DBT-Sponsored Workshop on AI in Modern Biology

(23rd – 25th 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[1] and Sarika Sahu[2])

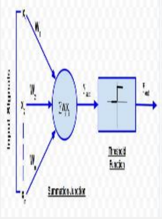
[1]Indian Council of Agricultural Research (ICAR), New Delhi
adg.pim@icar.gov.in; rao.cshl.work@gmail.com

[2]ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Artificial Intelligence – Landmarks

1943

First ANN (McCulloch & Pitt)



1948

First two robots: Elmer & Elsie

1955

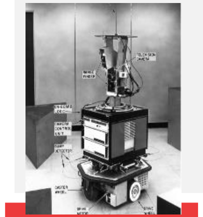
Artificial intelligence term

1964

First AI program - ELIZA

1970

First mobile robot - Shakey



1981

Commercialized AI: the expert systems

1989-96

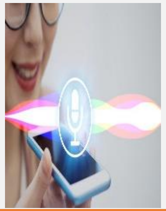
Chess Victories

2001

A.I Movie

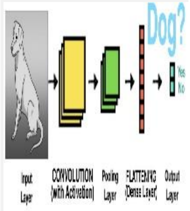
2011

The virtual assistant 'Siri'



2012

CNN

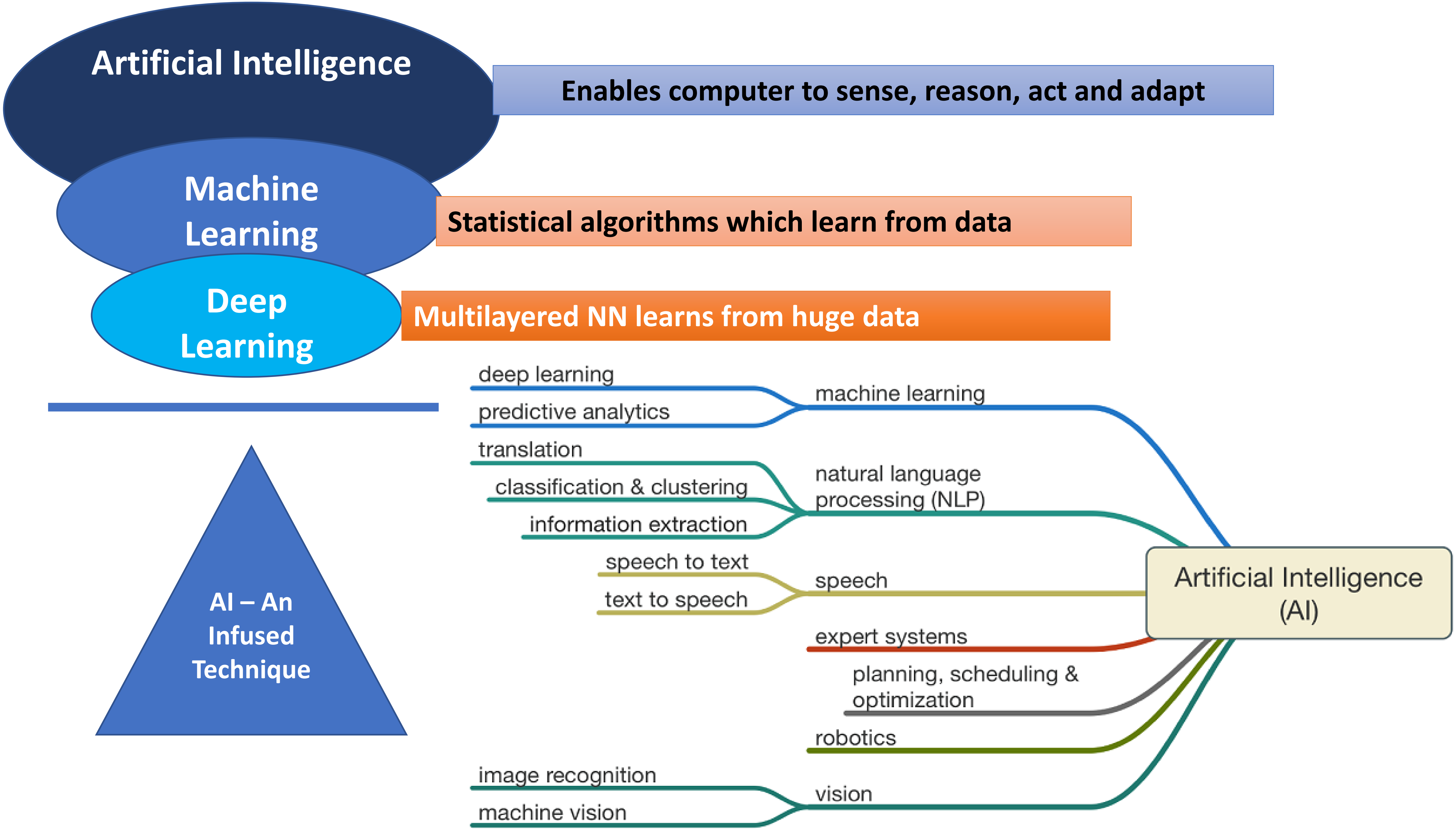


2016

Google DeepMind's Alpha Go

2019

Turing Award – DNNs as critical components of computing



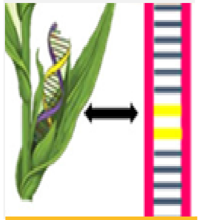
Artificial Intelligence
Enables computer to sense, reason, act and adapt
Machine Learning
Statistical algorithms which learn from data
Deep Learning
Multilayered NN learns from huge data
AI – An Infused Technique
deep learning
predictive analytics
machine learning
translation
classification & clustering
information extraction
natural language processing (NLP)
speech to text
text to speech
speech
Artificial Intelligence (AI)
expert systems
planning, scheduling & optimization
robotics
image recognition
machine vision
vision

Genomics – Landmarks

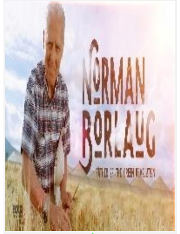
1866 – Laws of Genetics by Gregor Johann Mendel

1920 – Hybrid Breeding of 1st Commercial Maize

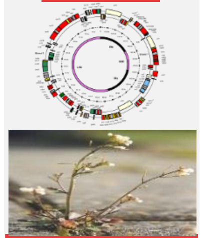
1928 – First Report of Mutation Breeding in Maize



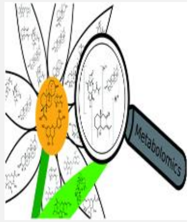
1960 – Green Revolution by Norman Borlaug



2000 – Arabidopsis thaliana genome sequenced



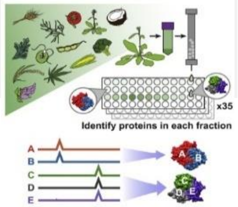
2001 – Plant Metabolomics



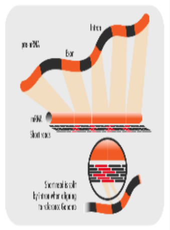
2002 – SNP Marker Development, Rice Genome Sequenced, Golden Rice

2005 – Development of NGS technologies (ePCR & Pyrosequencing)

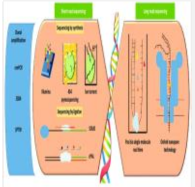
2007 – DNA-Protein interaction mapping using the ChiP-seq; Plant Pan-genomics & Genome Selection



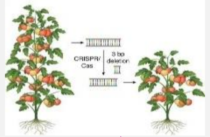
2008 – Transcriptomics by RNA-seq



2009 – Long read sequencing technologies [SMRT-seq & Nanopore]



2013 – 1st Event of Plant Genome Editing via CRISPR/Cas9 Multiplex Genome Editing



