



# DBT-Sponsored Workshop on AI in Modern Biology

(23<sup>rd</sup> – 25<sup>th</sup> August 2022)

**Theme: AI in Agriculture**

**Artificial Intelligence and Genomics – A blended approach for unravelling some  
underlying complex trait phenomena in plants and animals**

**(A.R.Rao<sup>1</sup> and Sarika Sahu<sup>2</sup>)**

<sup>1</sup>Indian Council of Agricultural Research (ICAR), New Delhi  
[adg.pim@icar.gov.in](mailto:adg.pim@icar.gov.in); [rao.cshl.work@gmail.com](mailto:rao.cshl.work@gmail.com)

<sup>2</sup>ICAR-Indian Agricultural Statistics Research Institute, New Delhi

# Artificial Intelligence – Landmarks

First two robots:  
Elmer & Elsie




First AI program  
- ELIZA



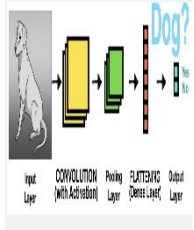
Commercialized AI:  
the expert systems



A.I. Movie



CNN

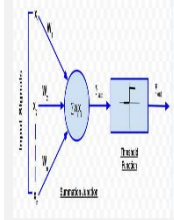


Turing Award  
– DNNs as critical components of computing



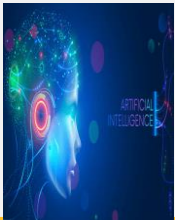
1943

First ANN (McCulloch & Pitt)




1948

Artificial intelligence term




1955

First mobile robot - Shakey




1964

Chess Victories




1970

The virtual assistant 'Siri'



1981

Google DeepMind's AlphaGo



1989-96

2001

2011

2012

2016

2019

# Artificial Intelligence

Enables computer to sense, reason, act and adapt

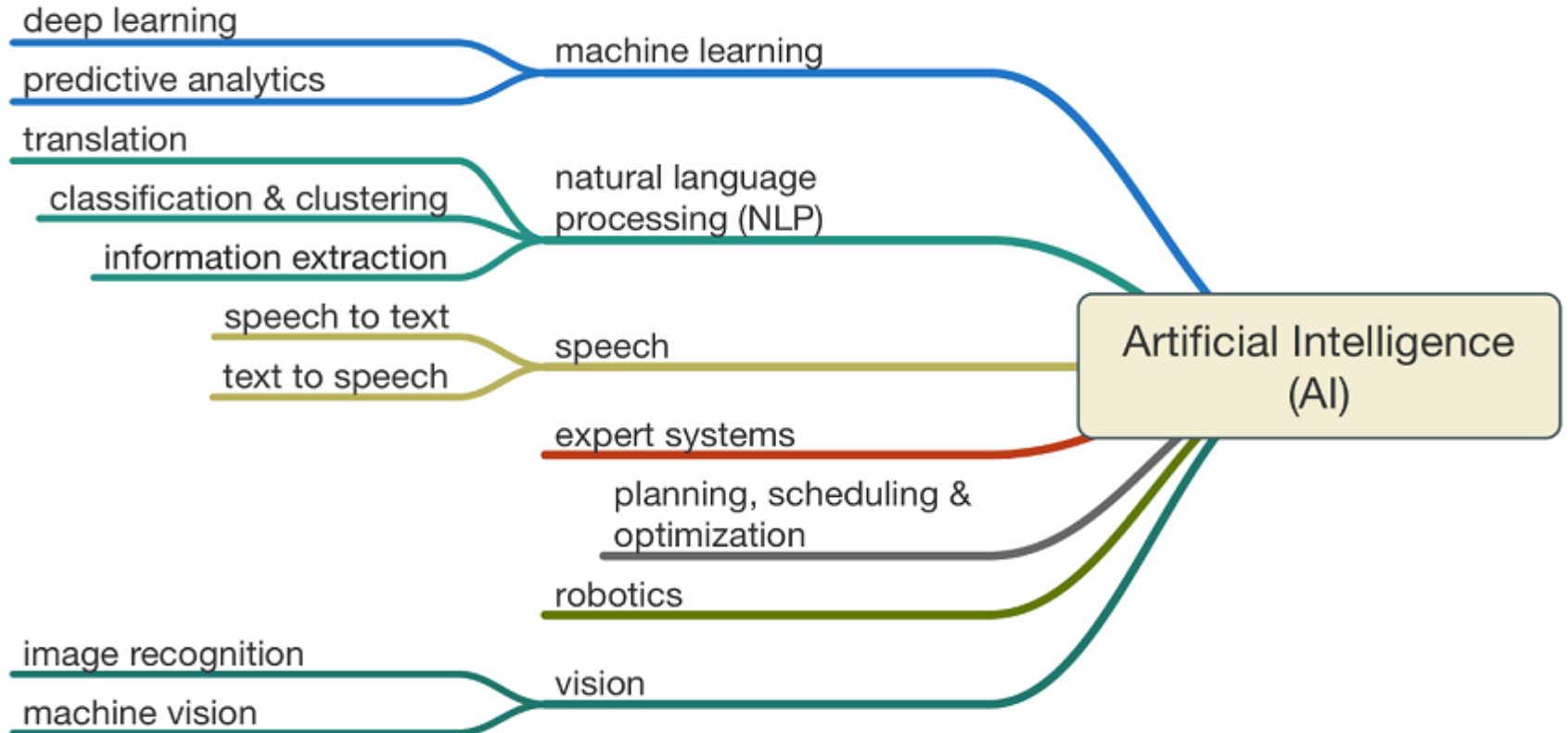
## Machine Learning

Statistical algorithms which learn from data

## Deep Learning

Multilayered NN learns from huge data

AI – An  
Infused  
Technique



# Genomics – Landmarks

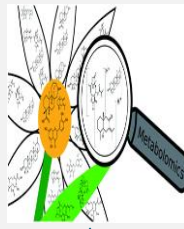
Hybrid  
Breeding of  
1<sup>st</sup>  
Commercial  
Maize



Green  
Revolution  
by Norman  
Borlaug



Plant  
Metabolomics

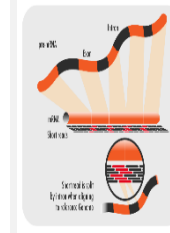


Development of  
NGS  
technologies  
(ePCR &  
Pyrosequencing)

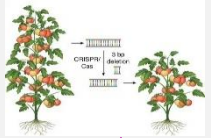
Next-Gen Sequencing  
Technologies



Transcripto  
mics by  
RNA-seq



1<sup>st</sup> Event of  
Plant Genome  
Editing via  
CRISPR/Cas9  
Multiplex  
Genome  
Editing



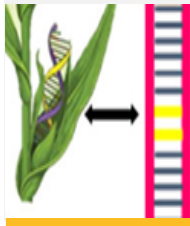
1866

Laws of  
Genetics by  
Gregor  
Johann  
Mendel



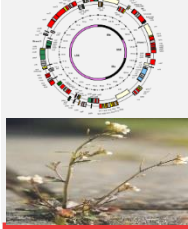
1920

First Report  
of Mutation  
Breeding in  
Maize



1928

Arabidopsis  
thaliana  
genome  
sequenced



1960

2000

SNP Marker  
Development,  
Rice Genome  
Sequenced,  
Golden Rice

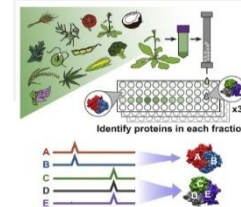


2001

2002

2005

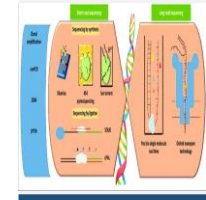
DNA-Protein  
interaction  
mapping using  
the ChIP-seq;  
Plant Pan-  
genomics &  
Genome  
Selection



2007

2008

Long read  
sequencing  
technologies  
[SMRT-seq &  
Nanopore]



2009

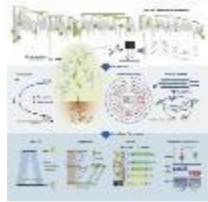
2013

Continued

# Genomics – Landmarks

Continued

Soyabean Pan-genome Assembly;  
Maize Pan-Transcriptome Construction;



2014

2017

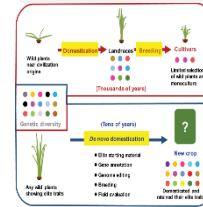
2018

2019

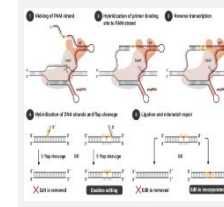
2020

Journey  
Contd...

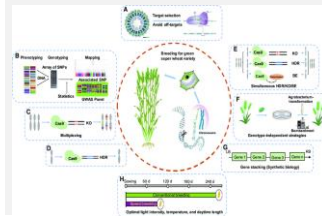
De-novo  
Domestication;  
Speed Breeding



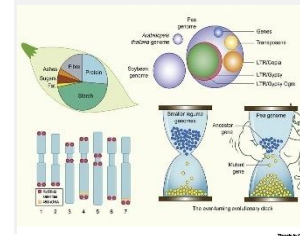
Prime Editing



DNA-free Editing & Base  
Editing in Wheat, Rice;  
Wheat genome Sequenced;  
Wheat Pan-genome  
Assembled



