Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 1826 - 1833
Received 25 April 2021; Accepted 08 May 2021.

Prediction and Analysis of Plant Growth Promoting Bacteria using Machine Learning for Millet Crops

[*]V. Indumathi[1], S.Santhana Megala[2], R.Padmapriya[3], M.Suganya[4] and B.Jayanthi[5]
[1, 2]Assistant Professor, [3,4,5] HOD(BCA, IT & CS) School of Computer Studies, RVS College of Arts and Science, Coimbatore, Tamilnadu, India
[1*]indhumathi@rvsgroup.com, [2]santhanamegala@rvsgroup.com
[3]padmapriya@rvsgroup.com, [4]suganya@rvsgroup.com, [5]jayanthi@rvsgroup.com

Abstract:

Objectives: Microorganisms present within the rhizosphere play important roles in ecological fitness of their plant host. Important microbial processes that are expected to occur within the rhizosphere include pathogenesis and its counterpart, along with plant protection and growth promotion. This paper deal's with predicting the Plant Growth Promoting Rhizobacteria from the microbes using Machine learning techniques to enhance the plant growth.
Methods: This paper presents some Machine Learning approaches such as LDA, KNN and SVM for analyzing the genomes and predicts the Rhizosphere molecular mechanisms to find the Growth promoting bacteria and recommending the secondary metabolic model for further growth. NCBI Dataset was used for the experimental purpose.
Findings: Among the Machine learning techniques used in this paper such as LDA, KNN and SVM, the best accuracy was obtained by SVM in predicting the Plant Promoting Rhizobacteria.
Novelty: Exploring these microorganisms by unraveling their possible relationships with plants has launched a new and fascinating area of investigations in the rhizosphere research. As a beginning, we tried using Machine Learning to predict the Rhizobacteria, which paves the way for further studies.

Keywords: Plant Growth Promiting Bacteria, Machine Learning, Rhizobacteria, Microbes, Genomes.

1 Introduction

Plant Growth Promoting Bacteria (PGPB) are bacteria which will enhance plant growth and protect plants from disease and abiotic stresses through a good sort of mechanisms; people who establish close associations with plants, like the en-dophytes, might be more successful in plant growth promotion[1]. Several important bacterial characteristics, like biological organic process , phosphate solubiliza-tion, ACC deaminase activity, and production of siderophores and phytohor-mones, are often assessed as plant growth promotion (PGP) traits. The plant growth-promoting bacteria (or PGPB) belong to a beneficial and heterogeneous group of microorganisms which will be found within the rhizosphere, on the basis surface or asso-ciated thereto , and are capable of enhancing the expansion of plants and protecting them from disease and abiotic stresses. Plant-microbe interactions within the rhizosphere are the determinants of plant health, productivity and soil fertility.

The rhizosphere may be a term, which was first introduced by a scientist L.Hiltner. It's the region that's a couple of distances (2-80 mm) extended from the basis system. It also can define as a zone, which favours the physical and chemical activity of the microorganisms and liable for the extreme microbial activity. The speed of microbial activity strongly influences the method of root

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 1826 - 1833
Received 25 April 2021; Accepted 08 May 2021.

exudation. The organic and inorganic wastes of the basis system are called root exudates. Rhizosphere zone is that the region of intense microbial activity, and it's isolated from the majority soil that always called as Edaphosphere or Non-rhizosphere[2].

A rhizosphere may be a region containing an incredible amount of rhizo deposits that creates it the foremost desirable zone for the microbial proliferation. we will remind the term, by understanding an easy incontrovertible fact that may be a sphere containing root secretions and diffusates which is in close vicinity to the rhizoplane or root surface.

Interactions between plants and bacteria occur through symbiotic, endophytic or associative processes with distinct degrees of proximity with the roots and sur-rounding soil denoted in Fig. 2. Endophytic PGPB are good inoculant candidates, because they colonize roots and make a positive environment for development and performance . Non-symbiotic endophytic relationships occur within the intercellu-lar spaces of plant tissues, which contain high levels of carbohydrates, amino ac-ids, and inorganic nutrients [3].

