ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

PhzPred – A Tool for Prediction of Phenazine Synthesizing Genes in Plant Growth Promoting *Pseudomonas* spp

Shilpa S[1], Anil Paul[1], Naganeeswaran S[1], Hemalatha N[2*], Rajesh M.K[1]

Bioinformatics Centre, Central Plantation Crops Research Institute, Kasaragod, Kerala, India[1]

Bioinformatics Centre, AIMIT, St. Aloysius College, Mangalore, Karnataka, India[2]

*Corresponding author

ABSTRACT: Phenazines are natural products produced by the bacterial strain of *Pseudomonas spp.* which possess anti-microbial activities and include more than 50 pigmented heterocyclic nitrogen containing secondary metabolites. Seven core phenazine biosynthetic genes have been identified in nearly all identified bacterial strains that produce phenazine compounds. In this study, a model has been developed to predict the phenazine biosynthetic genes from a set of protein sequences usingmachine learning algorithms from whole genomes of *Pseudomonas spp*. Initially, protein sequences from the *Pseudomonas spp.* were retrieved from public databases and used to train the WEKA models. To train the different classifiers in WEKA, three amino acid compositions were used: monomer amino acids, dipeptide amino acids, and a hybridmethod. The trained models were then used for the prediction of phenazine synthesizing gene in anuser submitted sequence. The best WEKA modules were selected based on the performance of different classifiers in training and testing. The performances of the classifier's were then evaluated based on 10-fold cross validationand independent data set validation techniques. In the proposed methodology, better performance was observed for the hybrid feature extraction method. The development of a genome wide prediction tool for phenazinesynthesizing genes will substantially have an impact on bacterial genome annotation and devising crop protection strategies using plant growth promoting rhizobacteria.

KEYWORDS: Plant growth promoting rhizobacteria, Genome, Prediction, Machine learning, Algorithms

I. INTRODUCTION

Plant growth promoting rhizobacteria(PGPR) are soil bacteria inhabiting around or on the plant root surface and which are known to be involved in promoting plant growth and development through production and secretion of various secondary metabolites in the vicinity of rhizosphere [1]. They facilitate plant growth by either (i) mobilizing soil nutrients or thereby directly assisting in the acquisition of minerals from the soil (ii) diminishing the inhibitory effects of various plant pathogens by production of secondary metabolites, or (iii) modulating plant hormone levels [2,3]. Many studies have reported improvement of crop health and productivity after application of PGPR and therefore, PGPR can play a vital role in decreasing reliance on harmful agricultural chemicals, whose application could potentially weaken crop-ecosystems[3].

Phenazines are pigmented, heterocyclic nitrogen-containing secondary metabolites synthesized by some strains of fluorescent *Pseudomonas* spp. which are known to possess anti-microbial activity [4]. More than 100 different structural derivatives of phenazine have been documented and the differences in the physical, chemical and biological properties of the individual phenazines have been attributed to the nature and position of substituents on the heterocyclic rings [4]. Phenazinescan perform multitude of functions, some of which include acting as electron shuttles and cell signals and modifying cellular redox states[6]. In addition, they play a part in biofilm formationand architecture, thus enhancing bacterial survival. Phenazines have diverse effects on hosts and they are capable of modifying many host cellular responses. Phenazinescan also influence plant growthand elicit induced systemic resistance[5].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

Data obtained from whole genome sequencing of microbial genomes have revealed knowledge of the phenazine biosynthesis operons, which include requires genes encoding the proteins PhzA, Phz B, PhzC, PhzD, PhzE,PhzFand PhzG. In fluorescent *Pseudomonas spp.*, phenazine biosynthesis commences late in bacterial growth phase and is known to be regulated by quorum sensing and twocomponentsensory transduction elements[6].

With the potential of phenazines in biological control of plant pathogens, it becomes imperative to explore and gain in depth insights into theevolution of phenazine biosynthesis as well as identification and future engineering of *Pseudomonas* strains capableof synthesizing novel and valuable phenazineswhich could find immense applications in environmental, industrial, and pharmaceutical industries.

