Methodology Article

Journal of Plantation Crops, 2014, 42(3):377-385

LTTRPred: A tool for prediction of LysR-type transcriptional regulator of pyoluteorin pathway in plant growth promoting *Pseudomonas* spp.

Anil Paul, N. Hemalatha[1] and M.K. Rajesh*

Bioinformatics Centre, Central Plantation Crops Research Institute, Kasaragod-671 124, Kerala, India
[1]St. Aloysius College (AIMIT), Mangalore-575 022, Karnataka, India

(Manuscript Received: 01-02-14, Revised: 05-09-14, Accepted: 22-09-14)

Abstract

Plant growth promoting *Pseudomonas* spp. produce an antifungal compound called pyoluteorin (Plt) that suppress diseases caused by phytopathogenic fungi. The pathway specific regulator PltR, a typical LysR-type transcriptional regulator (LTTR), is responsible for the transcriptional activation of the Plt biosynthetic operon. The LTTR family represents one of the largest classes of bacterial transcriptional regulatory proteins. A large number of LTTRs possess function as global transcriptional activators or repressors of unlinked genes or operons involved in metabolism, quinoline signal, virulence *etc*. The proposed method, LTTRPred, is an useful tool developed for identifying and predicting the LTTR, which is responsible for the activation of Plt transcription regulators, from whole genomes of various *Pseudomonas* spp. LTTRPred was developed using support vector machine (SVM) and Waikato Environment for Knowledge Analysis (WEKA) based on the composition of amino acid and amino acid pairs. Modules in SVM were developed using traditional amino acid, dipeptide (n+1) and hybrid amino acid composition modules and an overall accuracy of 100, 100 and 98 per cent respectively, was achieved. Modules in WEKA were also developed using the same modules and an overall accuracy of 100 per cent achieved for all. The performance of the tool was tested using various datasets of LTTR genes from different *Pseudomonas* spp. The best performing SVM and WEKA modules from the present investigation was implemented as a dynamic web server 'LTTRPred', which is freely available and can be accessed online (http://210.212.229.56/lttrpred/). This tool can be used for the functional annotation of the *Pseudomonas* spp. possessing LTTR genes.

Keywords: Antagonism, phosphate solubilizer, potassium solubilizing bacteria, tea soil

Introduction

Plant growth promoting rhizobacteria (PGPR) are a group of free-living bacteria that colonize the rhizosphere, live in a commensal relationship with plants and contribute to increased growth and yield of crop plants (Kloepper and Schroth, 1978; Paulsen *et al.,* 2005). PGPR can promote plant growth either directly or indirectly (Glick, 1995); however, the exact mechanisms by which PGPR promote plant growth are yet to be fully deciphered. Bacteria of diverse genera have been identified as PGPR of which *Bacillus* and *Pseudomonas* spp. are predominant (Podile and Kishore, 2006). PGPR can exert various effects on plants which may include improvement of soil structure, facilitation of nutrient acquisition, antagonism against phytopathogens and pests, alteration of plant physiological processes and degradation of xenobiotics and pollutants (Niranjan Raj *et al*., 2005; Paulsen *et al*., 2005). PGPR possess immense potential application in sustainable agriculture, especially with respect to long duration plantation crops like coconut (Bopaiah and Shetty, 1991), arecanut (Bopaiah, 1985), cocoa (Litty Thomas *et al*., 2011), tea (Chakraborty *et al*., 2013), coffee, spices and rubber (Hidayati *et al*., 2014).

