705	2.05E-37	44	%UG	28.23959	1.07E-07
14	CG	310.769907	1.48E-69	45	%UU	70.55004	4.49E-17
15	CU	0.036742262	0.847990749	46	pb	207.263	5.43E-47
16	GA	37.20174619	1.07E-09	47	Npb	0.122915	0.725894086
17	GC	0.890403448	0.345368001	48	mfe	3.836306	0.050153913
18	GG	189.1930252	4.77E-43	49	dG	2.977678	0.0844203
19	GU	125.0896939	4.86E-29	50	Q	89.08964	3.77E-21
20	UA	809.6357423	4.34E-178	51	NQ	0.127382	0.721161793

21	UC	108.3243821	2.28E-25	52	D	20.68904	5.40E-06
22	UG	12.46582748	0.000414465	53	ND	0.104828	0.746110736
23	UU	332.6416557	2.56E-74	54	nstem	843.8068	1.61E-185
24	%A	121.5987042	2.83E-28	55	MFE1	0.001372	0.970449495
25	%C	132.2983058	1.29E-30	56	MFE2	4.447836	0.034945495
26	%G	215.8669999	7.21E-49	57	MFE3	3.289021	0.069744578
27	%U	187.7672144	9.77E-43	58	MFE4	0.344789	0.557078147
28	%G+C	345.5943402	3.86E-77	59	total_base	207.263	5.43E-47
29	%A+U	308.212792	5.35E-69	60	n_stems	45.13261	1.84E-11
30	%AA	12.77394878	0.00035148	61	avg_bp	16.71504	4.34E-05
31	%AC	3.435804718	0.063797498				

Table 1: Chi-square test score and associated p-values of various parameters between insect and human.

2. Insect and Rice pre-miRNA:

Sl. No.	Parameter	Chi-Square_Value	p-value	Sl. No.	Parameter	Chi-Square_Value	p-value
1	Len	8518.613758	0	31	%AC	18.98931805	1.31E-05
2	A	2307.229605	0	32	%AG	16.28015807	5.46E-05
3	C	2066.64927	0	33	%AU	4.771120517	0.028940984
4	G	2129.1967	0	34	%CA	1.143615812	0.284889583
5	U	2058.687084	0	35	%CC	104.7864948	1.36E-24
6	G+C	4194.450049	0	36	%CG	12.28226568	0.000457283
7	A+U	4348.107302	0	37	%CU	17.77485732	2.49E-05
8	AA	702.2932784	9.48E-155	38	%GA	35.85843939	2.12E-09
9	AC	260.5913467	1.28E-58	39	%GC	30.00024056	4.32E-08
10	AG	697.189871	1.22E-153	40	%GG	71.75141101	2.44E-17
11	AU	701.7485559	1.25E-154	41	%GU	73.63160367	9.41E-18
12	CA	573.9178457	7.89E-127	42	%UA	89.12702049	3.70E-21
13	CC	735.3842893	6.05E-162	43	%UC	1.112525873	0.291533538
14	CG	145.0698359	2.07E-33	44	%UG	0.472404924	0.491882867
15	CU	725.5236512	8.42E-160	45	%UU	69.72440372	6.82E-17
16	GA	805.0788354	4.24E-177	46	pb	3722.559444	0
17	GC	555.5655847	7.75E-123	47	Npb	0.388650958	0.533009369
18	GG	716.1834163	9.05E-158	48	mfe	6251.447599	0
19	GU	202.5474056	5.81E-46	49	dG	4.909641721	0.02670718
20	UA	309.8333754	2.37E-69	50	Q	2693.255826	0
21	UC	593.5032465	4.33E-131	51	NQ	2.424620481	0.119442166
22	UG	645.5352737	2.09E-142	52	D	839.7556706	1.23E-184
23	UU	541.5159737	8.82E-120	53	ND	0.613197441	0.433586451
24	%A	3.859700074	0.049459169	54	nstem	8906.705803	0
25	%C	36.04177866	1.93E-09	55	MFE1	0.059657085	0.807038754
26	%G	27.36591139	1.68E-07	56	MFE2	0.470563654	0.492727975
27	%U	61.05506349	5.55E-15	57	MFE3	0.882888349	0.347411803
28	%G+C	62.83683004	2.25E-15	58	MFE4	0.209444435	0.647203199
29	%A+U	50.0354342	1.51E-12	59	total_base	3722.559444	0
30	%AA	2.255848395	0.133110487	60	n_stems	274.1005785	1.45E-61

				61	avg_bp	40.60739539	1.86E-10

Table 2: Chi-square test score and associated p-values of various parameters between insect and rice.

3. Insect and Rodent pre-miRNA:

Sl. No.	Parameters	Chi-Square_Value	p-value	Sl. No.	Parameters	Chi-Square_Value	p-value
1	Len	806.9565	1.66E-177	32	%AG	145.6287918	1.56E-33
2	A	651.7669	9.22E-144	33	%AU	420.0716086	2.35E-93
3	C	0.183035	0.668777571	34	%CA	62.7757414	2.32E-15
4	G	0.047975	0.826625541	35	%CC	360.601748	2.08E-80
5	U	890.1575	1.35E-195	36	%CG	467.0227331	1.42E-103
6	G+C	0.017765	0.893968877	37	%CU	202.9734995	4.69E-46
7	A+U	1540.251	0	38	%GA	1.527130271	0.216543651
8	AA	670.0491	9.75E-148	39	%GC	95.36211331	1.59E-22
9	AC	24.91625	5.99E-07	40	%GG	405.6588057	3.23E-90
10	AG	18.99368	1.31E-05	41	%GU	0.835827547	0.360592828
11	AU	651.2702	1.18E-143	42	%UA	303.4151586	5.94E-68
12	CA	0.7839	0.375950971	43	%UC	1.700764543	0.192188022
13	CC	118.9652	1.07E-27	44	%UG	105.2366888	1.08E-24
14	CG	719.9764	1.35E-158	45	%UU	673.5873469	1.66E-148
15	CU	34.42928	4.42E-09	46	pb	240.9011341	2.50E-54
16	GA	34.36372	4.57E-09	47	Npb	0.006534785	0.935570776
17	GC	3.277291	0.070244805	48	mfe	17.6357033	2.68E-05
18	GG	127.9262	1.16E-29	49	dG	2.021552023	0.155080563
19	GU	61.58661	4.24E-15	50	Q	6.583205734	0.010294533
20	UA	492.7646	3.57E-109	51	NQ	0.53523388	0.464414701
21	UC	71.75707	2.43E-17	52	D	0.908900553	0.340406303
22	UG	2.983826	0.084100235	53	ND	0.252538598	0.615293218
23	UU	1092.409	1.47E-239	54	nstem	807.2602825	1.42E-177
24	%A	159.6652	1.34E-36	55	MFE1	0.00373182	0.951288632

25	%C	198.4418	4.57E-45	56	MFE2	2.58736077	0.107719695
26	%G	241.6033	1.76E-54	57	MFE3	2.268204426	0.13205282
27	%U	249.856	2.79E-56	58	MFE4	0.382868262	0.536072174
28	%G+C	439.9476	1.11E-97	59	total_base	240.9011341	2.50E-54
29	%A+U	407.7895	1.11E-90	60	n_stems	27.81702522	1.33E-07
30	%AA	374.0611	2.44E-83	61	avg_bp	0.126086186	0.722525039
31	%AC	1.449584	0.228594643				

Table 3: Chi-square test score and associated p-values of various parameters between insect and rodent.

4. Insect and Cattle pre-miRNA:

Sl. No.	Parameter	Chi-Square_Value	p-value	Sl. No.	Parameter	Chi-Square_Value	p-value
1	Len	918.3214141	1.02E-201	32	%AG	48.93155	2.65E-12
2	A	651.7312622	9.39E-144	33	%AU	562.6417	2.24E-124
3	C	9.838195879	0.001709251	34	%CA	3.254882	0.071211072
4	G	2.664868515	0.102586302	35	%CC	456.6692	2.55E-101
5	U	832.2972211	5.13E-183	36	%CG	139.0578	4.28E-32
6	G+C	11.12231632	0.000852952	37	%CU	63.34503	1.73E-15
7	A+U	1483.717795	0	38	%GA	2.564301	0.109300872
8	AA	170.4881999	5.79E-39	39	%GC	49.33994	2.15E-12
9	AC	67.15952315	2.50E-16	40	%GG	560.143	7.83E-124
10	AG	0.586444529	0.4437973	41	%GU	16.41303	5.09E-05
11	AU	827.9536001	4.51E-182	42	%UA	592.9923	5.60E-131
12	CA	26.75923317	2.30E-07	43	%UC	14.85345	0.000116195
13	CC	163.3947705	2.05E-37	44	%UG	28.23959	1.07E-07
14	CG	310.769907	1.48E-69	45	%UU	70.55004	4.49E-17
15	CU	0.036742262	0.847990749	46	pb	207.263	5.43E-47
16	GA	37.20174619	1.07E-09	47	Npb	0.122915	0.725894086
17	GC	0.890403448	0.345368001	48	mfe	3.836306	0.050153913
18	GG	189.1930252	4.77E-43	49	dG	2.977678	0.0844203

19	GU	125.0896939	4.86E-29	50	Q	89.08964	3.77E-21
20	UA	809.6357423	4.34E-178	51	NQ	0.127382	0.721161793
21	UC	108.3243821	2.28E-25	52	D	20.68904	5.40E-06
22	UG	12.46582748	0.000414465	53	ND	0.104828	0.746110736
23	UU	332.6416557	2.56E-74	54	nstem	843.8068	1.61E-185
24	%A	121.5987042	2.83E-28	55	MFE1	0.001372	0.970449495
25	%C	132.2983058	1.29E-30	56	MFE2	4.447836	0.034945495
26	%G	215.8669999	7.21E-49	57	MFE3	3.289021	0.069744578
27	%U	187.7672144	9.77E-43	58	MFE4	0.344789	0.557078147
28	%G+C	345.5943402	3.86E-77	59	total_base	207.263	5.43E-47
29	%A+U	308.212792	5.35E-69	60	n_stems	45.13261	1.84E-11
30	%AA	12.77394878	0.00035148	61	avg_bp	16.71504	4.34E-05
31	%AC	3.435804718	0.063797498				

Table 4: Chi-square test score and associated p-values of various parameters between insect and cattle.

5. Insect and Aves pre-miRNA:

Sl. No.	Parameters	Chi-Square_Value	p-value	Sl. No.	Parameters	Chi-Square_Value	p-value
1	Len	61.76627219	3.87E-15	32	%AG	144.4826	2.79E-33
2	A	381.6699529	5.39E-85	33	%AU	579.0832	5.94E-128
3	C	139.4930765	3.44E-32	34	%CA	53.37173	2.76E-13
4	G	187.4562286	1.14E-42	35	%CC	356.5039	1.63E-79
5	U	450.262974	6.32E-100	36	%CG	163.7894	1.68E-37
6	G+C	326.38943	5.88E-73	37	%CU	219.8872	9.57E-50
7	A+U	831.8917453	6.29E-183	38	%GA	0.475384	0.490520867
8	AA	443.1704029	2.21E-98	39	%GC	388.4987	1.76E-86
9	AC	23.22547528	1.44E-06	40	%GG	436.8518	5.24E-97
10	AG	95.0952701	1.81E-22	41	%GU	18.29667	1.89E-05
11	AU	584.4577038	4.02E-129	42	%UA	588.3108	5.84E-130
12	CA	26.69141691	2.39E-07	43	%UC	12.62372	0.000380883
13	CC	232.7605994	1.49E-52	44	%UG	124.1619	7.76E-29

14	CG	220.1993829	8.18E-50	45	%UU	643.0437	7.28E-142
15	CU	154.7027339	1.63E-35	46	pb	15.13706	1.00E-04
16	GA	0.7905063	0.373947076	47	Npb	0.002045	0.963934201
17	GC	251.5284852	1.21E-56	48	mfe	161.0373	6.71E-37
18	GG	295.5107366	3.13E-66	49	dG	3.31838	0.068509312
19	GU	34.12439035	5.17E-09	50	Q	0.874532	0.3497037
20	UA	602.3266215	5.22E-133	51	NQ	0.122628	0.726201607
21	UC	20.99850791	4.60E-06	52	D	0.065907	0.797392103
22	UG	71.37481663	2.95E-17	53	ND	0.078789	0.778945017
23	UU	681.8512523	2.64E-150	54	nstem	89.02687	3.89E-21
24	%A	275.6314329	6.72E-62	55	MFE1	0.005797	0.939306717
25	%C	282.5181789	2.12E-63	56	MFE2	0.886497	0.346428226
26	%G	359.6920352	3.29E-80	57	MFE3	2.485008	0.114935762
27	%U	347.0449475	1.86E-77	58	MFE4	0.16074	0.688475856
28	%G+C	641.7339162	1.40E-141	59	total_base	15.13706	1.00E-04
29	%A+U	622.6445572	1.99E-137	60	n_stems	1.38889	0.238592666
30	%AA	376.4379186	7.42E-84	61	avg_bp	0.518863	0.471326912
31	%AC	10.67161946	0.001087918				

Table 5: Chi-square test score and associated p-values of various parameters between insect and aves.