In computational Biology, 'gene prediction' or 'gene finding' refers to the process of identifying regions of genome encoding proteins. While the genomes of many organisms have been sequenced over the last two decades, transforming such raw sequence data into knowledge remains a major bottleneck. A number of prediction programs have been developed that try to address one part of this problem which consists of locating the genes along a genome. The aim of the present study is to develop a tool, implementing machine learning algorithms, which couldpredictthefollowinggenesinvolvedinphenazinebiosynthesis in plant growth promoting *Pseudomonas spp.*

II. METHODOLOGY

2.1 Datasets

The most important task is the development of a prediction tool is the selection of dataset. Protein sequences of PhzA, Phz B, PhzC, PhzD, PhzE, PhzF and PhzG from plant growth promoting *Pseudomonas* spp. were downloaded from NCBI. For e.g., for PhzA, a total of 2181 protein sequences constituted the positive dataset. A total of 2598 non-PhzA proteins constituted the negative dataset. For training and testing, we used independent data test, where sequences in the training set and test set are entirely different.

2.2 WEKA

WEKA (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java [7]. The Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality.

2.3 Composition-based Methods

We have used three different approaches, based on amino acid composition, to train the different classifiers[8].

- *Amino-acid composition:* Amino-acid composition is the fraction of each amino acid in a given protein sequence. This has a feature dimension of 20.

- *Dipeptide composition:* Dipeptide composition, which gives a fixed pattern length of 400 (20x20), encompasses the information of the amino-acid composition along with the local order of amino acids.

- *Hybrid method:* The hybrid method was developed by combining amino-acid composition and dipeptide composition features of a protein sequences. The method created an input vector pattern of 420 (20 for amino acid and 400 for dipeptide composition) was created.

2.4 Measurement of performance of PhZPred

10-fold cross-validation and independent data set validation techniques were adopted for performance measurement. For 10- fold cross-validation, the relevant dataset was partitioned randomly into ten equally sized sets. The training and testing was carried out ten times with each distinct set used for testing and the remaining nine sets for training. None of the data to be tested occurs in the training dataset used to train the predictor in the independent dataset test. Here, the selection of data used for the testing dataset is quite arbitrary.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

2.5 Machine Learning

Naive Bayes

A Bayes classifier is a simple probabilistic classifier based onBayes' theorem with strong (naive) independence assumptions. Hence the name naive Bayes. “Independentfeature model" is a more descriptive term for the underlying probability model. In simple terms it can be defined as classifier that assumes the presence (or absence) of a particular feature of a class unrelated to the presence (or absence) of any other feature.

Logistic functions

A logistic function is a common "S" shape (sigmoid curve) with equation given as

$$f(x) = \frac{1}{1+e^{-x}} \tag{1}$$

where *e* is the natural logarithm base for values of *x* in the range of real numbers from $-\infty$ to $+\infty$. The function was named in 1844-1845 by Pierre François Verhulst, who studied this function in relation to population growth [9]. The initial stage of growth is approximately exponential from which saturation begins and as it begin the growth slows, and at maturity, growth stops. This function finds applications in various fields, including artificial neural networks, biology (ecology,biomathematics), chemistry, demography, economics, geosciences, mathematical psychology, probability, sociology, political science, and statistics.

Multi Layer Perceptron (MLP)

An MLP is a network consisting of simple *neurons* called perceptrons. The basic idea of a single perceptron was introduced by Rosenblatt in 1958. The perceptron calculates a single output from multiple inputs which are real-valued by forming a linear combination according to its input *weights* and then putting the output through some nonlinear activation function. Mathematically this can be written as

$$y = \varphi\left(\sum_{i=1}^{n} w_i x_i + b\right) = \varphi(\mathbf{w}^T\mathbf{x} + b) \tag{2}$$

where $\mathbf{w}$ denotes the vector of weights, $\mathbf{x}$ is the vector of inputs, b is the bias and φ is the activation function. A signal-flow graph of this operation is shown in Figure 1

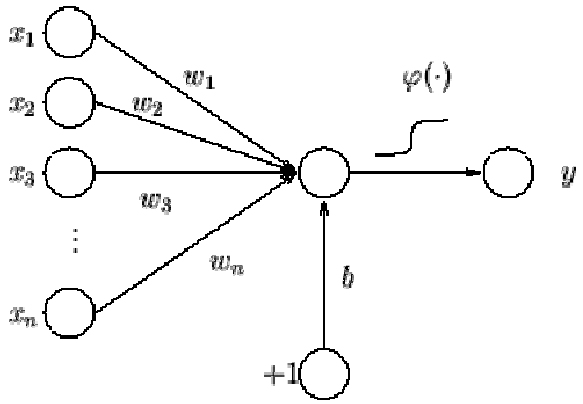

Figure 1**:** Signal-flow graph of the perceptron

Sequential Minimum Optimization (SMO)

SMO is an algorithm used for solving thequadratic programming (QP) problem that arises during the training ofsupport vector machines (SVM). It was found by John Platt in 1998. SMO is widely used for training support vector machines

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

and is an iterative algorithm for solving the optimization problem. This algorithm breaks the problem into a series of smallest possible sub-problems, which are then solved analytically. Hence it is much faster than SVM.