Pseudomonads possess the capacity to produce a wide array of metabolites, including antibiotics, which are toxic to phytopathogens (Haas and Keel, 2003; Raaijmakers *et al.,* 2002). Pyoluteorin (Plt),

*Corresponding Author: mkraju.cpcri@gmail.com

an antibiotic substance produced by certain strains of *Pseudomonas* spp., is composed of a bichlorinated pyrrole linked to a resorcinol moiety, which can inhibit phytopathogenic fungi, including *Pythium ultimum,* effectively (Maurhofer *et al.,* 1994), suppress plant disease caused by phytopathogenic fungus and in some instances, even contribute to the ecological competence of the producing strain within the rhizosphere (Dowling and O'Gara, 1994). The LysR-type transcriptional regulators (LTTRs) are considered to be the largest family of prokaryotic transcription factors. First described by Henikoff *et al.* (1988), they are known to be present in many bacterial genera, archaea and algal chloroplasts (Schell, 1993). The pathway specific regulator PltR, a typical LTTR, is responsible for the transcriptional activation of the Plt biosynthetic operon (Nowak-Thompson *et al*., 1999).

There are various methods for predicting the function of a given protein sequence. The similarity search-based tools have been used for functional annotation of proteins where a sequence is searched against an experimentally annotated database and a function is assigned to the protein. However, this approach fails when an unknown query protein does not have significant similarity to proteins in the database. Another way to predict the proteins is to identify sequence motifs such as signal peptide or nuclear localization signal. Many machine learning technique-based methods such as artificial neural networks and support vector machines (SVM) have been developed to predict the function of proteins. Recent advances in the prediction of protein sequences have stressed the need for organism-specific prediction tools (Schneider and Fechner, 2004). When compared with the general prokaryotic protein prediction methods, organism specific prediction methods are more accurate. To the best of our knowledge, there is no method currently available for predicting the LTTR genes, which are responsible for the transcriptional activation of the Plt biosynthetic operon. LTTRPred, the tool developed in the present study, predicts the transcriptional regulator of pyoluteorin pathway in *Pseudomonas* spp. using both SVM and Waikato Environment for Knowledge Analysis (WEKA) based approach. The performance of the models was evaluated using cross-validation techniques. A web-server was developed based on the best approach, based on accuracy, to provide help to the researchers working with LTTR proteins.

Materials and methods

Dataset

The selection of dataset is the most important concern during development of a prediction method. For training the model, both positive and negative data sets are required, which was retrieved from NCBI. There are 8500 LTTR protein sequences of *Pseudomonas* spp. available in NCBI and these sequences constituted the positive data set. The average length of the sequences was around 300 amino acids. The negative dataset was created by downloading another 5700 protein sequences (non-LTTR) from different *Pseudomonas* spp. of approximately the same length as the positive data set. The positive and negative sequences were divided for training and testing. For training and testing, independent datasets were used.

Support Vector Machine (SVM)

SVM is a machine learning algorithm introduced by Vapnik and co-workers (Cortes and Vapnik, 1995; Vapnik, 1995), based on the statistical and optimization theory and has been applied in many classification and regression problems. SVMs are becoming popular in a wide variety of biological applications like classifying objects as diverse as protein and DNA sequences, microarray expression profiles and mass spectra (Noble, 2004). In the present study, we have used SVM_light (Joachims, 1995), a freely downloadable package of SVM (http://svmlight.joachims.org/old/svm_light_v4.00.html), to predict LTTR protein sequences. During SVM training, a hyperplane in feature space is determined that gives the largest possible margin between the positive and negative class, thereby yielding an intuitively robust classifier. SVMs can learn accurate classifiers for data sets that cannot be linearly separated in the input space. This is achieved by the choice of a suitable kernel function to transform the input data into another feature space where it is easier to compute an accurate classification. By learning the optimal separating hyperplane in this feature space, a non-linear classifier can be learned in the original input space. SVM enables the user to define a number of parameters besides allowing a choice of

inbuilt kernel function including sigmoid, polynomial and radial basis function (RBF). In LTTRPred, we have used three different approaches based on amino acid composition to train the kernels.

Waikato Environment for Knowledge Analysis (WEKA)

WEKA is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand, which is freely available (http://www.cs.waikato.ac.nz/ml/weka/). 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. WEKA supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. Here too, as in SVM, we have used three different approaches, based on amino acid composition, to train the different classifiers in WEKA.