Annexure 4.2. Correlation matrix of all the features

	Len	A	C	G	U	G+C	A+U	AA	AC	AG	AU	CA	CC	CG	CU	GA	GC	GG	GU	UA	UC	UG	UU	%A	%C	%G	%U	%G+C	%A+U	%AA	%AC	%AG	%AU
Len	1	0.84 2693	0.77 3223	0.78 667	0.85 7063	0.81 5826	0.88 0224	0.60 6109	0.60 9466	0.67 9802	0.67 6074	0.67 3243	0.43 0497	0.39 9813	0.64 5105	0.73 0167	0.47 5787	0.44 5246	0.62 7139	0.57 3288	0.69 7544	0.67 9181	0.60 8799	0.09 1848	0.03 597	0.08 243	0.02 408	0.06 784	0.06 7836	0.07 3738	0.04 563	0.04 658	0.14 2622
A	0.84 2693	1	0.40 1339	0.39 4887	0.86 5012	0.41 6221	0.96 4367	0.86 512	0.66 5239	0.56 9978	0.83 8121	0.65 0467	0.06 2242	0.07 1578	0.38 3568	0.64 9063	0.08 0727	0.04 0433	0.50 6745	0.78 1318	0.49 0208	0.49 0926	0.70 97	0.56 6167	0.37 709	0.45 198	0.24 5089	0.47 592	0.47 592	0.45 0583	0.12 481	0.02 224	0.42 4447
C	0.77 3223	0.40 1339	1	0.82 8733	0.40 1684	0.95 3527	0.41 5786	0.16 262	0.46 095	0.54 5403	0.20 9582	0.57 562	0.81 2881	0.63 9673	0.67 8413	0.52 1836	0.75 6735	0.61 8457	0.34 9747	0.09 7207	0.66 5898	0.46 9932	0.12 5011	0.34 818	0.55 8062	0.25 0014	0.43 949	0.46 5245	0.46 525	0.27 731	0.03 943	0.02 453	0.25 954
G	0.78 667	0.39 4887	0.82 8733	1	0.43 915	0.95 8844	0.43 2263	0.14 1687	0.32 3259	0.65 5655	0.21 7189	0.44 4433	0.59 6174	0.62 9951	0.58 1627	0.65 0975	0.75 2724	0.81 7449	0.50 7886	0.10 7149	0.54 1549	0.61 7333	0.17 4407	0.36 017	0.26 5051	0.48 7251	0.37 103	0.43 1309	0.43 131	0.29 359	0.20 227	0.08 7618	0.24 029
U	0.85 7063	0.86 5012	0.40 1684	0.43 915	1	0.44 0227	0.96 6937	0.70 0344	0.51 5183	0.46 8529	0.83 1654	0.52 4792	0.06 3874	0.07 3156	0.50 324	0.56 2852	0.09 2324	0.09 6352	0.65 3003	0.76 5247	0.59 8032	0.63 4603	0.86 0337	0.30 0913	0.40 496	0.41 227	0.49 2418	0.46 937	0.46 9367	0.24 7353	0.05 381	0.16 953	0.41 6298
G+C	0.81 5826	0.41 6221	0.95 3527	0.95 8844	0.44 0227	1	0.44 3673	0.15 8787	0.40 79	0.62 9706	0.22 3263	0.53 1317	0.73 3386	0.66 3694	0.65 7332	0.61 523	0.78 9185	0.75 3878	0.45 0883	0.10 7006	0.62 9405	0.57 0784	0.15 7323	0.37 056	0.42 5831	0.38 9175	0.42 273	0.46 8253	0.46 825	0.29 876	0.12 891	0.03 4729	0.26 104
A+U	0.88 0224	0.96 4367	0.41 5786	0.43 2263	0.96 6937	0.44 3673	1	0.80 8985	0.60 9762	0.53 6745	0.86 4502	0.60 732	0.06 5315	0.07 4954	0.46 0304	0.62 6676	0.08 9711	0.07 1356	0.60 1878	0.80 0616	0.56 4487	0.58 4135	0.81 4356	0.44 6425	0.40 519	0.44 711	0.38 4215	0.48 938	0.48 9381	0.35 9438	0.03 5063	0.10 069	0.43 5237
AA	0.60 6109	0.86 512	0.16 262	0.14 1687	0.70 0344	0.15 8787	0.80 8985	1	0.44 564	0.31 6088	0.65 3717	0.37 9157	0.05 602	0.01 743	0.19 2535	0.41 8851	0.08 603	0.09 774	0.31 2312	0.63 0002	0.30 8581	0.25 8007	0.71 9128	0.64 2025	0.43 983	0.51 941	0.29 6308	0.55 067	0.55 067	0.76 6715	0.03 3072	0.12 462	0.34 0005
AC	0.60 9466	0.66 5239	0.46 095	0.32 3259	0.51 5183	0.40 79	0.60 9762	0.44 564	1	0.28 7369	0.44 3971	0.67 085	0.14 4689	0.19 1001	0.30 1562	0.40 1414	0.08 7015	0.06 1671	0.48 0049	0.51 0537	0.27 7319	0.39 5642	0.33 0701	0.31 743	0.02 464	0.28 693	0.01 122	0.17 801	0.17 8006	0.10 5226	0.72 1281	0.18 728	0.11 1195
AG	0.67 9802	0.56 9978	0.54 5403	0.65 5655	0.46 8529	0.62 9706	0.53 6745	0.31 6088	0.28 7369	1	0.25 5941	0.49 4605	0.30 2995	0.13 0618	0.59 8409	0.71 9429	0.42 3549	0.37 4597	0.35 4729	0.29 2816	0.50 0766	0.37 6285	0.27 6835	0.08 6765	0.00 836	0.11 4473	0.18 575	0.06 0505	0.06 051	0.05 006	0.18 234	0.66 1448	0.16 422
AU	0.67 6074	0.83 8121	0.20 9582	0.21 7189	0.83 1654	0.22 3263	0.86 4502	0.65 3717	0.44 3971	0.25 5941	1	0.51 7499	0.06 861	0.00 455	0.18 5277	0.47 301	0.04 323	0.08 305	0.43 4328	0.81 4254	0.41 3271	0.48 4206	0.65 563	0.49 9865	0.46 179	0.50 286	0.44 0181	0.55 392	0.55 392	0.29 4657	0.00 513	0.27 288	0.76 355
CA	0.67 3243	0.65 0467	0.57 562	0.44 4433	0.52 4792	0.53 1317	0.60 732	0.37 9157	0.67 085	0.49 4605	0.51 7499	1	0.23 2149	0.06 1811	0.35 1162	0.38 5278	0.35 0108	0.14 1026	0.43 6904	0.31 0361	0.44 983	0.58 141	0.28 2376	0.22 0751	0.05 5671	0.19 449	0.08 304	0.07 883	0.07 8832	0.01 0581	0.27 2333	0.02 0597	0.17 8215
CC	0.43 0497	0.06 2242	0.81 2881	0.59 6174	0.06 3874	0.73 3386	0.06 5315	0.05 602	0.14 4689	0.30 2995	0.06 861	0.23 2149	1	0.49 596	0.40 558	0.26 7945	0.55 8966	0.61 1725	0.09 3715	0.13 516	0.38 2278	0.17 9552	0.08 916	0.47 807	0.69 6197	0.34 8939	0.54 11	0.60 1551	0.60 155	0.31 638	0.15 317	0.01 979	0.37 276
CG	0.39 9813	0.07 1578	0.63 9673	0.62 9951	0.07 3156	0.66 3694	0.07 4954	0.01 743	0.19 1001	0.13 0618	0.00 455	0.06 1811	0.49 596	1	0.13 2521	0.30 2035	0.65 9198	0.49 1773	0.21 6146	0.05 164	0.31 5141	0.08 0748	0.03 053	0.39 385	0.45 0221	0.40 671	0.44 039	0.49 2361	0.49 236	0.22 861	0.06 295	0.22 169	0.23 275
CU	0.64 5105	0.38 3568	0.67 8413	0.58 1627	0.50 324	0.65 7332	0.46 0304	0.19 2535	0.30 1562	0.59 8409	0.18 5277	0.35 1162	0.40 558	0.13 2521	1	0.49 5706	0.47 1146	0.39 8405	0.24 3311	0.20 1231	0.69 7155	0.49 2977	0.23 1846	0.22 513	0.24 0064	0.05 7434	0.06 781	0.17 1522	0.17 152	0.18 584	0.13 611	0.16 9049	0.22 682
GA	0.73 0167	0.64 9063	0.52 1836	0.65 0975	0.56 2852	0.61 523	0.62 6676	0.41 8851	0.40 1414	0.71 9429	0.47 301	0.38 5278	0.26 7945	0.30 2035	0.49 5706	1	0.24 7446	0.36 1369	0.30 7214	0.26 9424	0.60 6223	0.45 719	0.38 8529	0.16 0728	0.10 608	0.03 8513	0.09 07	0.03 933	0.03 9327	0.04 6917	0.06 966	0.23 6785	0.09 7061
GC	0.47 5787	0.08 0727	0.75 6735	0.75 2724	0.09 2324	0.78 9185	0.08 9711	0.08 603	0.08 7015	0.42 3549	0.04 323	0.35 0108	0.55 8966	0.65 9198	0.47 1146	0.24 7446	1	0.56 564	0.12 8447	0.12 66	0.24 8673	0.40 3919	0.10 508	0.47 204	0.52 7676	0.49 3291	0.52 185	0.58 6536	0.58 654	0.34 872	0.24 949	0.08 1814	0.34 298
GG	0.44 5246	0.04 0433	0.61 8457	0.81 7449	0.09 6352	0.75 3878	0.07 1356	0.09 774	0.06 1671	0.37 4597	0.08 305	0.14 1026	0.61 1725	0.49 1773	0.39 8405	0.36 1369	0.56 564	1	0.17 0795	0.14 878	0.33 8256	0.26 8193	0.04 051	0.52 807	0.37 0052	0.67 096	0.48 455	0.59 6857	0.59 686	0.36 319	0.25 507	0.03 7257	0.38 641

	0.62	0.50	0.34	0.50	0.65	0.45	0.60	0.31	0.48	0.35	0.43	0.43	0.09	0.21	0.24	0.30	0.12	0.17		0.48	0.38	0.69	0.39	0.02	0.24	0.04	0.25	0.17	0.17	0.03	0.09	0.12	0.11
GU	7139	6745	9747	7886	3003	0883	1878	2312	0049	4729	4328	6904	3715	6146	3311	7214	8447	0795	1	265	3401	0466	8279	6291	883	682	9779	054	0539	83	81	757	9537
	0.57	0.78	0.09	0.10	0.76	0.10	0.80	0.63	0.51	0.29	0.81	0.31	0.13	0.05	0.20	0.26	0.12	0.14	0.48		0.20	0.27	0.62	0.52	0.49	0.52	0.46	0.58	0.58	0.30	0.14	0.11	0.56
UA	3288	1318	7207	7149	5247	7006	0616	0002	0537	2816	4254	0361	516	164	1231	9424	66	878	265	1	4467	7776	7221	2963	262	386	7549	372	3719	6214	5592	555	9616
	0.69	0.49	0.66	0.54	0.59	0.62	0.56	0.30	0.27	0.50	0.41	0.44	0.38	0.31	0.69	0.60	0.24	0.33	0.38	0.20		0.35	0.35	0.11	0.15	0.07	0.02	0.05	0.05	0.08	0.20	0.00	0.03
UC	7544	0208	5898	1549	8032	9405	4487	8581	7319	0766	3271	983	2278	5141	7155	6223	8673	8256	3401	4467	1	9051	3788	188	8674	2	4264	0605	06	213	97	324	0132
	0.67	0.49	0.46	0.61	0.63	0.57	0.58	0.25	0.39	0.37	0.48	0.58	0.17	0.08	0.49	0.45	0.40	0.26	0.69	0.27	0.35		0.32	0.05	0.13	0.04	0.14	0.05	0.05	0.11	0.03	0.16	0.16
UG	9181	0926	9932	7333	4603	0784	4135	8007	5642	6285	4206	141	9552	0748	2977	719	3919	8193	0466	7776	9051	1	7133	499	959	4845	5722	508	5083	094	213	707	9919
	0.60	0.70	0.12	0.17	0.86	0.15	0.81	0.71	0.33	0.27	0.65	0.28	0.08	0.03	0.23	0.38	0.10	0.04	0.39	0.62	0.35	0.32		0.36	0.51	0.49	0.61	0.57	0.57	0.43	0.08	0.17	0.34
UU	8799	97	5011	4407	0337	7323	4356	9128	0701	6835	563	2376	916	053	1846	8529	508	051	8279	7221	3788	7133	1	444	2	33	1482	748	7484	1861	627	733	8625
	0.09	0.56	0.34	0.36	0.30	0.37	0.44	0.64	0.31	0.08	0.49	0.22	0.47	0.39	0.22	0.16	0.47	0.52	0.02	0.52	0.11	0.05	0.36		0.69	0.77	0.43	0.84	0.84	0.79	0.32	0.04	0.62
%A	1848	6167	818	017	0913	056	6425	2025	743	6765	9865	0751	807	385	513	0728	204	807	6291	2963	188	499	444	1	685	193	7486	335	3347	1353	645	7088	1195
	0.03	0.37	0.55	0.26	0.40	0.42	0.40	0.43	0.02	0.00	0.46	0.05	0.69	0.45	0.24	0.10	0.52	0.37	0.24	0.49	0.15	0.13	0.51	0.69		0.51	0.78	0.87	0.87	0.55	0.01	0.02	0.62
%C	597	709	8062	5051	496	5831	519	983	464	836	179	5671	6197	0221	0064	608	7676	0052	883	262	8674	959	2	685	1	5574	103	2241	224	772	1319	1146	574
	0.08	0.45	0.25	0.48	0.41	0.38	0.44	0.51	0.28	0.11	0.50	0.19	0.34	0.40	0.05	0.03	0.49	0.67	0.04	0.52	0.07	0.04	0.49	0.77	0.51		0.70	0.86	0.86	0.62	0.29	0.22	0.63
%G	243	198	0014	7251	227	9175	711	941	693	4473	286	449	8939	671	7434	8513	3291	096	682	386	2	4845	33	193	5574	1	211	8767	877	231	949	2056	38
	0.02	0.24	0.43	0.37	0.49	0.42	0.38	0.29	0.01	0.18	0.44	0.08	0.54	0.44	0.06	0.09	0.52	0.48	0.25	0.46	0.02	0.14	0.61	0.43	0.78	0.70		0.85	0.85	0.36	0.04	0.27	0.60
%U	408	5089	949	103	2418	273	4215	6308	122	575	0181	304	11	039	781	07	185	455	9779	7549	4264	5722	1482	7486	103	211	1	217	2169	319	279	84	5669
	0.06	0.47	0.46	0.43	0.46	0.46	0.48	0.55	0.17	0.06	0.55	0.07	0.60	0.49	0.17	0.03	0.58	0.59	0.17	0.58	0.05	0.05	0.57	0.84	0.87	0.86	0.85			0.67	0.16	0.13	0.72
%G+C	784	592	5245	1309	937	8253	938	067	801	0505	392	883	1551	2361	1522	933	6536	6857	054	372	0605	508	748	335	2241	8767	217	1	-1	753	439	896	342
	0.06	0.47	0.46	0.43	0.46	0.46	0.48	0.55	0.17	0.06	0.55	0.07	0.60	0.49	0.17	0.03	0.58	0.59	0.17	0.58	0.05	0.05	0.57	0.84	0.87	0.86	0.85			0.67	0.16	0.13	0.72
%A+U	7836	592	525	131	9367	825	9381	067	8006	051	392	8832	155	236	152	9327	654	686	0539	3719	06	5083	7484	3347	224	877	2169	-1	1	7533	4392	896	3417
	0.07	0.45	0.27	0.29	0.24	0.29	0.35	0.76	0.10	0.05	0.29	0.01	0.31	0.22	0.18	0.04	0.34	0.36	0.03	0.30	0.08	0.11	0.43	0.79	0.55	0.62	0.36	0.67	0.67		0.05	0.13	0.33
%AA	3738	0583	731	359	7353	876	9438	6715	5226	006	4657	0581	638	861	584	6917	872	319	83	6214	213	094	1861	1353	772	231	319	753	7533	1	9263	597	2735
	0.04	0.12	0.03	0.20	0.05	0.12	0.03	0.03	0.72	0.18	0.00	0.27	0.15	0.06	0.13	0.06	0.24	0.25	0.09	0.14	0.20	0.03	0.08	0.32	0.01	0.29	0.04	0.16	0.16	0.05		0.20	0.01
%AC	563	481	943	227	381	891	5063	3072	1281	234	513	2333	317	295	611	966	949	507	81	5592	97	213	627	645	1319	949	279	439	4392	9263	1	795	3928
	0.04	0.02	0.02	0.08	0.16	0.03	0.10	0.12	0.18	0.66	0.27	0.02	0.01	0.22	0.16	0.23	0.08	0.03	0.12	0.11	0.00	0.16	0.17	0.04	0.02	0.22	0.27	0.13	0.13	0.13	0.20		0.37
%AG	658	224	453	7618	953	4729	069	462	728	1448	288	0597	979	169	9049	6785	1814	7257	757	555	324	707	733	7088	1146	2056	84	896	896	597	795	1	147
	0.14	0.42	0.25	0.24	0.41	0.26	0.43	0.34	0.11	0.16	0.76	0.17	0.37	0.23	0.22	0.09	0.34	0.38	0.11	0.56	0.03	0.16	0.34	0.62	0.62	0.63	0.60	0.72	0.72	0.33	0.01	0.37	
%AU	2622	4447	954	029	6298	104	5237	0005	1195	422	355	8215	276	275	682	7061	298	641	9537	9616	0132	9919	8625	1195	574	38	5669	342	3417	2735	3928	147	1
	0.02	0.08	0.04	0.12	0.08	0.04	0.00	0.05	0.31	0.03	0.05	0.67	0.09	0.26	0.14	0.15	0.03	0.21	0.00	0.10	0.05	0.12	0.16	0.22	0.11	0.19	0.14	0.04	0.04	0.06	0.43	0.08	0.10
%CA	73	2727	2547	558	834	603	453	476	3068	4611	7948	6361	865	067	02	745	135	173	6046	623	474	9475	602	2436	6513	394	404	336	3361	292	5356	562	6338
	0.03	0.31	0.45	0.21	0.32	0.34	0.33	0.31	0.13	0.02	0.36	0.09	0.83	0.29	0.08	0.07	0.30	0.40	0.21	0.37	0.04	0.16	0.35	0.58	0.81	0.43	0.63	0.71	0.71	0.38	0.14	0.00	0.49
%CC	318	39	0651	3808	993	3748	351	062	882	164	501	456	2889	3329	9108	783	942	9105	87	341	3644	457	53	173	8217	0527	609	8641	864	651	582	0724	586
	0.05	0.19	0.36	0.32	0.20	0.36	0.21	0.19	0.00	0.13	0.19	0.17	0.34	0.86	0.10	0.03	0.46	0.30	0.00	0.22	0.08	0.17	0.21	0.48	0.53	0.49	0.51	0.58	0.58	0.27	0.04	0.26	0.30
%CG	5492	99	4511	7049	606	1013	026	189	921	889	749	598	0551	1504	025	7007	4348	7516	658	211	6609	402	927	159	1581	1676	479	7882	788	665	868	903	759
	0.07	0.24	0.14	0.01	0.11	0.08	0.18	0.26	0.15	0.14	0.34	0.16	0.09	0.18	0.66	0.02	0.15	0.09	0.25	0.21	0.23	0.00	0.23	0.38	0.35	0.14	0.10	0.28	0.28	0.31	0.13	0.28	0.44
%CU	778	558	1248	8152	863	1435	737	938	387	2631	36	96	3106	609	4144	846	3492	2035	789	769	7425	377	179	016	081	6131	867	6147	615	706	229	6348	194
	0.04	0.10	0.03	0.11	0.04	0.04	0.02	0.01	0.01	0.32	0.02	0.11	0.04	0.01	0.06	0.67	0.14	0.04	0.19	0.15	0.16	0.03	0.03	0.16	0.12	0.12	0.16	0.00	0.00	0.00	0.04	0.40	0.00
%GA	2226	0567	482	3847	073	3629	9635	5214	767	5839	0128	872	964	2425	1026	4024	412	5188	509	537	1713	883	892	3857	348	9373	372	2484	248	6163	446	9011	2893