Instance Based learning
Instance-based learning (IBL) are an extension of k-Nearest Neighbour classification algorithms. They do not maintain a set of abstractions of model created from the instances. They also extend it with a significance test to work with noisy instances, since a lot of real-life datasets have training instances and k-NN algorithms do not work well with noise. These are supervised learning algorithms where instances are used to classify objects.

Random forest
Random forests are an ensemble learningmethod for classification. They output the class that is the mode of the classes output by individual trees. The algorithm for implementing random forest was developed by Leo Breiman[9].

2.6 Evaluation Parameters

We adopted five frequently considered measurements for evaluation– Accuracy (Ac), Sensitivity (Sn), Specificity (Sp), Precision (Pr) and Mathew's Correlation Coefficient (MCC)

$$\text{Sensitivity} = \frac{TP}{TP+FN} \times 100 \quad (1)$$

$$\text{Specificity} = \frac{TN}{FP+TN} \times 100 \quad (2)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (3)$$

$$\text{Precision} = \frac{TP}{TP+FP} \times 100 \quad (4)$$

$$\text{MCC} = \frac{(TPXTN)-(FPXFN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN}} \quad (5)$$

2.7 Development of web interface

A web interface was developed for the tool using HTML, PHP and JavaScript.

III. RESULTS

The methodology utilized for development of the prediction tool is given in Figure 2. Prediction of Phenazinewas carried out using computational approach, WEKA, implementing different classifiers.The training data consisted ofdifferent datasets for seven different genes invovled in phenazine biosynthesis. The training was done with as many classifiers as possible to know which classifier gave the highest accuracy level. The classifiers used were as follows:

•Bayes (NaiveBayes)
•Functions (Logistic, Multilayer Perceptron, RBF Network, Simple Logistic, SMO)
•Lazy (IB1, IBk, KStar)
•Trees (J48, Random Forest)

Aftertrainingthemodel,thetrainedmodelwassavedandlateritwasusedtotestthe'test' datasets.Withthehelpoftrainingmodel, thetestingdatawasmadetorunandto see which classifier provided the highest accuracyin classifyingallthe'test'datainto 'phenazine'or 'non-phenazine'.

10-fold cross-validation and independent data set validation techniques were then carried out to estimate how accurately a predictive model (training data) would perform in practice. Based upon the performance evaluation, classifiers with the best accuracy were selected. The crossvalidationoutputs of all the classifiers of training model aregivenbelow in Tables 1 and 2. As evident from Tables 1 and 2, the hybrid method out-performed the other two feature extraction methods.

The web interface was developed in user-friendly manner which allows the user to easily submit a sequence and retrieve the result. Users are provided with two ways for submitting the sequence which include pasting a valid FASTA sequence

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

in the text area provided or uploading a valid FASTA file. The tool can be run by clicking on the 'Submit button'. The result generated will be displayed in a new window (Fig. 2)

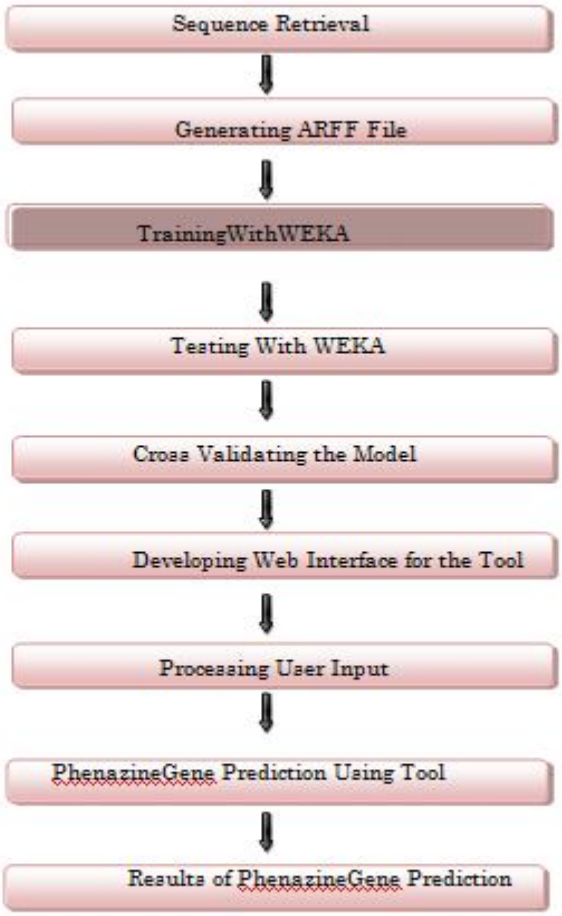

Figure 2. Flow chart /organization of this work

Table 1.The 10-fold cross-validation results showing sensitivity, specificity, accuracy and precision of the generated models and MCC showing the fitness function for model optimization.

