Indian Journal of Agricultural Sciences **83** (10): 1075–8, October 2013/Article

An SVM-based algorithm for the prediction and classification of enzymes involved in antibiotic biosynthetic pathways in plant growth promoting *Pseudomonas* species

G L SAIRAM[1], M K RAJESH[2], S NITHYA[3] and GEORGE V THOMAS[4]

Bioinformatics Centre, Central Plantation Crops Research Institute, Kasaragod, Kerala 671 124

Received 11 March 2012; Revised accepted: 2 September 2013

ABSTRACT

In this study, a tool has been developed for the prediction of enzymes involved in antibiotic biosynthetic pathways (2,4-diacetylphloroglucinol, phenazine, pyoluteorin and pyrrolnitrin) in plant growth promoting *Pseudomonas* species on the basis of amino acid and dipeptide composition by using the Support Vector Machines (SVM). The performance of the system was achieved by using a training set consisting of 330 non-redundant set of positively labeled enzymes involved in antibiotic biosynthetic pathway in *Pseudomonas* spp. and 309 non-redundant set of negatively labeled sequences from other organisms obtained from NCBI. First we developed a support vector machine based module using amino acid and dipeptide composition and achieved an overall accuracy of 87.00% and 91.00% respectively. Then, another SVM module was developed based on dipeptide composition for classifying the predicted enzymes into four main classes with accuracy 95%, 80%, and 75% 95% for 2,4-diacetylphloroglucinol, phenazine, pyoluteorin and pyrrolnitrin respectively. Based on the above method, a web server has been set up at http://210.212.229.59:8080/Prediction/home.jsp.

Key words: Antibiotics, Pattern classification, PGPR, Support Vector Machine

Plant growth promoting rhizobacteria (PGPR) are bacteria which colonize plant roots and exert beneficial mechanisms by either direct or indirect mechanisms. They play a significant role in crop protection and improve soil health. Common PGPR strains include *Pseudomonas*, *Bacillus*, *Azospirillum* and *Rhizobium*. One of the mechanisms of PGPR involves the production of antibiotics belonging to polyketides, heterocyclic nitrogenous and lipopeptides classes; these antibiotics possess broad-spectrum action against several phytopathogens. The synthesis of these antibiotics, such as phenazine-1-carboxyclic acid, 2,4-diacetyl phloroglucinol, oomycin, pyoluteorin, pyrrolnitrin, kanosamine, zwittermycin-A, and pantocin, are regulated by cascade of endogenous signals such as sigma factors, sensor kinases and N-acyl homoserine lactones (Fernando *et al*. 2005). Additionally, these antibiotics also operate as determinants in triggering induced systemic resistance (ISR) in plants (Fernando *et al*. 2005). A high degree of conservation, with respect to the sequence motifs, has been detected in the genes responsible for the synthesis of antibiotics.

[1] Senior Research Fellow (e mail: glsairam@gmail.com), [2] Senior Scientist (e mail: mkraju_cpcri@yahoo.com), [3] Senior Research Fellow (e mail: bioinfonithi87@gmail.com), [4] Director (e mail: georgevthomas@yahoo.com)

Recent microbial genome sequencing projects have generated enormous amounts of sequence data, the analysis of which can have profound implications on biological research. However, the analysis of these huge data is difficult and time consuming. Therefore mining of biological data is of great importance for researchers. Recently, SVM (Support Vector Machine) based approaches have been used to predict the sub-cellular localization of proteins (Garg *et al.* 2005), to recognize protein fold and super family (Iain *et al.* 2007), to predict the alpha turn types (Cai *et al.* 2003), for protein fold recognition (Yan *et al.* 2009) and β-lactam antibiotics (Smolander *et al.* 2009).

In this study, we have developed a method to predict the enzymes involved in antibiotic biosynthetic pathways in plant growth promoting *Pseudomonas* species based on amino acid composition and dipeptide composition. The performance of the method was evaluated by a 10 fold cross validation and the data set was obtained from NCBI (National Center for Biotechnology Information). Based on the proposed approach we have also developed a web server.