%GC	-0.01086	0.31 72	0.38 4275	0.35 0134	-0.32 297	0.38 3469	-0.33 152	0.35 369	-0.22 513	-0.08 9491	0.35 884	-0.03 3463	0.34 601	-0.45 039	-0.17 4074	-0.14 658	-0.80 3607	-0.32 6845	-0.21 751	-0.38 079	-0.10 72	-0.05 861	-0.39 635	-0.60 618	-0.65 3474	-0.63 213	-0.64 55	-0.73 8486	-0.73 849	-0.44 736	-0.28 24	-0.12 877	-0.48 41
%GG	-0.07008	0.37 496	0.20 5429	0.38 9683	-0.33 325	0.31 4025	-0.36 63	0.37 366	-0.25 46	-0.00 1934	0.40 412	-0.21 596	0.39 715	-0.24 9353	-0.06 422	-0.03 11	0.27 5762	0.79 0609	-0.18 472	-0.41 282	-0.01 7	-0.11 435	-0.33 141	-0.66 135	-0.44 3338	-0.83 1546	-0.57 903	-0.73 0873	-0.73 087	-0.44 935	-0.27 847	-0.05 4769	-0.52 193
%GU	-0.09284	-0.09 47	-0.24 007	-0.07 145	-0.06 7834	-0.16 026	-0.01 236	-0.12 894	-0.07 2276	-0.15 47	-0.03 751	-0.04 188	-0.25 798	-0.08 268	-0.26 117	-0.25 68	-0.26 298	-0.19 315	-0.67 7451	-0.11 5632	-0.13 045	-0.24 5015	-0.02 798	-0.04 306	-0.30 49	-0.00 148	-0.33 824	-0.17 707	-0.17 7068	-0.11 436	-0.17 7247	-0.12 956	-0.03 3579
%UA	-0.07 5633	0.36 9167	0.31 267	0.29 051	-0.35 6736	0.31 504	-0.37 5734	0.30 3857	-0.22 4842	-0.03 857	0.50 5388	-0.03 831	0.38 463	-0.25 822	-0.12 862	-0.10 735	-0.37 186	-0.40 271	-0.22 7108	-0.78 4438	-0.17 661	-0.03 671	-0.31 3905	-0.60 2641	-0.61 473	-0.61 417	-0.59 4337	-0.70 585	-0.70 5852	-0.31 2065	-0.21 5817	-0.10 982	-0.65 3658
%UC	-0.01 6099	0.10 863	0.16 5755	0.00 3419	-0.01 8279	0.08 5936	-0.04 557	0.13 099	-0.18 802	-0.04 9793	0.05 987	-0.02 512	0.09 0857	-0.06 2507	-0.32 6549	-0.13 7688	-0.10 59	-0.04 8346	-0.06 136	-0.23 369	-0.68 4838	-0.15 627	-0.07 33	-0.25 219	-0.26 1659	-0.02 548	-0.01 6862	-0.13 6675	-0.13 667	-0.19 33	-0.25 89	-0.05 8247	-0.09 866
%UG	-0.10059	0.16 277	0.17 351	0.00 27	0.00 0524	0.08 949	0.08 245	0.22 199	0.05 704	-0.20 715	0.00 306	0.08 2653	-0.21 662	-0.29 434	-0.01 308	-0.13 946	-0.02 2584	-0.12 216	-0.28 2764	-0.14 856	-0.23 789	-0.61 4604	-0.15 085	-0.15 501	-0.18 544	-0.13 0414	-0.20 5741	-0.03 275	-0.03 2753	-0.21 748	-0.02 1223	-0.20 871	-0.10 8101
%UU	-0.02 8703	0.22 2706	0.35 769	0.29 858	-0.40 2291	0.34 222	-0.32 5315	0.39 2843	-0.04 662	-0.12 868	0.26 2021	-0.12 559	0.37 607	-0.27 024	-0.16 939	-0.03 522	-0.40 508	-0.32 059	-0.04 4246	-0.28 1528	-0.05 638	-0.06 866	-0.74 5754	-0.38 9921	-0.64 695	-0.57 994	-0.79 9942	-0.70 494	-0.70 4936	-0.51 1967	-0.09 229	-0.20 545	-0.33 0712
pb	0.95 8345	0.80 9975	0.71 673	0.75 4173	0.83 8009	0.76 9669	0.85 3547	0.58 387	0.57 116	0.64 3669	0.66 3067	0.61 9752	0.38 5222	0.38 4816	0.60 2148	0.69 9668	0.44 428	0.42 2115	0.62 6229	0.57 3428	0.65 5022	0.66 1013	0.60 341	0.09 144	0.07 065	0.07 975	0.05 5464	0.08 635	0.08 6353	0.07 1103	0.05 479	0.05 47	0.15 6292
Npb	0.03 0902	0.03 5064	0.06 38	0.02 0071	0.09 058	0.02 156	0.06 5584	0.02 2159	0.02 083	0.01 196	0.08 2931	0.04 421	0.09 053	0.00 631	0.03 233	0.01 3753	0.02 829	0.00 677	0.11 2838	0.09 9615	0.02 634	0.07 6114	0.08 9275	0.03 1953	0.16 095	0.02 318	0.14 6698	0.10 625	0.10 6255	0.01 5945	0.04 378	0.05 43	0.10 3727
mfe	0.80 3806	0.56 7593	0.75 7472	0.76 9874	0.57 2126	0.79 8793	0.59 0158	0.34 8406	0.45 3339	0.57 8425	0.41 1145	0.51 6302	0.51 0387	0.48 8327	0.55 9713	0.58 5379	0.55 7708	0.52 9639	0.48 6033	0.33 9624	0.57 4452	0.53 0352	0.32 8006	0.11 169	0.16 2483	0.13 4804	0.17 715	0.17 0852	0.17 085	0.09 881	0.07 519	0.00 411	0.03 797
dG	-0.00 62	0.18 373	0.22 1177	0.21 4232	-0.19 525	0.22 7553	-0.19 633	0.23 416	-0.06 811	-0.04 3484	0.20 934	-0.02 785	0.28 0102	0.24 1812	0.07 0632	0.02 337	-0.28 1132	-0.29 0986	-0.04 859	0.19 452	-0.01 6786	-0.02 974	-0.27 039	-0.35 898	-0.37 9367	-0.39 3586	-0.39 326	-0.44 3905	-0.44 39	-0.30 737	-0.08 126	-0.06 3475	-0.29 037
Q	0.61 263	0.45 8268	0.50 7701	0.56 1731	0.48 7254	0.56 001	0.48 9841	0.30 6107	0.32 1491	0.43 2961	0.35 7488	0.37 8213	0.32 8687	0.28 3153	0.40 9096	0.47 7873	0.33 8949	0.40 2872	0.36 6532	0.26 8298	0.43 2976	0.42 8789	0.34 0034	-0.04 863	0.03 4316	0.08 1038	0.06 324	0.06 6099	0.06 61	0.02 662	0.08 882	0.01 116	0.01 1854
NQ	0.23 6295	0.10 8619	0.24 2641	0.30 6796	0.14 1129	0.28 8279	0.12 9622	0.04 3029	0.06 6303	0.18 6041	0.06 7097	0.10 3064	0.20 8091	0.15 9662	0.18 4316	0.20 5606	0.19 5184	0.29 4935	0.13 0166	0.01 907	0.17 3448	0.17 9731	0.07 9469	-0.15 247	0.08 7039	0.19 5134	0.12 21	0.16 1693	0.16 169	0.10 418	0.11 396	0.01 9266	0.08 925
D	0.61 4482	0.44 9789	0.51 4762	0.57 6298	0.48 309	0.57 1435	0.48 3336	0.29 4296	0.31 644	0.43 7113	0.34 8504	0.37 516	0.33 7495	0.28 8652	0.41 5673	0.47 9506	0.35 2248	0.42 0936	0.37 0017	0.25 9634	0.43 0867	0.43 8437	0.33 265	-0.06 196	0.04 0145	0.09 7597	0.07 172	0.07 8919	0.07 892	0.03 701	0.09 591	0.00 921	0.00 3199
ND	-0.20 0806	0.07 3409	0.21 7207	0.28 7032	0.10 7405	0.26 4734	0.09 3944	0.01 5431	0.04 0582	0.16 3059	0.03 769	0.07 6962	0.19 6139	0.14 6806	0.16 3535	0.17 9714	0.18 5731	0.28 9877	0.10 9982	0.00 727	-0.14 4889	0.16 2894	0.05 2739	-0.16 4	0.08 9138	0.21 0453	0.12 753	0.17 1649	0.17 165	0.11 159	0.02 796	0.09 0845	0.09 971
nste m	0.99 2131	0.83 5855	0.76 5255	0.78 1182	0.85 1416	0.80 8829	0.87 377	0.60 0915	0.60 2569	0.67 1405	0.67 4148	0.66 6273	0.42 4903	0.40 0835	0.63 3866	0.72 6016	0.47 3474	0.44 0286	0.62 3394	0.56 9648	-0.68 9217	0.67 4603	0.60 6916	-0.08 8968	0.03 892	0.07 833	0.02 5821	0.06 72	0.06 7199	0.07 1288	0.04 775	0.05 018	0.14 4484
MFE1	-0.06 9343	0.25 2102	0.17 214	0.15 747	-0.22 8656	0.17 211	-0.24 87	0.27 4802	-0.09 0331	0.01 832	-0.30 0442	0.02 823	-0.21 781	0.16 569	0.08 474	0.00 3562	-0.22 283	-0.21 458	0.08 9932	-0.35 4138	-0.03 537	0.00 038	0.25 248	0.37 7681	0.36 862	0.36 283	0.33 5304	0.42 015	0.42 0153	0.29 5786	0.05 0288	0.07 921	0.35 2985
MFE2	-0.33065	0.26 486	0.24 524	0.25 885	-0.30 597	0.26 378	-0.29 595	0.18 733	-0.18 657	-0.20 997	0.22 431	-0.21 946	0.11 944	-0.11 501	0.23 046	-0.24 82	-0.14 561	-0.12 821	-0.23 541	-0.16 548	-0.25 329	-0.27 187	-0.22 585	-0.03 086	-0.04 9752	-0.05 9137	-0.07 455	-0.06 2511	-0.06 251	-0.03 386	-0.03 235	-0.05 1743	-0.08 208
MFE3	-0.19202	0.29 511	0.01 9627	0.00 7741	-0.30 187	0.01 4125	-0.30 917	0.28 65	-0.16 616	-0.09 852	0.28 209	-0.15 547	-0.13 199	-0.10 2708	-0.06 436	-0.16 097	-0.11 1228	-0.12 6307	-0.13 844	-0.24 148	-0.11 19	-0.14 916	-0.31 75	-0.29 786	-0.30 2087	-0.31 5297	-0.30 327	-0.35 4552	-0.35 455	-0.25 755	-0.05 357	-0.05 5196	-0.24 855
MFE4	-0.65615	0.64 134	0.35 857	0.39 551	-0.69 573	0.39 486	-0.69 281	0.49 984	-0.42 458	-0.41 913	0.56 285	-0.45 019	-0.08 416	-0.09 748	0.39 652	-0.48 998	-0.14 632	-0.10 001	-0.49 831	-0.48 69	-0.44 659	-0.52 357	-0.56 315	-0.32 145	-0.34 0159	-0.34 9833	-0.35 011	-0.39 6274	-0.39 627	-0.25 809	-0.01 131	-0.09 3984	-0.34 351