Composition-based methods

Amino-acid composition

Amino-acid composition is the fraction of each amino acid in a given protein sequence. The fraction of all the natural 20 amino acids was calculated using the following equation:

$$\text{Fraction of amino acid (n)} = \frac{\text{Total number of amino acid n}}{\text{Total number of amino acids in the sequence}} \quad \text{(Eq 1)}$$

where, n can be any amino acid.

Dipeptide composition

Dipeptide composition, which gives a fixed pattern length of 400 (20 x 20), encompasses the information of the amino-acid composition along with the local order of amino acids. The fraction of each dipeptide was calculated according to the equation:

$$\text{Fraction of dipeptide (n+1)} = \frac{\text{Total number of dipeptide (n+1)}}{\text{Total number of all possible dipeptide}} \quad \text{(Eq 2)}$$

where, dipeptide (n+1) is one of the 400 dipeptides.

Hybrid method

The hybrid method was developed by combining amino-acid composition and dipeptide composition features of a protein sequence and calculated by Eq (1) and (2). The input vector pattern of 420 (20 for amino acid and 400 for dipeptide composition) was created.

Measurement of performance of LTTRPred in SVM

In our present work, we have adopted 10-fold cross-validation and independent data set validation techniques 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. In the independent dataset test, none of the data to be tested occurs in the training dataset used to train the predictor and the selection of data used for the testing dataset could be quite arbitrary.

Measurement of performance of LTTRPred in WEKA

Here too, we have performed 10-fold cross validation and independent data set validation techniques to evaluate the performance of LTTRPred. In 10-fold cross validation, nine parts were used for training and remaining one for testing. The procedure was repeated ten times so that each of the ten set was used for testing at least once. In the case of independent dataset test, training and test set are created such that each data in both the sets are unique. The classifiers were trained based on the three amino acid composition methods. Out of the 76 classifications and regression algorithms, 11 methods (*viz.,* Naïve Bayes, Logistic, Multilayer Perceptron, RBF Network, Simple Logistic, Voted Perceptron, IB1, IBk, KStar, J48 and Random Forest) were used for designing algorithms. Based on their performances, five each from amino-acid composition and hybrid method and four from dipeptide method were selected.

Evaluation parameters

We adopted five frequently considered measurements for evaluation – accuracy (Ac), sensitivity (Sn), specificity (Sp) precision (Pr) and

Mathew's Correlation Coefficient (MCC) (Hemalatha *et al*., 2013). Accuracy (Ac) defines the correct ratio between both positive (+) and negative (-) data sets. The sensitivity (Sn) and specificity (Sp) represent the correct prediction ratios of positive (+) and negative data (-) sets of LTTR proteins respectively. Precision (Pr) is the proportion of the predicted positive cases that were correct. When the numbers of positive and negative data differ too much from each other, the MCC should be included to evaluate the prediction performance of the developed tool. MCC is considered to be the most robust parameter of any class prediction method. The value of MCC ranges from -1 to 1, and a positive MCC value stands for better prediction performance. The measurements are expressed in terms of true positive (TP), true negative (TN), false positive (FP), false negative (FN):

$$\text{Sensitivity} = \frac{TP}{TP + FN} \times 100$$

$$\text{Specificity} = \frac{TN}{FP + TN} \times 100$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \times 100$$

Receiver operating characteristics (ROC) curve

The performance of a binary classifier can be explained with ROC curve which is a graphical plot drawn by varying threshold values. The analysis of ROC curve helps to characterize the prediction for individual locations (Swets, 1988; Zweig and Campbell, 1993). The graph is created 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 measures the classifier accuracy.

Results and discussion

Composition-based modules and hybrid approach in SVM

Models for SVM were created using polynomial, sigmoid and RBF kernels with different amino acid composition techniques and a hybrid approach. Performances of all the kernels with three composition-based modules were statistically evaluated.