Mendel's Genetic Model  
Pea Genome Annotated



## -omics, AI and Complex Traits

**MAGIC Triangle**

**Traits**  
**Biotic Stress**  
**Abiotic Stress**  
**Nutritional**  
**Agronomic**  
**Biochemical**

**Complex Phenomena**

**Next Generation**  
**AI infused**  
**Integrated**  
**multiomics**

**Genomics**  
**Epigenomics**  
**Transcriptomics**  
**Proteomics**  
**Metabolomics**

+

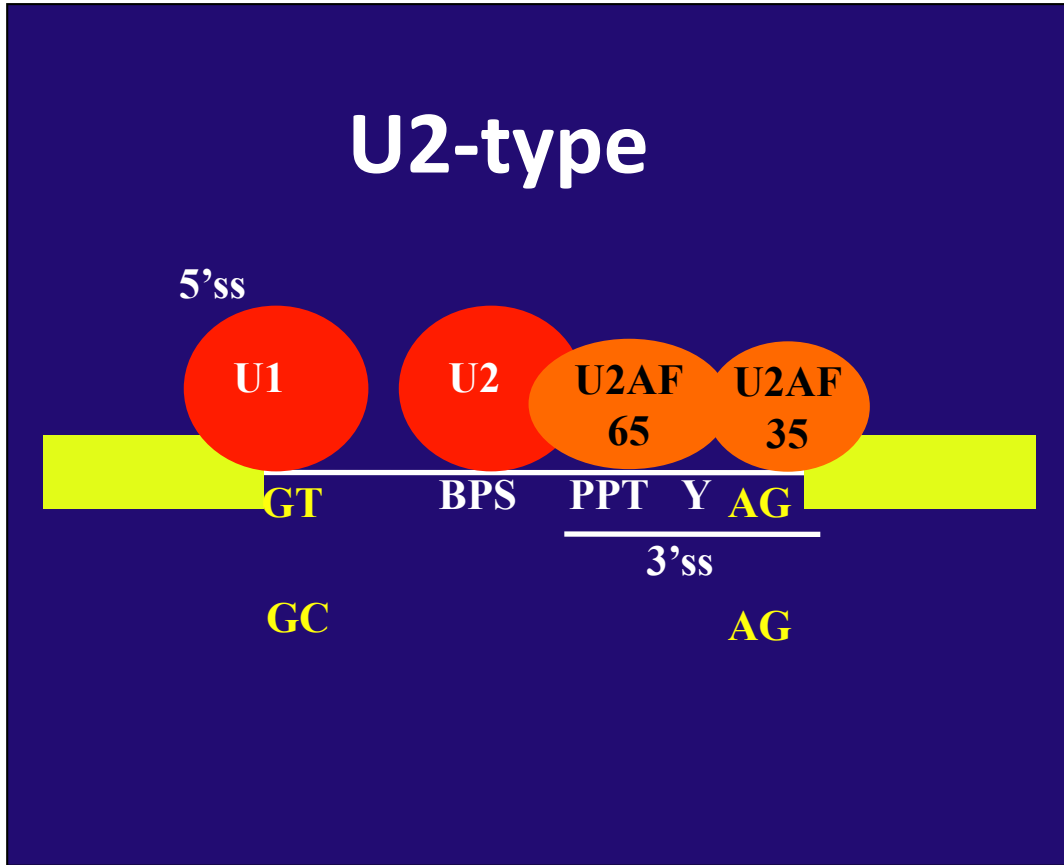
**Env. → Phenomics**

**AI**  
**Data Science**  
**Machine Learning**  
**Statistics**  
**Big Data**

- prediction of gene structure
  - Prediction of antimicrobial peptides
  - Discrimination of cRNAs from ncRNAs, and further classification of ncRNAs
  - Identification of nitrogen fixation genes, herbicide resistant genes, insecticide resistant proteins, heat shock proteins, etc.
- 
- DNA barcode based identification of microbial species
  - Identification of late blight susceptible genes
  - Prediction of multiple sub-cellular localization of genes
  - Abiotic stress responsive miRNA prediction
  - Spike recognition and counting from visual imaging
  - Genomic selection and prediction of genomic estimated breeding values.



# Gene Structure Prediction - Splicing



Position Weight Matrix

U2: GT\_AG 5'ss: Consensus Motif → CAG GTAAGT

$$\text{Splice Site motif score} = \sum_{i=1}^9 \log_2 \left( \frac{p_i}{0.25} \right)$$

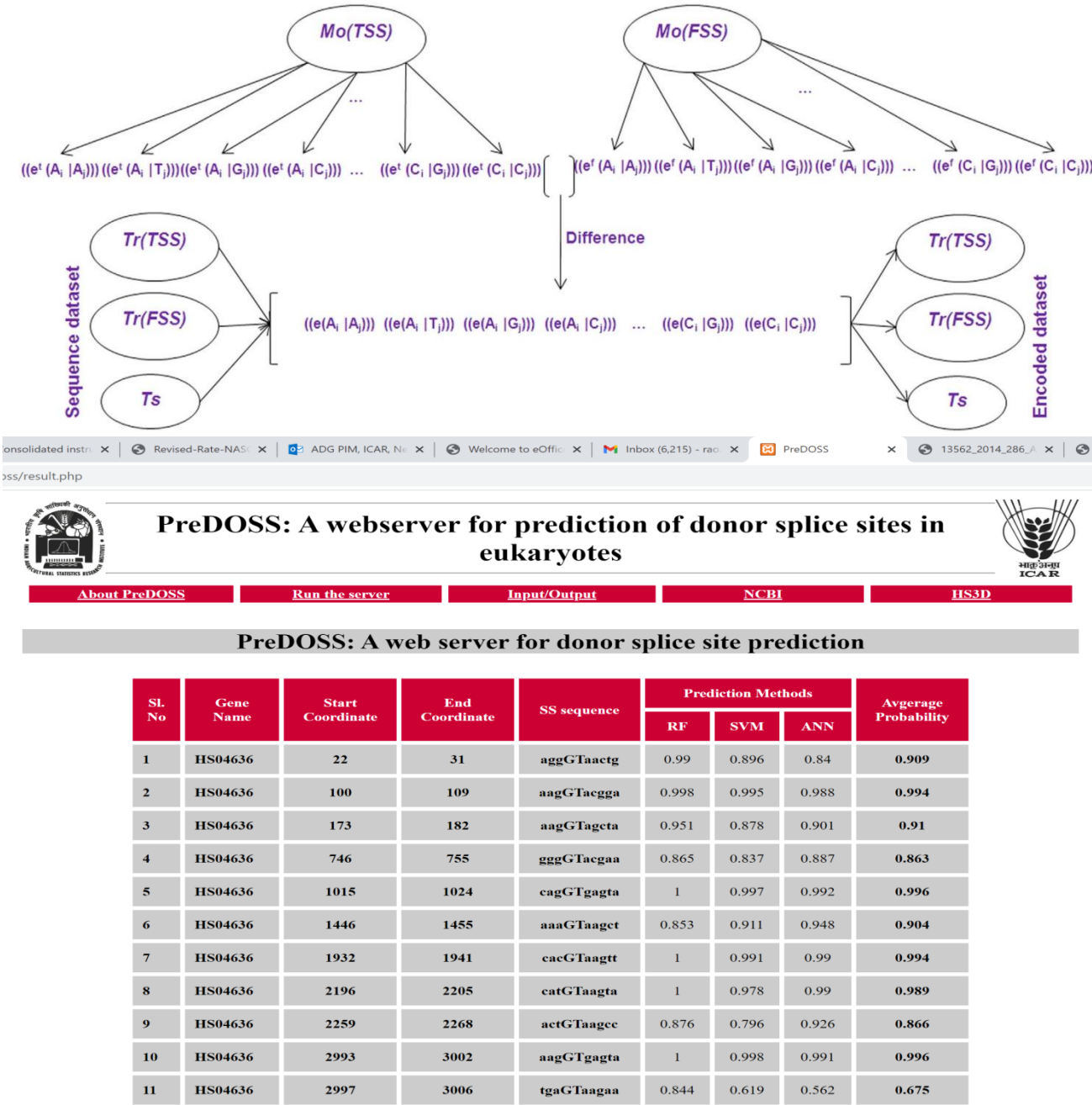
- U1 snRNP 5' splice site binding

	Exon	Intron
U1 snRNA:	G U C	C A U U C A
pre-mRNA sequence :	C A G	G T A A G T
Rank: 10		

	Exon	Intron
U1 snRNA:	G U C	C A U U C A
pre-mRNA sequence :	C A G	G T G A G G
Rank: 1		
		3 6

U1 snRNP splice site binding

- Research Gap: Are there mechanisms other than PWM exist?
- ML based approaches:
  - Determination of window size
  - Encoding of nucleotide dependencies into numeric vectors
  - Numeric vectors as input in ML classifiers for prediction of donor splice sites
    - Random Forest (RF)
    - Support Vector Machines (SVM)
    - Artificial Neural Network (ANN)
    - Bagging, Boosting
    - Logistic regression
    - kNN
    - Naïve Bayes classifiers
- Rice, Maize, Barley, Cattle, H3SD
- Highest prediction accuracy for SVM (balanced data) and ANN (imbalanced data)





# Improved Prediction of Antimicrobial Peptides

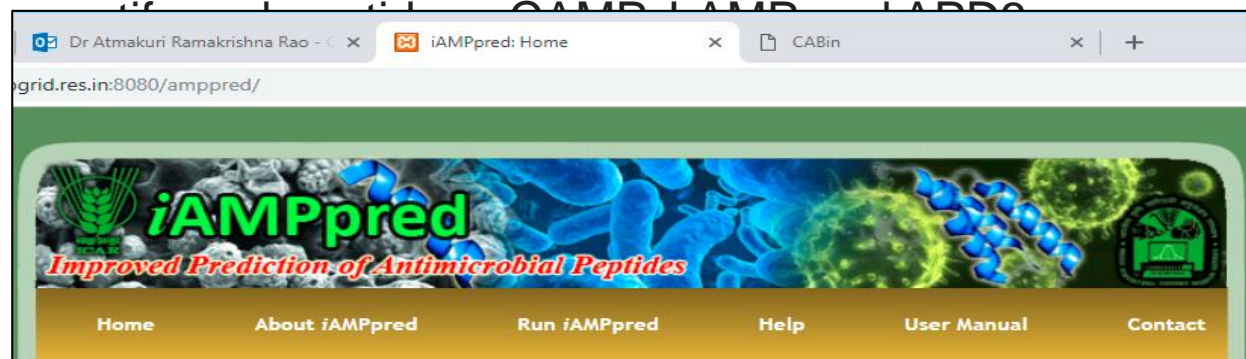
- AMPs gaining attention due to growing resistance of microbes against conventional antibiotics
- AMPs kill target cells without affecting host cells

## Research Gap

- Identification and designing of AMPs through wet lab experiments is resource intensive.
- *In silico* identification may supplement already identified and designed new antimicrobial agents.