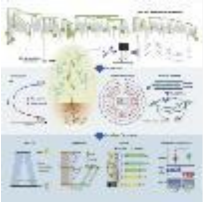
Continued

Genomics – Landmarks

Continued

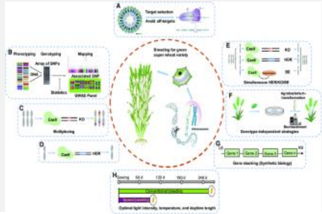
2014

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



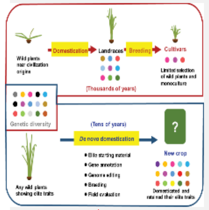
2017

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



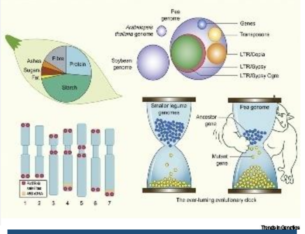
2018

De-novo Domestication; Speed Breeding



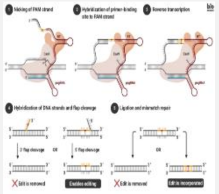
2019

Mendel's Genetic Model Pea Genome Annotated

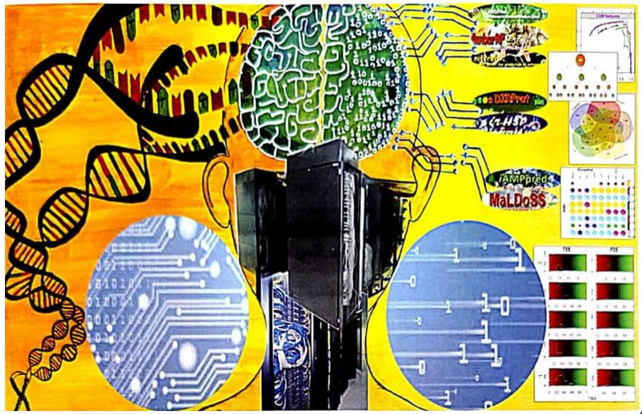


2020

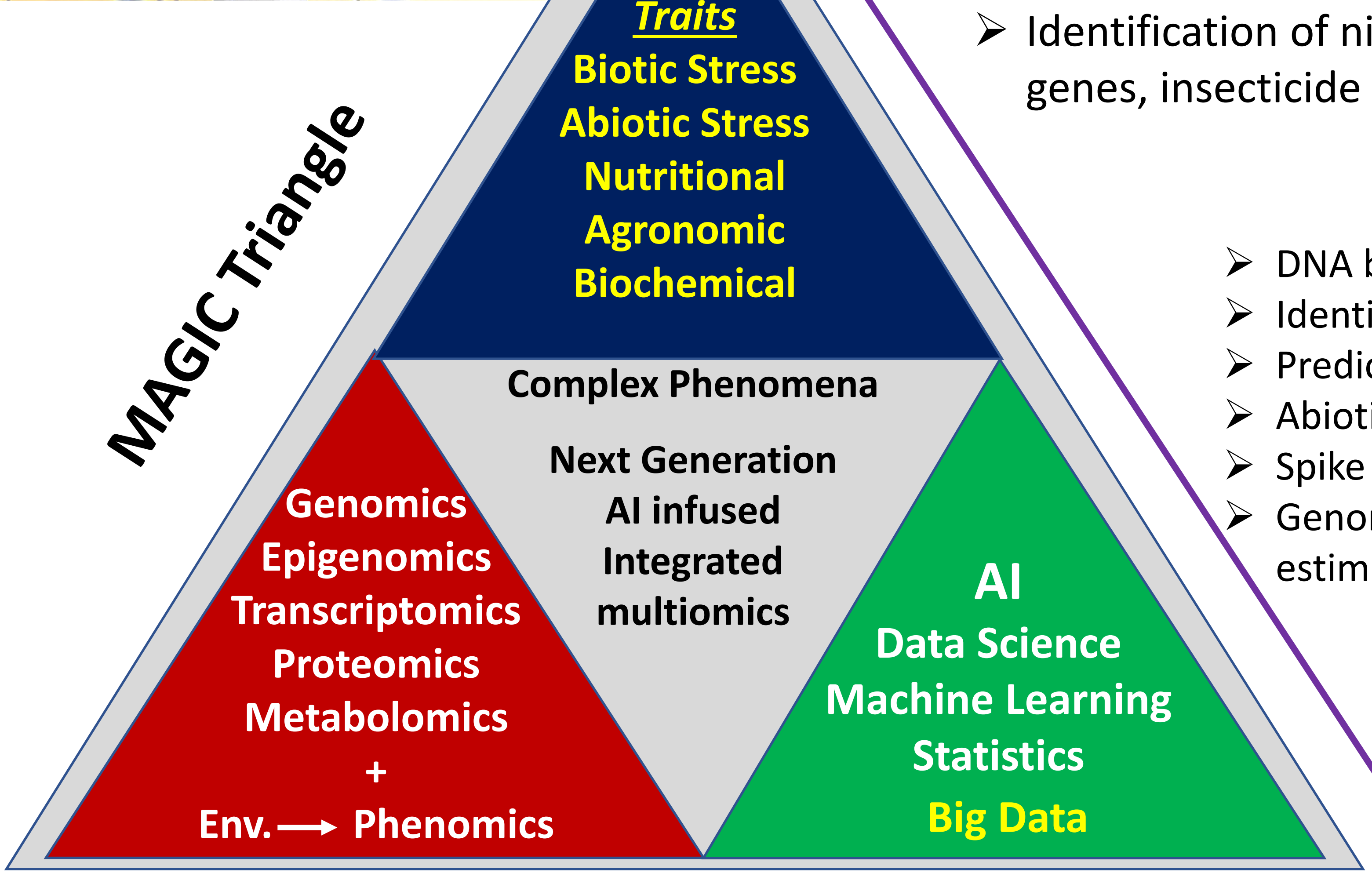
Prime Editing



Journey Contd...



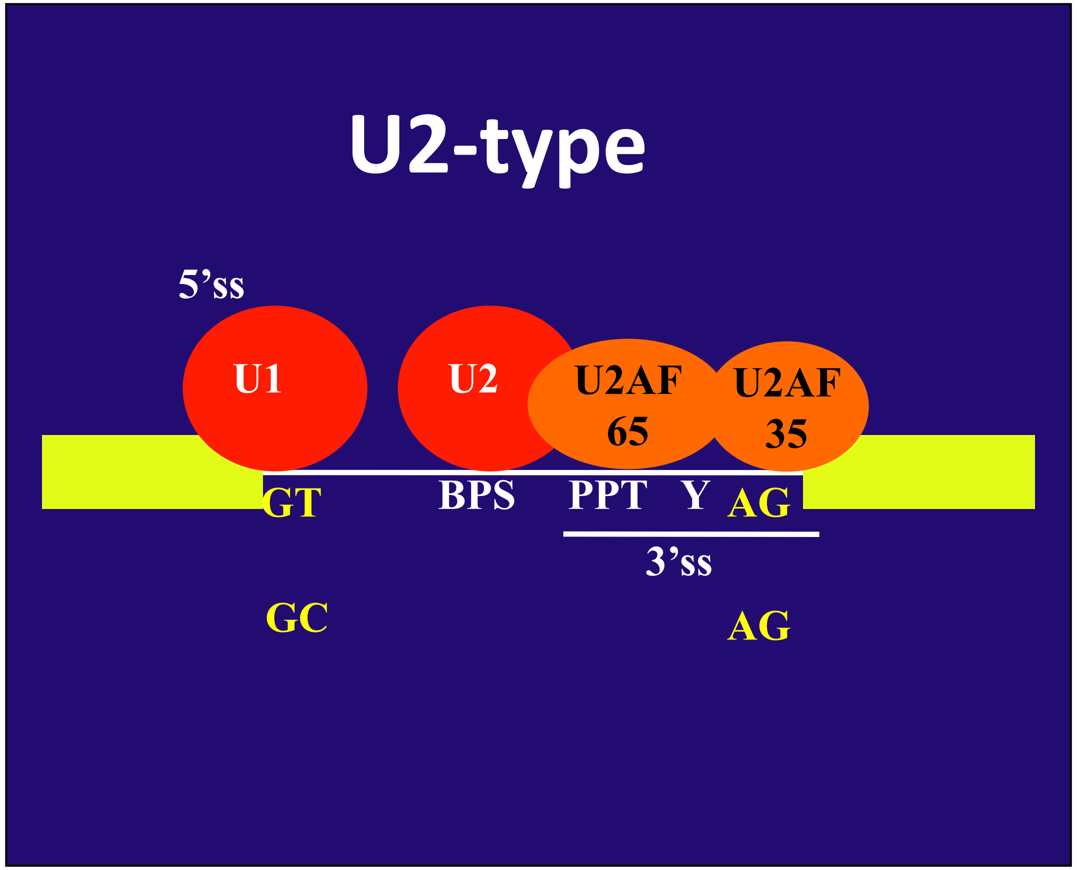
-omics, AI and Complex Traits



- 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

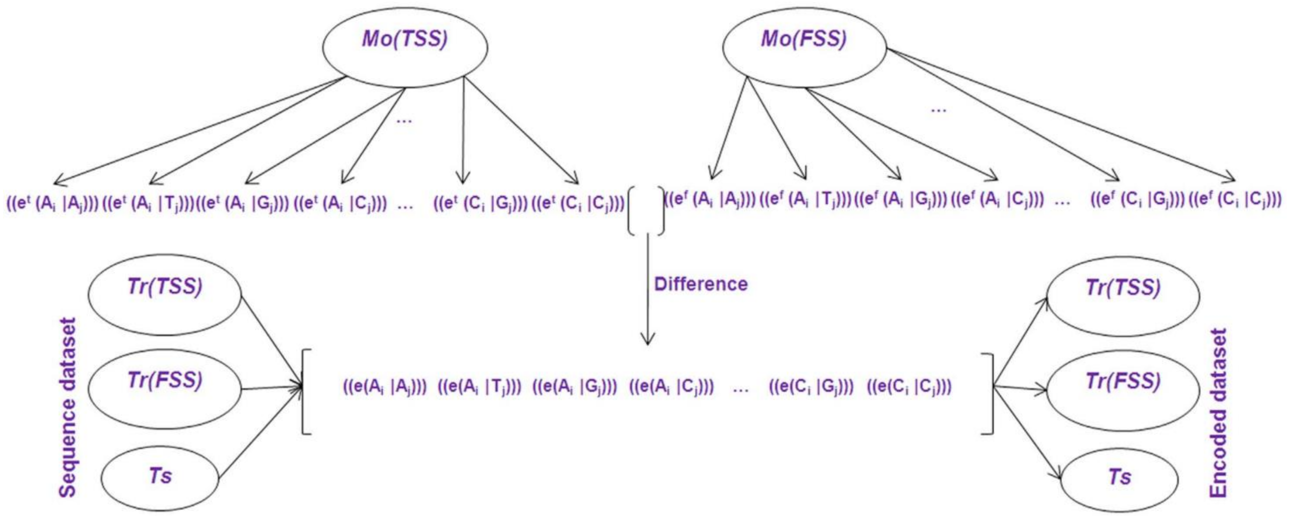
	Exon	Intron
U1 snRNA:	G U C	C A U U C A
	| | |	| | | | |
pre-mRNA sequence :	C A G	G T G A G G
		3 6