The independent data test results of amino-acid composition and traditional dipeptide composition based module achieved overall accuracy of 100 per cent for all the three kernels *viz*., sigmoid, polynomial and radial basis function. The detailed performance of independent data test results of LTTR proteins with SVM is listed in Table 1. It was observed from the table that the hybrid approach has less accuracy and MCC when compared with other two.

The detailed performance of 10-fold cross-validation results of LTTR proteins with SVM is

Table 1. Comparison of the prediction performance of three kernels of SVM with different composition techniques using independent data test validation

Composition	Algorithm	Sn (%)	Sp (%)	Ac (%)	Pr (%)	MCC
Amino acid	Polynomial	100	100	100	100	1
	RBF	100	100	100	100	1
	Sigmoid	100	100	100	100	1
Dipeptide	Polynomial	100	100	100	100	1
	RBF	100	100	100	100	1
	Sigmoid	100	100	100	100	1
Hybrid	Polynomial	100	100	100	100	1
	RBF	100	96	98	96.15	0.961
	Sigmoid	100	96	98	96.15	0.961

Table 2. Comparison of the prediction performance of three kernels of SVM with different composition techniques using 10-fold cross validation

Composition	Classifier	Sn (%)	Sp (%)	Ac (%)	Pr (%)	MCC
Amino acid	Polynomial	99.37	91.57	96.25	94.65	0.92
	RBF	99.37	91.57	96.25	94.65	0.92
	Sigmoid	99.37	91.57	96.25	94.65	0.92
Dipeptide	Polynomial	99.62	94.54	97.59	96.47	0.95
	RBF	99.63	94.54	97.59	96.47	0.95
	Sigmoid	99.61	94.54	97.58	96.47	0.95
Hybrid	Polynomial	99.75	94.43	97.62	96.41	0.95
	RBF	99.70	93.88	97.37	96.07	0.95
	Sigmoid	99.70	93.86	97.36	96.05	0.95

listed in Table 2. The 10-fold cross validation of amino acid composition-based module gave an accuracy of 96.3 per cent with MCC of 0.92 with all the three kernels. In the 10-fold cross validation of dipeptide method, using the polynomial and RBF kernels, an accuracy of 97.6 per cent was achieved with MCC 0.95. With the sigmoid kernel in the dipeptide method, an accuracy of 97.6 per cent was achieved with MCC 0.95. In the 10-fold cross validation of hybrid-based approach, the polynomial kernel had maximum accuracy of 97.6 and MCC 0.95. Use of the other two kernels gave less accuracy and MCC in 10-fold cross validation. Therefore, we selected the hybrid polynomial kernel for the tool development.

Composition-based modules and hybrid approach in WEKA

Models for WEKA were created using 11 classifiers and features were extracted using two different amino acid composition techniques and a hybrid approach. The performance of all the classifiers with three methods was then statistically evaluated. From these 11 classifiers, Multi Layer Perceptron, IB1, IBK, KStar, Random Forest, Logistic and Simple Logistic classifiers gave best

Table 3. Comparisons of the prediction performance of classifiers of WEKA with different composition techniques using independent data test validation

Composition	Classifier	Sn (%)	Sp (%)	Ac (%)	Pr (%)	MCC
Amino acid	Multi Layer Perceptron	100	100	100	100	1
	IB1	100	100	100	100	1
	IBK	100	100	100	100	1
	Kstar	100	100	100	100	1
	Random Forest	100	100	100	100	1
Dipeptide	Logistic	100	100	100	100	1
	Simple Logistic	100	100	100	100	1
	IB1	100	100	100	100	1
	IBK	100	100	100	100	1
Hybrid	Logistic	100	100	100	100	1
	Simple Logistic	100	100	100	100	1
	IB1	100	100	100	100	1
	IBK	100	100	100	100	1
	Random Forest	100	100	100	100	1

Table 4. Comparisons of the prediction performance of classifiers of WEKA with different composition techniques using 10-fold cross validation

