Machine learning, an impetus approach for molecular functional annotation in plants

Isha Joshi[1], Anuraj Nayarisseri[1,2]

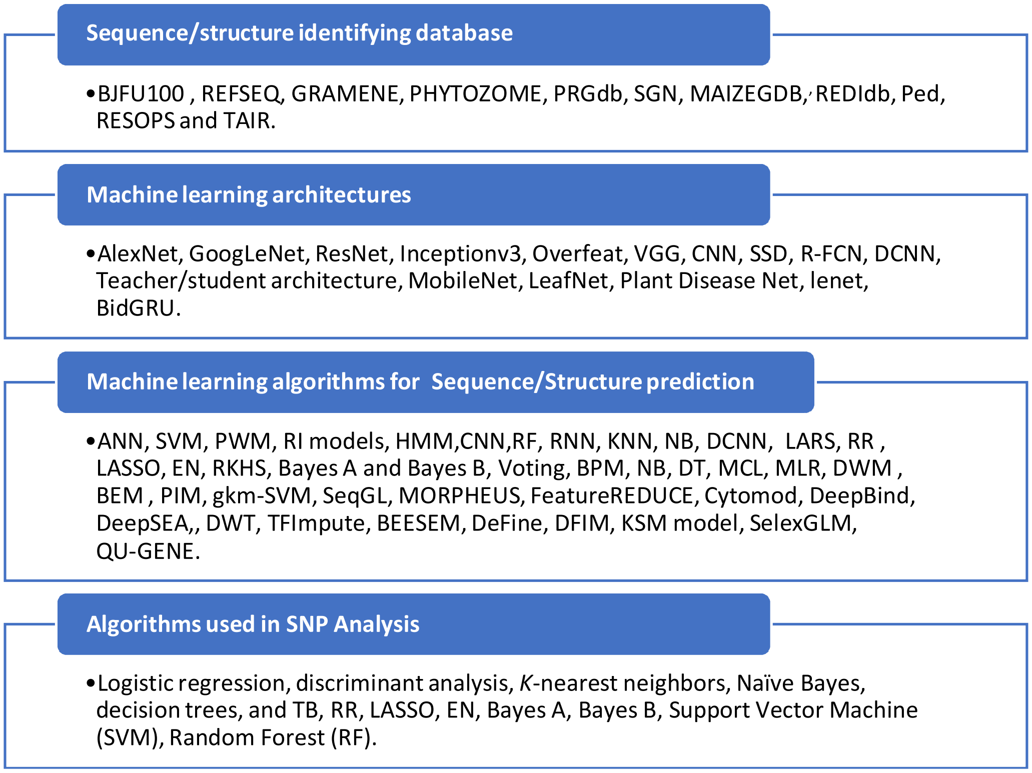
*Corresponding author: Dr. Anuraj Nayarisseri; Email: anuraj@eminentbio.com

1. *In silico* Research Laboratory, Eminent Biosciences, 91, Sector-A, Mahalakshmi Nagar, Indore – 452010, Madhya Pradesh, India.
2. Bioinformatics Research Laboratory, LeGene Biosciences Pvt Ltd, 91, Sector-A, Mahalakshmi Nagar, Indore – 452010, Madhya Pradesh, India.

Traditional agriculture research programs have used classical breeding and molecular biology approaches for crop improvement. Besides, they are proved inadequate to deal collectively with a major number of problems. High throughput sequencing has shown a way towards overcoming those barriers along with storing and evaluating various big scale datasets on experimental basis. Artificial intelligence with Machine and deep learning techniques uses a training dataset as a calibrator for performing identification, classification, quantification and prediction. Different algorithms can interpret the same data to different desirable outputs; the output includes a simpler solution for the complex problems in link with a given dataset. Its application has moved research towards less biased and high precision results which are extensively accepted on a global level [1-3].

The sophisticated application of AI and machine learning is prevalent in genomics, transcriptomics, proteomics, metabolomics and systems biology[4]. The approach of Interpreting a given dataset with deep learning algorithms mentioned in figure 1 has been used for predicting translational initiation site recognition[5],signal peptide prediction[6], subcellular localisation[7], plant effectors[8], fungal effectors[9], promoter recognition[10], mRNA based alternative splicing[11], m5cap[12], poly A site[13], RNA editing[14], epistatic state[15], gene[16] and protein function and interaction[17],mutational analysis[18], epigenetic interaction[19], gene expression analysis[20], transcription factor binding[21], Chromatin signature[21], gene–environment interactions[22], SNP detection for QTL and interactome analysis[23-25].

