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

# A Machine Learning Approach for Discovery of Novel Non- Ribosomal Peptide Synthetases (NRPS) in genomes of Plant Growth Promoting *Pseudomonas* Spp

Philip Job N.[1], Jamshinath T.P.[1], Hemalatha N.[2*], Rajesh M.K.[1]

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

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

Corresponding author[*]

**ABSTRACT**: Non-ribosomal peptide synthetases (NRPSs) are multi-modular megasynthasespossessing the ability to catalyze biosynthesis of small bioactive peptides through a thiotemplate mechanismwhich is independent of ribosomes. These enzymes are invovled in production of a wide range of chemical products of broad structural and biological activity. The present study was performed with an aim to develop a gene prediction tool using a machine learning work bench called WEKA (Waikato Environment for Knowledge Analysis) for NRPS in plant growth promoting *Pseudomonas* spp.First, a model was developed using the training data which was generated using many classifiers. The trained model was then used for the prediction of NRPS in a given set of unknown sequences. Cross-validation results showed that the 'Logisticof Functions' was the best classifier when compared to others, showing high accuracy and performance in classifying the instances. We hope that the tool will aid in discovering of novel NRPS by predicting them from sequence data obtained by whole genome sequencing of bacteria or metagenomics.

**KEYWORDS**: *Pseudomonas fluorescens*, pyoverdine, Non Ribosomal Peptide Synthetase, Machine learning, WEKA.

## I. INTRODUCTION

Plant Growth-Promoting Rhizobacteria (PGPR)exert a positive effect on plant growth either through direct or indirect interaction with their plant hosts, some of which include production of plant growth regulators, solubilization of minerals, production of anti-microbial secondary metabolites and siderophores[1]. The application of PGPRas crop inoculants for biocontrol, biofertilization and phytostimulation serves as an attractive alternative touse of chemicals for crop protection fertilizers which can cause severe environmentalpollution over a long period [2]. *Pseudomonas* spp., a widespread bacteria in agriculturalsoils, comprise of gram negative, motile, rodshaped bacteria with a multitude of crop growth promoting activities include production of siderophores,proteases, anti-microbials, phosphate solubilizing enzymesand HCN[3].

Non-ribosomal peptide synthetases (NRPSs)are large,multi-modular enzymes, found in bacteriaand fungi, and involved in the synthesis of a widearray of secondary metabolites[4]. These secondary metabolites possessdifferent biologicalroles, for e.g., iron sequestration, antimicrobial, insecticidal,and antiviral activity [5]. NRPSs synthesize peptides by a multiple carrierthiotemplate mechanism. In general,NRPSs are modular, with each module catalyzing the incorporationof one amino acid substrate into the growing peptide [5].NRPS modules are, in turn, made up of independently foldingfunctional domains that catalyze the individual reactions ofpeptide synthesis.

## II. RELATED WORK

There is a great potential for discovery of novel NRPS for pharmacological and biotechnological uses, especially for new drugs and bioactive compounds. The present genomics era has facilitated the exponential growth of sequenced NRPS. Researchers have utilized machinelearning algorithms to build classifiersfor prediction of proteins based on20 amino acid residues and on the physico-chemical properties of aminoacids [6]. Additionally, a host of tools and software's have been developed for prediction of NRPS, some of them of them being NORINE[7], NRPS-PKS [8], NRPS-PKS [9],NP.Searcher[10], PKS/NRPS Analysis [11], NRPSPredictor2 [12]and NRPS ToolBox[13].In this study, we present a more accurate prediction tool for NRPS developed using a machine learning platform called WEKA (Waikato Environment for Knowledge Analysis) workbench, containinga collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality [14].

## III. METHODOLOGY

### *Sequence Retrieval and generating ARFF file*

Nucleotide sequences coding for NRPS proteinsand non-NRPS proteins wereretrieved from NCBI. The coding sequences were then obtained using ORF Finder and saved in FASTA format. The bacterial species with number of NRPS sequences selected for this study is given in Table 1.