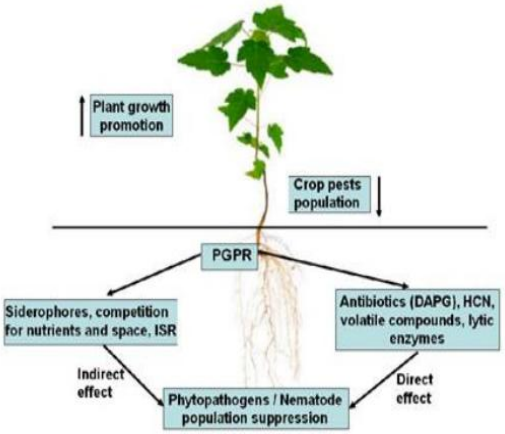

Fig. 1. Overall Activity of Plant Growth Promotion

The ability to get complete genome sequences from bacteria in environmental samples, like soil samples from the rhizosphere, has highlighted the microbial diversity and complexity of environmental communities. However, new algorithms to research genome sequence information within the context of community structure are needed to reinforce our understanding of the precise ecological roles of those organisms in soil environments. We present a machine learning approach using LDA, KNN and SVM to analyse the genomes including outputs of metabolic and transportomic computational models for identifying the foremost predictive molecular mechanisms indicative of a rhizosphere. Computational predictions of niche were highly accurate overall with models trained on transportomic model output being the foremost accurate. The strongest predictive molecular mechanism features for rhizosphere niche overlap with many previously reported analyses of Pseudomonad interactions within the rhizosphere, suggesting that this approach successfully informs a system-scale level understanding of how Pseudomonads sense and interact with their environments. The observation that an organism's transportome is very predictive of its niche may be a novel discovery and should have implications in our understanding microbial ecology. The framework developed here are often generalized to the analysis of any bacteria across a good range of environments and ecological niches making this approach a strong tool for providing insights into functional predictions from bacterial genomic data.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 1826 - 1833
Received 25 April 2021; Accepted 08 May 2021.

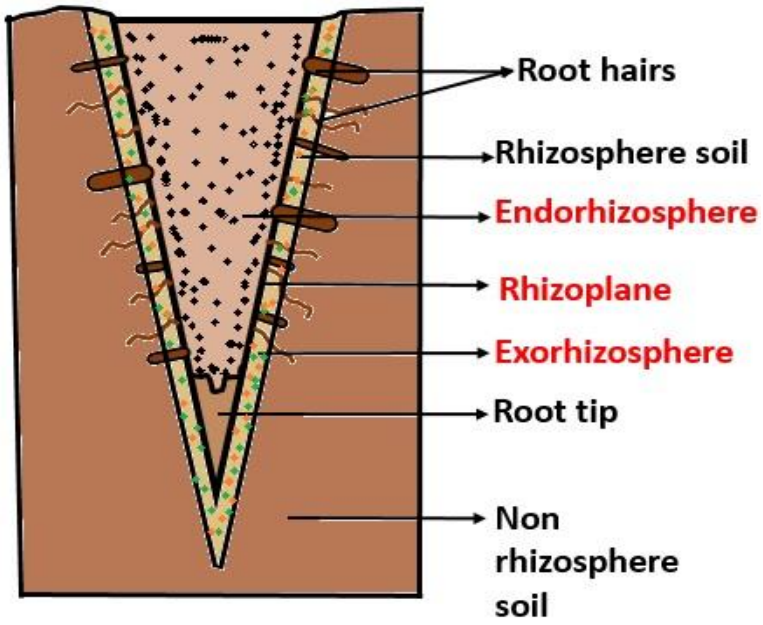

Fig. 2. Structure of Rhizosphere

2 Methodology

The objective of this paper is to extend the accuracy of prediction model by using different parameters for future agriculture. the info set has been taken from NCBI and applied to the processing model[4]. To process the info , map reduce is employed to attenuate the time of execution and further classification is completed . Preprocessing, Feature selection and extraction are important steps in classification systems. The proposed approach is split into three phase, which was denoted in Fig. 3.

2.1 Data Pre-Processing:

Data pre-processing is deemed to be the most step within the data processing method and machine learning projects. In effective data collection can subsequently cause improper combination, weak control and incorrect values. It moreover provides misleading results if the improper data is in the dataset. Therefore, it is essentially important to hold over quality and representation of knowledge at the initial stage by removing the null values and improper values.