Single nucleotide polymorphism is one of the major molecular markers for the indication of genetic diversity for crop improvement programs. It is majorly used for the assessment of genomic breeding values. Approaches like NGS are used to locate SNP in economic improvement traits, for the easy and early domestication of beneficial crops. However, the error-prone fashion of the available NGS analysis tools is still a big concern which can lead to false-positive results. Machine learning methods have paved a way towards more precise SNP screening from the sequenced data available in large natural population [23-25]. Fig.1 depicts the available machine learning algorithm used in SNP detection. In addition to it, “Integrated SNP Mining and Utilization” (ISMU) Pipeline [26] and “SNP machine learning” (SNP-ML)[27] are two of the ML based models presently in use for SNP based QTL analysis. Use of molecular marker datasets with machine learning algorithm holds promising results in genetic analysis and hybrid breeding [28].



Sequence/structure identifying database

- BJFU100 , REFSEQ, GRAMENE, PHYTOZOME, PRGdb, SGN, MAIZEGDB, REDIdb, Ped, RESOPS and TAIR.

Machine learning architectures

- AlexNet, GoogLeNet, ResNet, Inceptionv3, Overfeat, VGG, CNN, SSD, R-FCN, DCNN, Teacher/student architecture, MobileNet, LeafNet, Plant Disease Net, lenet, BidGRU.

Machine learning algorithms for Sequence/Structure prediction

- ANN, SVM, PWM, RI models, HMM,CNN,RF, RNN, KNN, NB, DCNN, LARS, RR , LASSO, EN, RKHS, Bayes A and Bayes B, Voting, BPM, NB, DT, MCL, MLR, DWM , BEM , PIM, gkm-SVM, SeqGL, MORPHEUS, FeatureREDUCE, Cytomod, DeepBind, DeepSEA,, DWT, TFImpute, BEESEM, DeFine, DFIM, KSM model, SelexGLM, QU-GENE.

Algorithms used in SNP Analysis

- Logistic regression, discriminant analysis, *K*-nearest neighbors, Naïve Bayes, decision trees, and TB, RR, LASSO, EN, Bayes A, Bayes B, Support Vector Machine (SVM), Random Forest (RF).

Figure 1- Machine Learning tools in Plant Biology.

References

1. Nayarisseri, A. (2019). Machine Learning, Deep Learning and Artificial Intelligence approach for predicting CRISPR for the Cancer treatment. DOI: 10.3390/mol2net-05-06258 Ma, C., Zhang, H.H. and Wang, X., 2014. Machine learning for Big Data analytics in plants. *Trends in plant science*, *19*(12), pp.798-808.
2. Khandelwal, R., & Nayarisseri, A. (2019). A Machine learning approach for the prediction of efficient iPSC modeling. DOI: 10.3390/mol2net-05-06376
3. Nayarisseri, A., Khandelwal, R., Madhavi, M., Selvaraj, C., Panwar, U., Sharma, K., & Singh, S. K. (2020). Shape-based machine learning models for the potential novel COVID-19 protease inhibitors assisted by molecular dynamics simulation. *Current topics in medicinal chemistry*, *20*(24), 2146-2167.
4. Pedersen, A.G. and Nielsen, H., 1997, June. Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. In *Ismb* (Vol. 5, pp. 226-233).
5. Petersen, T.N., Brunak, S., Von Heijne, G. and Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature methods*, *8*(10), pp.785-786.
6. Lingner, T., Kataya, A.R., Antonicelli, G.E., Benichou, A., Nilssen, K., Chen, X.Y., Siemsen, T., Morgenstern, B., Meinicke, P. and Reumann, S., 2011. Identification of novel plant peroxisomal targeting signals by a combination of machine learning methods and in vivo subcellular targeting analyses. *The Plant Cell*, *23*(4), pp.1556-1572.
7. Sperschneider, J., Dodds, P.N., Singh, K.B. and Taylor, J.M., 2018. ApoplastP: prediction of effectors and plant proteins in the apoplast using machine learning. *New Phytologist*, *217*(4), pp.1764-1778.