**Table 1.NRPS sequences from *Pseudomonas*spp. used to generate training and testing data sets.**

| Species | No. of NRPS sequences |
|---|---|
| *P.fluorescens* | 9 |
| *P.aeruginosa*16 | |
| *P.putida*7 | |
| *P.entomophila*6 | |
| *P.brassicacearum*5 | |
| *P.syringae*6 | |

Data sets for developing training and testing was done by considering the FASTA sequences of 49 sequences from *Pseudomonas* spp. and 49non-NRPS sequences. Each of these sequences were divided randomly and stored as training and testing data. The training data consisted of 64 (32+32) sequences from both NRPS and non-NRPS sequences.The same process was used for testing data which consisted of 34 (17+17) sequences. These datasets were converted using Perl script into binary form containing relation,attributes and data (ARFF format).

### *Model development and evaluation using cross-validation techniques*

We trained the data containing 32 positive and 32 negative sets using five different classifiers *viz.*, Naïve Bayes, SMO, IBK, Bagging, J48 and Logistic. Based on the best performing algorithm, the training model was generated. The model was evaluated using sub-sampling test's (three-fold cross-validation and eight-fold cross-validation) and Leave One Outcross-validation techniques [15].Among these cross-validation techniques, Leave one out method is the most suitable method because it uses more training data and less test data. The model that we developed showed high performance in predicting the instances correctly.

### *Performance Evaluation*

The performance of various models using the five classifiers developed in this study was computed by using sensitivity, specificity, accuracy and Matthew's correlation coefficient. The measurements are expressed in terms of true positive (TP), true negative (TN), false positive (FP), false negative (FN). Sensitivity (Sn) is the parameter which allows the computation of percentage of correctly predicted NRPS sequences. Specificity (Sp) parameter allows the computation of percentage of correctly predicted non-NRPS genes. Accuracy (Ac) shows the percentage of correctly predicted NRPS and non-NRPS genes. Matthew's correlation coefficient(MCC) is a statistical parameter which measures the quality of the NRPS and non-NRPS classifications. MCC with value 1 is indicates the best possible prediction while MCC with 0 value indicates the worst possible prediction scheme [16].

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{FP + TN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

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

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

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

## IV. PROPOSED ALGORITHM

We used 64 datasets of both NRPS and non-NRPS data as training data to generate the model. We generated many models using five different classifiers (Naïve Bayes, SMO, IBK, Bagging, J48 and Logistic). The performance of each model was evaluated using different cross-validation techniques such as sub-samplingtest's (three-fold cross-validation and eight-fold cross-validation) and Leave One Out cross-validation methods. Three-fold cross-validation showed that the Logistic Classifier had the best accuracy of 96.87% and MCC having 0.9279 (Table 2). The eight-fold cross-validation method showed that the 'Logistic Classifier' had the highest percentage of accuracy of 95.31% and MCC of 0.9105 (Table 3). The Leave One Out cross-validation method also showed that 'Logistic Classifier' had the highest accuracy of 98.43% and MCC having 0.9699 (Table 4). MCC of 1 is regarded as a perfect prediction, whereas 0 is regarded as random prediction.In this investigation, we finally came to a conclusion that 'Logisitc Classifier' was the best model showing the highest percentage of accuracy in all these cross validation techniques.

Table 2.The three-fold cross-validation results showing sensitivity,specificity, accuracy of the generated models and MCC showing the fitness function for model optimization.

| **Algorithm** | **Sn (%)** | **Sp (%)** | **Ac (%)** | **MCC** |
|---|---|---|---|---|
| Logistic | 100 | 93.70 | 96.87 | 0.9279 |
| Naïve Bayes | 90.60 | 100 | 95.31 | 0.9105 |
| SMO | 96.90 | 93.70 | 95.31 | 0.9070 |
| IBK | 93.70 | 90.60 | 92.18 | 0.8440 |
| Bagging | 87.50 | 93.70 | 90.62 | 0.8138 |
| J 48 | 91.14 | 96.90 | 93.75 | 0.8767 |