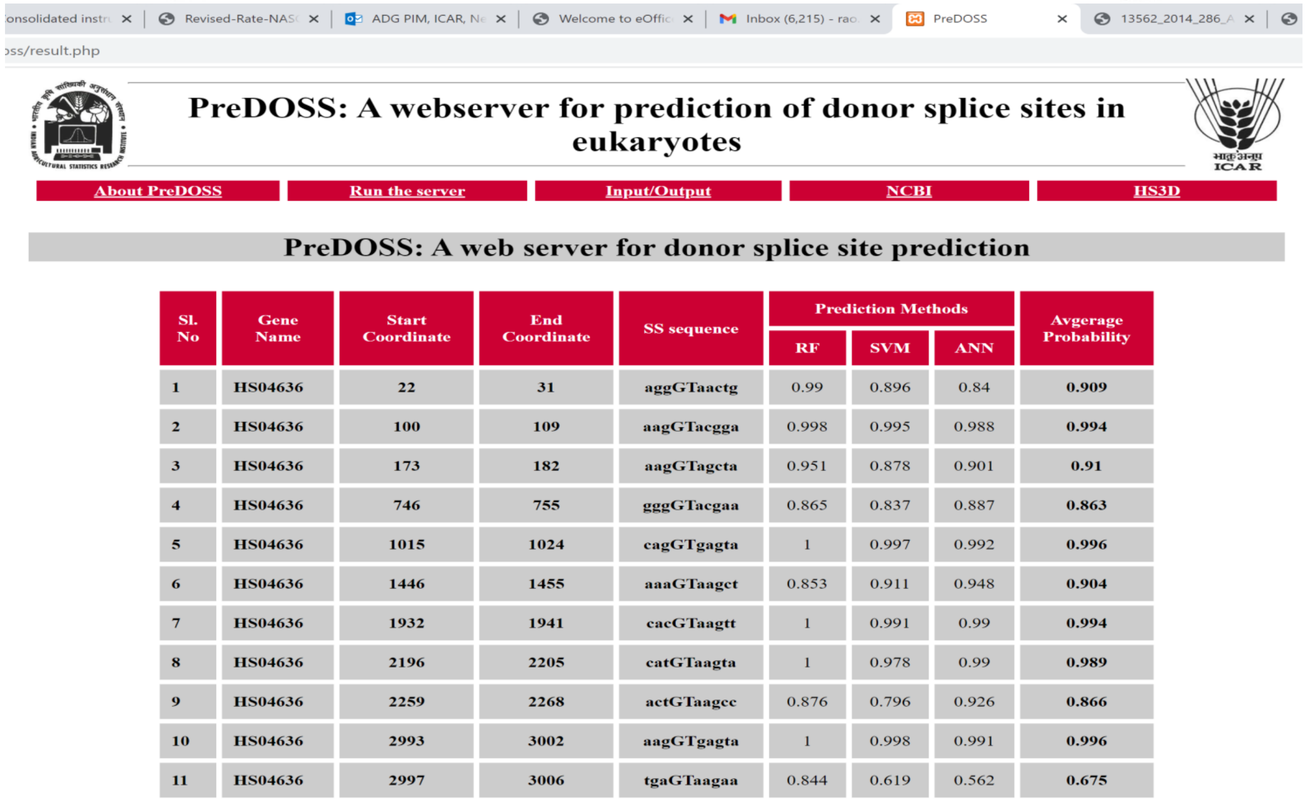
Rank: 1

U1 snRNP splice site binding

Sheth et al (2006). *Nucleic Acid Research*. 34 (14)

- 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)





Sl. No	Gene Name	Start Coordinate	End Coordinate	SS sequence	Prediction Methods RF	SVM	ANN	Avgerage Probability
1	HS04636	22	31	aggGTaactg	0.99	0.896	0.84	0.909
2	HS04636	100	109	aagGTacgga	0.998	0.995	0.988	0.994
3	HS04636	173	182	aagGTagcta	0.951	0.878	0.901	0.91
4	HS04636	746	755	gggGTacgaa	0.865	0.837	0.887	0.863
5	HS04636	1015	1024	cagGTgagta	1	0.997	0.992	0.996
6	HS04636	1446	1455	aaaGTaagct	0.853	0.911	0.948	0.904
7	HS04636	1932	1941	cacGTaagtt	1	0.991	0.99	0.994
8	HS04636	2196	2205	catGTaagta	1	0.978	0.99	0.989
9	HS04636	2259	2268	actGTaagcc	0.876	0.796	0.926	0.866
10	HS04636	2993	3002	aagGTgagta	1	0.998	0.991	0.996
11	HS04636	2997	3006	tgaGTaagaa	0.844	0.619	0.562	0.675

Journal of Theoretical Biology (2016), **404**, 285-294; *Algorithms for Molecular Biology* (2016), **11**:6; *BioData Mining* (2016), **9**:4; *Jour. Plant Biochem. and Biotech.* (2015), **24**(4), 385-392

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;
antifungal peptides - CAMP, LAMP and APD3

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

Compositional, Structural and Physico-chemical features (66)

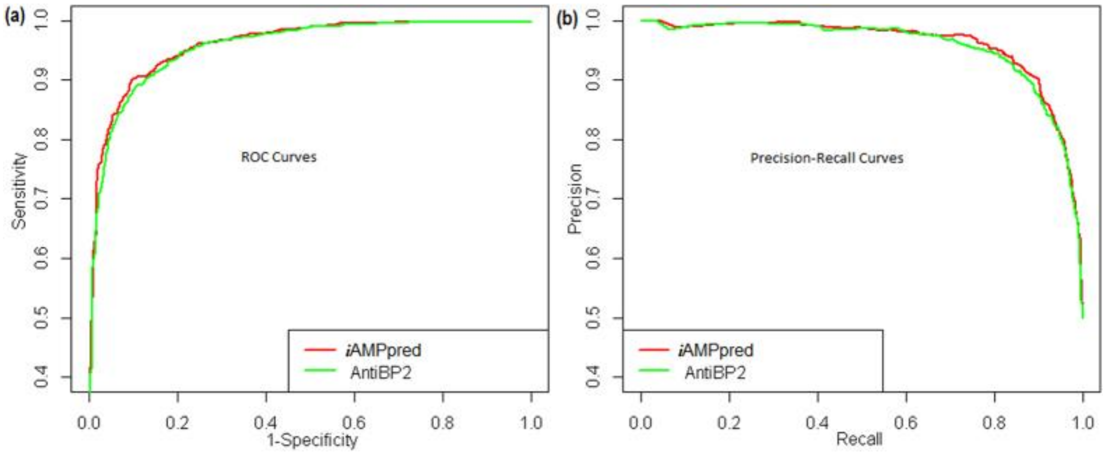
AAC, NAAC and PAAC

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, lincRNA, 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

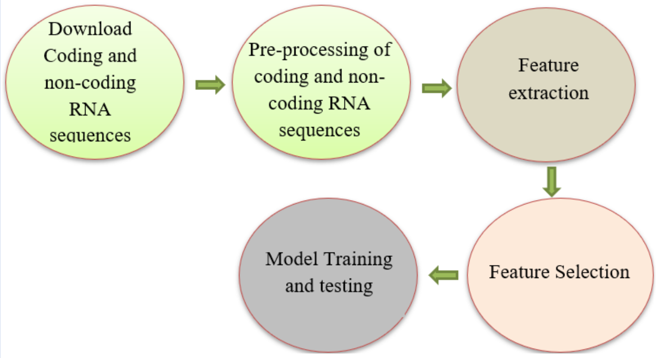
<u>Features</u>

Transcript length, ORF length, ORF coverage

Peptide length, K-mer frequencies, BLAST features

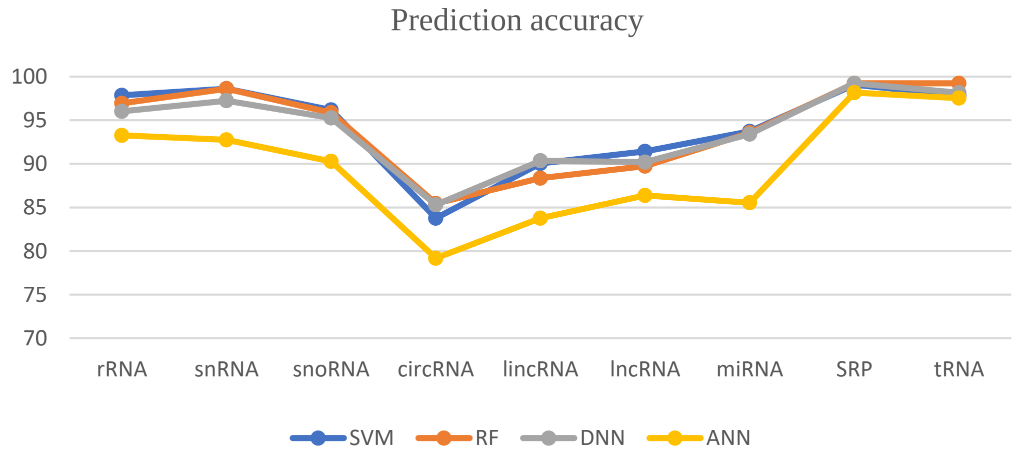
Amino acid composition, Molecular weight,

Isoelectric point, GC%, Codon Bias Indices, RSCU



Workflow

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

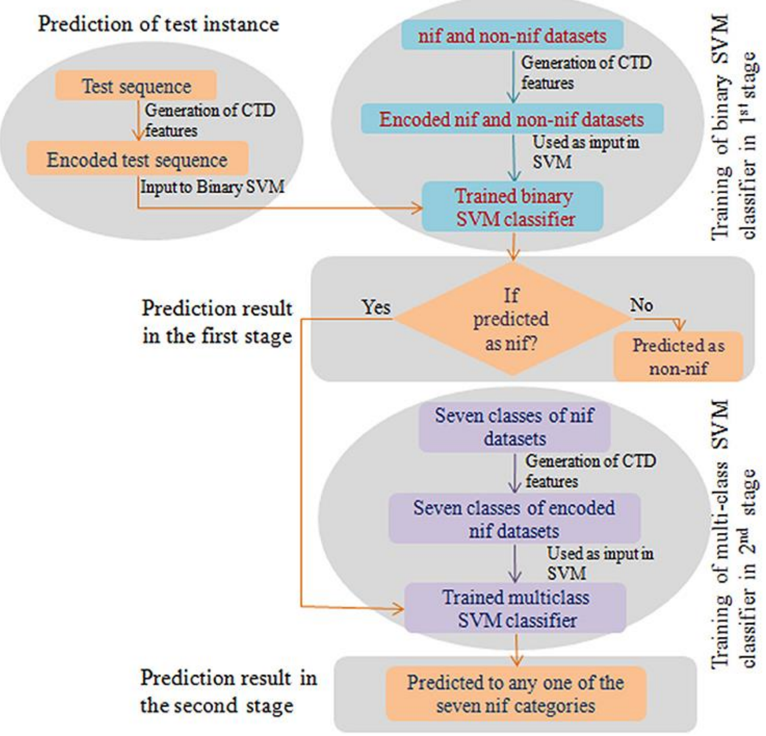
- Nitrogen fixing microbes depend on nitrogenase enzyme complex, consists of structural genes: nifH, nifD, nifK, nifH, nifE, nifN and nifB genes, essential in characterized systems (diaztrophs)
- Identification of nif proteins are essential
- SVM classifier + kernels (linear, polynomial, sigmoidal, radial)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