total_	0.95	0.80	0.71	0.75	0.83	0.76	0.85	0.58	0.57	0.64	0.66	0.61	0.38	0.38	0.60	0.69	0.44	0.42	0.62	0.57	0.65	0.66	0.60	0.09	0.07	0.07	0.05	0.08	0.08	0.07	0.05	0.05	0.15
base	8345	9975	673	4173	8009	9669	3547	387	116	3669	3067	9752	5222	4816	2148	9668	428	2115	6229	3428	5022	1013	341	144	065	975	5464	635	6353	1103	479	47	6292
n_ste	0.80	0.60	0.68	0.70	0.64	0.72	0.64	0.39	0.44	0.55	0.47	0.53	0.41	0.38	0.55	0.59	0.46	0.44	0.49	0.36	0.58	0.57	0.42	0.03	0.06	0.03	0.06	0.05	0.05	0.03	0.07	0.02	0.03
ms	1152	5191	17	6468	1042	6213	5606	6984	8747	9616	2798	176	6212	3568	0948	9094	6191	6475	9076	6425	1016	1068	7047	705	3999	8556	265	9	9	215	235	762	9963
avg_b	0.00	0.08	0.08	0.07	0.06	0.08	0.07	0.09	0.03	0.00	0.08	0.00	0.09	0.08	0.04	0.01	0.09	0.09	0.02	0.12	0.04	0.01	0.08	0.15	0.15	0.14	0.13	0.17	0.17	0.12	0.03	0.00	0.11
p	3548	1661	876	772	6924	688	6793	2692	2643	62	996	312	851	174	78	066	761	532	5063	6241	588	46	6498	7926	869	18	5143	265	2647	105	6783	76	9111
	0.22	0.18	0.18	0.18	0.19	0.18	0.19	0.12	0.12	0.12	0.17	0.12	0.13	0.09	0.14	0.17	0.07	0.12	0.14	0.12	0.18	0.15	0.12	0.01	0.01	0.02	0.02	0.02	0.02	0.00	0.00	0.06	0.06
SEQ	714	074	157	146	794	982	623	15	399	435	188	92	273	198	772	704	597	932	292	614	763	999	65	161	4553	7176	871	3915	392	399	6609	092	421

CHAPTER 5

Conclusion

CHAPTER 5: Conclusion

miRNAs are transcribed by RNA pol II and III as pri-miRNA which are hairpin loops formed in the nucleus. They are processed to form pre-miRNAs which are further processed by Dicer along with other enzymes to produce mature miRNAs. The role of miRNA is crucial in insects which has been reported to participate in a wide range of biological activities including metamorphosis, reproduction, immune response, etc.

Several tools have been designed using machine learning algorithms for the prediction of pre-miRNA which share the common training features. These training features included sequential as well as thermodynamic features such as length, MFE, GC%, Shannon entropy, etc. Although miRNAs are believed to be conserved, however, these pre-miRNA features vary greatly among different class of organisms. The length of insect pre-miRNA is more than human pre-miRNA but is less than plant pre-miRNA. GC content of insect is lower than plant and humans. Since the previous tools were based on these features, they are not suitable for efficient prediction of insect pre-miRNA.

This thesis is based on the application of machine learning algorithms for prediction of precursor miRNA in insects. The work presents a roadmap for creation of an intelligent system using AI and serve it over the internet as a web-based tool which we named here RNAinsecta. The training of algorithms, development of predictive model and design of API were done in python programming language. Supervised ml algorithms were used for training where a label is provided for each class. Algorithms like SVM, Random Forest, Logistic Regression and k-Nearest Neighbours were trained using binary class label of true and pseudo insect pre-miRNA. Two class imbalance handling methods namely, Near-Miss and SMOTE were used for undersampling the majority class and oversampling the minority class respectively. Independent test datasets were used for evaluating the performance of these trained models which were also compared with the already existing tools. The performance of the trained classifier on SVM and Random Forest outperformed all the other classifiers and tools for which they were selected for implementing in a web server.

The web-tool is hosted over Linux platform in a cloud-based server which uses NGINX as the reverse proxy server. Front end of the web tool was developed using HTML, CSS and JavaScript. Char.js along with JQuery was used for creating the dynamic graphs. SSL certificated for the website was created using chartbot which is licensed by “Let’s Encrypt”. The exceptions of missing or error input was handled using regular expression. The tool

further enables users to pre-process their predicted true pre-miRNA that will produce mature miRNA. Users can use these miRNA sequences to find the target of their miRNA in the genome of the model organism *Drosophila melanogaster*. The resultant page gives the miRanda score for target match and along with hyperlink to Flybase containing the gene name and also hyperlink to miRbase for the miRNA which is responsible for regulation of that gene which has been experimentally verified.

This thesis also highlights the features which differ in insects as compared to other groups of organisms. The difference was confirmed with chi-square test for these features between insects and other groups of organisms. Further, an XGBoost based classifier was trained that was capable of classifying insect, rice and sauria pre-miRNA. PCA was used for dimensional reduction along with "permutation feature importance" to calculated and analyse the weight of each feature involved in the training process.

The developed tool can be run through any shell script in Linux which in future can be easily implemented in any pipeline. There are many RNA-Seq based NGS tool that gives the differentially expressed genes in a transcriptome. These models in future can be implemented in such a pipeline for the prediction of total pre-miRNA in transcriptome of a given insect. This will enable researchers to find fluctuations in pre-miRNA levels in any given biological condition of an insect or in timescale data.

CHAPTER 6

Reference

Ahmad, F., Farooq, A., & Khan, M. U. G. (2020). Deep Learning Model for Pathogen Classification Using Feature Fusion and Data Augmentation. *Current Bioinformatics*, *16*(3), 466–483. https://doi.org/10.2174/1574893615999200707143535

Allmer, J., & Yousef, M. (2012). Computational methods for ab initio detection of microRNAs. *Frontiers in Genetics*, *3*. https://doi.org/10.3389/fgene.2012.00209

Ambros, V. (2004). The functions of animal microRNAs. *Nature*, *431*(7006), 350–355. https://doi.org/10.1038/nature02871

Avila-Bonilla, R. G., Yocupicio-Monroy, M., Marchat, L. A., Pérez-Ishiwara, D. G., Cerecedo-Mercado, D. A., del Ángel, R. M., & Salas-Benito, J. S. (2020). miR-927 has pro-viral effects during acute and persistent infection with dengue virus type 2 in C6/36 mosquito cells. *Journal of General Virology*, *101*(8), 825–839. https://doi.org/10.1099/JGV.0.001441/CITE/REFWORKS

Backes, C., Kehl, T., Stöckel, D., Fehlmann, T., Schneider, L., Meese, E., Lenhof, H. P., & Keller, A. (2017). miRPathDB: a new dictionary on microRNAs and target pathways. *Nucleic Acids Research*, *45*(D1), D90–D96. https://doi.org/10.1093/NAR/GKW926

Bandyopadhyay, S., & Mitra, R. (2009). TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinformatics*, *25*(20), 2625–2631. https://doi.org/10.1093/BIOINFORMATICS/BTP503

Batuwita, R., & Palade, V. (2009). microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics*, *25*(8), 989–995. https://doi.org/10.1093/bioinformatics/btp107

Belles, X., Cristino, A. S., Tanaka, E. D., Rubio, M., & Piulachs, M.-D. (2012). Insect MicroRNAs. In *Insect Molecular Biology and Biochemistry* (pp. 30–56). Elsevier. https://doi.org/10.1016/B978-0-12-384747-8.10002-9

Bernstein, E., Kim, S. Y., Carmell, M. A., Murchison, E. P., Alcorn, H., Li, M. Z., Mills, A. A., Elledge, S. J., Anderson, K. V., & Hannon, G. J. (2003). Dicer is essential for mouse development. *Nature Genetics*, *35*(3), 215–217. https://doi.org/10.1038/NG1253

Bi, Y., Xiang, D., Ge, Z., Li, F., Jia, C., & Song, J. (2020). An Interpretable Prediction Model for Identifying N7-Methylguanosine Sites Based on XGBoost and SHAP. *Molecular Therapy - Nucleic Acids*, *22*, 362–372. https://doi.org/10.1016/j.omtn.2020.08.022

Biesiada, J., & Duch, W. (2007). Feature selection for high-dimensional data - A

pearson redundancy based filter. *Advances in Soft Computing*, *45*, 242–249. https://doi.org/10.1007/978-3-540-75175-5_30/COVER

Blaszczyk, J., Tropea, J. E., Bubunenko, M., Routzahn, K. M., Waugh, D. S., Court, D. L., & Ji, X. (2001). Crystallographic and modeling studies of RNase III suggest a mechanism for double-stranded RNA cleavage. *Structure*, *9*(12), 1225–1236. https://doi.org/10.1016/S0969-2126(01)00685-2

Borchert, G. M., Lanier, W., & Davidson, B. L. (2006). RNA polymerase III transcribes human microRNAs. *Nature Structural & Molecular Biology*, *13*(12), 1097–1101. https://doi.org/10.1038/nsmb1167

Brazzolotto, X., Pierrel, F., & Pelosi, L. (2014). Three conserved histidine residues contribute to mitochondrial iron transport through mitoferrins. *Biochemical Journal*, *460*(1), 79–92. https://doi.org/10.1042/BJ20140107

Broughton, J. P., Lovci, M. T., Huang, J. L., Yeo, G. W., & Pasquinelli, A. E. (2016). Pairing beyond the Seed Supports MicroRNA Targeting Specificity. *Molecular Cell*, *64*(2), 320–333. https://doi.org/10.1016/j.molcel.2016.09.004

Bushati, N., & Cohen, S. M. (2007). microRNA Functions. *Annual Review of Cell and Developmental Biology*, *23*(1), 175–205. https://doi.org/10.1146/annurev.cellbio.23.090506.123406

Büssing, I., Slack, F. J., & Großhans, H. (2008). let-7 microRNAs in development, stem cells and cancer. *Trends in Molecular Medicine*, *14*(9), 400–409. https://doi.org/10.1016/J.MOLMED.2008.07.001

Cai, X., Hagedorn, C. H., & Cullen, B. R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA (New York, N.Y.)*, *10*(12), 1957–1966. https://doi.org/10.1261/RNA.7135204

Calin, G. A., Liu, C.-G., Sevignani, C., Ferracin, M., Felli, N., Dumitru, C. D., Shimizu, M., Cimmino, A., Zupo, S., Dono, M., Dell'Aquila, M. L., Alder, H., Rassenti, L., Kipps, T. J., Bullrich, F., Negrini, M., & Croce, C. M. (2004). MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias. *Proceedings of the National Academy of Sciences*, *101*(32), 11755–11760. https://doi.org/10.1073/pnas.0404432101

Caygill, E. E., & Johnston, L. A. (2008). Temporal Regulation of Metamorphic Processes in Drosophila by the let-7 and miR-125 Heterochronic MicroRNAs. *Current Biology : CB*, *18*(13), 943. https://doi.org/10.1016/J.CUB.2008.06.020

Chen, J., Wang, X., & Liu, B. (2016). iMiRNA-SSF: Improving the Identification

of MicroRNA Precursors by Combining Negative Sets with Different Distributions. *Scientific Reports*, *6*(1), 19062. https://doi.org/10.1038/srep19062

Chen, Y., & Wang, X. (2020). miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Research*, *48*(D1), D127–D131. https://doi.org/10.1093/nar/gkz757

Chendrimada, T. P., Gregory, R. I., Kumaraswamy, E., Norman, J., Cooch, N., Nishikura, K., & Shiekhattar, R. (2005). TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature*, *436*(7051), 740–744. https://doi.org/10.1038/NATURE03868

Cheng, Y., Dong, L., Zhang, J., Zhao, Y., & Li, Z. (2018). Recent advances in microRNA detection. *Cite This: Analyst*, *143*, 1758. https://doi.org/10.1039/c7an02001e

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1). https://doi.org/10.1186/s12864-019-6413-7

Chou, C. H., Chang, N. W., Shrestha, S., Hsu, S. Da, Lin, Y. L., Lee, W. H., Yang, C. D., Hong, H. C., Wei, T. Y., Tu, S. J., Tsai, T. R., Ho, S. Y., Jian, T. Y., Wu, H. Y., Chen, P. R., Lin, N. C., Huang, H. T., Yang, T. L., Pai, C. Y., … Huang, H. Da. (2016). miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Research*, *44*(D1), D239–D247. https://doi.org/10.1093/NAR/GKV1258

Condrat, C. E., Thompson, D. C., Barbu, M. G., Bugnar, O. L., Boboc, A., Cretoiu, D., Suciu, N., Cretoiu, S. M., & Voinea, S. C. (2020). miRNAs as Biomarkers in Disease: Latest Findings Regarding Their Role in Diagnosis and Prognosis. *Cells 2020, Vol. 9, Page 276*, *9*(2), 276. https://doi.org/10.3390/CELLS9020276

Cullen, B. R. (2009). Viral and cellular messenger RNA targets of viral microRNAs. *Nature*, *457*(7228), 421–425. https://doi.org/10.1038/nature07757

Czimmerer, Z., Hulvely, J., Simandi, Z., Varallyay, E., Havelda, Z., Szabo, E., Varga, A., Dezso, B., Balogh, M., Horvath, A., Domokos, B., Torok, Z., Nagy, L., & Balint, B. L. (2013). A Versatile Method to Design Stem-Loop Primer-Based Quantitative PCR Assays for Detecting Small Regulatory RNA Molecules. *PLOS ONE*, *8*(1), e55168. https://doi.org/10.1371/JOURNAL.PONE.0055168

Davis, T. H., Cuellar, T. L., Koch, S. M., Barker, A. J., Harfe, B. D., McManus, M.