Composition	Algorithm	Sn	Sp	Accuracy	Precision	MCC
Monomer	MLP	95.64	100	97.02	100	0.95
	IBk	88.64	100	94.32	100	0.89
	IB1	88.64	100	94.32	100	0.89
Dipeptide	IB1	99.86	98.85	99.35	98.8	0.98
	SMO	99.31	99.63	99.47	99.63	0.98
	Random Forest	99.54	98.89	99.22	98.90	0.98
Hybrid	IB1	99.92	99.88	99.25	99.88	0.98
	IBk	99.92	99.88	99.25	99.88	0.98
	SMO	99.69	100	97.82	100	0.99

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

Table 2.Independentdata test results showing sensitivity, specificity, accuracy, precision and MCC of the generated models and MCC showing the fitness function for model optimization.

Composition	Algorithm	Sn	Sp	Accuracy	Precision	MCC
Monomer	MLP	97.15	99.54	99.35	99.53	0.97
	IB1	97.69	99.54	99.61	99.54	0.98
	IBk	98.69	99.54	99.61	99.54	0.98
Dipeptide	Logistic	98.54	99.77	99.65	99.77	0.98
	SMO	98.69	99.77	99.73	99.77	0.98
	Simple Logistics	98.61	99.54	99.57	99.54	0.98
Hybrid	SMO	99.69	99.77	99.73	99.77	0.99
	Simple Logistics	99.61	99.54	99.57	99.54	0.99
	Logistics	99.54	99.77	99.65	99.77	0.99

Figure 3. Web interface of phenazine prediction tool

IV. CONCLUSION

Properties of phenazines are of great interest in biological and biotechnological applications. The present genomics era has facilitated the exponential growth of sequenced bacterial genome; however, the annotation of these genomes has proved to be a challenging area. Researchers have utilized machine learning algorithms to build classifiers for prediction of proteins from unannotated genomes. The results of the present study have revealed that hybrid feature extraction algorithm performs better for prediction of phenazine from genome of PGPR.

REFERENCES

1. Lugtenberg, B. And Kamilova, F. 'Plant-growth-promoting rhizobacteria', Annual Review of Microbiology, Vol. 63, pp. 541–556, 2009.
2. Glick, B. R. 'Plant Growth-Promoting Bacteria: Mechanisms and Applications,'Scientifica, Article ID 963401, doi.org/10.6064/2012/9634012012.
3. Ahemada, M. and , Kibretb, M, 'Mechanisms and applications of plant growth promoting rhizobacteria: Current perspective', Journal of King Saud University - Science, Volume 26, Issue 1, pp. 1–20, 2014.
4. Mavrodi, D.V., Blankenfeldt, W., Thomashow, L.S. and Mentel, M. Phenazine compounds in fluorescent *Pseudomonas* spp.: Biosynthesis and regulation,' Annual Review of Phytopathology, Vol .44, pp. 417–445, 2006.
5. Pierson, L.S. III, and Pierson, E.A. 'Metabolism and function of phenazines in bacteria: impacts on the behavior of bacteria in the environment and biotechnological processes,' Applied Microbiology and Biotechnology, Vol86, pp. 1659-1670, 2010.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

6. Pierson, L.S. III, Lam, S., Gaffney, T. and Gong, F.C. 'Molecular analysis of genes encoding phenazine biosynthesis in the biological control bacterium *Pseudomonas aureofaciens* 30-84, FEMS Microbiology Letters, Vol. 34, pp. 299–307, 1995.
7. Frank, E., Hall M., Trigg, L., Holmes, G. and Ian H. Witten, I.H., 'Data mining in bioinformatics using Weka,' Bioinformatics, Vol. 20, pp. 2479-2481, 2004.
8. Hemalatha, N., Rajesh, M. K. and Narayanan, N. K. 'Machine learning approaches for prediction of expansin gene family in indica rice,' Agricultural Research, Vol. 2, pp. 309-318, 2013.
9. Breiman, Leo (2001). "Random Forests". Machine Learning 45 (1): 5–32.