Composition	Classifier	Sn (%)	Sp (%)	Ac (%)	Pr (%)	MCC
Amino acid	Multi LayerPerceptron	99.65	99.17	99.46	99.45	0.988
	IB1	99.84	99.37	99.65	99.58	0.992
	IBK	99.84	99.37	99.65	99.58	0.992
	Kstar	99.91	98.6	99.39	99.07	0.987
	Random Forest	99.73	99.44	99.62	99.63	0.992
Dipeptide	Logistic	99.66	99.51	99.60	99.67	0.991
	Simple Logistic	99.71	99.46	99.61	99.64	0.991
	IB1	98.9	99.94	99.32	99.96	0.985
	IBK	98.9	99.94	99.32	99.96	0.985
Hybrid	Logistic	99.6	99.46	99.55	99.64	0.990
	Simple Logistic	99.66	99.51	99.6	99.67	0.991
	IB1	99.08	99.94	99.42	99.96	0.988
	IBK	99.08	99.94	99.42	99.96	0.988
	Random Forest	99.8	99.48	99.67	99.65	0.993

results during testing. All the classifiers used in amino-acid composition, traditional dipeptide composition and hybrid approach achieved overall accuracy of 100 per cent in independent data test results (Table 3). The detailed performance of 10-fold cross validation results of LTTR proteins with WEKA is listed in Table 4. The 10-fold cross validation of all the amino acid methods give an accuracy of 99 per cent with MCC of 0.99 and 0.98 with all the classifiers used. In the 10-fold cross validation of hybrid-based approach, use of Random Forest classifier resulted in maximum accuracy of 99.7 and MCC 0.993.

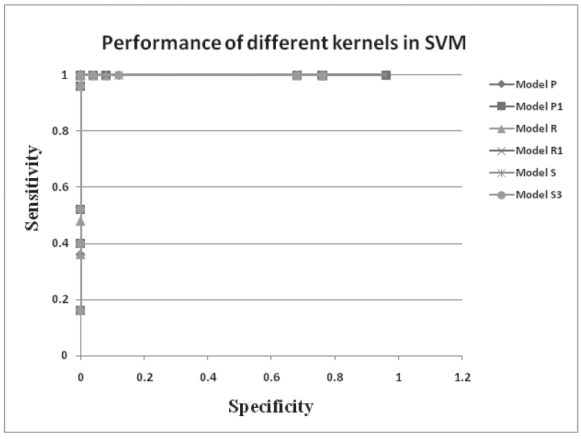

Fig.1. ROC curve for different compositions in SVM

ROC curves

SVM

The ROC curve is a measure which represents the relationship between sensitivity and specificity for a class. We have plotted the ROC curves based on the performance of the various compositions. From the ROC curve (Fig.1), it is clear that all the methods used to train the three kernels represent a perfect classifier since the curve represents an inverted 'L'. This is a desirable characteristic of an ROC curve. Each point on the ROC curve was plotted based on different threshold scores. The figure also depicts area under the curve (AUC = 1) value for all the methods (Hosmer and Lemeshow, 2000). The AUC shows the probability that when one positive and negative sample are drawn at random, the decision function assigns a higher value to the positive than to the negative sample.

WEKA

From the ROC curve (Fig. 2), it is clear that all the classifiers used for hybrid approach represents perfect classifiers since the curve represents an inverted 'L', which is a desirable characteristic of an ROC curve. The figure also depicts area under the curve (AUC = 1) value for all the classifiers

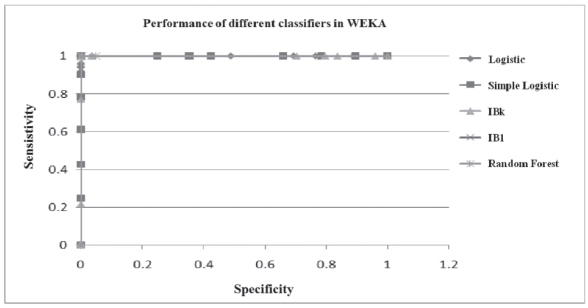

Fig. 2. ROC curve for different classifiers in WEKA

used. Here too, as in SVM, the AUC shows the probability that when one positive and negative sample are drawn at random, the decision function assigns a higher value to the positive than to the negative sample.