## Data

antibacterial peptides - CAMP, APD3 and AntiBP2;  
antiviral peptides - CAMP, APD3, LAMP and  
AVPpred;



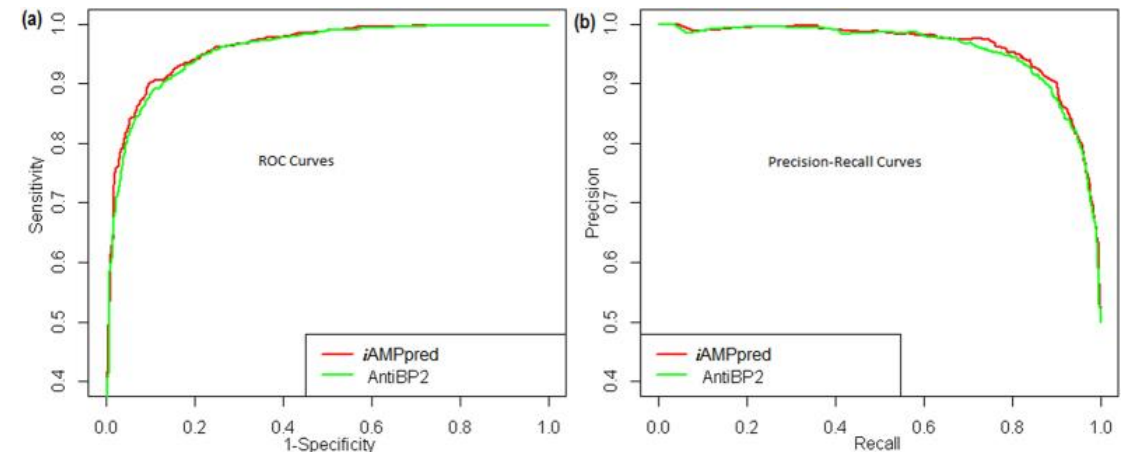
<http://cabgrid.res.in:8080/amppred/>

Compositional, Structural and Physico-chemical features (66)  
AAC, NAAC and PAAC

Alpha helix propensity, Beta-sheet propensity, Turn propensity  
Iso-electric point, Hydrophobicity, Net-charge

## SVM with Gaussian RBF

$$k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$$



## % Prediction accuracy

Anti-bacterial = 94.69

Anti-viral = 90.09

Anti-fungal = 93.35

Meher, P.K., Sahu, T.K., Saini, V. and Rao, A.R.(2017). *Scientific Reports* 7:42362,

# Multi-class classification of RNAs

- Binary classification of coding and non coding RNAs
- Multi-class classification of ncRNAs: snRNA, snoRNA, miRNA, lncRNA, circRNA, tRNA, rRNA, SRP
- Deep Neural Network (DNN), Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN)
- Transcript sequences of 63 plant species, covering cereals, pulses, oilseeds, fruits and forestry trees
- <https://plants.ensembl.org/info/data/ftp/index.html>
- PNRD, PlantCircBase, 5SRNAdb, CANTATAdb 2.0

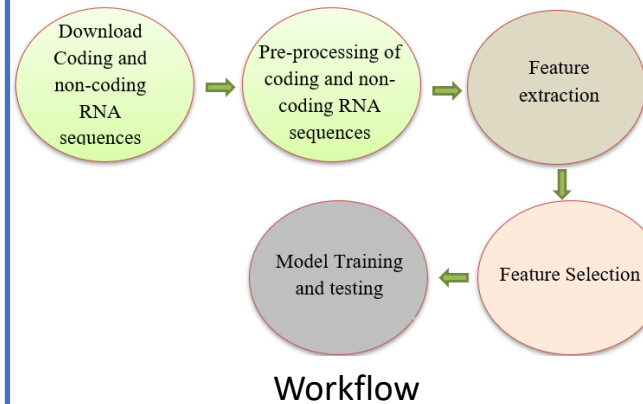
## Features

Transcript length, ORF length, ORF coverage

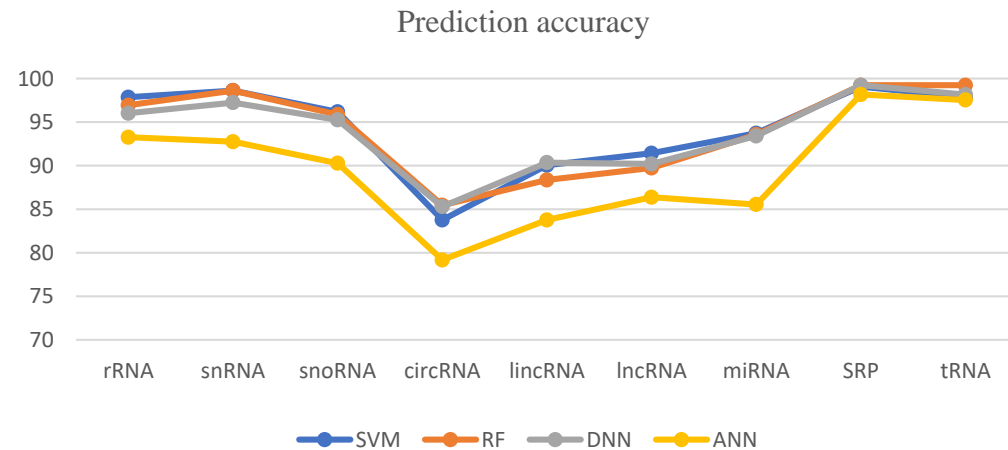
Peptide length, K-mer frequencies, BLAST features

Amino acid composition, Molecular weight,

Isoelectric point, GC%, Codon Bias Indices, RSCU



Method	Accuracy	Sensitivity	Specificity
RF (proposed)	99.803	99.7	99.9
LncRNA-ID (RF)	95.78	96.28	95.28
LncRNApred (RF)	94.3	95.27	93.48
PLncPRO (RF)	83-99.5	-	-
SVM (proposed)	97.364	97.9	92.2
CPC1 (SVM)	93.2	99.5	87.3
CPC2 (SVM)	96.1	95.2	97
DNN (proposed)	99.519	99.4	99.6
DeepLNC (DNN)	98.07	98.98	97.19



Comparison of multiclass classifiers based on performance metrics using independent test data

Method	Accuracy (%)	Sensitivity	Specificity	Precision	F1-score	MCC
SVM	94.283	0.762	0.966	0.795	0.770	0.741
RF	94.113	0.757	0.965	0.803	0.766	0.741
DNN	93.908	0.726	0.966	0.726	0.742	0.692
ANN	89.653	0.511	0.945	0.535	0.523	0.465

# Identification of nif Genes

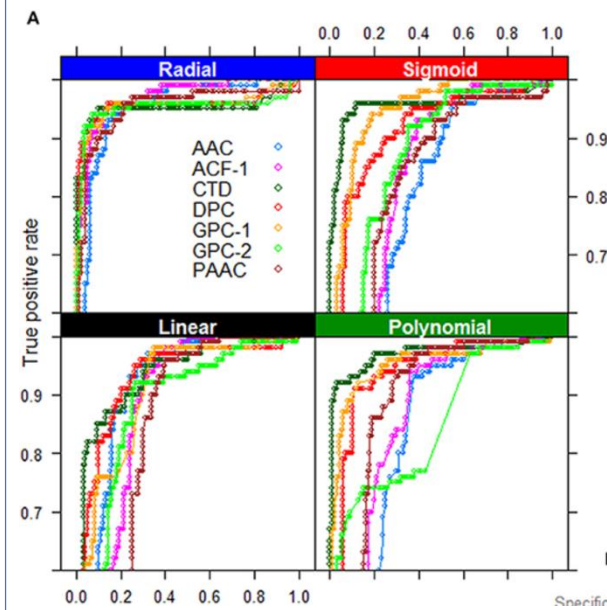
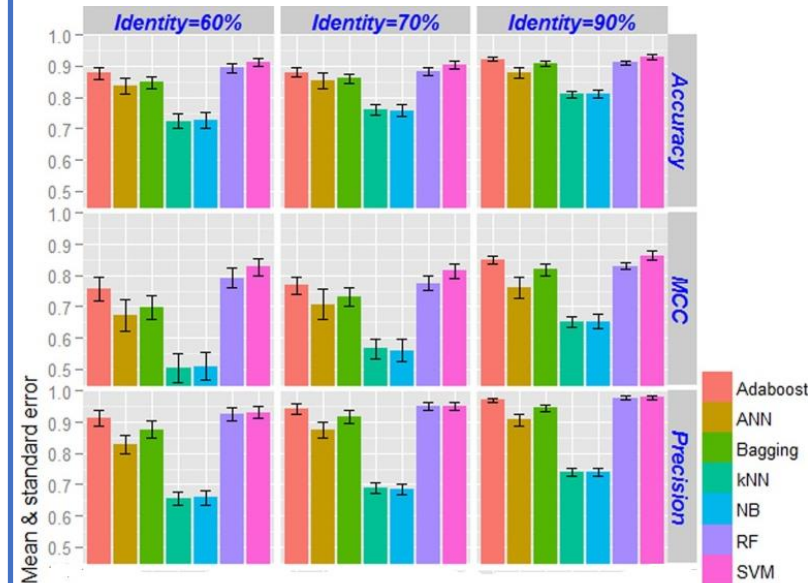
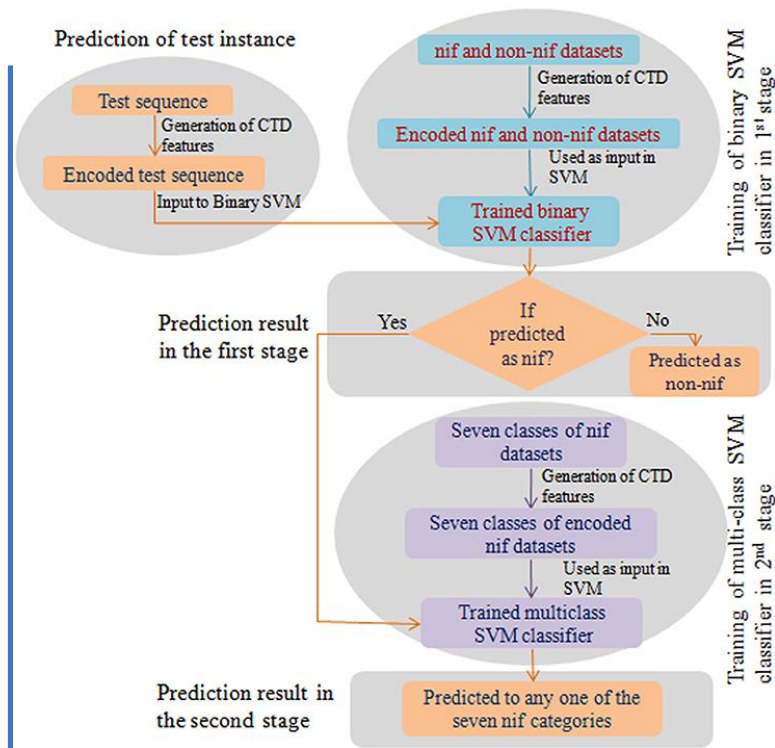
- Nitrogen fixing microbes depend on nitrogenase enzyme complex, consists of structural genes: nifH, nifD, nifK, nifH, nifE, nifN and nifB genes, essential in characterized systems (diazotrophs)
- Identification of nif proteins are essential
- SVM classifier + kernels (linear, polynomial, sigmoidal, radial)
- Binary classification for nif and non-nif proteins, multi-class classification for categorization of nif proteins

## Data

82 diazotrophs (UniProtKB); +ve and -ve datasets for training classifier

## Features

1. Amino acid composition (AAC)
2. Di-peptide composition (DPC)
3. Gap-pair composition (GPC)
4. Pseudo amino acid composition (PseAAC)
5. Composition-transition-distribution (CTD)
6. Auto-correlation function (ACF)



**A Test set-I**

	non-nif	nifB	nifD	nifE	nifH	nifK	nifN
non-nif	3	0	3	1	1	0	0
nifB	4	2	3	0	6	55	4
nifD	0	0	0	0	65	4	0
nifE	0	0	0	81	0	0	0
nifH	0	3	65	1	1	6	0
nifK	0	70	0	0	0	0	0
nifN	68	0	0	0	0	0	0

**B Test set-II**

	non-nif	nifB	nifD	nifE	nifH	nifK	nifN
non-nif	109	0	2	753	3	8	0
nifB	64	0	10	7	1	689	0
nifD	0	0	0	0	979	2	0
nifE	0	0	0	1973	0	0	0
nifH	0	0	979	3	0	36	0
nifK	0	1007	0	0	0	0	0
nifN	1304	0	0	1	0	0	0

Meher, P.K., Sahu, T.K., Mohanty, J., Gahoi, S., Purru, S., Grover, M. and Rao, A.R. (2018). nifPred: *Frontiers in Microbiology*, 9: 1100.