2.2 Feature selection:

Feature selection is that the process of choosing a subset of relevant features to be used in model construction. Feature selection is beneficial if the info contains many features that are either redundant or irrelevant, and may thus be removed without incurring much loss of data . Feature selection reduces over fitting by eliminating redundant data which causes noise. This leads to improvement in accuracy and faster training. During this study, supervised filter method is applied on the microbes data for identifying important features[5].

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 1826 - 1833
Received 25 April 2021; Accepted 08 May 2021.

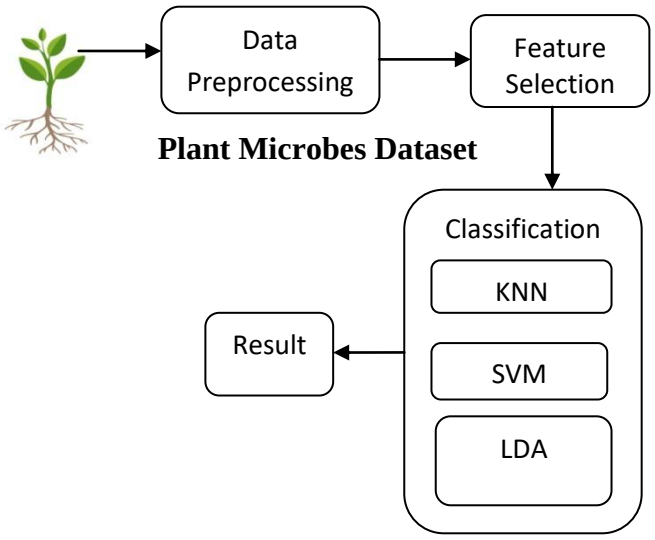

Fig. 3. Proposed Architecture for Predicting the Plant Growth Promotion

2.3 Classification:

2.3.1 K-Neighrest Neighbour

A refinement of the k-NN classification algorithm is to weigh the contribution of every of the k neighbors consistent with their distance to the query point xq, giving greater weight wi to closer neighbors [6]. It is given by

$$F(x_q) = \frac{\sum_{i=1}^{k} w_i f(x_i)}{\sum_{i=1}^{k} w_i} \qquad (1)$$

Where the weight is,

$$w_i = \frac{1}{d(x_q, x_i)^2}$$

In case x_q exactly matches one among x_i in order that the denominator becomes zero, we assign $f(x_q)$ equals $f(x_i)$ during this case. It is sensible to use all training examples not just k if weighting is employed, the algorithm then becomes a worldwide one. The main disadvantage is that the run time of this algorithm is bit longer.

2.3.2 Support Vector Machine

SVM may be a supervised machine learning algorithm which works supported the concept of decision planes that defines decision boundaries. a choice boundary separates the thing s of 1 class from the object of another class [7]. Support vectors are the info points which are nearest to the hyper-plane. Kernel function is employed to separate non-linear data by transforming input to a better dimensional space. Gaussian radial basis function kernel is employed in our proposed method.

$$k(x_i, x_j) = e^{-\|x_i, x_j\|^2 / 2\emptyset^2} \quad (2)$$

Where, $k(x_i, x_j)$ = Feature vectors in input space, $\|x_i, x_j\|^2$ = High dimensional space of X and Y coordinate, and σ is a free parameter.

The constructed SVM based rhizosphere classification model is evaluated using trained model developed based on the training dataset and evaluated against the test dataset. A random split of 70% of knowledge for training and 30% of knowledge for testing is performed. The trained model isn't exposed to the test data set during training and hence, the predictions made on

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 1826 - 1833
Received 25 April 2021; Accepted 08 May 2021.

the test dataset are indicative of the performance of the model.

2.3.3 Linear Discriminant Analysis (LDA):

Linear Discriminant Analysis (LDA) may be a classification method originally developed in 1936 by R. A. Fisher. it's simple, mathematically robust and sometimes produces models whose accuracy surpasses more complex methods. This linear classification model doesn't require multiple passes over the info for optimization.