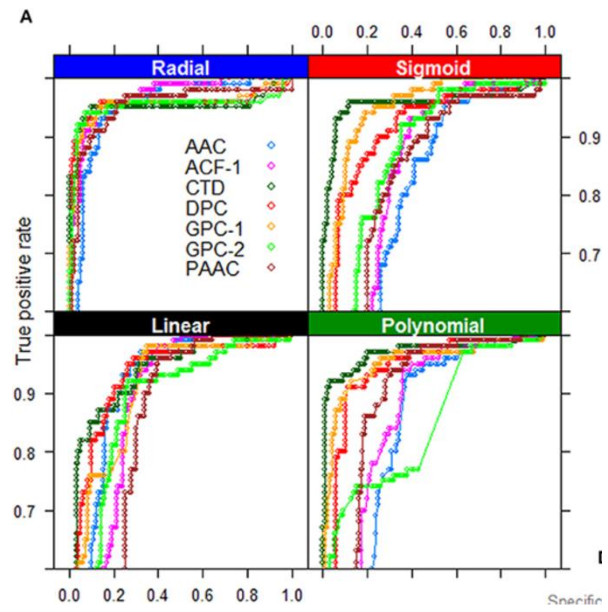
<u>Data</u>

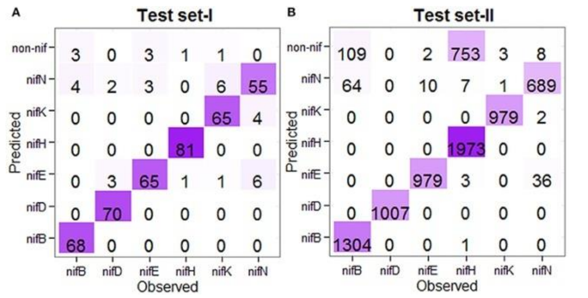
82 diaztrophs (UniProtKB); +ve and –ve datasets for training classifier

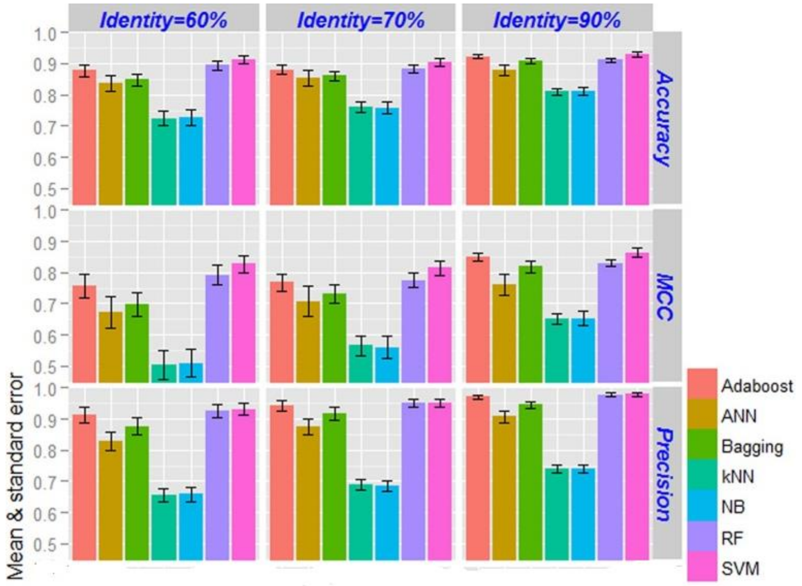
<u>Features</u>

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)

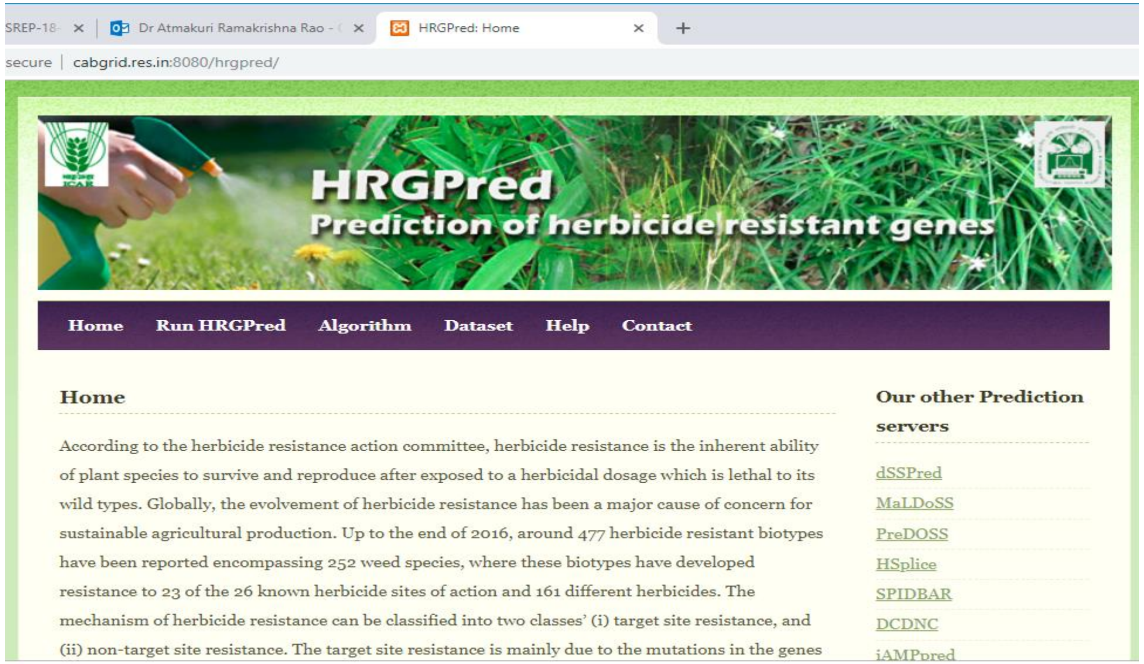








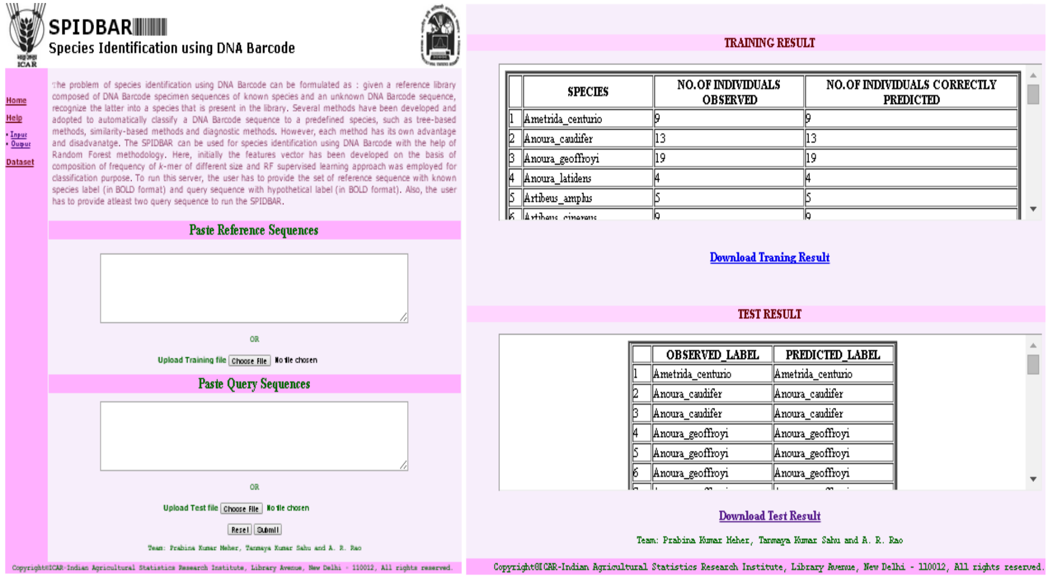
Meher, P.K., Sahu, T.K., Mohanty, J., Gahoi, S., Purru, S., Grover, M. and Rao, A.R. (2018). nifPred: *Frontiers in Microbiology*, 9: 1100.