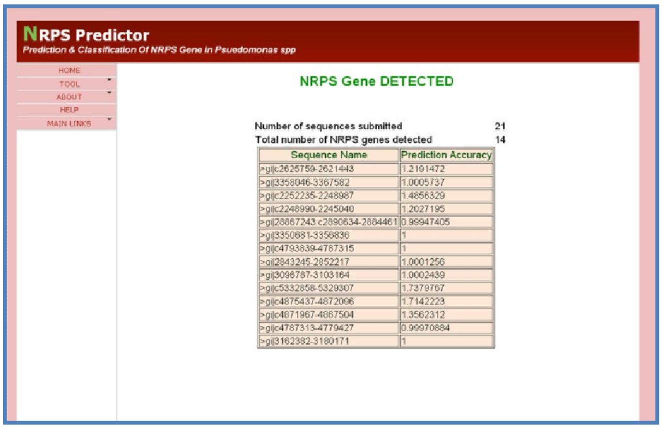
Table 3. The eight-fold cross-validation results showing sensitivity, specificity, accuracy of the generated models and MCC showing the fitness function for model optimization.

| **Algorithm** | **Sn (%)** | **Sp(%)** | **Ac (%)** | **MCC** |
|---|---|---|---|---|
| Logistic | 90.60 | 100 | 95.31 | 0.9105 |
| NaiveBayes | 93.70 | 90.60 | 92.18 | 0.8440 |
| SMO | 87.50 | 87.50 | 87.50 | 0.7500 |
| IBK | 87.50 | 87.50 | 87.50 | 0.7500 |
| Bagging | 90.60 | 96.90 | 93.75 | 0.8767 |
| J 48 | 93.70 | 90.60 | 92.18 | 0.8440 |

Table 4. The Leave One Out cross-validation (LOO CV) results showing sensitivity, specificity, accuracy of the generated models and MCC showing the fitness function for model optimization.

| **Algorithm** | **Sn (%)** | **Sp (%)** | **Ac (%)** | **MCC** |
|---|---|---|---|---|
| Logistic | 100% | 96.90% | 98.43% | 0.9699 |
| NaiveBayes | 93.70% | 90.60% | 92.18% | 0.8440 |
| SMO | 90.60% | 100% | 95.31% | 0.9105 |
| IBK | 100% | 93.70% | 96.87% | 0.9279 |
| Bagging | 90.60% | 100% | 95.31% | 0.9105 |
| J 48 | 87.50% | 87.50% | 87.50% | 0.7500 |

Development of a tool for prediction of NRPS sequence in *Pseudomonas spp* was developed using the machine learning algorithm WEKA. This was achieved using the Perl Language, HTML and Java script. The tool was named 'NRPS Predictor'. While all the programs required to process the user input was written using Perl, HTML codes were used to create a user interface. The web interface allows users to submit a query. The sequence submitted will be processed and the output will be displayed in a new window. The interface was developed in a user friendly manner and contains web pages that serve different purposes (Figure 1).



NRPS Predictor

Prediction & Classification Of NRPS Gene in Psuedomonas spp

NRPS Gene DETECTED

Number of sequences submitted 21
Total number of NRPS genes detected 14

| Sequence Name | Prediction Accuracy |
|---|---|
| >gi\|c2625759-2621443 | 1.2191472 |
| >gi\|3358046-3367582 | 1.0005737 |
| >gi\|c2252235-2248987 | 1.4856329 |
| >gi\|c2248990-2245040 | 1.2027195 |
| >gi\|28867243:c2890634-2884461 | 0.99947405 |
| >gi\|3350681-3356836 | 1 |
| >gi\|c4793839-4787315 | 1 |
| >gi\|2843245-2852217 | 1.0001256 |
| >gi\|3096787-3103164 | 1.0002439 |
| >gi\|c5332858-5329307 | 1.7379767 |
| >gi\|c4875437-4872096 | 1.7142223 |
| >gi\|c4871967-4867504 | 1.3562312 |
| >gi\|c4787313-4779427 | 0.99970884 |
| >gi\|3162382-3180171 | 1 |

Figure 1. Web interface of NRPS Predictor

## V. CONCLUSION

In this paper, we introduce a new gene prediction tool for classifying NRPS gene sequences and non-NPRS gene sequences. We obtained 64 gene and non gene sequences from NCBI out of which 32 genes of both NRPS and non-NRPS were used as training set. Our tool introduces a technically simple method for predicting NRPS without compromising the accuracy of the process. We developed a Logistic based approach, NRPS predictor tool for classifying NRPS. Our model showed very high prediction accuracies on the training and testing datasets which increases the chances of accurately predicting the genes. Successful prediction of NRPS shows that the method followed has significant merit as an approach for successful prediction of NRPS in other organisms sharing a close evolutionary relationship with the *Pseudomonasspp*.