SREP-18 x Dr Atmakuri Ramakrishna Rao x HRGPred: Home x +

secure | cabgrid.res.in:8080/hrgpred/



**HRGPred**  
Prediction of herbicide resistant genes

Home Run HRGPred Algorithm Dataset Help Contact

**Home**

According to the herbicide resistance action committee, herbicide resistance is the inherent ability of plant species to survive and reproduce after exposed to a herbicidal dosage which is lethal to its wild types. Globally, the evolution of herbicide resistance has been a major cause of concern for sustainable agricultural production. Up to the end of 2016, around 477 herbicide resistant biotypes have been reported encompassing 252 weed species, where these biotypes have developed resistance to 23 of the 26 known herbicide sites of action and 161 different herbicides. The mechanism of herbicide resistance can be classified into two classes' (i) target site resistance, and (ii) non-target site resistance. The target site resistance is mainly due to the mutations in the genes

**Our other Prediction servers**

- dSSPpred
- MaLDoSS
- PreDOSS
- HSsplice
- SPIDBAR
- DCDNC
- iAMPpred

<http://cabgrid.res.in:8080/hrgpred/> *Scientific Reports* (2019), 9: 778.  
DOI:10.1038/s41598-018-37309-9



**DIRProt**

Home About DIRProt Run DIRProt Help Contact

**OUR OTHER PREDICTION SERVERS**

- dSSPpred
- MaLDoSS
- PreDOSS
- HSsplice
- SPIDBAR
- DCDNC

**OTHER USEFUL LINKS**

- NCBI

Paste the protein sequences in fasta format

```
>NP_476907.2
MFLVIGAILASALFVGLLLYHLKFKRLIDLIISYMPGPPVPLVGHGHHF IGKPPHEHVKKIFEFMETYSK
DQVLKVVWLGPELNVLMGNPKDVEVVLGTLRFNDKAGEYKALEPWLEGLLVSRGRKWHKRRKIITPAHF
KILDQFVEVFEKGSRDLLRNMEQDRLEKHGDSGFSLYDWINLCTMDTICETANGVSINAQSNADSEYVQAV
KTIISMVLHKRMFNILYRFDLT YMLTPLARAEEKALNVLHQFTEKIIIVQRREELIREGSSQESNDADVG
AKRKMFLDILLQSTVDERPLSNLDIREEDVTFMFEGHDTTSSALMFFFYNIATHPEAQKKCFEIRSVV
GNDKSTPVS YELLNLQHYVDLCVKETLRMYPSPVLLGRKVLDECEINGKLIPAGTNGIGISPLYLGRREEL
FSEPHIFKPERFDVVTAEKLN PYAYIPFSAGPRNCIGQKFAMLEIKAIVANVLRHYEVDVFGDSSEPPV
LIAELILRTKEPLMFKVRRVY
```

[Load Example Data](#) [Clear Textarea](#)

<http://cabgrid.res.in:8080/dirprot/> *BMC Bioinformatics* (2017) 18:190

x Dr Atmakuri Ramakrishna Rao x ir-HSP x CABIN x +

cabgrid.res.in:8080/ir-hsp/



**ir-HSP**  
improved recognition of HSPs and their families

Home Run ir-HSP Dataset Help Contact Us

**Home**

Heat shock proteins (HSPs) are one of the largest groups of molecular chaperones that assist in correct folding of partially folded or denatured proteins, establishment of proper protein conformation and prevention of unalterable aggregation of damaged proteins. Besides chaperonine activities, HSPs are also involved in other functions like modulation of their synthesis, participation in signal transduction pathways, RNA processing etc. HSPs also play vital role in maintaining the overall cellular protein homeostasis. Due to broad range of activities, they have received a considerable attention of the researchers. Keeping in view the wide range of functions of HSPs, we developed this server for prediction of HSPs, their families (HSP20, HSP40, HSP60, HSP70, HSP90 and HSP100) and sub-types of DnaJ proteins (Type I, Type II, Type III, Type IV). The ir-HSP achieved higher accuracy as compared to the existing approaches, and thus believed to supplement the existing efforts for annotation of protein sequences.




**Our other prediction servers**

- dSSPpred
- MaLDoSS
- PreDOSS
- HSsplice
- SPIDBAR
- DCDNC
- iAMPpred
- DIRprot

**Useful Links**

- NCBI
- HSPiR
- iHSP-PseRAAAC
- JPred

<http://cabgrid.res.in:8080/ir-hsp/> *Frontiers in Genetics: Bioinformatics and Computational Biology* (2018). 8, 235



**SPIDBAR**  
Species Identification using DNA Barcode

Home Help About Dataset

The problem of species identification using DNA Barcode can be formulated as : given a reference library composed of DNA Barcode specimen sequences of known species and an unknown DNA Barcode sequence, recognize the latter into a species that is present in the library. Several methods have been developed and adopted to automatically classify a DNA Barcode sequence to a predefined species, such as tree-based methods, similarity-based methods and diagnostic methods. However, each method has its own advantage and disadvantage. The SPIDBAR can be used for species identification using DNA Barcode with the help of Random Forest methodology. Here, initially the features vector has been developed on the basis of composition of frequency of k-mer of different size and RF supervised learning approach was employed for classification purpose. To run this server, the user has to provide the set of reference sequence with known species label (in BOLD format) and query sequence with hypothetical label (in BOLD format). Also, the user has to provide atleast two query sequence to run the SPIDBAR.

**Paste Reference Sequences**

OR

Upload Training file  No file chosen

**Paste Query Sequences**

OR

Upload Test file  No file chosen

Team: Prabina Kumar Mohar, Tanmay Kumar Sahu and A. R. Rao

Copyright © 2018 Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi - 110012. All rights reserved.

**TRAINING RESULT**

	SPECIES	NO.OF INDIVIDUALS OBSERVED	NO.OF INDIVIDUALS CORRECTLY PREDICTED
1	Ametrida_centurio	9	9
2	Anoura_caulifer	13	13
3	Anoura_geoffroyi	19	19
4	Anoura_latidens	4	4
5	Arthous_amphus	5	5
6	Arthous_rimansens	6	6

[Download Training Result](#)

**TEST RESULT**

	OBSERVED LABEL	PREDICTED LABEL
1	Ametrida_centurio	Ametrida_centurio
2	Anoura_caulifer	Anoura_caulifer
3	Anoura_caulifer	Anoura_caulifer
4	Anoura_geoffroyi	Anoura_geoffroyi
5	Anoura_geoffroyi	Anoura_geoffroyi
6	Anoura_geoffroyi	Anoura_geoffroyi

[Download Test Result](#)

Team: Prabina Kumar Mohar, Tanmay Kumar Sahu and A. R. Rao

Copyright © 2018 Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi - 110012. All rights reserved.

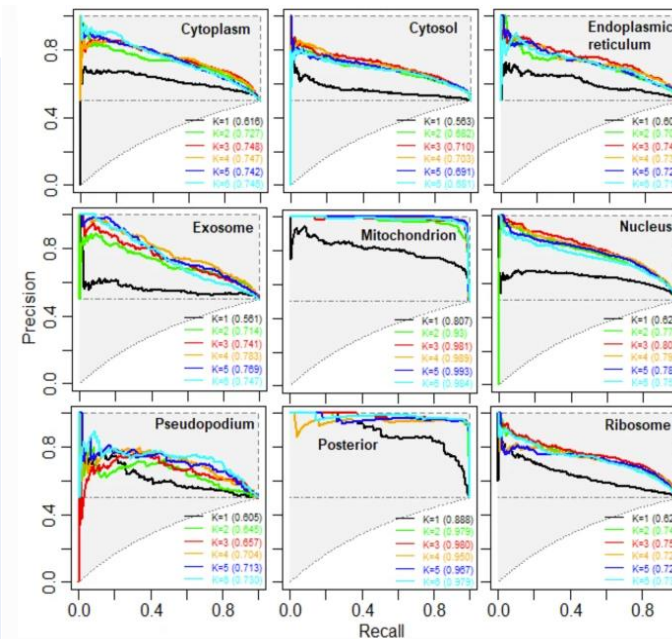
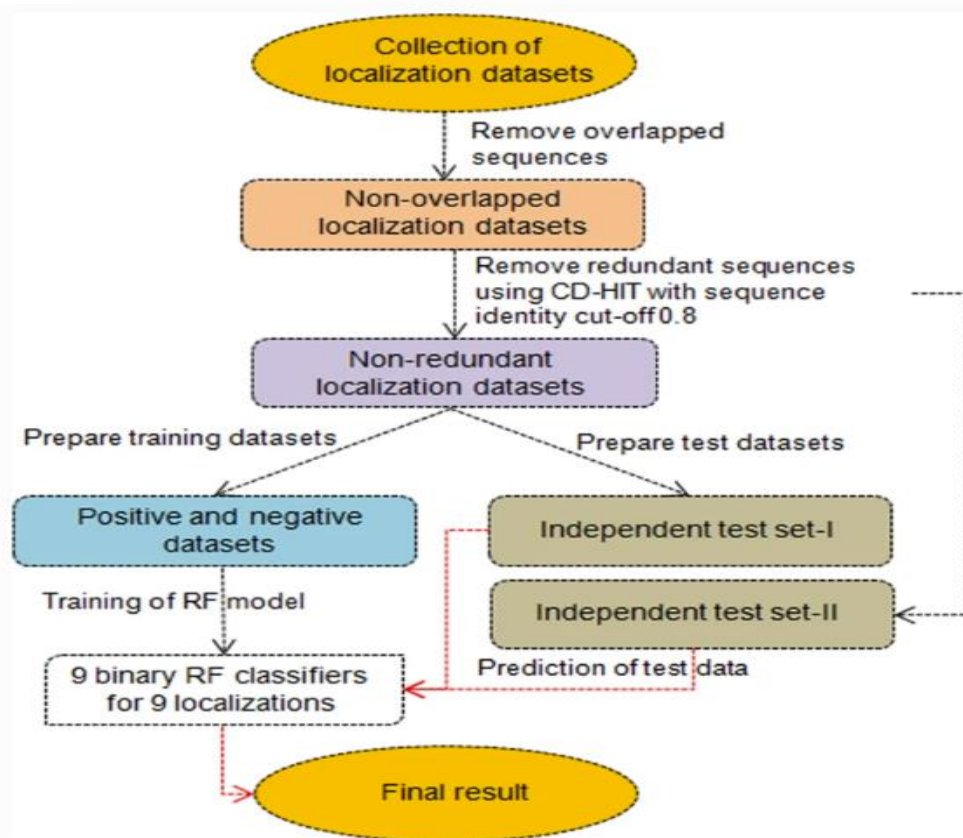
(a) <http://cabgrid.res.in:8080/spidbar/> *Gene* (2016), 592(2), 316-324