MATERIALS AND METHODS

Full length coding sequences of enzymes involved in biosynthesis of 2,4-diacetylphloroglucinol (71), phenazine (187), pyoluteorin (59), and pyrrolnitrin (13) in plant growth

promoting *Pseudomonas* spp. were obtained from NCBI. To reduce the redundancy in our data, we used ExPASy sequence alignment tool Decrease Redundancy (http://expasy.org/tools/redundancy/) with a criteria that no two sequence had >90% sequence identity to any other sequence in the data set (Saha and Raghava 2006). The final data set contained 330 non-redundant set of positively labeled *Pseudomonas* enzymes and 309 non-redundant set of sequences from other organisms.

Support Vector Machine (SVM) is a set of related supervised learning methods used for classification and regression (Vapnik 1995). SVM modules were implemented by using SVM^{light} version 6.01 (http://svmlight.joachims.org). The software enables users to define a number of parameters as well as inbuilt kernel functions such as linear kernel, radial basis function and polynomial kernel of a given degree. In order to develop the prediction method, we trained SVM using enzymes involved in biosynthesis of 2,4-diacetylphloroglucinol, phenazine, pyoluteorin, and pyrrolnitrin as positive labels and other bacterial proteins as negative labels. The training was carried out by using RBF kernel with a gamma value of 0.01 as better results were obtained with RBF kernel compared to the other two kernels.

Amino Acid Composition (AAC) is the amount of each amino acid present in the protein sequence, which transforms the protein sequence into an input vector of 20 dimensions. If Qi is the frequency of occurrence of an amino acid i, then the amino acid composition is AACi = Qi/L, where i is any amino acid, L is the length of protein.

$$\text{Fraction of aa(i)} = \frac{\text{Total number of amino acid of type (i)}}{\text{Total number of amino acids in protein}}$$

For Dipeptide Composition (DC), the protein sequence was represented into an input vector of 400 (20 * 20) dimensions. L is the total number of all possible dipeptide in protein P (L = 400). Then Qij be a fraction of any pair of amino acids ij. ij is any amino acid from 1 to 20. The dipeptide composition DCij = Qij/L.

$$\text{Fraction of dipeptide(i)} = \frac{\text{Total number of dipeptide of type (ij)}}{\text{Total number of all possible dipeptide}}$$

A 10-fold LOOCV (Leave-One-Out Cross Validation) was carried out to evaluate the performance of SVM in the training set. In cross validation experiment, the data set was divided into 10 equal sized samples and for each experiment, 10-1 samples were used for training and the remaining samples for testing. The accuracy was calculated as the average accuracy over 10 samples.

The training set, containing 330 positively labeled sequences, was used to evaluate the performance of SVM. A confusion matrix with True Positive, True Negative, False Positive, and False Negative was created to determine the performance of SVM on training set (Fig 1).

We also calculated sensitivity, specificity, accuracy and Mathew's correlation coefficient (MCC). Sensitivity is the percentage of positive instances correctly classified as positive; specificity is the percentage of negative instances correctly classified as negative and accuracy is the percentage of correctly classified instances. If the value of MCC is 1, the prediction is perfect and if the value of MCC is 0, then it is assumed to be a random prediction. These parameters were calculated by using the following equations:

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

$$\text{Specificity} = \frac{TP}{TP + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{P + N}$$

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

The performance of the classifier was evaluated constructing a ROC curve by plotting the fraction of false positives (FPR) against true positives (TPR) at various threshold settings. The area under the curve (AUC) represented in the ROC curve further gives a measure of the accuracy of the classifier.

		Actual value: Positive	Actual value: Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

Fig 1 Classification of a prediction into true positive (TP), true negative (TN), False Positive (FP) and False Negative (FN). TP known Positive data predicted as positive; TN known negative data predicted as negative; FP known negative data predicted as positive; FN known positive data predicted as negative

RESULTS AND DISCUSSION

The work flow for the development of SVM modules to predict and classify the enzymes involved in antibiotic biosynthetic pathways in plant growth promoting *Pseudomonas* species using amino acid composition and dipeptide composition is given in Fig 2.