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

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

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

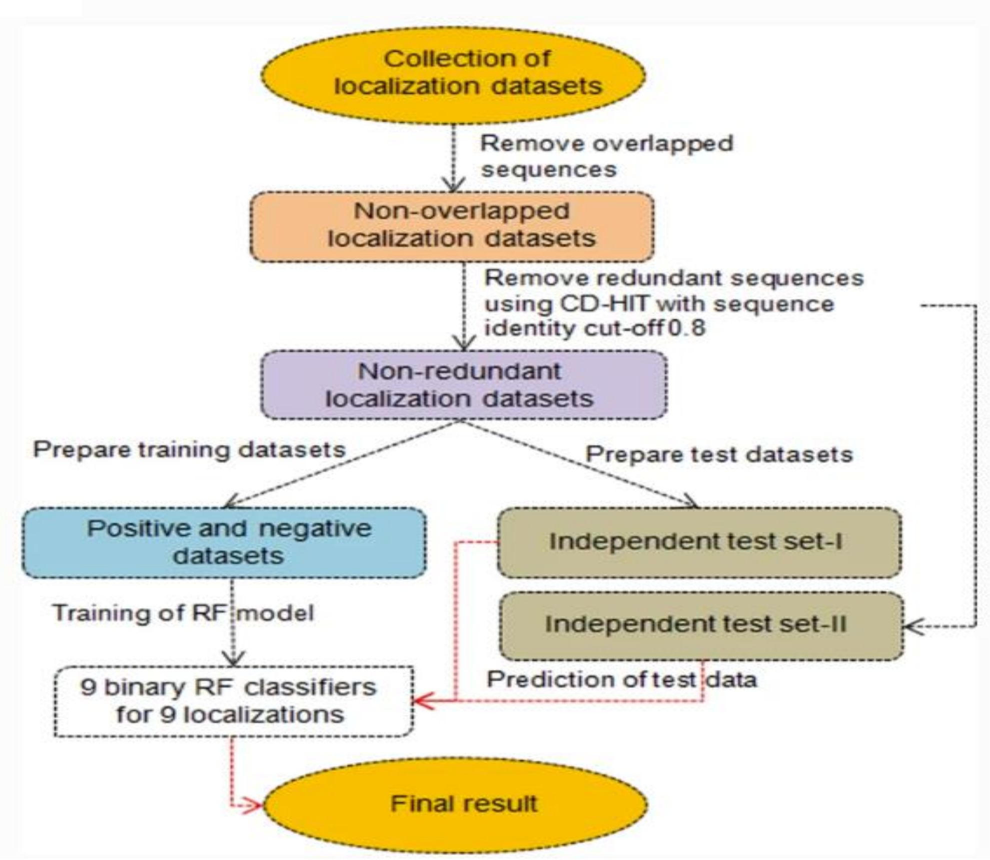


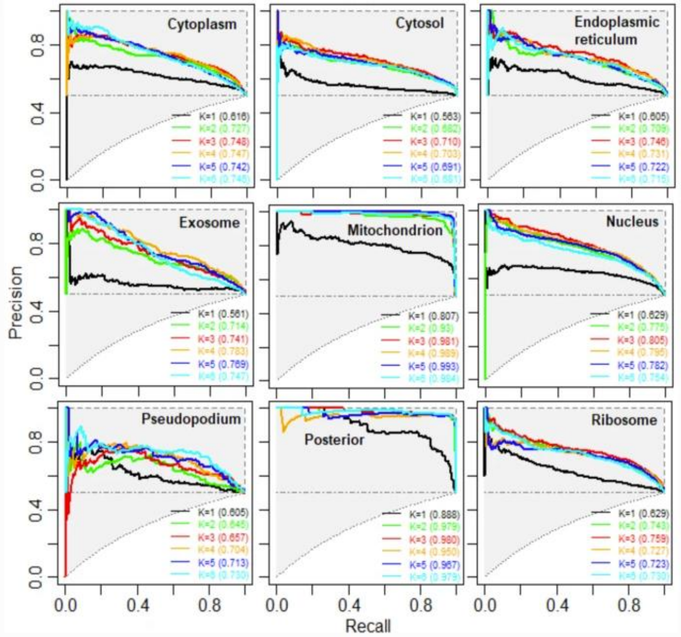
(a) (b)

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

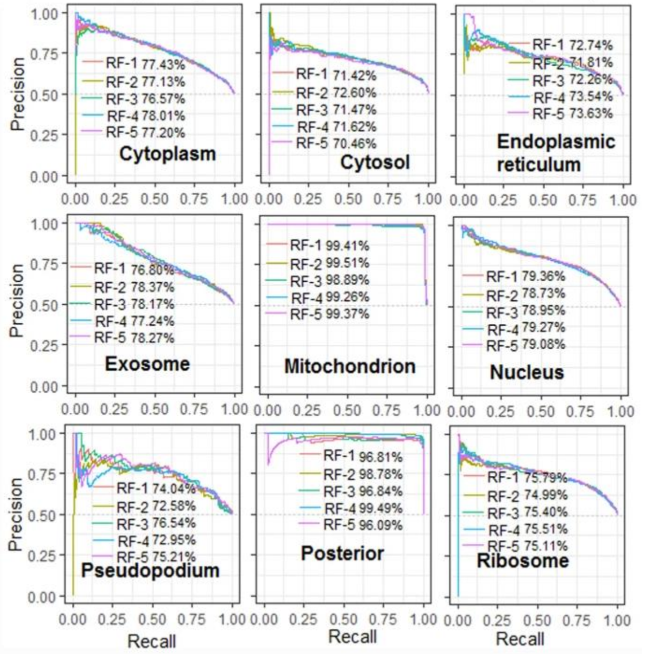
Multiple sub-cellular mRNA localization

- Localization dataset: RNALOCTE 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

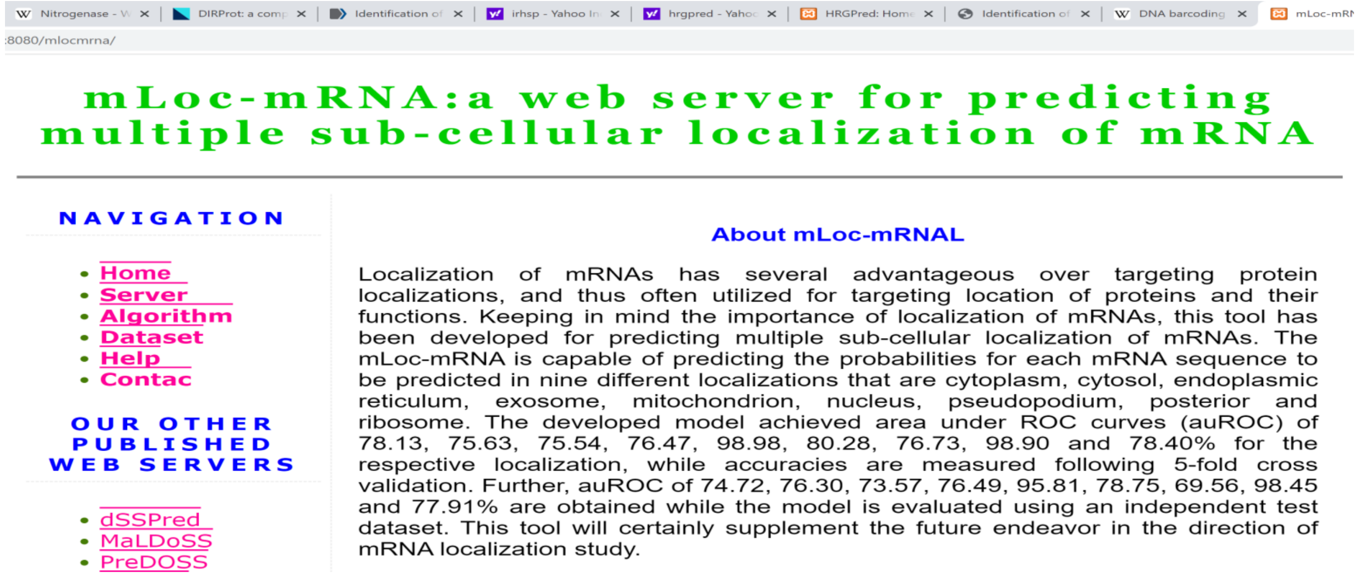




Precision-recall curve for different K-mer



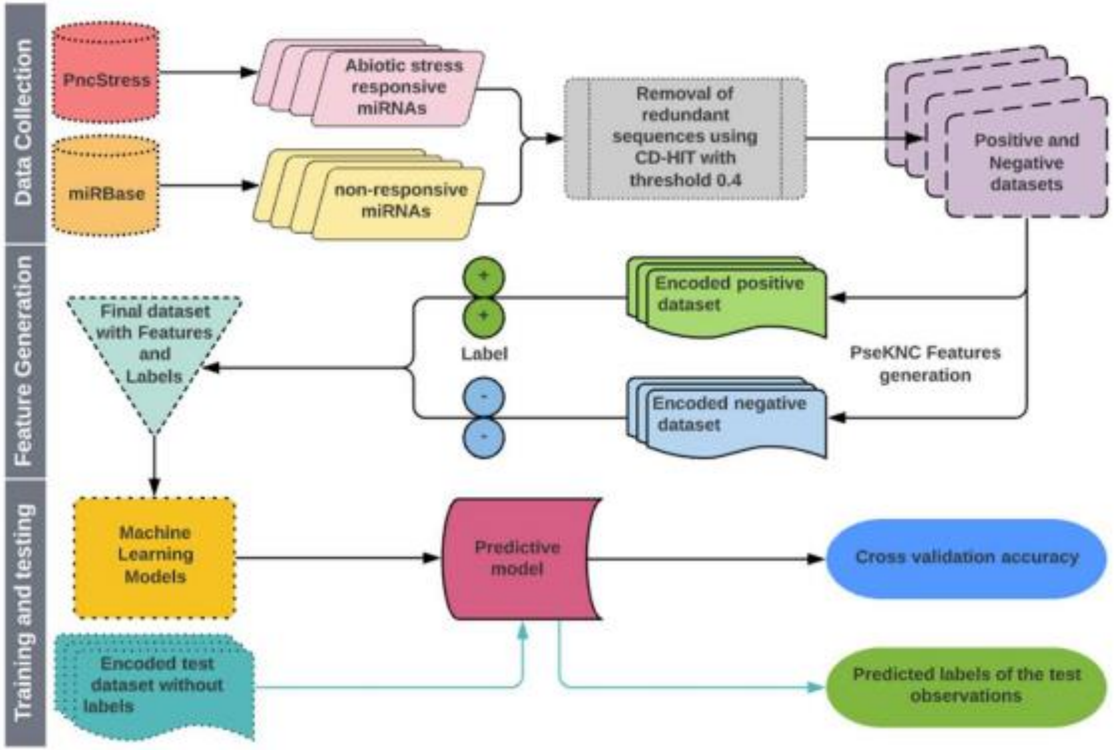
Cross-validation accuracy



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)