(b)

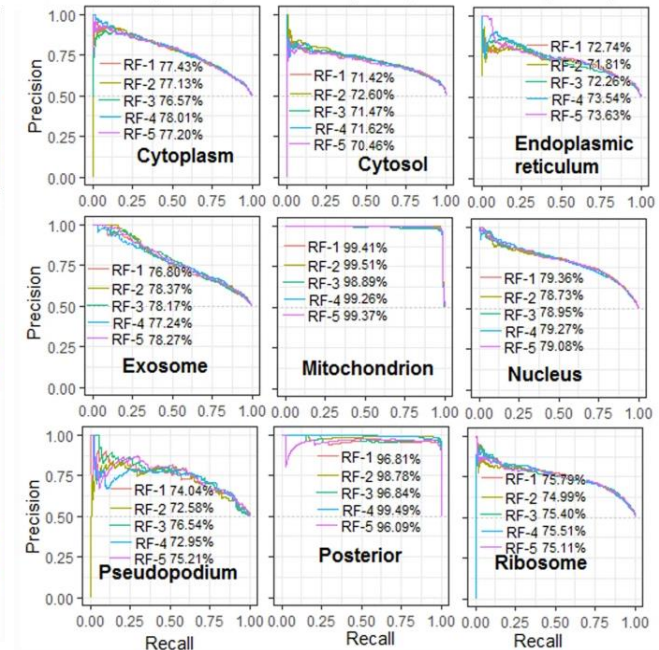


# Multiple sub-cellular mRNA localization

- Localization dataset: RNALOC database (9 localization)
- K-mer features: 5460 (k-mer size 1 to 6)
- Feature selection: Elastic Net algorithm (1812 features selected)
- Prediction algorithm: Random Forest



Precision-recall curve for different K-mer



Cross-validation accuracy

W Nitrogenase - V x DIRProt: a com x Identification of x lrhsp - Yahoo In x hrp - Yahoo x HRGPred: Hom x Identification of x DNA barcoding x mLoc-mRNA

## mLoc-mRNA: a web server for predicting multiple sub-cellular localization of mRNA

### NAVIGATION

- [Home](#)
- [Server](#)
- [Algorithm](#)
- [Dataset](#)
- [Help](#)
- [Contac](#)

### OUR OTHER PUBLISHED WEB SERVERS

- [dSSPred](#)
- [MaLDoSS](#)
- [PreDOSS](#)

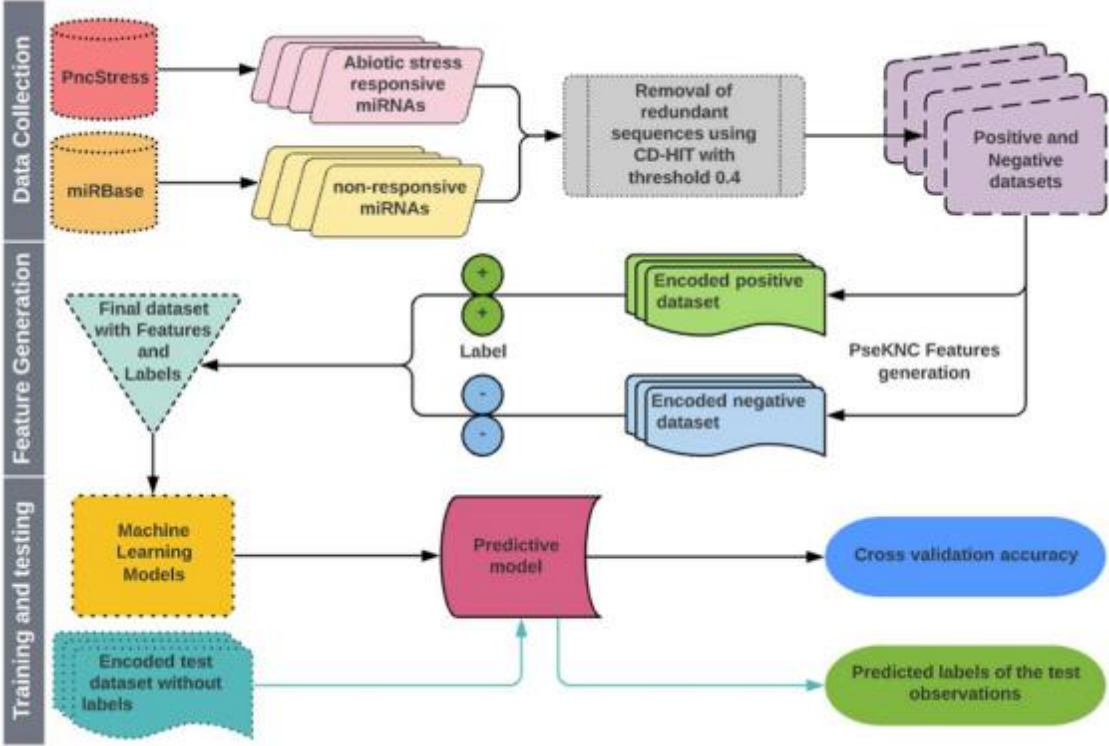
### About mLoc-mRNA

Localization of mRNAs has several advantages over targeting protein localizations, and thus often utilized for targeting location of proteins and their functions. Keeping in mind the importance of localization of mRNAs, this tool has been developed for predicting multiple sub-cellular localization of mRNAs. The mLoc-mRNA is capable of predicting the probabilities for each mRNA sequence to be predicted in nine different localizations that are cytoplasm, cytosol, endoplasmic reticulum, exosome, mitochondrion, nucleus, pseudopodium, posterior and ribosome. The developed model achieved area under ROC curves (auROC) of 78.13, 75.63, 75.54, 76.47, 98.98, 80.28, 76.73, 98.90 and 78.40% for the respective localization, while accuracies are measured following 5-fold cross validation. Further, auROC of 74.72, 76.30, 73.57, 76.49, 95.81, 78.75, 69.56, 98.45 and 77.91% are obtained while the model is evaluated using an independent test dataset. This tool will certainly supplement the future endeavor in the direction of mRNA localization study.

Meher, P.K., Rai, A. and **Rao, A.R.** (2021). *BMC Bioinformatics* **22**, 342 (2021)

# Abiotic stress-responsive miRNA prediction

- Dataset: Abiotic stress associated miRNA and Pre-miRNA
- Feature: Pseudo K-tuple nucleotide compositional features (1372)
- Feature selection: SVM-RFE (SVM-recursive feature elimination)
- Prediction algorithm: SVM, XGB, ADB, RF



Cross-validation accuracy for SVM

Dataset	Sen (%)	Spe (%)	Acc (%)	Pre (%)	F-Score (%)	auROC (%)	auPRC (%)
miRNA	66.13	64.53	65.33	65.09	65.61	70.21	69.96
Pre-miRNA	69.20	63.60	66.40	65.53	67.31	69.71	65.64
Pre-miRNA + miRNA	74.00	68.80	71.40	70.34	72.12	77.94	77.32

Prediction with other learning algorithms

Dataset	Method	Sen (%)	Spe (%)	Acc (%)	Pre (%)	F-Score (%)	auROC (%)	auPRC (%)
miRNA	RF	55.20	58.13	56.66	56.86	56.02	58.88	58.25
	XGB	51.21	56.00	53.61	53.78	52.46	54.79	56.03
	ADB	52.26	57.06	54.67	54.91	53.55	57.45	57.01
PremiRNA	RF	65.60	58.50	62.20	61.42	63.44	64.25	58.03
	XGB	55.61	56.40	56.00	56.04	55.82	58.26	54.91
	ADB	58.01	60.00	59.00	59.18	58.58	62.28	57.86
Pre-miRNA + miRNA	RF	63.20	62.00	62.60	62.45	62.82	64.63	60.28
	XGB	62.20	61.60	62.00	61.90	62.15	62.56	59.64
	ADB	61.60	59.60	60.60	60.39	60.99	63.55	59.96

Independent test set prediction

Dataset	#Sequences		Performance Metrics		
	Positive	Negative	Sensitivity (%)	Specificity (%)	Accuracy (%)
miRNA	72	100	66.66	58.00	62.33
Pre-miRNA	70	100	65.71	64.00	64.85
miRNA + Pre-miRNA	70	100	71.42	67.00	69.21