T., & Ullian, E. M. (2008). Conditional loss of Dicer disrupts cellular and tissue morphogenesis in the cortex and hippocampus. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *28*(17), 4322–4330. https://doi.org/10.1523/JNEUROSCI.4815-07.2008

Dick, S. (2019). Issue 1.1, Summer 2019. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608F92.92FE150C

Dubey, S. K., Shrinet, J., Jain, J., Ali, S., & Sunil, S. (2017). Aedes aegypti microRNA miR-2b regulates ubiquitin-related modifier to control chikungunya virus replication. *Scientific Reports 2017 7:1*, *7*(1), 1–10. https://doi.org/10.1038/s41598-017-18043-0

Dweep, H., & Gretz, N. (2015). miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nature Methods 2015 12:8*, *12*(8), 697–697. https://doi.org/10.1038/nmeth.3485

Etebari, K., Afrad, M. H., Tang, B., Silva, R., Furlong, M. J., & Asgari, S. (2018). Involvement of microRNA miR-2b-3p in regulation of metabolic resistance to insecticides in Plutella xylostella. *Insect Molecular Biology*, *27*(4), 478–491. https://doi.org/10.1111/imb.12387

Etebari, Kayvan, & Asgari, S. (2013). Conserved microRNA miR-8 blocks activation of the Toll pathway by upregulating Serpin 27 transcripts. *Https://Doi.Org/10.4161/Rna.25481*, *10*(8), 1356–1364. https://doi.org/10.4161/RNA.25481

Feng, Y., Zhang, X., Graves, P., & Zeng, Y. (2012). A comprehensive analysis of precursor microRNA cleavage by human Dicer. *RNA (New York, N.Y.)*, *18*(11), 2083–2092. https://doi.org/10.1261/RNA.033688.112

Foda, M. F., Huang, L., Shao, F., & Han, H. Y. (2014). Biocompatible and highly luminescent near-infrared CuInS2/ZnS quantum dots embedded silica beads for cancer cell imaging. *ACS Applied Materials and Interfaces*, *6*(3), 2011–2017. https://doi.org/10.1021/AM4050772/SUPPL_FILE/AM4050772_SI_001.PDF

Freyhult, E., Gardner, P. P., & Moulton, V. (2005). A comparison of RNA folding measures. *BMC Bioinformatics*, *6*(1), 241. https://doi.org/10.1186/1471-2105-6-241

Friedman, R. C., Farh, K. K. H., Burge, C. B., & Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, *19*(1), 92–105. https://doi.org/10.1101/GR.082701.108

Froschauer, E. M., Schweyen, R. J., & Wiesenberger, G. (2009). The yeast mitochondrial carrier proteins Mrs3p/Mrs4p mediate iron transport across

the inner mitochondrial membrane. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, *1788*(5), 1044–1050. https://doi.org/10.1016/J.BBAMEM.2009.03.004

Fu, X., Zhu, W., Cai, L., Liao, B., Peng, L., Chen, Y., & Yang, J. (2019). Improved Pre-miRNAs Identification Through Mutual Information of Pre-miRNA Sequences and Structures. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00119

Gan, Y. bo, Zhou, Z. jing, An, L. jun, Bao, S. jie, & Forde, B. G. (2011). A Comparison Between Northern Blotting and Quantitative Real-Time PCR as a Means of Detecting the Nutritional Regulation of Genes Expressed in Roots of Arabidopsis thaliana. *Agricultural Sciences in China*, *10*(3), 335–342. https://doi.org/10.1016/S1671-2927(11)60012-6

Gerlach, D., Kriventseva, E. V., Rahman, N., Vejnar, C. E., & Zdobnov, E. M. (2009). miROrtho: computational survey of microRNA genes. *Nucleic Acids Research*, *37*(suppl_1), D111–D117. https://doi.org/10.1093/NAR/GKN707

Giraldez, A. J., Cinalli, R. M., Glasner, M. E., Enright, A. J., Thomson, J. M., Baskerville, S., Hammond, S. M., Bartel, D. P., & Schier, A. F. (2005). MicroRNAs Regulate Brain Morphogenesis in Zebrafish. *Science*, *308*(5723), 833–838. https://doi.org/10.1126/science.1109020

Gkirtzou, K., Tsamardinos, I., Tsakalides, P., & Poirazi, P. (2010). MatureBayes: A Probabilistic Algorithm for Identifying the Mature miRNA within Novel Precursors. *PLoS ONE*, *5*(8), e11843. https://doi.org/10.1371/journal.pone.0011843

Gomes, C. P. C., Cho, J.-H., Hood, L., Franco, O. L., Pereira, R. W., & Wang, K. (2013). A Review of Computational Tools in microRNA Discovery. *Frontiers in Genetics*, *4*. https://doi.org/10.3389/fgene.2013.00081

Gomez-Orte, E., & Belles, X. (2009). MicroRNA-dependent metamorphosis in hemimetabolan insects. *Proceedings of the National Academy of Sciences*, *106*(51), 21678–21682. https://doi.org/10.1073/pnas.0907391106

Gomez-Orte, Eva, & Belles, X. (2009). MicroRNA-dependent metamorphosis in hemimetabolan insects. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(51), 21678–21682. https://doi.org/10.1073/PNAS.0907391106/SUPPL_FILE/0907391106SI.PDF

Gregory, R. I., Chendrimada, T. P., Cooch, N., & Shiekhattar, R. (2005). Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. *Cell*, *123*(4), 631–640. https://doi.org/10.1016/J.CELL.2005.10.022

Gudnason, H., Dufva, M., Bang, D. D., & Wolff, A. (2007). Comparison of

multiple DNA dyes for real-time PCR: effects of dye concentration and sequence composition on DNA amplification and melting temperature. *Nucleic Acids Research*, *35*(19), e127–e127. https://doi.org/10.1093/NAR/GKM671

Gudyś, A., Szcześniak, M. W., Sikora, M., & Makałowska, I. (2013). HuntMi: an efficient and taxon-specific approach in pre-miRNA identification. *BMC Bioinformatics*, *14*(1), 83. https://doi.org/10.1186/1471-2105-14-83

Gulhane, P., Nimsarkar, P., Kharat, K., & Singh, S. (2022). Deciphering miR-520c-3p as a probable target for immunometabolism in non-small cell lung cancer using systems biology approach. *Oncotarget*, *13*(1), 725. https://doi.org/10.18632/ONCOTARGET.28233

Guo, L., & Chen, F. (2014). A challenge for miRNA: multiple isomiRs in miRNAomics. *Gene*, *544*(1), 1–7. https://doi.org/10.1016/J.GENE.2014.04.039

Guo, Q., Huang, Y., Zou, F., Liu, B., Tian, M., Ye, W., Guo, J., Sun, X., Zhou, D., Sun, Y., Ma, L., Shen, B., & Zhu, C. (2017). The role of miR-2~13~71 cluster in resistance to deltamethrin in Culex pipiens pallens. *Insect Biochemistry and Molecular Biology*, *84*, 15–22. https://doi.org/10.1016/J.IBMB.2017.03.006

Ha, M., & Kim, V. N. (2014). Regulation of microRNA biogenesis. *Nature Reviews Molecular Cell Biology*, *15*(8), 509–524. https://doi.org/10.1038/nrm3838

Han, J., Lee, Y., Yeom, K. H., Kim, Y. K., Jin, H., & Kim, V. N. (2004a). The Drosha-DGCR8 complex in primary microRNA processing. *Genes & Development*, *18*(24), 3016–3027. https://doi.org/10.1101/GAD.1262504

Han, J., Lee, Y., Yeom, K. H., Kim, Y. K., Jin, H., & Kim, V. N. (2004b). The Drosha-DGCR8 complex in primary microRNA processing. *Genes and Development*, *18*(24), 3016–3027. https://doi.org/10.1101/gad.1262504

Han, J., Lee, Y., Yeom, K. H., Nam, J. W., Heo, I., Rhee, J. K., Sohn, S. Y., Cho, Y., Zhang, B. T., & Kim, V. N. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell*, *125*(5), 887–901. https://doi.org/10.1016/J.CELL.2006.03.043

Han, J., Pedersen, J. S., Kwon, S. C., Belair, C. D., Kim, Y. K., Yeom, K. H., Yang, W. Y., Haussler, D., Blelloch, R., & Kim, V. N. (2009). Posttranscriptional crossregulation between Drosha and DGCR8. *Cell*, *136*(1), 75–84. https://doi.org/10.1016/J.CELL.2008.10.053

Hertel, J., & Stadler, P. F. (2006). Hairpins in a Haystack: recognizing microRNA

precursors in comparative genomics data. *Bioinformatics*, *22*(14), e197–e202. https://doi.org/10.1093/bioinformatics/btl257

Hong, C., Baek, A., Hah, S. S., Jung, W., & Kim, D. E. (2016). Fluorometric Detection of MicroRNA Using Isothermal Gene Amplification and Graphene Oxide. *Analytical Chemistry*, *88*(6), 2999–3003. https://doi.org/10.1021/ACS.ANALCHEM.6B00046/ASSET/IMAGES/LARGE/AC-2016-00046J_0005.JPEG

Hosseinzadeh, E., Ravan, H., Mohammadi, A., Mohammad-rezaei, R., Norouzi, A., & Hosseinzadeh, H. (2018). Target-triggered three-way junction in conjugation with catalytic concatemers-functionalized nanocomposites provides a highly sensitive colorimetric method for miR-21 detection. *Biosensors and Bioelectronics*, *117*, 567–574. https://doi.org/10.1016/J.BIOS.2018.06.051

Hsu, S. Da, Chu, C. H., Tsou, A. P., Chen, S. J., Chen, H. C., Hsu, P. W. C., Wong, Y. H., Chen, Y. H., Chen, G. H., & Huang, H. Da. (2008). miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes. *Nucleic Acids Research*, *36*(suppl_1), D165–D169. https://doi.org/10.1093/NAR/GKM1012

Hsu, Paul W.C., Huang, H. Da, Hsu, S. Da, Lin, L. Z., Tsou, A. P., Tseng, C. P., Stadler, P. F., Washietl, S., & Hofacker, I. L. (2006). miRNAMap: genomic maps of microRNA genes and their target genes in mammalian genomes. *Nucleic Acids Research*, *34*(suppl_1), D135–D139. https://doi.org/10.1093/NAR/GKJ135

Hsu, Paul Wei Che, Lin, L. Z., Hsu, S. Da, Hsu, J. B. K., & Huang, H. Da. (2007). ViTa: prediction of host microRNAs targets on viruses. *Nucleic Acids Research*, *35*(suppl_1), D381–D385. https://doi.org/10.1093/NAR/GKL1009

Hsu, S.-D., Chu, C.-H., Tsou, A.-P., Chen, S.-J., Chen, H.-C., Hsu, P. W.-C., Wong, Y.-H., Chen, Y.-H., Chen, G.-H., & Huang, H.-D. (2007). miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes. *Nucleic Acids Research*, *36*(Database), D165–D169. https://doi.org/10.1093/nar/gkm1012

Huang, H.-Y., Lin, Y.-C.-D., Li, J., Huang, K.-Y., Shrestha, S., Hong, H.-C., Tang, Y., Chen, Y.-G., Jin, C.-N., Yu, Y., Xu, J.-T., Li, Y.-M., Cai, X.-X., Zhou, Z.-Y., Chen, X.-H., Pei, Y.-Y., Hu, L., Su, J.-J., Cui, S.-D., … Huang, H.-D. (2019). miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz896

Huang, H. Y., Lin, Y. C. D., Cui, S., Huang, Y., Tang, Y., Xu, J., Bao, J., Li, Y.,

Wen, J., Zuo, H., Wang, W., Li, J., Ni, J., Ruan, Y., Li, L., Chen, Y., Xie, Y., Zhu, Z., Cai, X., … Huang, H. Da. (2022). miRTarBase update 2022: an informative resource for experimentally validated miRNA–target interactions. *Nucleic Acids Research*, *50*(D1), D222–D230. https://doi.org/10.1093/NAR/GKAB1079

Huang, K.-Y., Lee, T.-Y., Teng, Y.-C., & Chang, T.-H. (2015). ViralmiR: a support-vector-machine-based method for predicting viral microRNA precursors. *BMC Bioinformatics*, *16*(Suppl 1), S9. https://doi.org/10.1186/1471-2105-16-S1-S9

Huang, T.-H., Fan, B., Rothschild, M. F., Hu, Z.-L., Li, K., & Zhao, S.-H. (2007). MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. *BMC Bioinformatics*, *8*(1), 341. https://doi.org/10.1186/1471-2105-8-341