8. Sperschneider, J., Dodds, P.N., Gardiner, D.M., Manners, J.M., Singh, K.B. and Taylor, J.M., 2015. Advances and challenges in computational prediction of effectors from plant pathogenic fungi. *PLoS Pathog*, *11*(5), p.e1004806.
9. Umarov, R.K. and Solovyev, V.V., 2017. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PloS one*, *12*(2), p.e0171410.
10. Rogers, M.F., Thomas, J., Reddy, A.S. and Ben-Hur, A., 2012. SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. *Genome biology*, *13*(1), p.R4.
11. Song, J., Zhai, J., Bian, E., Song, Y., Yu, J. and Ma, C., 2018. Transcriptome-wide annotation of m5C RNA modifications using machine learning. *Frontiers in plant science*, *9*, p.519.
12. Gao, X., Zhang, J., Wei, Z. and Hakonarson, H., 2018. DeepPolyA: a convolutional neural network approach for polyadenylation site prediction. *IEEE Access*, *6*, pp.24340-24349.
13. Giudice, C.L., Hernández, I., Ceci, L.R., Pesole, G. and Picardi, E., 2019. RNA editing in plants: A comprehensive survey of bioinformatics tools and databases. *Plant Physiology and Biochemistry*, *137*, pp.53-61.
14. Wang, D., El-Basyoni, I.S., Baenziger, P.S., Crossa, J., Eskridge, K.M. and Dweikat, I., 2012. Prediction of genetic values of quantitative traits with epistatic effects in plant breeding populations. *Heredity*, *109*(5), pp.313-319.
15. Mahood, E.H., Kruse, L.H. and Moghe, G.D., 2020. Machine learning: A powerful tool for gene function prediction in plants. *Applications in Plant Sciences*, *8*(7), p.e11376.
16. Xu, F., Li, G., Zhao, C., Li, Y., Li, P., Cui, J., Deng, Y. and Shi, T., 2010. Global protein interactome exploration through mining genome-scale data in Arabidopsis thaliana. *BMC genomics*, *11*(S2), p.S2.
17. Lloyd, J.P., Seddon, A.E., Moghe, G.D., Simenc, M.C. and Shiu, S.H., 2015. Characteristics of plant essential genes allow for within-and between-species prediction of lethal mutant phenotypes. *The Plant Cell*, *27*(8), pp.2133-2147.
18. Sinha, P., Singh, V.K., Saxena, R.K., Kale, S.M., Li, Y., Garg, V., Meifang, T., Khan, A.W., Do Kim, K., Chitikineni, A. and Saxena, K.B., 2020. Genome-wide analysis of epigenetic and transcriptional changes associated with heterosis in pigeonpea. *Plant Biotechnology Journal*, pp.1-14.
19. Dondelinger, F., Husmeier, D. and Lèbre, S., 2012. Dynamic Bayesian networks in molecular plant science: inferring gene regulatory networks from multiple gene expression time series. *Euphytica*, *183*(3), pp.361-377.
20. Lai, X., Stigliani, A., Vachon, G., Carles, C., Smaczniak, C., Zubieta, C., Kaufmann, K. and Parcy, F., 2019. Building transcription factor binding site models to understand gene regulation in plants. *Molecular plant*, *12*(6), pp.743-763.
21. Chapman, S., Cooper, M., Podlich, D. and Hammer, G., 2003. Evaluating plant breeding strategies by simulating gene action and dryland environment effects. *Agronomy Journal*, *95*(1), pp.99-113.
22. Shikha, M., Kanika, A., Rao, A.R., Mallikarjuna, M.G., Gupta, H.S. and Nepolean, T., 2017. Genomic selection for drought tolerance using genome-wide SNPs in maize. *Frontiers in plant science*, *8*, p.550.
23. Zhao, N., Han, J.G., Shyu, C.R. and Korkin, D., 2014. Determining effects of non-synonymous SNPs on protein-protein interactions using supervised and semi-supervised learning. *PLoS Comput Biol*, *10*(5), p.e1003592.
24. Korani, W., Clevenger, J.P., Chu, Y. and Ozias-Akins, P., 2019. Machine learning as an effective method for identifying true single nucleotide polymorphisms in polyploid plants. *The Plant Genome*, *12*(1), pp.1-10.
25. Azam, S., Rathore, A., Shah, T.M., Telluri, M., Amindala, B., Ruperao, P., Katta, M.A. and Varshney, R.K., 2014. An integrated SNP mining and utilization (ISMU) pipeline for next generation sequencing data. *PLoS One*, *9*(7), p.e101754.

26. Bhardwaj, A. and Bag, S.K., 2019. PLANET-SNP pipeline: PLants based ANnotation and Establishment of True SNP pipeline. *Genomics*, *111*(5), pp.1066-1077.
27. Ornella, L. and Tapia, E., 2010. Supervised machine learning and heterotic classification of maize (Zea mays L.) using molecular marker data. *Computers and electronics in agriculture*, *74*(2), pp.250-257.
28. Limaye, A., & Nayarisseri, A. (2019). Machine learning models to predict the precise progression of Tay-Sachs and Related Disease. DOI: 10.3390/mol2net-05-06180
29. Udhwani, T., & Nayarisseri, A. (2019). A Machine Learning approach for the identification of CRISPR/Cas9 nuclease off-target for the treatment of Hemophilia. DOI: 10.3390/mol2net-05-06179