Description of the web server

The best performing SVM and WEKA modules from the present investigation was implemented on the World Wide Web as a dynamic web server 'LTTRPred', which is freely available and can be accessed online (http://210.212.229.56/lttrpred/). All the CGI scripts of LTTRPred were written in PERL and the interface was designed using HTML to assess user queries. The overall architecture of the 'LTTRPred' web server is shown in Figure 3. The web server gives the option to user for selection of the modules in SVM or WEKA. Then it allows users to submit their proteins sequence in one of the standard formats such as FASTA or plain text (Figure 4). Users can type or paste the sequence in the box, or upload the sequence through a file. The prediction result will be displayed in a user friendly format on the screen within few seconds.

With advances in genome sequencing technologies and rapid availability of whole genome sequences, tools and resources need to be developed to deduce the information contained in these genome sequences. A major difficulty with prokaryotic genome annotation is the lack of accurate gene prediction programs. Similar to all completed genomes, *Pseudomonas* has a substantial number of genes that are hypothetical since they are predicted solely on the basis of gene prediction programs. This necessitates the need for refinement and improvement of the quality of gene prediction programs for *Pseudomonas*. An era of biological revolution has begun during which a huge amount of information on microbial genetics will be accumulated at a fast pace. Thus, the availability of systems/tools that can predict location from sequence is essential to the full characterization of expressed proteins. Computational tools provide faster and accurate access to predictions for any organism. Identification of LTTR proteins from sequence databases is difficult due to poor sequence similarity. In this work, we present a new method for LTTR prediction based on SVM and WEKA,