Hussain, M., Walker, T., O'Neill, S. L., & Asgari, S. (2013). Blood meal induced microRNA regulates development and immune associated genes in the Dengue mosquito vector, Aedes aegypti. *Insect Biochemistry and Molecular Biology*, *43*(2), 146–152. https://doi.org/10.1016/J.IBMB.2012.11.005

Jaiswal, S., Iquebal, M. A., Arora, V., Sheoran, S., Sharma, P., Angadi, U. B., Dahiya, V., Singh, R., Tiwari, R., Singh, G. P., Rai, A., & Kumar, D. (2019). Development of species specific putative miRNA and its target prediction tool in wheat (Triticum aestivum L.). *Scientific Reports*, *9*(1), 3790. https://doi.org/10.1038/s41598-019-40333-y

Jayachandran, B., Hussain, M., & Asgari, S. (2013). An insect trypsin-like serine protease as a target of microRNA: Utilization of microRNA mimics and inhibitors by oral feeding. *Insect Biochemistry and Molecular Biology*, *43*(4), 398–406. https://doi.org/10.1016/J.IBMB.2012.10.004

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., & Lu, Z. (2007). MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Research*, *35*(suppl_2), W339–W344. https://doi.org/10.1093/nar/gkm368

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., & Marks, D. S. (2004). Human MicroRNA Targets. *PLOS Biology*, *2*(11), e363. https://doi.org/10.1371/JOURNAL.PBIO.0020363

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., & Marks, D. S. (2005). Correction: Human MicroRNA Targets. *PLoS Biology*, *3*(7), e264. https://doi.org/10.1371/journal.pbio.0030264

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

https://doi.org/10.1126/SCIENCE.AAA8415/ASSET/AB2EF18A-576D-464D-B1B6-1301159EE29A/ASSETS/GRAPHIC/349_255_F5.JPEG

Kadri, S., Hinman, V., & Benos, P. V. (2009). HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. *BMC Bioinformatics*, *10*(Suppl 1), S35. https://doi.org/10.1186/1471-2105-10-S1-S35

Kakumani, P. K., Chinnappan, M., Singh, A. K., Malhotra, P., Mukherjee, S. K., & Bhatnagar, R. K. (2015). Identification and characteristics of microRNAs from army worm, Spodoptera frugiperda cell line Sf21. *PloS One*, *10*(2), e0116988. https://doi.org/10.1371/journal.pone.0116988

Kambhatla, N., & Leen, T. K. (1997). Dimension Reduction by Local Principal Component Analysis. *Neural Computation*, *9*(7), 1493–1516. https://doi.org/10.1162/NECO.1997.9.7.1493

Kans, J. (2020). Entrez direct: E-utilities on the UNIX command line. In *Entrez Programming Utilities Help [Internet]*. National Center for Biotechnology Information (US).

Kehl, T., Kern, F., Backes, C., Fehlmann, T., Stöckel, D., Meese, E., Lenhof, H. P., & Keller, A. (2020). miRPathDB 2.0: a novel release of the miRNA Pathway Dictionary Database. *Nucleic Acids Research*, *48*(D1), D142–D147. https://doi.org/10.1093/NAR/GKZ1022

Kozomara, A., Birgaoanu, M., & Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Research*, *47*(D1), D155–D162. https://doi.org/10.1093/nar/gky1141

Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M., & Rajewsky, N. (2005). Combinatorial microRNA target predictions. *Nature Genetics*, *37*(5), 495–500. https://doi.org/10.1038/ng1536

Krüger, J., & Rehmsmeier, M. (2006). RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Research*, *34*(suppl_2), W451–W454. https://doi.org/10.1093/NAR/GKL243

Kwok, C. K., Sahakyan, A. B., & Balasubramanian, S. (2016). Structural Analysis using SHALiPE to Reveal RNA G-Quadruplex Formation in Human Precursor MicroRNA. *Angewandte Chemie*, *128*(31), 9104–9107. https://doi.org/10.1002/ANGE.201603562

Lai, E. C., Tomancak, P., Williams, R. W., & Rubin, G. M. (2003). Computational identification of Drosophila microRNA genes. *Genome Biology*, *4*(7), 1–20. https://doi.org/10.1186/GB-2003-4-7-R42/FIGURES/7

Larkin, A., Marygold, S. J., Antonazzo, G., Attrill, H., dos Santos, G., Garapati, P. V, Goodman, J. L., Gramates, L. S., Millburn, G., Strelets, V. B., Tabone, C. J., Thurmond, J., Perrimon, N., Gelbart, S. R., Agapite, J., Broll, K., Crosby, M., dos Santos, G., Falls, K., … Lovato, T. (2021). FlyBase: updates to the Drosophila melanogaster knowledge base. *Nucleic Acids Research*, *49*(D1), D899–D907. https://doi.org/10.1093/nar/gkaa1026

Lawrie, C. H., Gal, S., Dunlop, H. M., Pushkaran, B., Liggins, A. P., Pulford, K., Banham, A. H., Pezzella, F., Boultwood, J., Wainscoat, J. S., Hatton, C. S. R., & Harris, A. L. (2008). Detection of elevated levels of tumour-associated microRNAs in serum of patients with diffuse large B-cell lymphoma. *British Journal of Haematology*, *141*(5), 672–675. https://doi.org/10.1111/J.1365-2141.2008.07077.X

Le, B. H., Nguyen, T. V. T., Joo, H. N., & Seo, Y. J. (2018). Large-Stokes-shift-based folded DNA probing systems targeting DNA and miRNA 21 with signal amplification. *Bioorganic & Medicinal Chemistry*, *26*(17), 4881–4885. https://doi.org/10.1016/J.BMC.2018.08.027

Lee, C. T., Risom, T., & Strauss, W. M. (2007). Evolutionary Conservation of MicroRNA Regulatory Circuits: An Examination of MicroRNA Gene Complexity and Conserved MicroRNA-Target Interactions through Metazoan Phylogeny. *Https://Home.Liebertpub.Com/Dna*, *26*(4), 209–218. https://doi.org/10.1089/DNA.2006.0545

Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, *75*(5), 843–854. https://doi.org/10.1016/0092-8674(93)90529-Y

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S., & Kim, V. N. (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature*, *425*(6956), 415–419. https://doi.org/10.1038/NATURE01957

Lee, Y., Hur, I., Park, S. Y., Kim, Y. K., Mi, R. S., & Kim, V. N. (2006). The role of PACT in the RNA silencing pathway. *EMBO Journal*, *25*(3), 522–532. https://doi.org/10.1038/sj.emboj.7600942

Lee, Y., Kim, M., Han, J., Yeom, K. H., Lee, S., Baek, S. H., & Kim, V. N. (2004). MicroRNA genes are transcribed by RNA polymerase II. *The EMBO Journal*, *23*(20), 4051–4060. https://doi.org/10.1038/SJ.EMBOJ.7600385

Lei, Z., Lv, Y., Wang, W., Guo, Q., Zou, F., Hu, S., Fang, F., Tian, M., Liu, B., Liu, X., Ma, K., Ma, L., Zhou, D., Zhang, D., Sun, Y., Shen, B., & Zhu, C. (2015). MiR-278-3p regulates pyrethroid resistance in Culex pipiens pallens. *Parasitology Research*, *114*(2), 699–706. https://doi.org/10.1007/S00436-014-

4236-7/FIGURES/8

Li, C., Li, Z., Jia, H., & Yan, J. (2011). One-step ultrasensitive detection of microRNAs with loop-mediated isothermal amplification (LAMP). *Chemical Communications*, *47*(9), 2595–2597. https://doi.org/10.1039/C0CC03957H

Li, H., Shi, L., Gao, W., Zhang, Z., Zhang, L., Zhao, Y., & Wang, G. (2022). dPromoter-XGBoost: Detecting promoters and strength by combining multiple descriptors and feature selection using XGBoost. *Methods*, *204*, 215–222. https://doi.org/10.1016/j.ymeth.2022.01.001

Li, J. H., Liu, S., Zhou, H., Qu, L. H., & Yang, J. H. (2014). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Research*, *42*(D1), D92–D97. https://doi.org/10.1093/NAR/GKT1248

Li, R. D., Yin, B. C., & Ye, B. C. (2016). Ultrasensitive, colorimetric detection of microRNAs based on isothermal exponential amplification reaction-assisted gold nanoparticle amplification. *Biosensors and Bioelectronics*, *86*, 1011–1016. https://doi.org/10.1016/J.BIOS.2016.07.042

Li, W., & Ruan, K. (2009). MicroRNA detection by microarray. *Analytical and Bioanalytical Chemistry*, *394*(4), 1117–1124. https://doi.org/10.1007/S00216-008-2570-2/FIGURES/3

Li, X., Guo, L., Zhou, X., Gao, X., & Liang, P. (2015). MiRNAs regulated overexpression of ryanodine receptor is involved in chlorantraniliprole resistance in Plutella xylostella (L.). *Scientific Reports*, *5*. https://doi.org/10.1038/SREP14095

Li, X., Ren, X., Liu, Y., Smagghe, G., Liang, P., & Gao, X. (2020). MiR-189942 regulates fufenozide susceptibility by modulating ecdysone receptor isoform B in Plutella xylostella (L.). *Pesticide Biochemistry and Physiology*, *163*, 235–240. https://doi.org/10.1016/J.PESTBP.2019.11.021

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., & Cui, Q. (2014). HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Research*, *42*(D1), D1070–D1074. https://doi.org/10.1093/NAR/GKT1023

Lim, L. P., Lau, N. C., Weinstein, E. G., Abdelhakim, A., Yekta, S., Rhoades, M. W., Burge, C. B., & Bartel, D. P. (2003). The microRNAs of Caenorhabditis elegans. *Genes & Development*, *17*(8), 991–1008. https://doi.org/10.1101/GAD.1074403

Ling, L., Ge, X., Li, Z., Zeng, B., Xu, J., Aslam, A. F. M., Song, Q., Shang, P., Huang, Y., & Tan, A. (2014). MicroRNA Let-7 regulates molting and

metamorphosis in the silkworm, Bombyx mori. *Insect Biochemistry and Molecular Biology*, *53*, 13–21. https://doi.org/10.1016/j.ibmb.2014.06.011

Ling, L., Kokoza, V. A., Zhang, C., Aksoy, E., & Raikhel, A. S. (2017). MicroRNA-277 targets insulin-like peptides 7 and 8 to control lipid metabolism and reproduction in Aedes aegypti mosquitoes. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(38), E8017–E8024. https://doi.org/10.1073/PNAS.1710970114/SUPPL_FILE/PNAS.201710970SI.PDF

Liu, B., Tian, M., Guo, Q., Ma, L., Zhou, D., Shen, B., Sun, Y., & Zhu, C. (2016). MiR-932 Regulates Pyrethroid Resistance in Culex pipiens pallens (Diptera: Culicidae). *Journal of Medical Entomology*, *53*(5), 1205–1210. https://doi.org/10.1093/JME/TJW083

Liu, R., Wang, Q., Li, Q., Yang, X., Wang, K., & Nie, W. (2017). Surface plasmon resonance biosensor for sensitive detection of microRNA and cancer cell using multiple signal amplification strategy. *Biosensors and Bioelectronics*, *87*, 433–438. https://doi.org/10.1016/J.BIOS.2016.08.090

Liu, S., Xia, Q., Zhao, P., Cheng, T., Hong, K., & Xiang, Z. (2007). Characterization and expression patterns of let-7 microRNA in the silkworm (Bombyx mori). *BMC Developmental Biology*, *7*(1), 1–17. https://doi.org/10.1186/1471-213X-7-88/FIGURES/11

Lucas, K. J., Zhao, B., Roy, S., Gervaise, A. L., & Raikhel, A. S. (2015). Mosquito-specific microRNA-1890 targets the juvenile hormone-regulated serine protease JHA15 in the female mosquito gut. *RNA Biology*, *12*(12), 1383–1390. https://doi.org/10.1080/15476286.2015.1101525/SUPPL_FILE/KRNB_A_1101525_SM0186.PDF

Lukasik, A., Wójcikowski, M., & Zielenkiewicz, P. (2016). Tools4miRs – one place to gather all the tools for miRNA analysis. *Bioinformatics*, *32*(17), 2722–2724. https://doi.org/10.1093/BIOINFORMATICS/BTW189

Ma, L., Liu, L., Zhao, Y., Yang, L., Chen, C., Li, Z., & Lu, Z. (2020). JNK pathway plays a key role in the immune system of the pea aphid and is regulated by microRNA-184. *PLOS Pathogens*, *16*(6), e1008627. https://doi.org/10.1371/JOURNAL.PPAT.1008627

Maniataki, E., & Mourelatos, Z. (2005). A human, ATP-independent, RISC assembly machine fueled by pre-miRNA. *Genes & Development*, *19*(24), 2979–2990. https://doi.org/10.1101/GAD.1384005

Marco, A., Hui, J. H. L., Ronshaugen, M., & Griffiths-Jones, S. (2010). Functional

shifts in insect microRNA evolution. *Genome Biology and Evolution*, *2*, 686–696. https://doi.org/10.1093/gbe/evq053

Mencía, A., Modamio-Høybjør, S., Redshaw, N., Morín, M., Mayo-Merino, F., Olavarrieta, L., Aguirre, L. A., Del Castillo, I., Steel, K. P., Dalmay, T., Moreno, F., & Moreno-Pelayo, M. A. (2009). Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. *Nature Genetics 2009 41:5*, *41*(5), 609–613. https://doi.org/10.1038/ng.355

Mendes, N. D., Freitas, A. T., & Sagot, M.-F. (2009). Current tools for the identification of miRNA genes and their targets. *Nucleic Acids Research*, *37*(8), 2419–2433. https://doi.org/10.1093/nar/gkp145

Mitchell, P. S., Parkin, R. K., Kroh, E. M., Fritz, B. R., Wyman, S. K., Pogosova-Agadjanyan, E. L., Peterson, A., Noteboom, J., O'Briant, K. C., Allen, A., Lin, D. W., Urban, N., Drescher, C. W., Knudsen, B. S., Stirewalt, D. L., Gentleman, R., Vessella, R. L., Nelson, P. S., Martin, D. B., & Tewari, M. (2008). Circulating microRNAs as stable blood-based markers for cancer detection. *Proceedings of the National Academy of Sciences*, *105*(30), 10513–10518. https://doi.org/10.1073/pnas.0804549105