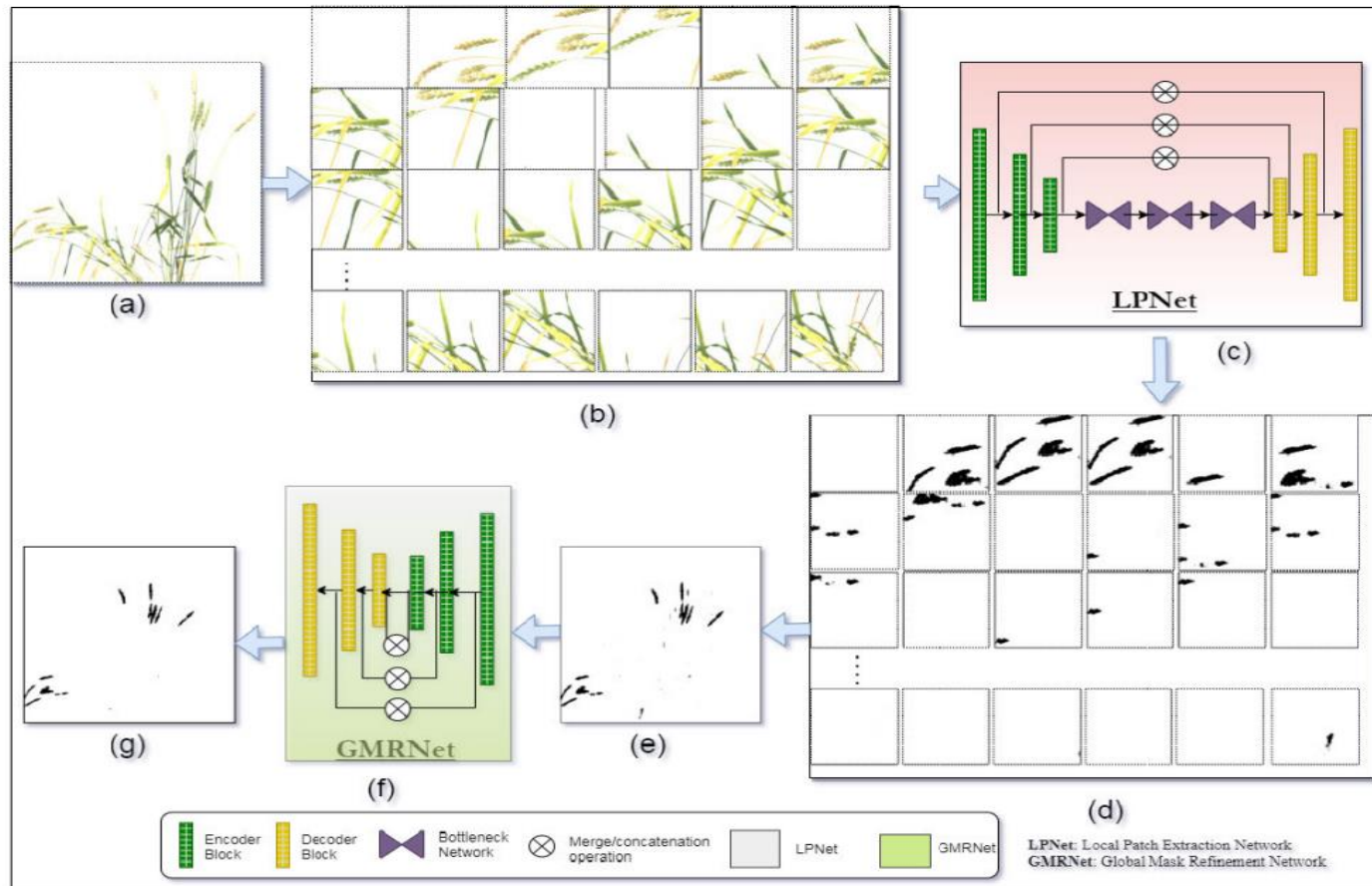
<http://cabgrid.res.in:8080/asrmirna/>

Meher, P.K., Begam, S., Sahu, T.K., Gupta, A., Kumar, A., Kumar, U., Rao, A.R., Singh, K.P., and Dhankher, O.P. (2022). *International Journal of Molecular Sciences*, **23**, 1612

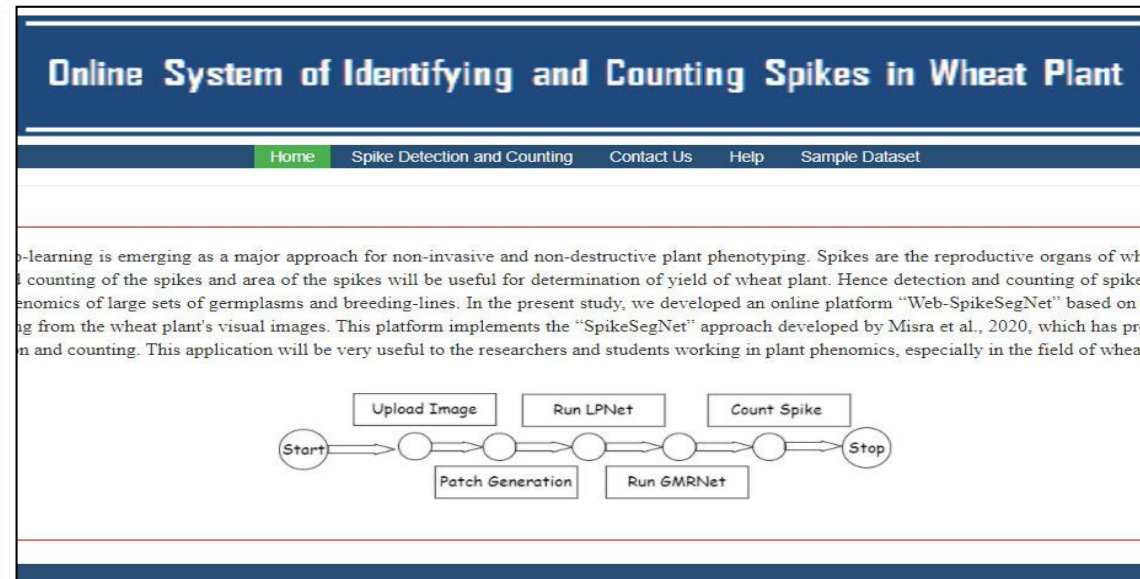


# Spike recognition and counting in wheat plants from visual imaging

- Computer vision emerging as a significant approach for non-invasive and non-destructive plant phenotyping.
- Detection and counting of spikes - critical to determine yield
- Object detection from the digital images – Challenge
- Deep learning network – Local Patch extraction Network (LPNet) and Global Mask refinement Network (GMRNet)



**FIGURE 1.** Flow diagram of SpikeSegNet: Here, input is visual image of wheat plant of size 1656\*1356. The input image is divided into patches of size 256\*256 before entering into the LPNet. The output of LPNet are patch-by-patch segmented mask images which are then combined to form the mask image as per the size of the input visual image. This image may contain some sort of inaccurate segmentation of the object (or, spikes) and are refined at global level using GMRNet network. The output of GMRNet network is nothing but the refined mask image containing spike regions only.



Misra, T., Arora, A., Marwaha, S., Chinnusamy, V., **Rao, A.R.**, Jain, R., Sahoo, R.N., Ray, M., Kumar, S., Raju, D. and Jha, R.R. (2020). *Plant Methods*, **16**:40

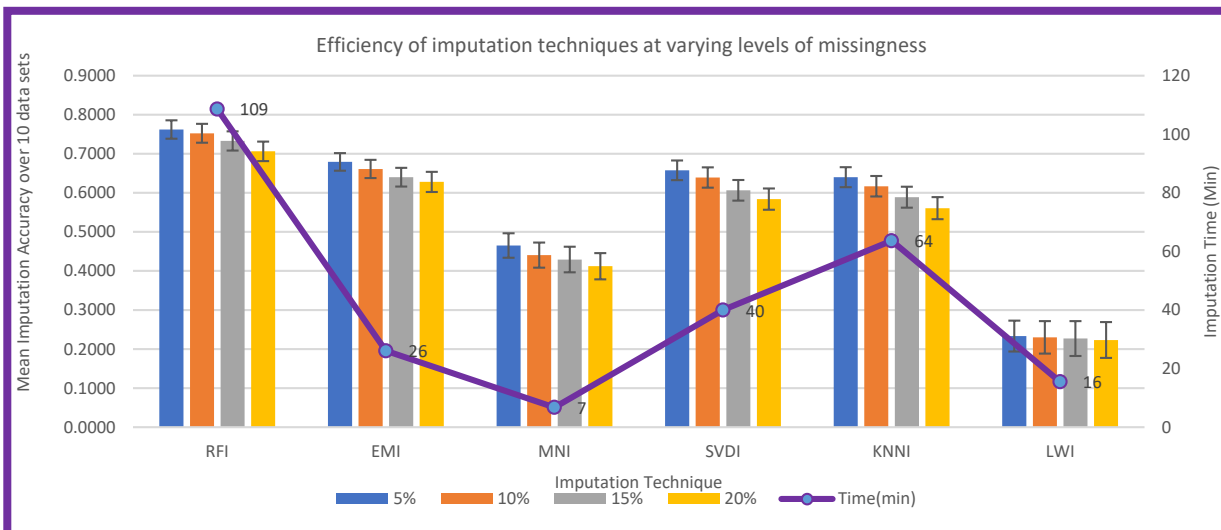
Misra, T., Arora, A., Marwaha, S., Jha, R.R., Ray, M., Jain, R., **Rao, A.R.**, Varghese, E., Kumar, S., Kumar, S., Nigam, A., Sahoo, R.N., and Chinnusamy, V. (2021). *IEEE Access*, **9**, 76235-76247.

# Genomic Selection and AI

- Suitable imputation method against missing observations in GBS data
- robust GS model against missing SNP genotyping data
- Estimation of GEBVs in presence of missing observations

## Imputation techniques

1. Mean allele frequency Imputation (MNI)
2. Locally weighted linear Regression Imputation (LWI)
3. k- Nearest Neighbour Imputation (k-NNI)
4. Single Value Decomposition Imputation (SVDI)
5. Expectation-Maximization Imputation(EMI)
6. Random Forest Imputation(RFI)



## I. BLUP based Models

1. G-BLUP (Genomic BLUP)
2. EG-BLUP (Epistatic Genomic BLUP)

## II. Models Based on Penalization :

1. Ridge Regression (RR)
2. Least Absolute Shrinkage and Selection Operator (LASSO)
3. Elastic Net (EN)

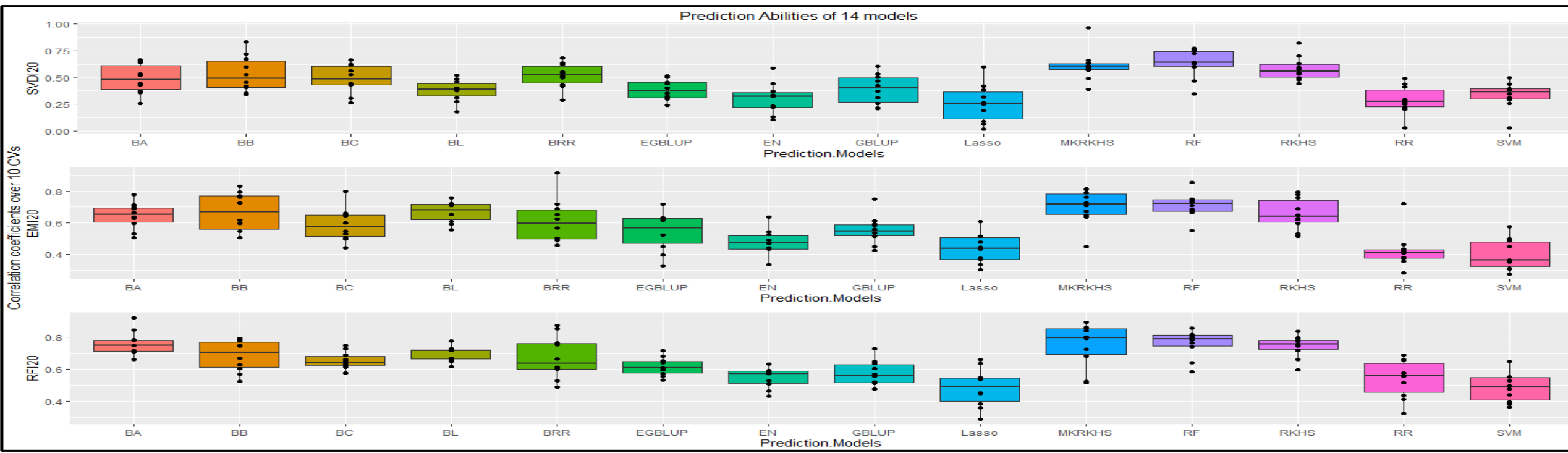
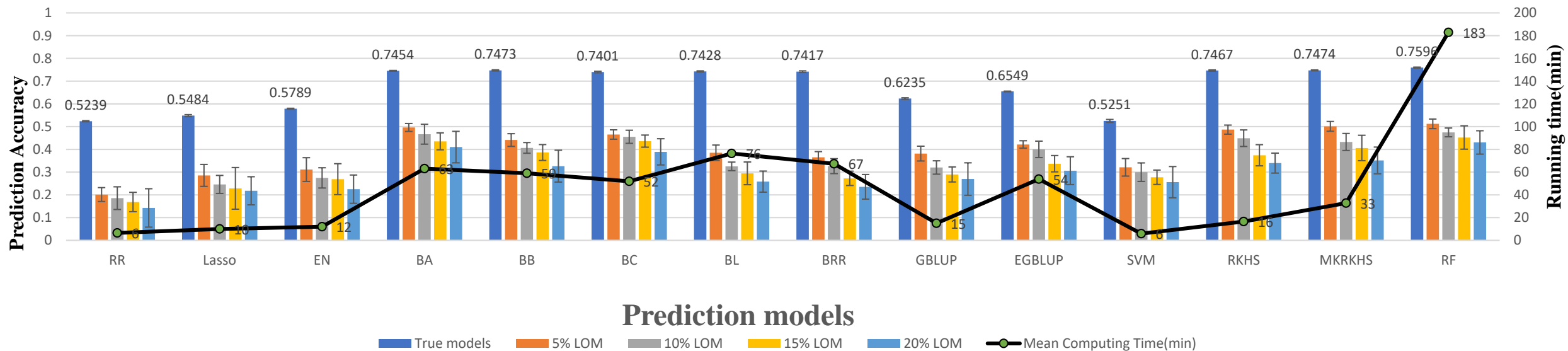
## III. Models Based on Bayesian Approach