Algorithm:

1. Fisher's Linear Discriminant Analysis (LDA) builds j = min(k1,p) discriminant functions that estimate Discriminant scores (D ji) for each of i = 1,...,n subjects (soil samples) classified into k groups, from p linearly independent predictor variables (x) as Dji=wi1X1i+wi2X2i +…+wipXpi , [i=1,…,n and j=1,…,min(k-1,p)]
2. Discriminant weights (w ij) are estimated by ordinary least squares so that the ratio of the variance within the k groups to the variance between the k groups is minimal.
3. Classification functions of the type Cji=cjo+cj1X1i+cj2X2i+…+cjpXpi for each of the j = 1,...,k groups can therefore be constructed from the Discriminant scores.
4. The coefficients of the classification function for the j th group are estimated from the within sum of squares matrices (W) of the Discriminant scores for each group and from the vector of the p Discriminant predictors means in each of the classifying groups (M) as C j = W-1M with cjo=logp-12/CjMj .
5. A subject is then classified into the group for which its classification function score is higher.

3. EXPERIMENTAL RESULS & DISCUSSION:

Confusion matrix is one among the foremost intuitive metrics used for locating the correctness and accuracy of the model. The performance of the study made during this paper for predicting plant Growth model is assessed supported the confusion matrices. The performance metrics used to find the efficiency of the system were Accuracy, Kappa Coefficient, Precision, Recall and F-Measure.

Accuracy is defined because the fraction of instances that are correctly classified samples. The Kappa statistic may be a measure of agreement between the predictions and therefore the actual class labels and the value ranges from 0 and 1. İt is a measure of comparison of overall accuracy to the expected random chance accuracy. The upper the Kappa metric of the classifier, the higher is that the classifier as compared to a random chance classifier. Precision is that the ratio of correctly predicted positive observations to the entire positive predicted observations (TP/(TP+FP)). Recall is defined because the ratio of correctly predicted positive observations to the entire observations within the actual class (TP/(TP+FN)). F measure is that the weighted average of Precision and Recall. It takes both False Positives and False Negatives under consideration . The derived overall performance metrics of the classification model based on the confusion matrix are shown in Table 1.

Table 1: Comparison of Classification methods based on the experimental results

	Accuracy	**Kappa**	**Precision**	**Recall**	**F-measure**
KNN	82	0.79	0.84	0.71	0.84
LDA	75	0.61	0.72	0.69	0.76
SVM	97	0.95	0.97	0.91	0.95

The SVM methods got good accuracy and Kappa of the classification model such as 97% and

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 1826 - 1833
Received 25 April 2021; Accepted 08 May 2021.

0.956 respectively.

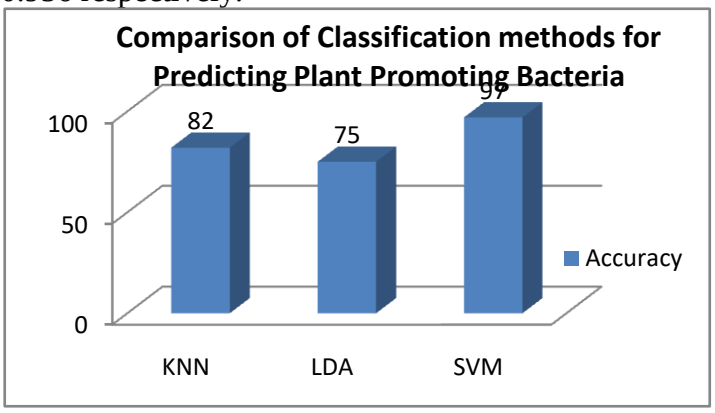

Fig. 4. Comparison of Classification methods for Predicting Plant Promoting Bacteria based on Accuracy

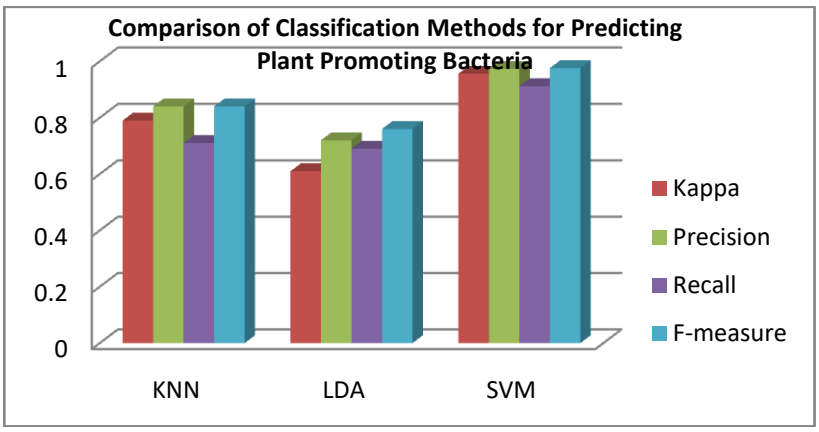

Fig. 5. Comparison of Classification methods for Predicting Plant Promoting Bacteria based on Kappa Coefficient, Precision, Recall and F-Measure

The Precision, Recall and F-measure for SVM are observed to be 0.97, 0.91 and 0.95 respectively. These results show that SVM model are often used more effectively to predict the plant Growth bacteria using microbes for facilitating improved crop and soil management.