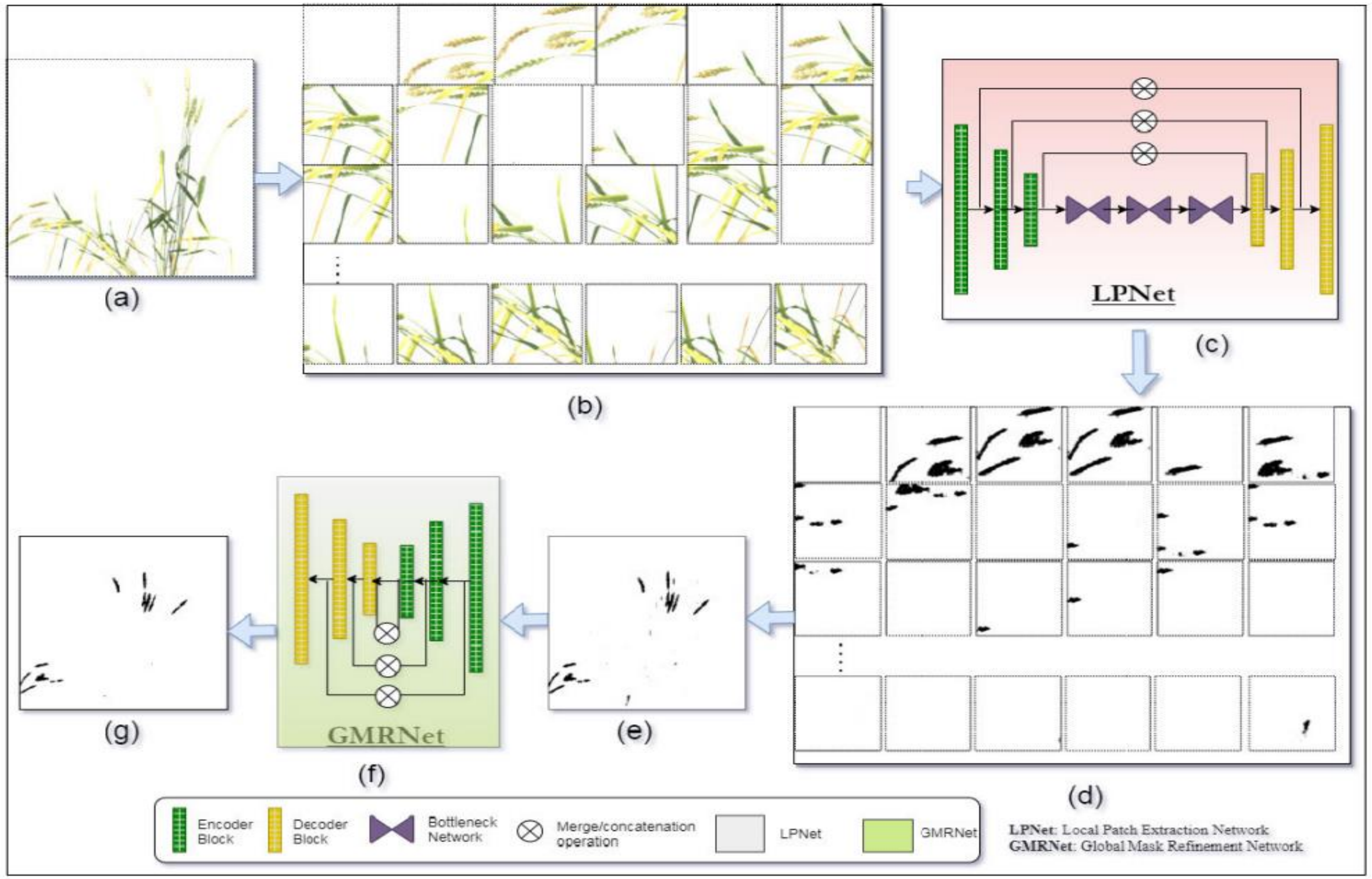
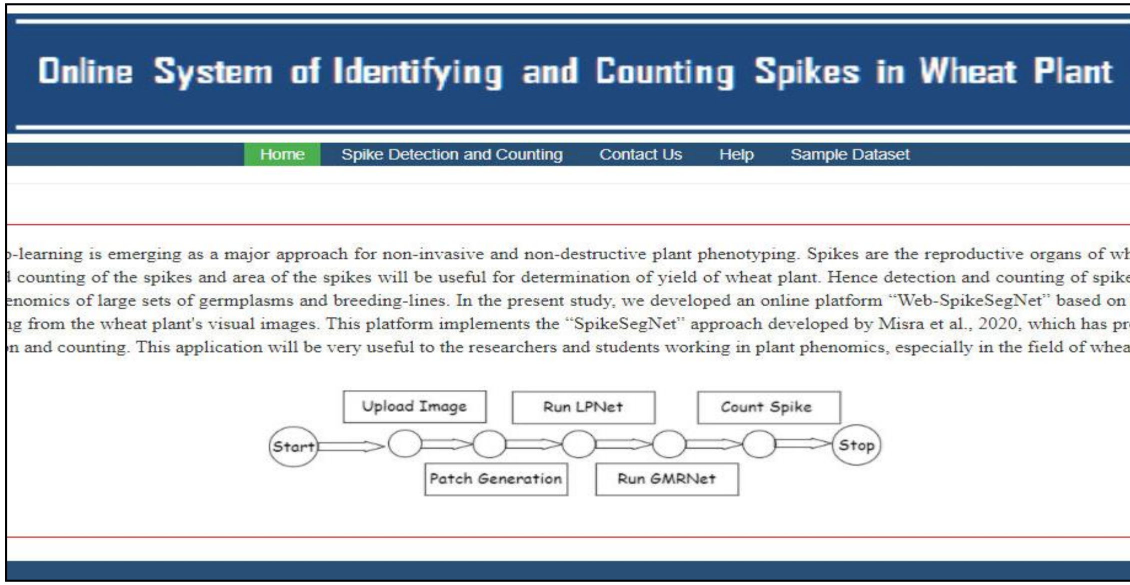


FIGURE 1. Flow diagram of SpikeSegNet: Here, input is visual image of wheat plant of size 1656*1356. The input image is divided into patches of size 256*256 before entering into the LPNet. The output of LPNet are patch-by-patch segmented mask images which are then combined to form the mask image as per 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

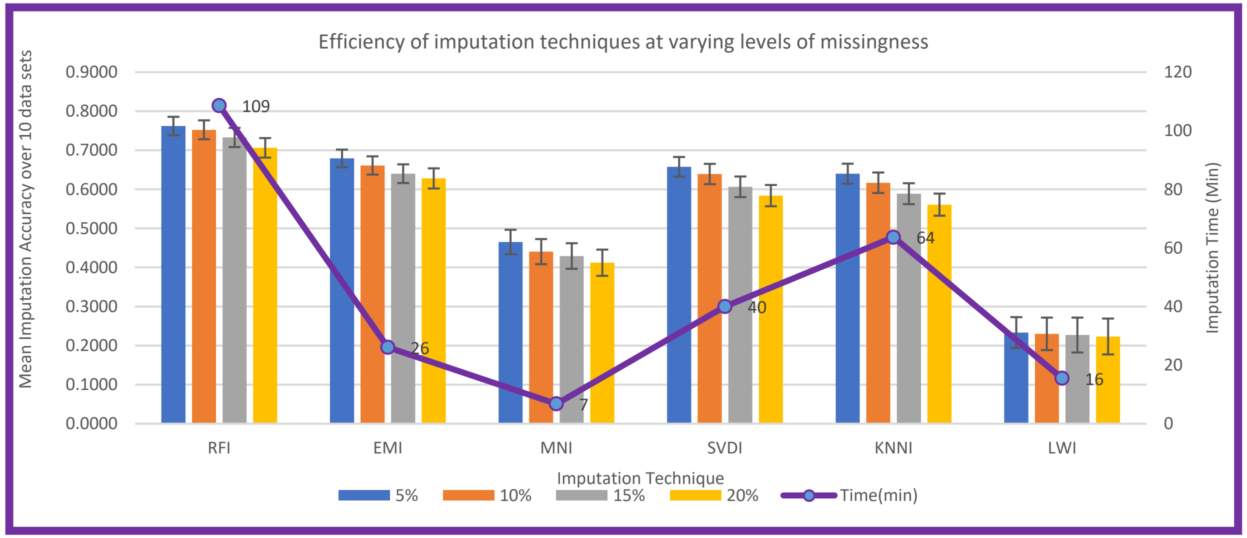
Misra, T., Arora, A., Marwaha, S., Jha, R.R.., Ray, M., Jain, R., **Rao, A.R.**, Varghese, E., Kumar, S., Kumar, S., Nigam, A., Sahoo, R.N., and Chinnusamy, V. (2021). *IEEE Access*, **9**, 76235-76247.