1. Bayes A (BA)
2. Bayes B (BB)
3. Bayes C (BC)
4. Bayesian Ridge Regression (BRR)
5. Bayesian LASSO (BL)

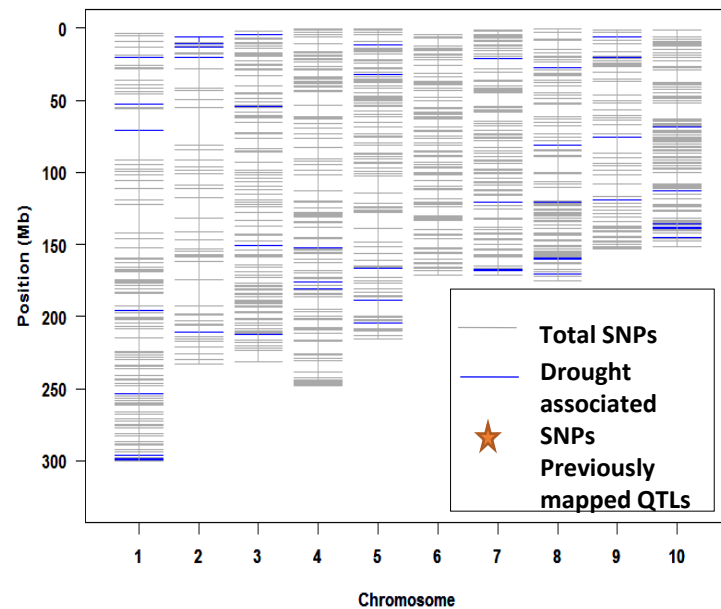
## IV. Models based on Machine learning algorithms

1. Support Vector Machines (SVM)
2. Reproducing Kernel Hilbert Space (RKHS)
3. Multi Kernel RKHS (MKRKHS)
4. Random Forest (RF)

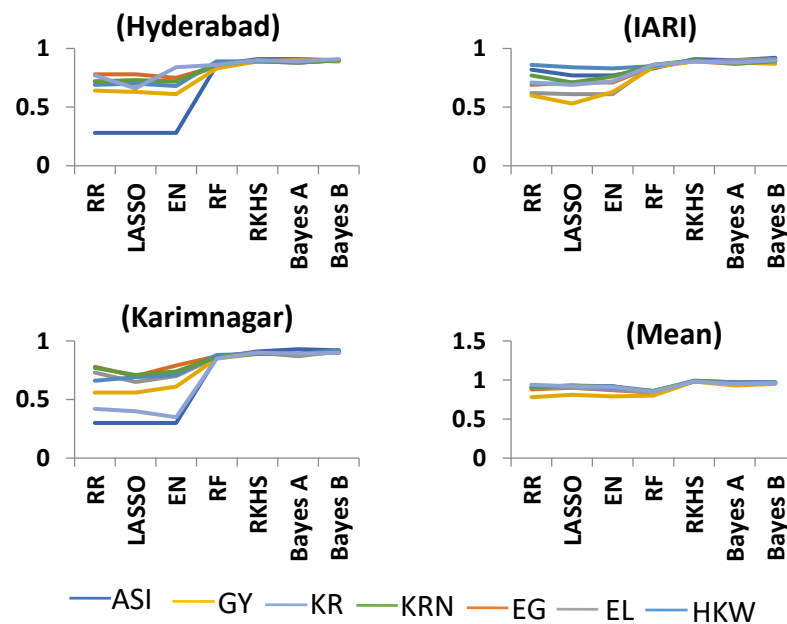
# Efficacy of GS models in case of complete as well as incomplete SNP data



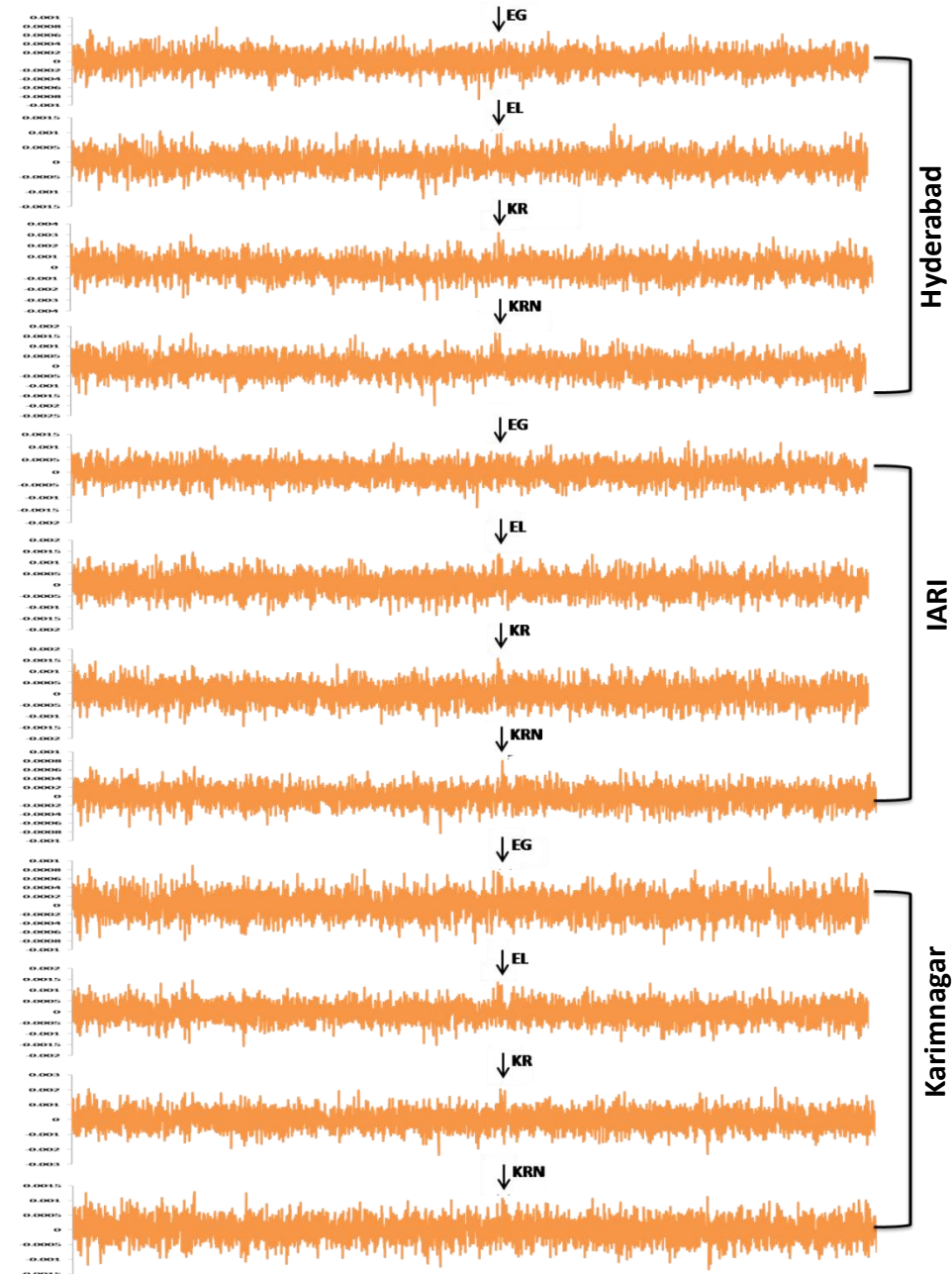
## Top 1053 SNPs detected by Bayes B



## Prediction accuracies of GS models



## Marker effect consistency from Bayes B





# Metagenome & Machine Learning

## ❖ Main Challenges

- Assessment of molecular diversity and density
- Accurate Binning
- Assessment of unknown microbes into different categories

## ❖ Data

- ✓ Contaminated sediment samples from the Ganga and Yamuna Rivers.
- ✓ Locations – Kanpur, Farakka for Ganga river  
- Delhi for Yamuna river

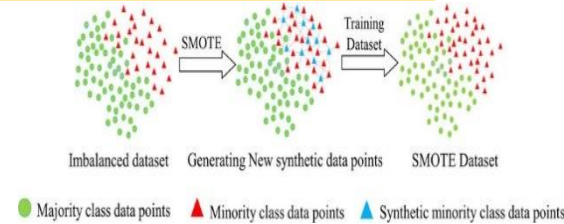
Dataset by Location	Sites	Total Sequences
Farakka	F1	2,91,28,182
	F2	5,44,69,302
Kanpur	K1	2,81,58,772
	K2	3,30,84,931
Delhi	D1	6,38,16,159
	D2	6,36,60,637