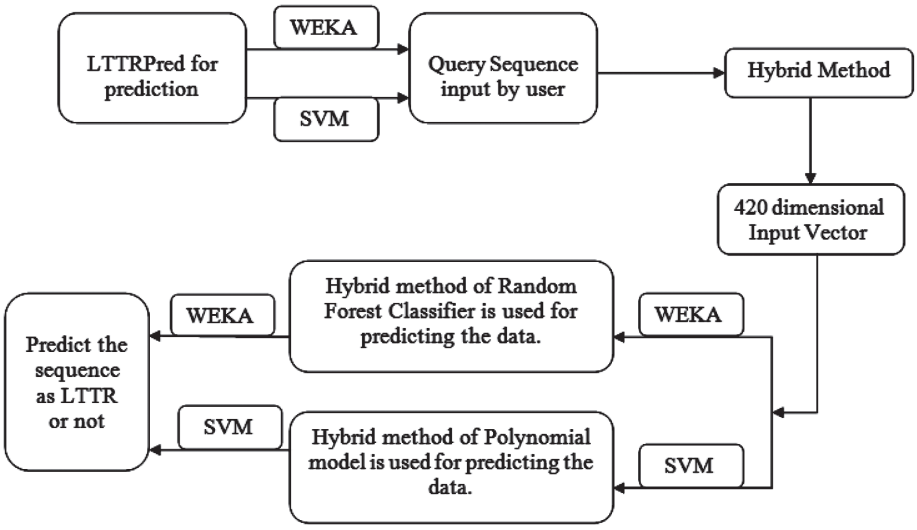

Fig. 3. Overall architecture of the LTTRPred web server

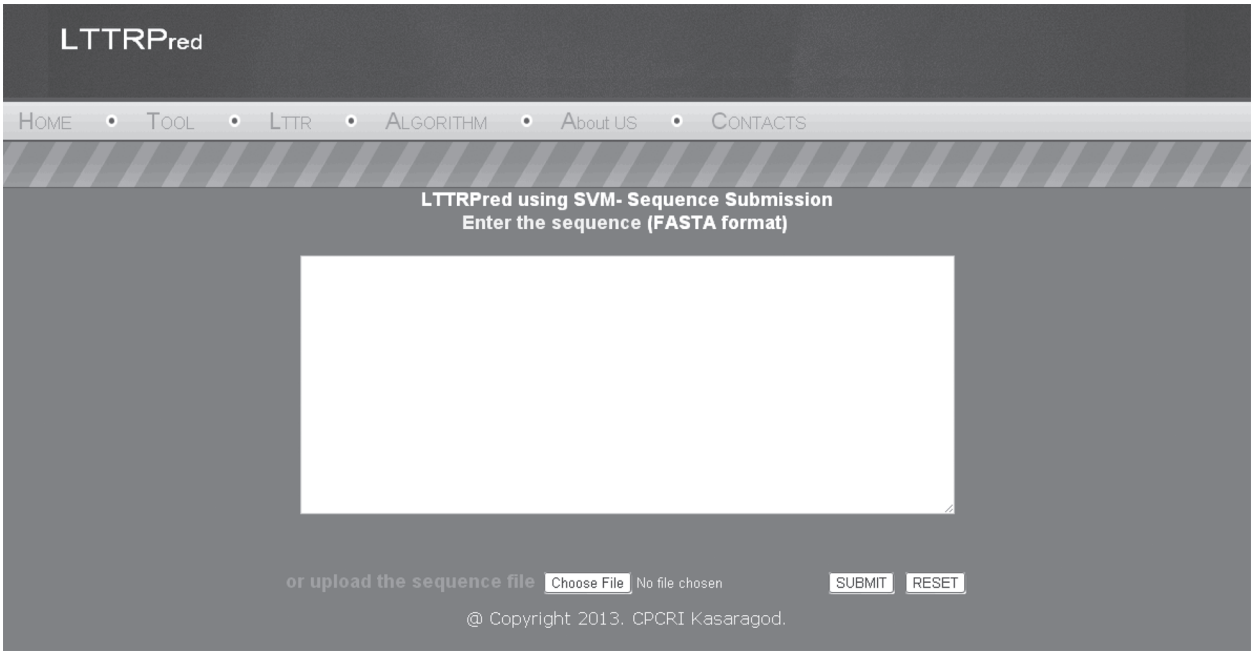

Fig. 4. An overview of the online LTTRPred web server

the performance of which was found to be highly satisfactory. Very high prediction accuracies for the validation tests show that LTTRPred is a potentially useful tool for the prediction of LTTR proteins in *Pseudomonas spp.*

Acknowledgment

The authors are thankful to Dr. George V. Thomas, Director, CPCRI for all support, encouragement and facilities provided. This work was supported by a grant from Department of Biotechnology (BTISnet) and Department of Information Technology (ABPC), Government of India.

References

Bopaiah, B.M. 1985. Occurrence of phosphate solubilising microorganisms in the root region of arecanut palms. *Journal of Plantation Crops* **13**: 60-62.

Bopaiah, B.M. and Shetty, H.S. 1991. Soil microflora and biological activities in the rhizospheres and root regions of coconut-based multistoreyed cropping and coconut monocropping systems. *Soil Biology and Biochemistry* **23**: 89-94.

Chakraborty, U., Chakraborty, B.N., Chakraborty, A.P., Sunar, K. and Dey, P.L. 2013. Plant growth promoting rhizobacteria mediated improvement of health status of tea plants. *Indian Journal of Biotechnology* **12**: 20-31.

Cortes, C. and Vapnik, V. 1995. Support vector networks. *Machine Learning* **20**: 273-293.

Dowling, D.N. and O'Gara, F. 1994. Metabolites of *Pseudomonas* involved in the biocontrol of plant disease. *Trends in Biotechnology* **12**: 33-144.

Glick, B.R. 1995. The enhancement of plant-growth by free-living bacteria. *Canadian Journal of Microbiology* **41**: 109-117.

Haas, D. and Keel, C. 2003. Regulation of antibiotic production in root colonizing *Pseudomonas* spp., and relevance for biological control of plant disease. *Annual Review of Phytopathology* **79**: 117-153.