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)

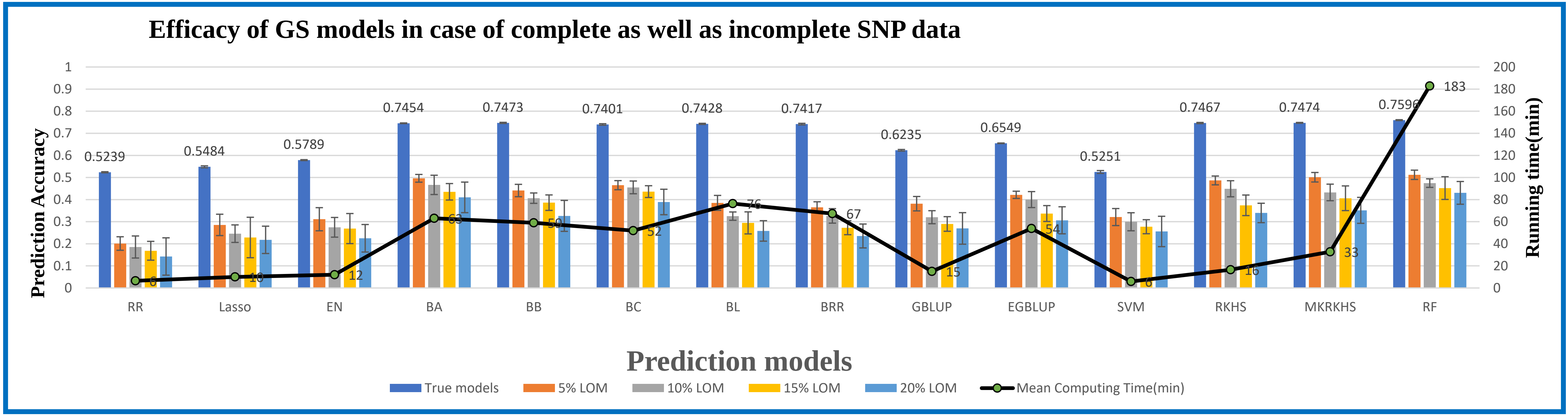
III. Models Based on Bayesian Approach

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

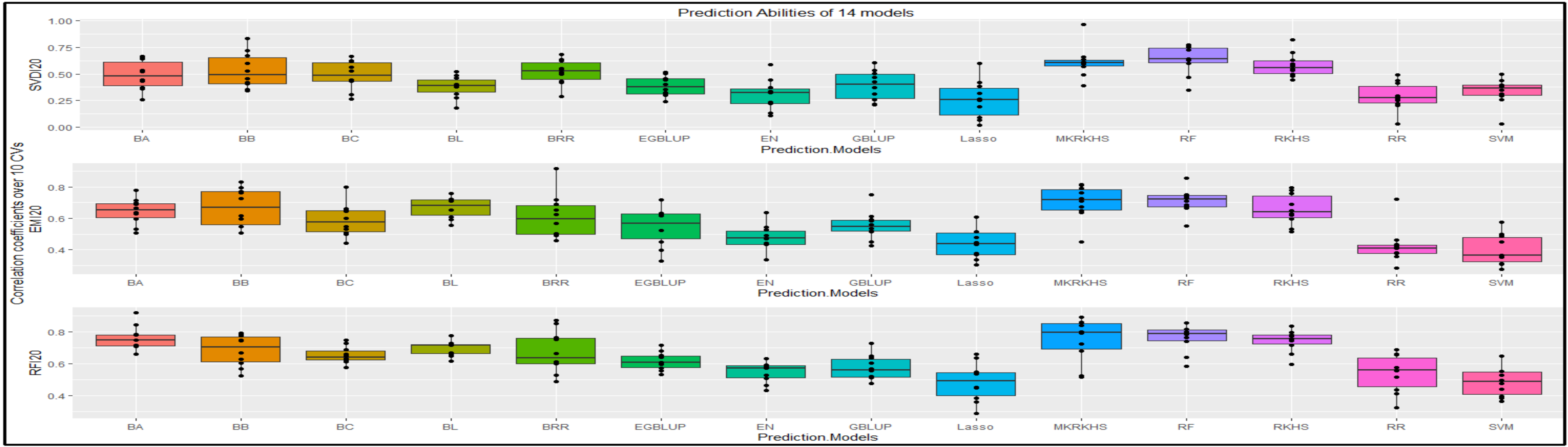
IV. Models based on Machine learning algorithms

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

https://www.panzea.org/data

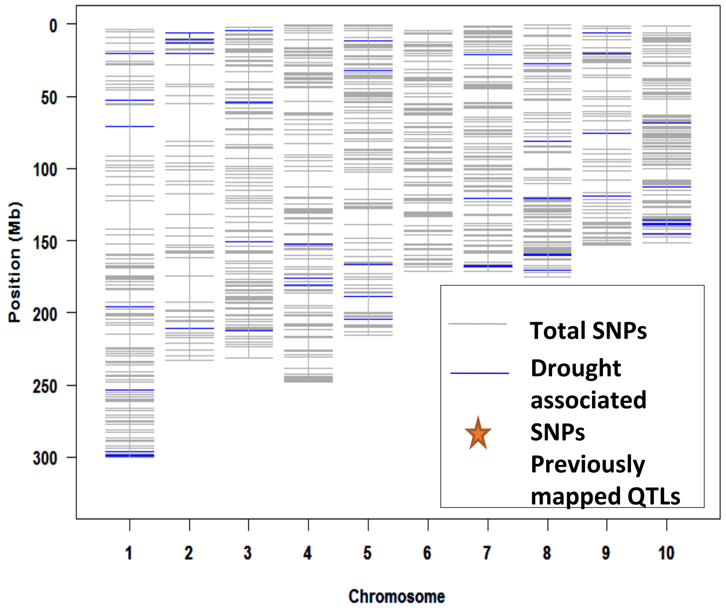


Efficacy of GS models in case of complete as well as incomplete SNP data
Prediction Accuracy
Running time(min)
0.5239
0.5484
0.5789
0.7454
0.7473
0.7401
0.7428
0.7417
0.6235
0.6549
0.5251
0.7467
0.7474
0.7596
6
10
12
63
59
52
76
67
15
54
6
16
33
183
RR
Lasso
EN
BA
BB
BC
BL
BRR
GBLUP
EGBLUP
SVM
RKHS
MKRKHS
RF
Prediction models
True models
5% LOM
10% LOM
15% LOM
20% LOM
Mean Computing Time(min)

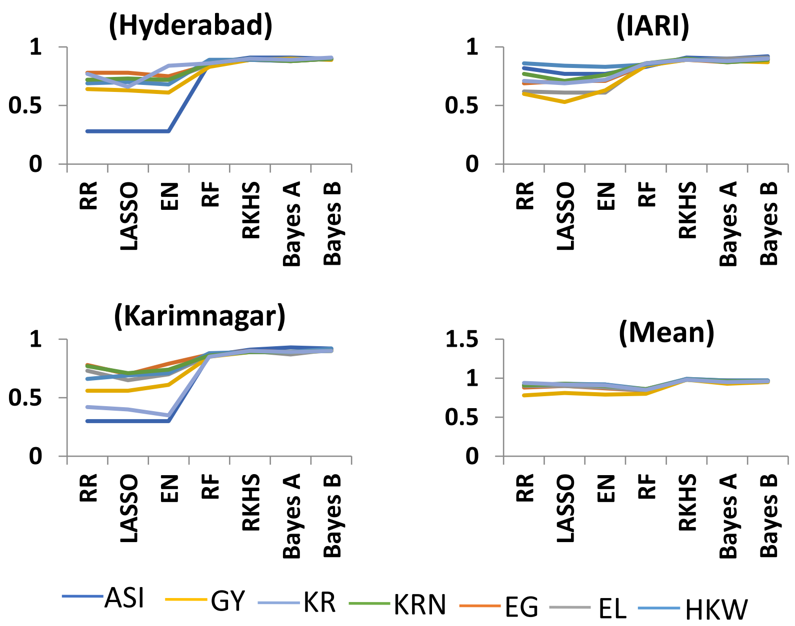


Prediction Abilities of 14 models
Correlation coefficients over 10 CVs
SVDI20
EMI20
RFI20
BA
BB
BC
BL
BRR
EGBLUP
EN
GBLUP
Lasso
MKRKHS
RF
RKHS
RR
SVM
Prediction.Models