All SVM modules were trained using 10-fold Leave-one-out cross validation. The SVM modules developed using amino acid composition shows a classification accuracy of 87% with a Matthew's correlation (MCC) value of 0.74 with 86% specificity and 88% sensitivity (Table 1). The SVM modules developed using dipeptide composition shows a classification accuracy of 91% with a MCC value of 0.83 with 82% specificity and 100% sensitivity (Table 1). The

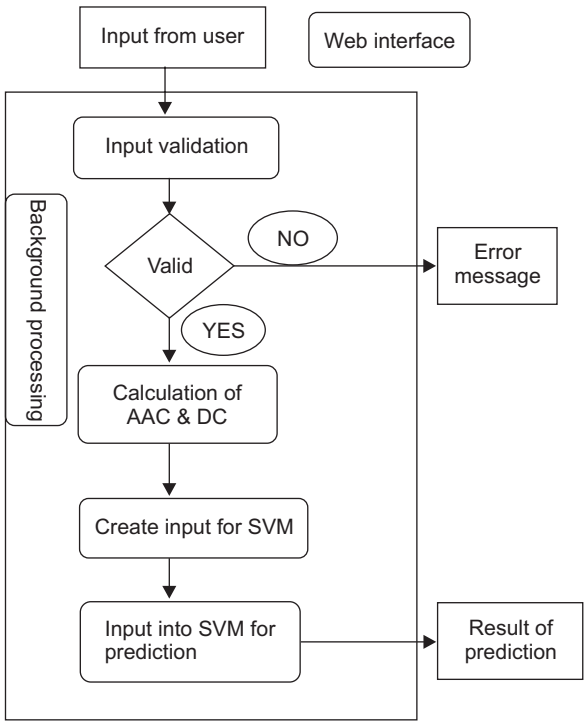

Fig 2 The work flow of web server developed

Table 1 Detailed accuracy for different prediction methods obtained from *Pseudomonas fluorescens*

Prediction Method	MCC	Accuracy	Sensitivity	Specificity
Amino acid composition	0.74	87%	88%	86%
Dipeptide composition	0.83	91%	100%	82%

results show that dipeptide composition-based approaches have more classification accuracy when compared to amino acid composition-based approach because dipeptide composition provides more information than simple amino acid composition.

Four different SVM modules were also developed to classify the predicted pseudomonas enzymes into four different classes, like 2, 4-diacetylphloroglucinol, phenazine, pyoluteorin and pyrrolnitrin. It was found that the composition of these proteins vary significantly (Fig 3). Earlier studies have revealed that the prediction methods based on diverse compositional features offer more accuracy when compared to homology-based searching in prediction of functional roles of proteins, secondary structures and sub-cellular localization (Bhasin and Raghava 2004, Bhasin *et al*. 2005, Garg *et al*. 2005). Therefore, in this study, first amino acid composition-based modules were developed and an overall accuracy of 75% (Table 2) was achieved. Dipeptide composition-based modules were then developed, which gave an overall accuracy of 86.25% (Table 2). The results show

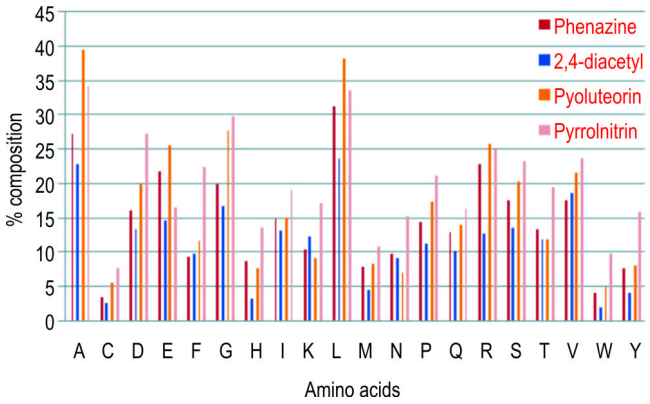

Fig 3 Average amino acid composition of four different classes of proteins

Table 2 Detailed classification accuracy of four different classes of enzymes

Enzyme Class	Accuracy (%)	
	Amino acid composition	Dipeptide composition
2,4-diacetylphloro-glucinol	55%	95%
Phenazine	65%	80%
Pyoluteorin	80%	75%
Pyrrolnitrin	100%	95%
Average	75%	86.25%

that dipeptide composition-based method possessed more classification accuracy than amino acid based method. From the ROC curve (Fig 4), it is clear that the didpeptide composition module represents a better classifier than the amino acid classifier.