Mohammadi-Yeganeh, S., Paryan, M., Mirab Samiee, S., Soleimani, M., Arefian, E., Azadmanesh, K., Mostafavi, E., Mahdian, R., & Karimipoor, M. (2013). Development of a robust, low cost stem-loop real-time quantification PCR technique for miRNA expression analysis. *Molecular Biology Reports*, *40*(5), 3665–3674. https://doi.org/10.1007/S11033-012-2442-X/TABLES/5

Morin, R. D., O'Connor, M. D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A. L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M., Eaves, C. J., & Marra, M. A. (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Research*, *18*(4), 610–621. https://doi.org/10.1101/GR.7179508

Morlando, M., Ballarino, M., Gromak, N., Pagano, F., Bozzoni, I., & Proudfoot, N. J. (2008). Primary microRNA transcripts are processed co-transcriptionally. *Nature Structural & Molecular Biology*, *15*(9), 902–909. https://doi.org/10.1038/NSMB.1475

Nedelcu, C. (2013). *Nginx HTTP server (2nd ed)*. http://117.3.71.125:8080/dspace/handle/DHKTDN/7049

Ng, K. L. S., & Mishra, S. K. (2007). De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics*, *23*(11), 1321–1330. https://doi.org/10.1093/bioinformatics/btm026

Niu, Y., Zhang, L., Qiu, H., Wu, Y., Wang, Z., Zai, Y., Liu, L., Qu, J., Kang, K., & Gou, D. (2015). An improved method for detecting circulating microRNAs with S-Poly(T) Plus real-time PCR. *Scientific Reports 2015 5:1*, *5*(1), 1–10. https://doi.org/10.1038/srep15100

O'Brien, J., Hayder, H., Zayed, Y., & Peng, C. (2018). Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. *Frontiers in Endocrinology*, *9*, 402. https://doi.org/10.3389/fendo.2018.00402

Oishi, M., Sugiyama, S., Oishi, M., & Sugiyama, S. (2016). An Efficient Particle-Based DNA Circuit System: Catalytic Disassembly of DNA/PEG-Modified Gold Nanoparticle–Magnetic Bead Composites for Colorimetric Detection of miRNA. *Small*, *12*(37), 5153–5158. https://doi.org/10.1002/SMLL.201601741

Ozenne, B., Subtil, F., & Maucort-Boulch, D. (2015). The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of Clinical Epidemiology*, *68*(8), 855–859. https://doi.org/10.1016/j.jclinepi.2015.02.010

Pall, G. S., Codony-Servat, C., Byrne, J., Ritchie, L., & Hamilton, A. (2007). Carbodiimide-mediated cross-linking of RNA to nylon membranes improves the detection of siRNA, miRNA and piRNA by northern blot. *Nucleic Acids Research*, *35*(8), e60–e60. https://doi.org/10.1093/NAR/GKM112

Pan, M., Liang, M., Sun, J., Liu, X., & Wang, F. (2018). Lighting Up Fluorescent Silver Clusters via Target-Catalyzed Hairpin Assembly for Amplified Biosensing. *Langmuir*, *34*(49), 14851–14857. https://doi.org/10.1021/ACS.LANGMUIR.8B01576/ASSET/IMAGES/LARGE/LA-2018-015768_0003.JPEG

Pandey, S., Agarwala, P., Jayaraj, G. G., Gargallo, R., & Maiti, S. (2015). The RNA Stem-Loop to G-Quadruplex Equilibrium Controls Mature MicroRNA Production inside the Cell. *Biochemistry*, *54*(48), 7067–7078. https://doi.org/10.1021/ACS.BIOCHEM.5B00574/ASSET/IMAGES/LARGE/BI-2015-005747_0009.JPEG

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & others. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Peng, W., Yu, S., Handler, A. M., Tu, Z., Saccone, G., Xi, Z., & Zhang, H. (2020). miRNA-1-3p is an early embryonic male sex-determining factor in the Oriental fruit fly Bactrocera dorsalis. *Nature Communications 2020 11:1*,

11(1), 1–11. https://doi.org/10.1038/s41467-020-14622-4

Persano, S., Guevara, M. L., Wolfram, J., Blanco, E., Shen, H., Ferrari, M., & Pompa, P. P. (2016). Label-Free Isothermal Amplification Assay for Specific and Highly Sensitive Colorimetric miRNA Detection. *ACS Omega*, *1*(3), 448–455. https://doi.org/10.1021/ACSOMEGA.6B00109/ASSET/IMAGES/LARGE/AO-2016-00109Y_0005.JPEG

Qin, Y., & Hurley, L. H. (2008). Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions. *Biochimie*, *90*(8), 1149–1171. https://doi.org/10.1016/J.BIOCHI.2008.02.020

Qiu, X., Zhang, H., Yu, H., Jiang, T., & Luo, Y. (2015). Duplex-specific nuclease-mediated bioanalysis. *Trends in Biotechnology*, *33*(3), 180–188. https://doi.org/10.1016/J.TIBTECH.2014.12.008

Raad, J., Bugnon, L. A., Milone, D. H., & Stegmayer, G. (2022). MiRe2e: A full end-to-end deep model based on transformers for prediction of pre-miRNAs. *Bioinformatics*, *38*(5), 1191–1197. https://doi.org/10.1093/BIOINFORMATICS/BTAB823

Rahman, M. E., Islam, R., Islam, S., Mondal, S. I., & Amin, M. R. (2012). MiRANN: A reliable approach for improved classification of precursor microRNA using Artificial Neural Network model. *Genomics*, *99*(4), 189–194. https://doi.org/10.1016/j.ygeno.2012.02.001

Rahman, R.-U., Gautam, A., Bethune, J., Sattar, A., Fiosins, M., Magruder, D. S., Capece, V., Shomroni, O., & Bonn, S. (2018). Oasis 2: improved online analysis of small RNA-seq data. *BMC Bioinformatics*, *19*(1). https://doi.org/10.1186/s12859-018-2047-z

Ray, S. (2019). A Quick Review of Machine Learning Algorithms. *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019*, 35–39. https://doi.org/10.1109/COMITCON.2019.8862451

Raymond, C. K., Roberts, B. S., Garrett-Engele, P., Lim, L. P., & Johnson, J. M. (2005). Simple, quantitative primer-extension PCR assay for direct monitoring of microRNAs and short-interfering RNAs. *RNA*, *11*(11), 1737–1744. https://doi.org/10.1261/RNA.2148705

Riffo-Campos, Á. L., Riquelme, I., & Brebi-Mieville, P. (2016). Tools for Sequence-Based miRNA Target Prediction: What to Choose? *International Journal of Molecular Sciences 2016, Vol. 17, Page 1987*, *17*(12), 1987. https://doi.org/10.3390/IJMS17121987

Romero-Cordoba, S. L., Salido-Guadarrama, I., Rodriguez-Dorantes, M., & Hidalgo-Miranda, A. (2014). miRNA biogenesis: Biological impact in the development of cancer. *Https://Doi.Org/10.4161/15384047.2014.955442*, *15*(11), 1444–1455. https://doi.org/10.4161/15384047.2014.955442

Ruvkun, G. B. (n.d.). The tiny RNA world. *Harvey Lectures*, *99*, 1–21. http://www.ncbi.nlm.nih.gov/pubmed/15984549

Saidi, R., Bouaguel, W., & Essoussi, N. (2019). Hybrid feature selection method based on the genetic algorithm and pearson correlation coefficient. *Studies in Computational Intelligence*, *801*, 3–24. https://doi.org/10.1007/978-3-030-02357-7_1/COVER

Salahandish, R., Ghaffarinejad, A., Omidinia, E., Zargartalebi, H., Majidzadeh-A, K., Naghib, S. M., & Sanati-Nezhad, A. (2018). Label-free ultrasensitive detection of breast cancer miRNA-21 biomarker employing electrochemical nano-genosensor based on sandwiched AgNPs in PANI and N-doped graphene. *Biosensors and Bioelectronics*, *120*, 129–136. https://doi.org/10.1016/J.BIOS.2018.08.025

Salmena, L., Poliseno, L., Tay, Y., Kats, L., & Pandolfi, P. P. (2011). A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language? *Cell*, *146*(3), 353–358. https://doi.org/10.1016/J.CELL.2011.07.014

Sethupathy, P., Corda, B., & Hatzigeorgiou, A. G. (2006). TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA*, *12*(2), 192–197. https://doi.org/10.1261/RNA.2239606

Shi, C., Liu, Q., Ma, C., & Zhong, W. (2014). Exponential strand-displacement amplification for detection of micrornas. *Analytical Chemistry*, *86*(1), 336–339. https://doi.org/10.1021/AC4038043/SUPPL_FILE/AC4038043_SI_001.PDF

Shi, R., & Chiang, V. L. (2005). Facile means for quantifying microRNA expression by real-time PCR. *BioTechniques*, *39*(4), 519–524. https://doi.org/10.2144/000112010/ASSET/IMAGES/LARGE/FIGURE3.JPEG

Singh, J., & Nagaraju, J. (2008). In silico prediction and characterization of microRNAs from red flour beetle (Tribolium castaneum). *Insect Molecular Biology*, *17*(4), 427–436. https://doi.org/10.1111/j.1365-2583.2008.00816.x

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. *AAAI Workshop - Technical Report*, *WS-06-06*, 24–29. https://doi.org/10.1007/11941439_114/COVER

Song, J., Li, W., Zhao, H., Gao, L., Fan, Y., & Zhou, S. (2018). The microRNAs

let 7 and mir 278 regulate insect metamorphosis and oogenesis by targeting the juvenile hormone early response gene Krüppel homolog. *Development (Cambridge)*, *145*(24). https://doi.org/10.1242/DEV.170670/264814/AM/MICRORNA-LET-7-AND-MIR-278-REGULATE-INSECT

Song, J., Li, W., Zhao, H., & Zhou, S. (2019). Clustered miR-2, miR-13a, miR-13b and miR-71 coordinately target Notch gene to regulate oogenesis of the migratory locust Locusta migratoria. *Insect Biochemistry and Molecular Biology*, *106*, 39–46. https://doi.org/10.1016/J.IBMB.2018.11.004

Song, J., & Zhou, S. (2019). Post-transcriptional regulation of insect metamorphosis and oogenesis. *Cellular and Molecular Life Sciences 2019 77:10*, *77*(10), 1893–1909. https://doi.org/10.1007/S00018-019-03361-5

Stegmayer, G., Di Persia, L. E., Rubiolo, M., Gerard, M., Pividori, M., Yones, C., Bugnon, L. A., Rodriguez, T., Raad, J., & Milone, D. H. (2019). Predicting novel microRNA: a comprehensive comparison of machine learning approaches. *Briefings in Bioinformatics*, *20*(5), 1607–1620. https://doi.org/10.1093/bib/bby037

Stegmayer, G., Yones, C., Kamenetzky, L., & Milone, D. H. (2017). High Class-Imbalance in pre-miRNA Prediction: A Novel Approach Based on deepSOM. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *14*(6). https://doi.org/10.1109/TCBB.2016.2576459

Sun, X. H., Xu, N., Xu, Y., Zhou, D., Sun, Y., Wang, W. J., Ma, L., Zhu, C. L., & Shen, B. (2019). A novel miRNA, miR-13664, targets CpCYP314A1 to regulate deltamethrin resistance in Culex pipiens pallens. *Parasitology*, *146*(2), 197–205. https://doi.org/10.1017/S0031182018001002

T Mitchell, B Buchanan, G DeJong, T Dietterich, P Rosenbloom, and, & Waibel, A. (2003). Machine Learning. *Https://Doi.Org/10.1146/Annurev.Cs.04.060190.002221*, *4*(1), 417–433. https://doi.org/10.1146/ANNUREV.CS.04.060190.002221

Tang, C. K., Tsai, C. H., Wu, C. P., Lin, Y. H., Wei, S. C., Lu, Y. H., Li, C. H., & Wu, Y. L. (2021). MicroRNAs from Snellenius manilae bracovirus regulate innate and cellular immune responses of its host Spodoptera litura. *Communications Biology 2021 4:1*, *4*(1), 1–11. https://doi.org/10.1038/s42003-020-01563-3

Tariq, K., Peng, W., Saccone, G., & Zhang, H. (2016). Identification, characterization and target gene analysis of testicular microRNAs in the oriental fruit fly Bactrocera dorsalis. *Insect Molecular Biology*, *25*(1), 32–43. https://doi.org/10.1111/imb.12196

Tariq, Kaleem, Metzendorf, C., Peng, W., Sohail, S., & Zhang, H. (2016a). miR-8-3p regulates mitoferrin in the testes of Bactrocera dorsalis to ensure normal spermatogenesis. *Scientific Reports 2016 6:1*, *6*(1), 1–9. https://doi.org/10.1038/srep22565

Tariq, Kaleem, Metzendorf, C., Peng, W., Sohail, S., & Zhang, H. (2016b). miR-8-3p regulates mitoferrin in the testes of Bactrocera dorsalis to ensure normal spermatogenesis. *Scientific Reports*, *6*(1), 22565. https://doi.org/10.1038/srep22565

Thum, T., Galuppo, P., Wolf, C., Fiedler, J., Kneitz, S., van Laake, L. W., Doevendans, P. A., Mummery, C. L., Borlak, J., Haverich, A., Gross, C., Engelhardt, S., Ertl, G., & Bauersachs, J. (2007). MicroRNAs in the Human Heart. *Circulation*, *116*(3), 258–267. https://doi.org/10.1161/CIRCULATIONAHA.107.687947

Tian, W., Li, P., He, W., Liu, C., & Li, Z. (2019). Rolling circle extension-actuated loop-mediated isothermal amplification (RCA-LAMP) for ultrasensitive detection of microRNAs. *Biosensors and Bioelectronics*, *128*, 17–22. https://doi.org/10.1016/J.BIOS.2018.12.041

Tiefenau, C., Von Zezschwitz, E., Häring, M., Krombholz, K., & Smith, M. (2019). A usability evaluation of let's encrypt and CertBot: Usable security done right. *Proceedings of the ACM Conference on Computer and Communications Security*, 1971–1988. https://doi.org/10.1145/3319535.3363220