### ACKNOWLEDGEMENT

This work was supported by a grant from Department of Biotechnology (BTISnet), New Delhi, India.

## REFERENCES

1. Lugtenberg, B. and ‚Kamilova, F., ‘Plant-growth-promoting rhizobacteria’, Annual Review of Microbiology, Vol. 63, pp. 541–556, 2009.
2. Noori, M.S.S. and Saud, H.M., ‘Potential plant growth-promoting activity of *Pseudomonas* sp.isolated from paddy soil in Malaysia as biocontrol agent’, Journal of Plant Pathology and Microbiology, Vol. 3, Issue 2, doi:10.4172/2157-7471.1000120, 2012.
3. Preston, G. M., ‘Plant perceptions of plant growth-promoting *Pseudomonas*’, Philosophical Transactions of the Royal Society London B, Vol. 359, pp. 907–918, 2004.
4. Finking, R. and Marahiel, M.A., ‘Biosynthesis of non-ribosomal peptides,’ Annual Review of Microbiology, Vol. 58, pp. 453–488, 2004.
5. Verne Lee, T., Johnson, L. J., Johnson, R. D., Koulman, A., Lane, G. A., Lott, J. S. and Arcus V. L., ‘Structure of a eukaryotic non-ribosomal peptide synthetaseadenylation domain that activates a large hydroxamate amino acid in siderophore biosynthesis,’ The Journal of Biological Chemistry, Vol. 285, No. 4, pp. 2415–2427, 2010.
6. Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W. and Huson, D. H., ‘Specificity prediction of adenylation domains in non-ribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs),’ Nucleic Acids Research, Vol. 33, pp. 5799–5808, 2005.
7. Caboche, S., Pupin, M., Leclère, V., Fontaine, A., Jacques, P. and Kucherov, G., ‘NORINE: a database of non-ribosomal peptides,’ Nucleic Acids Research, Vol. 36, pp. D326-D331, 2008.
8. Anand, S., Prasad, M.V., Yadav, G., Kumar, N., Shehara, J., Ansari, M.Z., Mohanty, D., ‘SBSPKS: structure based sequence analysis of polyketide synthases,’ Nucleic Acids Research, Vol. 38, pp. W487–W496, 2010.
9. Ansari, M.Z., Yadav, G., Gokhale, R.S. andMohanty D., ‘NRPS-PKS: a knowledge-based resource for analysis of NRPS/ PKS megasynthases,’ Nucleic Acids Research, Vol. 32, pp. W405–W413, 2004.
10. Li, M.H., Ung, P.M., Zajkowski, J., Garneau-Tsodikova, S. and Sherman, D.H., ‘Automated genome mining for natural products,’BMCBioinformatics, Vol. **10**, doi: 10.1186/1471-2105-10-185, 2009.
11. Bachmann, B. O. and Ravel, J., ‘*In Silico* prediction of microbial secondary metabolic pathways from DNA sequence data,’ Methods in Enzymology, Vol. 458, pp. 181-217, 2009.

12. Röttig, M., Medema, M,H., Blin, K., Weber, T., Rausch, C., and Kohlbacher, O., ‘NRPSpredictor2 - a web server for predicting NRPS adenylation domain specificity,’ Nucleic Acids Research, Vol. 39, pp. W362–W367, 2011.
13. Pupin, M., Smaïl-Tabbone, M., Jacques, P., Marie-Dominique, D. and Leclère, V. , ‘NRPS toolbox for the discovery of new nonribosomal peptides and synthetases,’ JournéesOuvertes en Biologie, l'Informatique et les Mathématiques, pp. 89-93, 2012.
14. 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.
15. Chou, K.C. andZhang, C. T., ‘Prediction of protein structural classes,’ Critical Reviews in Biochemistry and Molecular Biology, Vol. 30, pp. 275-349, 1995.
16. Mishra, K.N.,Agarwal, S.and Raghava, G.P.S., ‘Prediction of cytochrome P450 isoform responsible for metabolizing a drug molecule,’ BMC Pharmacology, Vol. 10, doi:10.1186/1471-2210-10-8, 2010.