4. CONCLUSION

This paper proposed a model for predicting Plant Promoting Bacteria and providing suitable Metobolic combination for that specific plant. The Experimental results has been done using NCBI Dataset. The model has been tested by applying different sorts of machine learning algorithm. LDA and K-NN shows good accuracy but among all the three classifiers, SVM has given the most best accuracy in predicting Plant Promoting Bacteria. The proposed model is justified by a standard dataset and machine learning algorithms. Accurate prediction of Plant Promoting Bacteria and also the recommendation of metabolic combination for specific plant are more appropriate than many existing methods. In future, providing fertilizer recommendation is our concern, also data of other districts are going to be added to form this model more reliable and accurate.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 1826 - 1833
Received 25 April 2021; Accepted 08 May 2021.

References

1. Khabbaz, S.E.; Ladhalakshmi, D.; Babu, M.; Kandan, A.; Ramamoorthy, V.; Saravanakumar, D.;Al-Mughrabi, T.; Kandasamy, S. Plant Growth Promoting Bacteria (PGPB)—A Versatile Tool for Plant Health Management. Can. J. Pestic. Pest Manag. 2019, 1(1), 1–25; doi:10.34195/can.j.ppm.2019.05.001.
2. Nakkeeran, S.; Fernando, D.W.G.; Siddiqui, Z.A. Plant growth promoting rhizobacteria formulations and its Scope in commercialization for the management of pests and diseases. In PGPR: Biocontrol and Biofertilization; Siddiqui, Z.A., Ed.; Springer: Dordrecht, The Netherlands, 2011; pp. 257–296.
3. Mhatre, P.H.; Karthik, C.; Kadirvelu, K.; Divya, K.L.; Venkatasalam, E.P.; Srinivasan, S.; Ramkumar, G.; Saranya, C.; Shanmuganathan, R. Plant growth promoting rhizobacteria (PGPR): A potential alternative tool for nematodes bio-control. Biocatal. Agric. Biotechnol. 2019, 17, 119–128. [CrossRef]
4. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov
5. Larsen, Peter E et al. "Predicting Ecological Roles in the Rhizosphere Using Metabolome and Transportome Modeling." PloS one, vol. 10, Is: 9, 2015, DOİ:10.1371/journal.pone.0132837
6. Ruchi Srivastava, Alok K. Srivastava, Promod W. Ramteke, Vijai K. Gupta, Anchal K. Srivastava,Metagenome dataset of wheat rhizosphere from Ghazipur region of Eastern Uttar Pradesh,Data in Brief,Volume 28,2020,https://doi.org/10.1016/j.dib.2019.105094.
7. K. Radhika and D. Madhavi Latha, Machine learning model for automation of soil texture classification, Indıan Journal Of Agrıcultural Research, Volume 53 Issue 1, 2019, DOI: 10.18805/IJARe.A-5053
8. Sels J, Mathys J, De Coninck BM, Cammue BP, De Bolle MF. Plant pathogenesis-related (PR) proteins: a focus on PR peptides. Plant physiology and biochemistry: PPB / Societe francaise de physiologie vegetale. [Research Support, Non-U.S. Gov'tReview]. 2008. November;46(11):941–50.
9. R.M. Leggette, M.D. Clark, A world of opportunities with Nanopore sequencing, Journal of Experimental Botany, Vol:68, Is: 20, pp. 5419-5429, 2017,
10. Kudoyarova GR, Vysotskaya LB, Arkhipova TN, et al. Effect of auxin producing and phosphate solubilizing bacteria on mobility of soil phosphorus, growth rate, and P acquisition by wheat plants. Acta Physiol Plant. Vol:39, Is:253, 2017.