Torres, A. G., Fabani, M. M., Vigorito, E., & Gait, M. J. (2011). MicroRNA fate upon targeting with anti-miRNA oligonucleotides as revealed by an improved Northern-blot-based method for miRNA detection. *RNA*, *17*(5), 933–943. https://doi.org/10.1261/RNA.2533811

Tran, V. D. T., Tempel, S., Zerath, B., Zehraoui, F., & Tahi, F. (2015). miRBoost: boosting support vector machines for microRNA precursor classification. *RNA*, *21*(5), 775–785. https://doi.org/10.1261/RNA.043612.113

Troczka, B., Zimmer, C. T., Elias, J., Schorn, C., Bass, C., Davies, T. G. E., Field, L. M., Williamson, M. S., Slater, R., & Nauen, R. (2012). Resistance to diamide insecticides in diamondback moth, Plutella xylostella (Lepidoptera: Plutellidae) is associated with a mutation in the membrane-spanning domain of the ryanodine receptor. *Insect Biochemistry and Molecular Biology*, *42*(11), 873–880. https://doi.org/10.1016/J.IBMB.2012.09.001

Varkonyi-Gasic, E., Wu, R., Wood, M., Walton, E. F., & Hellens, R. P. (2007). Protocol: A highly sensitive RT-PCR method for detection and quantification of microRNAs. *Plant Methods*, *3*(1), 1–12.

https://doi.org/10.1186/1746-4811-3-12/FIGURES/7

Vejnar, C. E., & Zdobnov, E. M. (2012). miRmap: Comprehensive prediction of microRNA target repression strength. *Nucleic Acids Research*, *40*(22), 11673–11683. https://doi.org/10.1093/nar/gks901

Ventura, A., & Jacks, T. (2009). MicroRNAs and Cancer: Short RNAs Go a Long Way. *Cell*, *136*(4), 586–591. https://doi.org/10.1016/j.cell.2009.02.005

Vlachos, I. S., Paraskevopoulou, M. D., Karagkouni, D., Georgakilas, G., Vergoulis, T., Kanellos, I., Anastasopoulos, I. L., Maniou, S., Karathanou, K., Kalfakakou, D., Fevgas, A., Dalamagas, T., & Hatzigeorgiou, A. G. (2015). DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Research*, *43*(D1), D153–D159. https://doi.org/10.1093/NAR/GKU1215

Wang, Xiaowei. (2008). miRDB: A microRNA target prediction and functional annotation database with a wiki interface. *RNA*, *14*(6), 1012–1017. https://doi.org/10.1261/RNA.965408

Wang, Xingliang, & Wu, Y. (2012). High Levels of Resistance to Chlorantraniliprole Evolved in Field Populations of Plutella xylostella. *Journal of Economic Entomology*, *105*(3), 1019–1023. https://doi.org/10.1603/EC12059

Weber, J. A., Baxter, D. H., Zhang, S., Huang, D. Y., Huang, K. H., Lee, M. J., Galas, D. J., & Wang, K. (2010). The microRNA spectrum in 12 body fluids. *Clinical Chemistry*, *56*(11), 1733–1741. https://doi.org/10.1373/CLINCHEM.2010.147405

Wei, X., Zheng, C., Peng, T., Pan, Y., Xi, J., Chen, X., Zhang, J., Yang, S., Gao, X., & Shang, Q. (2016). miR-276 and miR-3016-modulated expression of acetyl-CoA carboxylase accounts for spirotetramat resistance in Aphis gossypii Glover. *Insect Biochemistry and Molecular Biology*, *79*, 57–65. https://doi.org/10.1016/J.IBMB.2016.10.011

Wienholds, E., Koudijs, M. J., Van Eeden, F. J. M., Cuppen, E., & Plasterk, R. H. A. (2003). The microRNA-producing enzyme Dicer1 is essential for zebrafish development. *Nature Genetics*, *35*(3), 217–218. https://doi.org/10.1038/ng1251

Wienholds, E., & Plasterk, R. H. A. (2005). MicroRNA function in animal development. *FEBS Letters*, *579*(26), 5911–5922. https://doi.org/10.1016/j.febslet.2005.07.070

Wightman, B., Ha, I., & Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in

C. elegans. *Cell*, *75*(5), 855–862. https://doi.org/10.1016/0092-8674(93)90530-4

Willmann, M. R., & Poethig, R. S. (2007). Conservation and evolution of miRNA regulatory programs in plant development. *Current Opinion in Plant Biology*, *10*(5), 503–511. https://doi.org/10.1016/J.PBI.2007.07.004

Xavier Leitão, B., Guedes, Á. L. V., & Colcher, S. (2020). Toward Web Templates Support in Nested Context Language. *Communications in Computer and Information Science*, *1202 CCIS*, 16–30. https://doi.org/10.1007/978-3-030-56574-9_2

Xu, H., Wu, D., Zhang, Y., Shi, H., Ouyang, C., Li, F., Jia, L., Yu, S., & Wu, Z. S. (2018). RCA-enhanced multifunctional molecule beacon-based strand-displacement amplification for sensitive microRNA detection. *Sensors and Actuators B: Chemical*, *258*, 470–477. https://doi.org/10.1016/J.SNB.2017.09.050

Xu, H., Zhang, Y., Zhang, S., Sun, M., Li, W., Jiang, Y., & Wu, Z. S. (2019). Ultrasensitive assay based on a combined cascade amplification by nicking-mediated rolling circle amplification and symmetric strand-displacement amplification. *Analytica Chimica Acta*, *1047*, 172–178. https://doi.org/10.1016/J.ACA.2018.10.004

Xu, J.-H., Li, F., & Sun, Q.-F. (2008). Identification of MicroRNA Precursors with Support Vector Machine and String Kernel. *Genomics, Proteomics & Bioinformatics*, *6*(2), 121–128. https://doi.org/10.1016/S1672-0229(08)60027-3

Xu, Y., Zhou, X., & Zhang, W. (2008). MicroRNA prediction with a novel ranking algorithm based on random walks. *Bioinformatics*, *24*(13), i50–i58. https://doi.org/10.1093/bioinformatics/btn175

Xue, C., Li, F., He, T., Liu, G.-P., Li, Y., & Zhang, X. (2005). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, *6*(1), 310. https://doi.org/10.1186/1471-2105-6-310

Yang, H., Yang, W., Dao, F.-Y., Lv, H., Ding, H., Chen, W., & Lin, H. (2020). A comparison and assessment of computational method for identifying recombination hotspots in Saccharomyces cerevisiae. *Briefings in Bioinformatics*, *21*(5), 1568–1580. https://doi.org/10.1093/bib/bbz123

Yang, J. H., Li, J. H., Shao, P., Zhou, H., Chen, Y. Q., & Qu, L. H. (2011). starBase: a database for exploring microRNA–mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic Acids Research*, *39*(suppl_1), D202–D209. https://doi.org/10.1093/NAR/GKQ1056

Yang, J., Tang, M., Diao, W., Cheng, W., Zhang, Y., & Yan, Y. (2016). Electrochemical strategy for ultrasensitive detection of microRNA based on MNAzyme-mediated rolling circle amplification on a gold electrode. *Microchimica Acta*, *183*(11), 3061–3067. https://doi.org/10.1007/S00604-016-1958-5/FIGURES/4

Yang, Q., Qiu, C., Yang, J., Wu, Q., & Cui, Q. (2011). miREnvironment Database: providing a bridge for microRNAs, environmental factors and phenotypes. *Bioinformatics*, *27*(23), 3329–3330. https://doi.org/10.1093/BIOINFORMATICS/BTR556

Ye, J., Xu, M., Tian, X., Cai, S., & Zeng, S. (2019). Research advances in the detection of miRNA. *Journal of Pharmaceutical Analysis*, *9*(4), 217–226. https://doi.org/10.1016/J.JPHA.2019.05.004

Yi, R., Qin, Y., Macara, I. G., & Cullen, B. R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes & Development*, *17*(24), 3011–3016. https://doi.org/10.1101/GAD.1158803

Yin, C., Li, M., Hu, J., Lang, K., Chen, Q., Liu, J., Guo, D., He, K., Dong, Y., Luo, J., Song, Z., Walters, J. R., Zhang, W., Li, F., & Chen, X. (2018). The genomic features of parasitism, Polyembryony and immune evasion in the endoparasitic wasp Macrocentrus cingulum. *BMC Genomics*, *19*(1), 420. https://doi.org/10.1186/s12864-018-4783-x

Zalucki, M. P., Shabbir, A., Silva, R., Adamson, D., Liu, S. S., & Furlong, M. J. (2012). Estimating the Economic Cost of One of the World's Major Insect Pests, Plutella xylostella (Lepidoptera: Plutellidae): Just How Long Is a Piece of String? *Journal of Economic Entomology*, *105*(4), 1115–1129. https://doi.org/10.1603/EC12107

Zen, K., & Zhang, C. Y. (2012). Circulating microRNAs: a novel class of biomarkers to diagnose and monitor human cancers. *Medicinal Research Reviews*, *32*(2), 326–348. https://doi.org/10.1002/MED.20215

Zeng, Y., & Cullen, B. R. (2003). Sequence requirements for micro RNA processing and function in human cells. *RNA (New York, N.Y.)*, *9*(1), 112–123. https://doi.org/10.1261/RNA.2780503

Zeng, Y., & Cullen, B. R. (2004). Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. *Nucleic Acids Research*, *32*(16), 4776–4785. https://doi.org/10.1093/NAR/GKH824

Zhang, J., Dong, Y., Wang, M., Wang, H., Yi, D., Zhou, Y., & Xu, Q. (2021). MicroRNA-315-5p promotes rice black-streaked dwarf virus infection by targeting a melatonin receptor in the small brown planthopper. *Pest Management Science*, *77*(7), 3561–3570. https://doi.org/10.1002/PS.6410

Zhang, Q., Dou, W., Pan, D., Chen, E.-H., Niu, J.-Z., Smagghe, G., & Wang, J.-J. (2019). Genome-Wide Analysis of MicroRNAs in Relation to Pupariation in Oriental Fruit Fly. *Frontiers in Physiology*, *10*(MAR), 301. https://doi.org/10.3389/fphys.2019.00301

Zhang, Q., Dou, W., Taning, C. N. T., Smagghe, G., & Wang, J. J. (2021). Regulatory roles of microRNAs in insect pests: prospective targets for insect pest control. *Current Opinion in Biotechnology*, *70*, 158–166. https://doi.org/10.1016/J.COPBIO.2021.05.002

Zhang, T., & Yang, B. (2016). Big Data Dimension Reduction Using PCA. *Proceedings - 2016 IEEE International Conference on Smart Cloud, SmartCloud 2016*, 152–157. https://doi.org/10.1109/SMARTCLOUD.2016.33

Zhang, X., Zheng, Y., Jagadeeswaran, G., Ren, R., Sunkar, R., & Jiang, H. (2014). Identification of conserved and novel microRNAs in Manduca sexta and their possible roles in the expression regulation of immunity-related genes. *Insect Biochemistry and Molecular Biology*, *47*, 12–22. https://doi.org/10.1016/j.ibmb.2014.01.008

Zhang, Y., Feng, K., Hu, J., Shi, L., Wei, P., Xu, Z., Shen, G., Li, M., Xu, Q., & He, L. (2018). A microRNA-1 gene, tci-miR-1-3p , is involved in cyflumetofen resistance by targeting a glutathione S-transferase gene, TCGSTM4 , in Tetranychus cinnabarinus. *Insect Molecular Biology*, *27*(3), 352–364. https://doi.org/10.1111/imb.12375

Zhang, Yang, Zhao, B., Roy, S., Saha, T. T., Kokoza, V. A., Li, M., & Raikhel, A. S. (2016). microRNA-309 targets the Homeobox gene SIX4 and controls ovarian development in the mosquito Aedes aegypti. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(33), E4828-36. https://doi.org/10.1073/pnas.1609792113

Zhao, W., Yu, J., Jiang, F., Wang, W., Kang, L., & Cuiid, F. (2021). Coordination between terminal variation of the viral genome and insect microRNAs regulates rice stripe virus replication in insect vectors. *PLOS Pathogens*, *17*(3), e1009424. https://doi.org/10.1371/JOURNAL.PPAT.1009424

Zhu, B., Sun, X., Nie, X., Liang, P., & Gao, X. (2020). MicroRNA-998–3p contributes to Cry1Ac-resistance by targeting ABCC2 in lepidopteran insects. *Insect Biochemistry and Molecular Biology*, *117*, 103283. https://doi.org/10.1016/J.IBMB.2019.103283

Zubakov, D., Boersma, A. W. M., Choi, Y., Van Kuijk, P. F., Wiemer, E. A. C., & Kayser, M. (2010). MicroRNA markers for forensic body fluid identification obtained from microarray screening and quantitative RT-PCR confirmation. *International Journal of Legal Medicine*, *124*(3), 217–226.

https://doi.org/10.1007/S00414-009-0402-3

Publications:

- **Adhiraj Nath**, & Utpal Bora. “RNAinsecta: A tool for prediction of precursor microRNA in insects and search for their target in the model organism *Drosophila melanogaster*”. PloS one (2023) doi:10.1371/journal.pone.0287323.

- **Adhiraj Nath**, & Utpal Bora. “Comparative analysis of sequential and thermodynamic features of insect pre-miRNA with other groups of organisms”.(Ready for submission)

Collaboration:

- Kabiraj, D., Chetia, H., **Nath, A.**, Sharma, P., Mosahari, P. V., Singh, D., ... & Bora, U. (2022). Mitogenome-wise codon usage pattern from comparative analysis of the first mitogenome of Blepharipa sp.(Muga uzifly) with other Oestroid flies. *Scientific Reports*, *12*(1), 1-33.

- Dharitri Saikia, Pulakeswar Basumatary, **Adhiraj Nath**, Jon Jyoti Kalita, Kartik Neog, Mihir Kumar Purkait, Utpal Bora**.** "Metagenomic and metatranscriptomic insights into the structure and function of gut microbial community of *Antheraea assamensis* Helfer".

Conferences:

- Poster presentation at SeriTech 2017, IIT Guwahati.
- Submitted abstract and poster at the conference on “Evolutionary Emergence of Life Cycles”, Max Planck Institute for Evolutionary Biology, Germany; October 2018