Henikoff, S., Haughn, G.W., Calvo, J.M. and Wallace, J.C. 1988. A large family of bacterial activator proteins. *Proceeding of National Academy of Sciences USA* **85**: 6602-6066.

Hidayati, U., Chaniago, I.A., Munif, A., Siswanto and Santosa, D.A. 2014. Potency of plant growth promoting endophytic bacteria from rubber plants (*Hevea brasiliensis* Mull. Arg.). *Journal of Agronomy* **13:** 147-152.

Hosmer, D.W. and Lemeshow, S. 2000. *Applied Logistic Regression*, Ed. 2. John Wiley and Sons, New York. pp. 156-164.

Joachims, T. 1999. Making large-scale SVM learning practical. In: *Advances in Kernel Methods - Support Vector Learning*. (Eds.) Schölkopf, B., Burges, C.J.C. and Smola, A. J. MIT Press, Cambridge, Massachusetts. pp. 169-184.

Kloepper, J.W. and Schroth, M.N. 1978. Plant growth-promoting rhizobacteria on radishes. *Proceedings of the 4th International Conference on Plant Pathogenic Bacteria.* INRA, Angers, France. pp. 879 – 882.

Litty Thomas, Alka Gupta, Murali Gopal, Chandramohanan, R., Priya George and George V. Thomas. 2011. Evaluation of rhizospheric and endophytic *Bacillus* spp. and fluorescent *Pseudomonas* spp. isolated from *Theobroma cacao* L. for antagonistic reaction to *Phytophthora palmivora*, the causal organism of black pod disease of cocoa. *Journal of Plantation Crops* **39**: 370-376.

Maurhofer, M., Keel, C., Haas, D. and Défago, G. 1994. Pyoluteorin production by *Pseudomonas fluorescens* strain CHA0 is involved in the suppression of *Pythium* damping-off of cress but not of cucumber. *European Journal of Plant Pathology* **100**: 221-232.

Niranjan Raj, S., Shetty, H.S. and Reddy, M.S. 2005. Plant growth-promoting rhizobacteria: potential green alternative for plant productivity. In: *PGPR: Biocontrol and Biofertilization.* (Ed.) Siddiqui, Z. A. Springer, Dordrecht, The Netherlands. pp 197-216.

Noble, W.S. 2004. Support vector machine applications in computational biology. In: *Kernel Methods in Computational Biology*. (Eds.) Scholkopf, B., Tsuda, K. and Vert, J. P. MIT Press, Cambridge, Massachusetts. pp. 71–92.

Nowak-Thompson, B., Chaney, N., Wing, J.S., Gould, S.J. and Loper, J.E. 1999. Characterization of the pyoluteorin biosynthetic gene cluster of *Pseudomonas fluorescens* Pf-5. *Journal of Bacteriology* **181**: 2166-2174.

Paulsen, I.T., Press, C.M., Ravel, J., Kobayashi, D.Y., Myers, G.S. 2005. Complete genome sequence of the plant commensal *Pseudomonas fluorescens* Pf-5. *Nature Biotechnology* **23**: 873-878.

Podile, A.R. and Kishore, G.K. 2006. Plant Growth Promoting Rhizobacteria. In: *Plant Associated Bacteria.* (Ed.) Gnanamanickam, S.S. Springer, Dordrecht, The Netherlands. pp. 195-230.

Raaijmakers, J.M, Vlami, M. and de Souza, J.T. 2002. Antibiotic production by bacterial biocontrol agents. *Antonie Van Leeuwenhoek* **81**: 537-547.

Schell, M.A. 1993. Molecular biology of the LysR family of transcriptional regulators. *Annual Review of Microbiology* **47**: 597-626.

Schneider, G. and Fechner, U. 2004. Advances in the prediction of protein targeting signals. *Proteomics* **4**:1571-1580.

Swets, J.A. 1988. Measuring the accuracy of diagnostic systems. *Science* **240**: 1285-1293.

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

Zweig, M.H. and Campbell, G. 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39**: 561-577.