Based on the modules developed a web server was also developed using Jsp/Java. The server accepts FASTA,

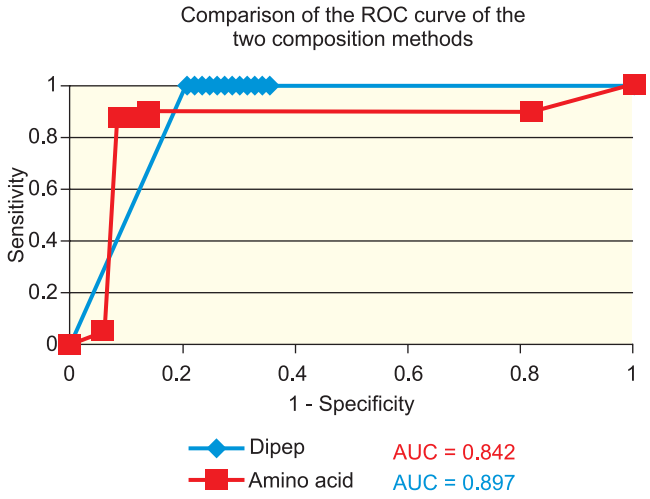

Fig 4 ROC curve for amino acid and dipeptide composition in SVM

GenBank, and EMBL formatted sequence. The user can enter multiple sequences at a time. The result will be displayed separately for each protein. First the server predicts the sequence and then classifies it into any of the four groups. The web server is accessible at http://210.212.229.59:8080/Prediction/home.jsp.

Thus, the classification system developed in this study can be used to screen larger sets of antibiotic classes. The accuracy and reliability of the classification suggests that this method can be used in the future to probe for enzymes involved in antibiotic production in other strains of plant growth promoting rhizobacteria. To the best of our knowledge, this is the first report of use of SVM-based algorithm for the prediction and classification of enzymes involved in antibiotic biosynthetic pathways in plant growth promoting *Pseudomonas* species.

ACKNOWLEDGMENT

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

REFERENCES

Bhasin M, Garg A, and Raghava G P S. 2005 PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics* **21**: 2 522–44.

Bhasin M, Raghava G P. 2004. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Research*, **32**: W414–9.

Garg A, Bhasin M and Raghava G P S. 2005. SVM-based method for subcellular localization of human proteins using amino acid composition, their order and similarity search. *Journal of Biological Chemistry* **280:** 14 427–32.

Cai Y D, Feng K Y, Li Y X and Chou K C. 2003. Support vector machine for predicting alpha-turn types. *Peptides* **24**: 629–30.

Fernando W G D, Nakkeeran S and Zhang S. 2005. Biosynthesis of antibiotics by PGPR and its relation in biocontrol of plant diseases. *PGPR: Biocontrol and Biofertilization*, pp 67–109. Siddiqui ZA (Ed). Springer, Dordrecht, The Netherlands.

Iain M, Eugene le, Rui K, Jason W, William S N and Chrostina L. 2007. SVM-Fold: a tool for discriminative multi class protein fold and superfamily recognition. *BMC Bioinformatics* **8:** (Suppl 4):S2

Saha S and Raghava G P S. 2006. VICMpred: An SVM-based method for the prediction of functional proteins of gram-negative bacteria using amino acid patterns and composition. *Genomics, Proteomics, and Bioinformatics* **4**: 42–7.

Smolander O-P, Riberio AS, Yli-Harja O and Karp M. 2009. Identification of β-lactam antibiotics using bioluminescent *Escherichia coli* and a support vector machine classifier algorithm. *Sensors and Actuators* **141**: 604–9.

Vapnik V N. 1995. *The Nature of Statistical Learning Theory,* 188 pp. Springer, New York.

Yan R X, Si J N, Wang C and Zhang Z. 2009. DescFold: a web server for protein fold recognition. *BMC Bioinformatics* **10**: 416.