## ❖ Result

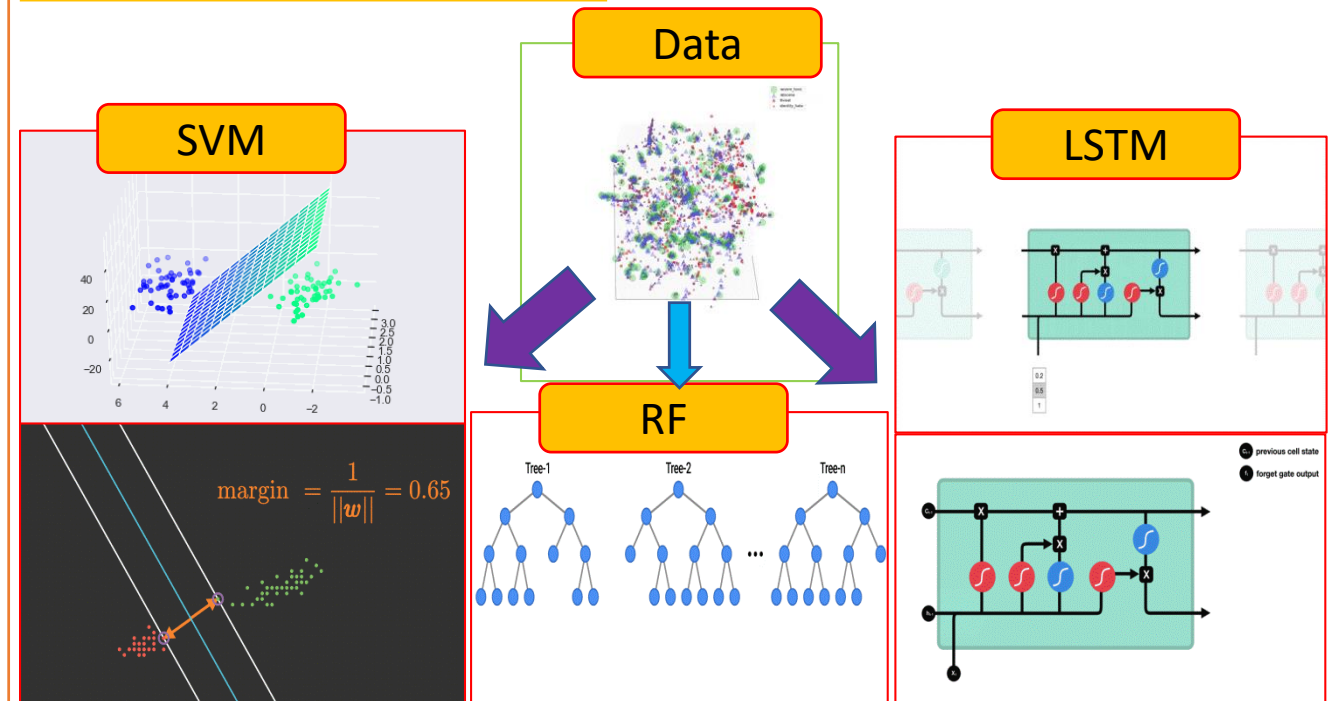
Microbial Diversity		Ganga	Yamuna	NIRS
Cellular	Bacteria	50,305	53,506	1,03,811
	Archaea	1,039	3,254	4,293
	Eukaryota	10,799	15,985	26,784
Non-cellular	Virus	4,55	2,48	7,03

## ❖ Kmeans Clustering

## ❖ SMOT Analysis

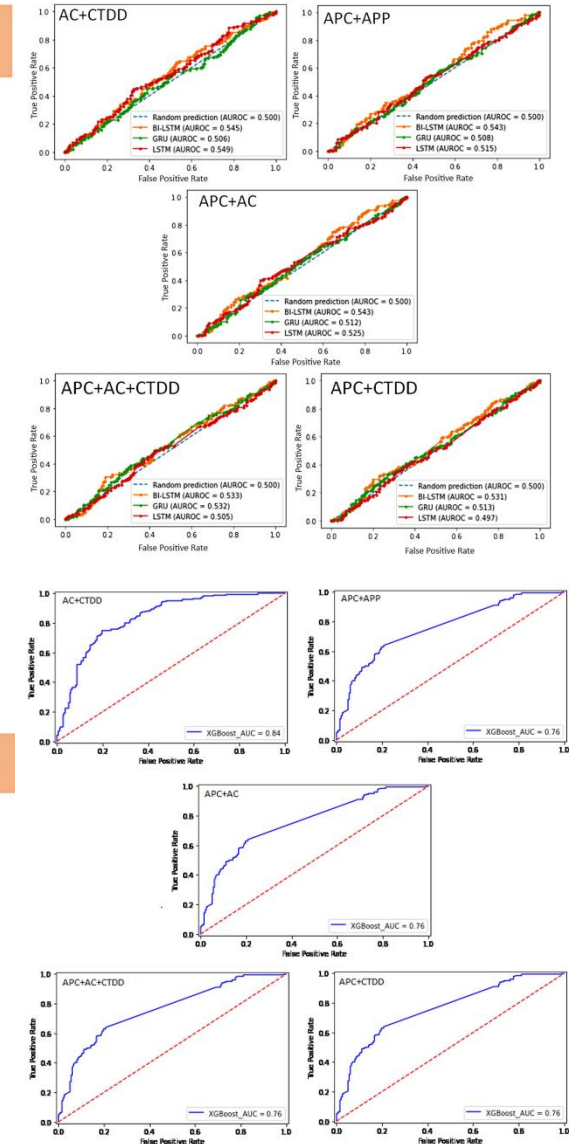
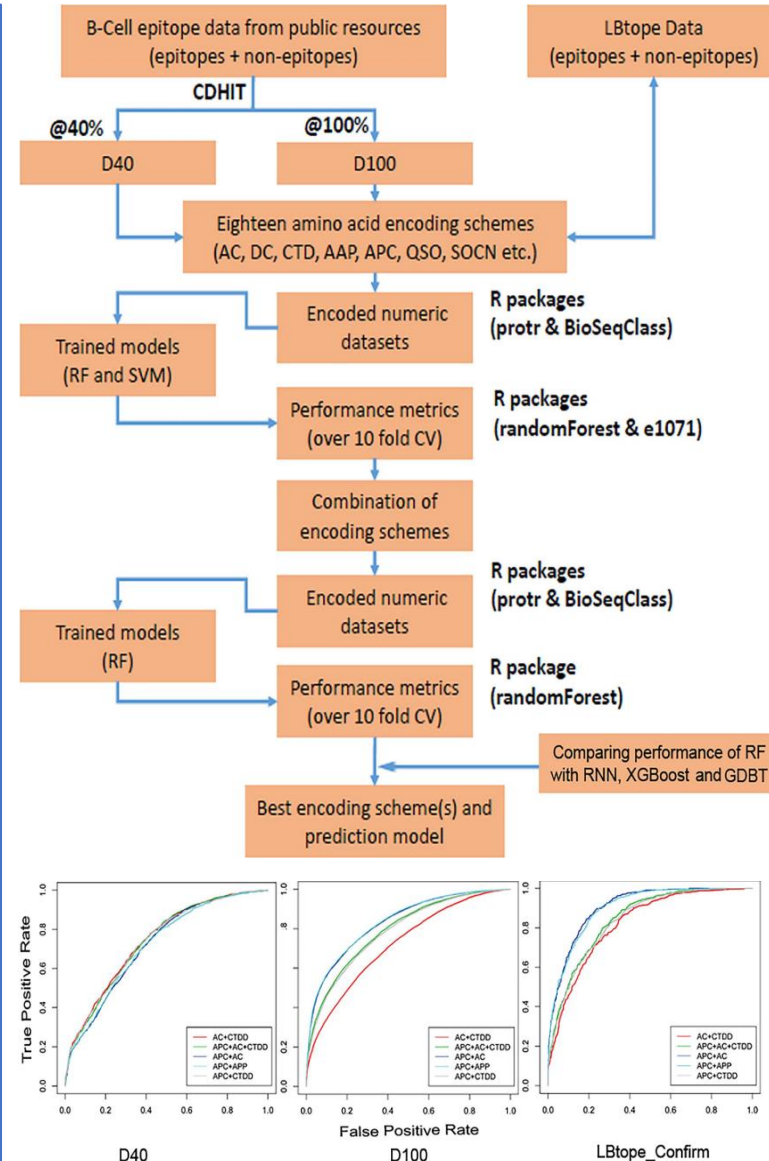


## ❖ Multiclass Classification



# Comparative analysis of B-Cell epitope Prediction Tools

- B-cell epitopes have a prominent role in development of peptide-based vaccines and disease diagnosis.
- High variability in the length of these epitopes is the major reason for low accuracy in their prediction.
- We have analyzed the performance of machine learning approaches (MLA) with eighteen different amino acid encoding schemes in the prediction of flexible length linear B-cell epitopes.
- The APC encoding scheme was found suitable for homogeneous and longer flexible length B-cell epitopes, while its combination with DC, AC, and CTDD encoding schemes is likely to improve accuracy.
- The CTDD feature set can be opted for heterogeneous dataset and shorter flexible length B-cell epitopes and its combination with AC is favorable for an enhanced prediction performance.
- Besides APC and CTDD, DC and APP encoding schemes were found more appropriate for homogeneous B-cell epitopes whereas AC was found suitable for heterogeneous B-cell epitopes.
- Two combinations of peptide encoding schemes *i.e.*, APC+AC and APC+APP were identified to have improved performance over the state-of-the-art tools for flexible length linear B-cell epitope prediction.





## Future Strategy

- Design guide RNA sequence with minimum off-target effects and high on-target efficiency
- Develop efficient algorithms – in terms of time and space complexity
- Explore Functional-PCA, Functional-Classification and regression, *etc.* in Phenomics
- Very Fast Decision Tree (VFDT) – Construction of Hoeffding Trees
- Phenome Wide Association Study (PheWAS)
- Assessment of performance of various classifiers with different kernels for prediction purposes
- Specific-stress responsive miRNA prediction
- Estimation of yield in field crops from visual imaging
- Search for more data scientists – rare hybrids

## Conclusion

- Artificial Narrow Intelligence has been successfully implemented in Genomics for revealing hidden mechanisms of complex trait expression and their improvement
- Real application of Next Generation Artificial Intelligence in integrated multi-omics is essential for crop improvement
- More focus required for image recognition and machine vision in phenotyping
- Multiclass classification of a greater number of non-coding RNAs in RNome needs attention
- Preparation of high standard data sets, transformation to numeric vectors, choice of competent prediction algorithm, validation, server development are essential for successful hybridization of AI and Genomics

## Acknowledgements

- Funding agencies: CABin Scheme, CRP Genomics, NASF of ICAR
- Dr. P.K. Meher, Scientist, IASRI; Dr. Tanmaya Kumar Sahu, Project Scientist-II, NBPGR; Dr. Sarika Sahu, Scientist, IASRI
- Students – Dr. Priyanka Guha Majumdar, Dr. Srikanth Bairi, Sh. Nailini Kanta Choudhary
- Collaborators of Partner Institutes under CABin Scheme, CRP Genomics
- Organizers of Symposium
- Indian Agricultural Statistics Research Institute
- Dr. T. Mohapatra, Secy. (DARE) & DG, ICAR

THANKS !!!