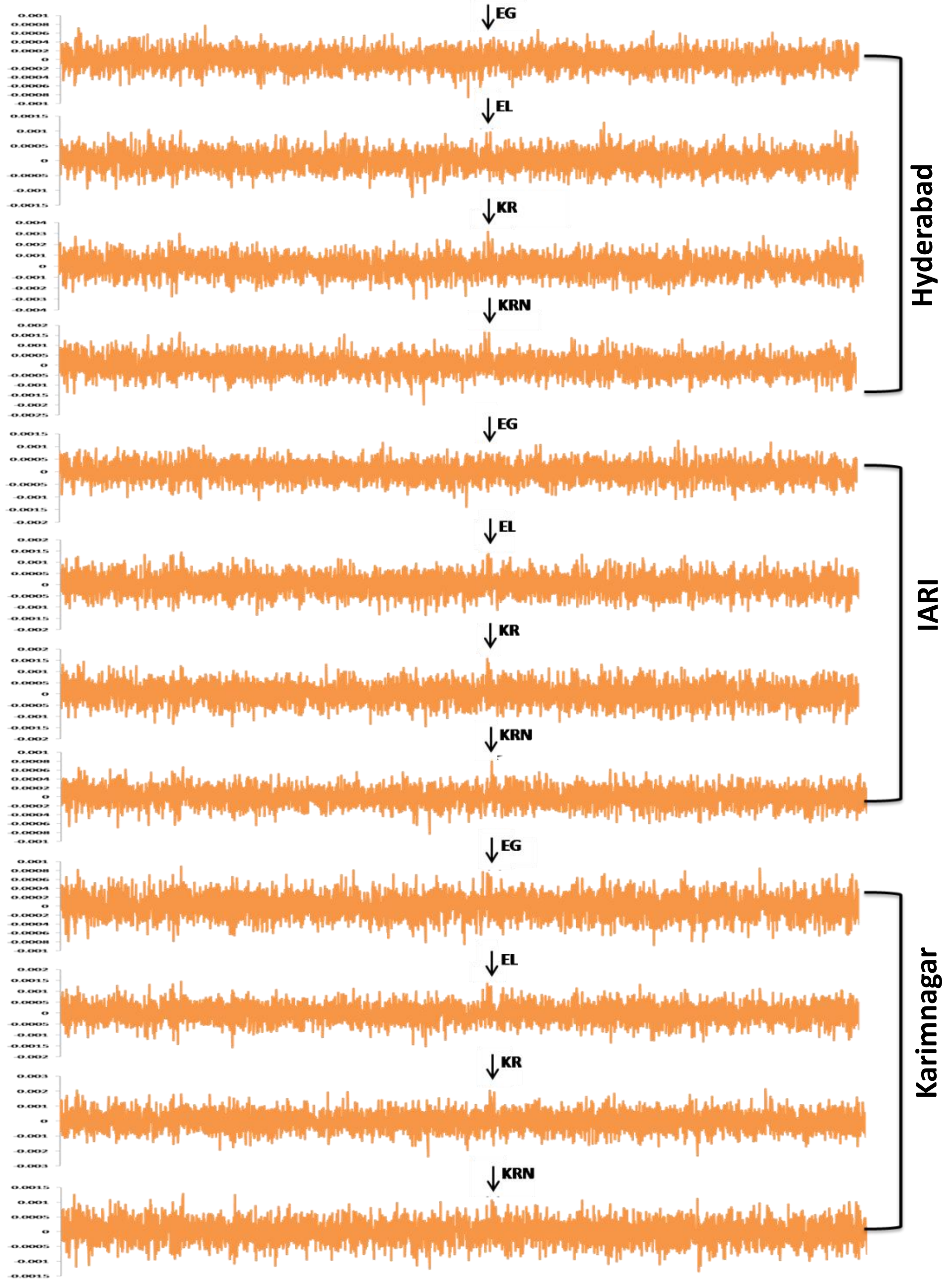
Top 1053 SNPs detected by Bayes B



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

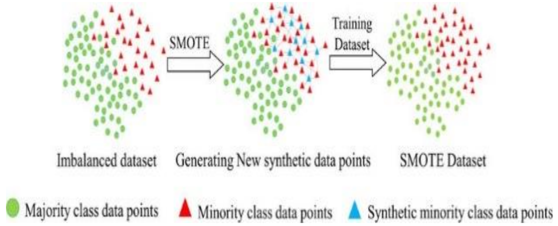
Dataset by Location	Sites	Total Sequences
Farakka	F1	2,91,28,182
	F2	5,44,69,302
Kanpur	K1	2,81,58,772
	K2	3,30,84,931
Delhi	D1	6,38,16,159
	D2	6,36,60,637

❖ 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

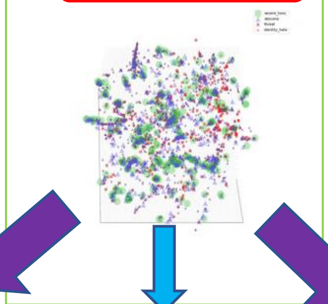
❖ Kmeans Clustering

❖ SMOT Analysis

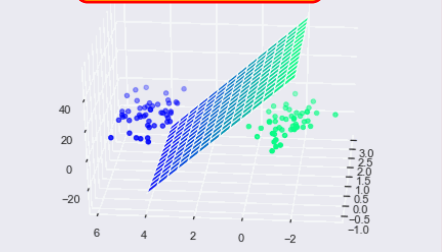


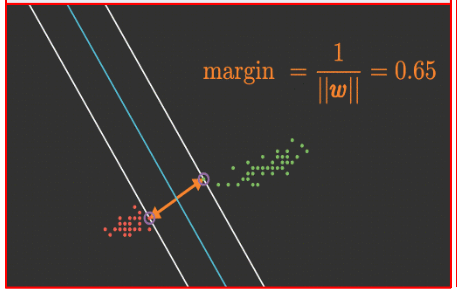
❖ Multiclass Classification

Data

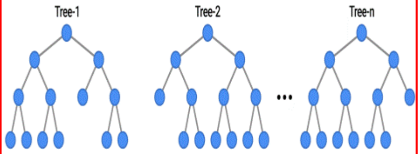


SVM

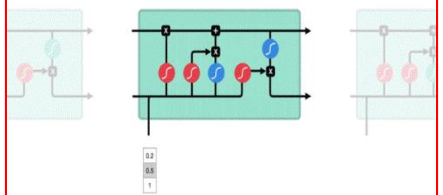


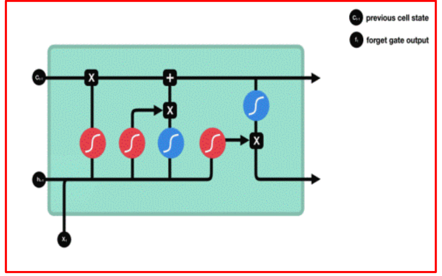


RF



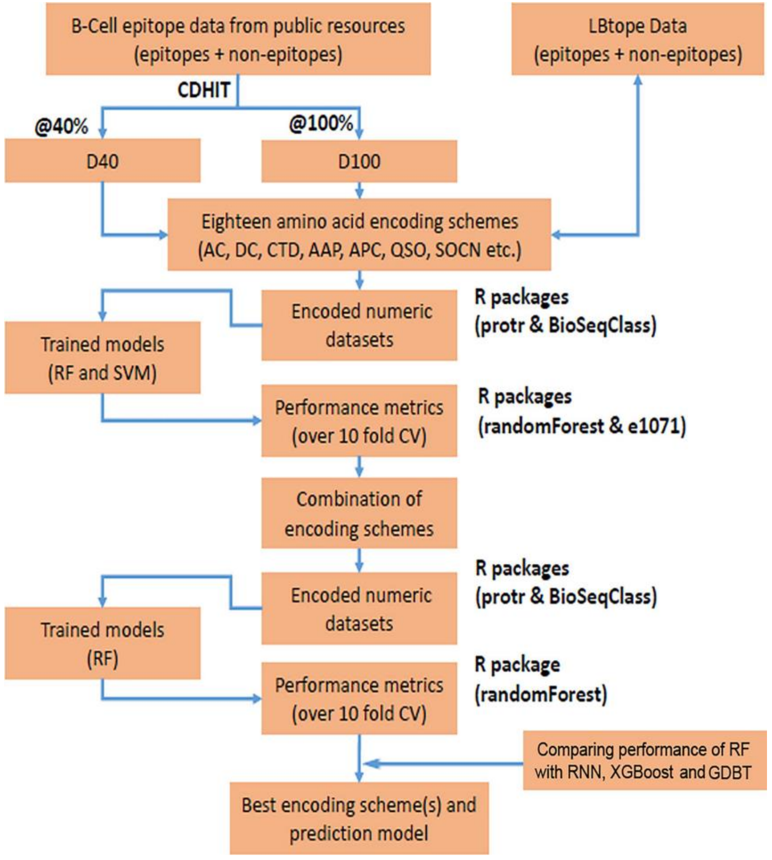
LSTM

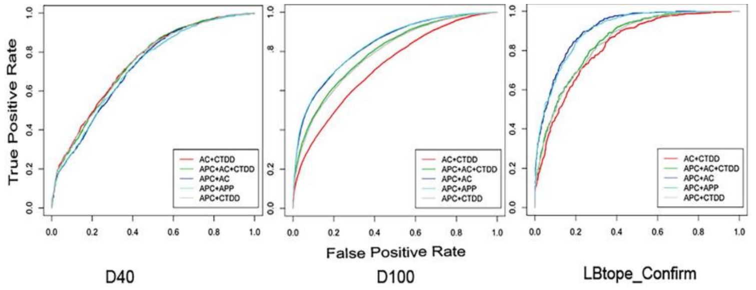


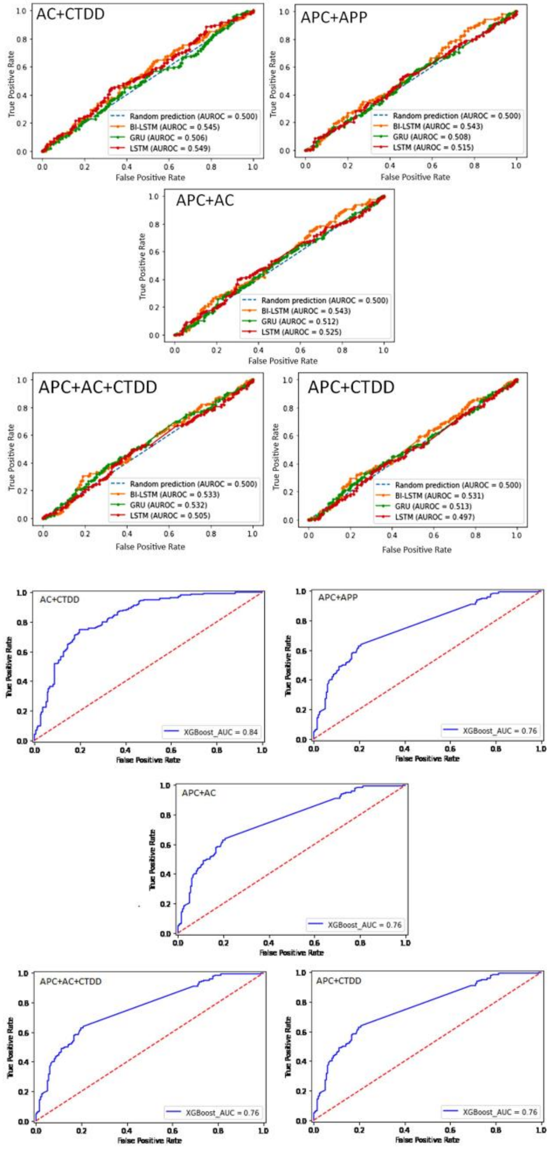


Comparative analysis of B-Cell epitope Prediction Tools

- 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 analyze 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 longmer 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 shortmer 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.







Sahu TK, Meher PK, Choudhury NK and Rao AR, (2022) A comparative analysis of amino acid encoding schemes for the prediction of flexible length linear B-cell epitopes. Briefings in Bioinformatics (proof read completed).

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 !!!