Int. J. Agricult. Stat. Sci., Vol. 9, No. 2, pp. 717-728, 2013 *ISSN : 0973-1903*

SUPPORT VECTOR MACHINE FOR PREDICTION OF ANTIMICROBIAL PEPTIDES IN LEGUMES

Sarika*, M. A. Iquebal, Anil Rai and Anshika[1]

Centre for Agricultural Bioinformatics, I. A. S. R. I., PUSA, New Delhi - 110 012, India.
[1]Jaipur National University, Jagatpura, Jaipur - 302 025, India.
E-mail: sarika@iasri.res.in

Abstract

Resistance to chemical antibiotics is an unsolved and growing problem. A new generation of native peptide molecules, also known as antimicrobial peptides (AMPs) may be a natural alternative to chemical antibiotics and a potential area of research under applied biotechnology. In the present study, a systematic attempt has been made to develop a direct method for predicting AMPs of legumes using Support Vector Machine (SVM). The SVM based method with polynomial kernel function with degree 2 was found to be the best model for classification of legume AMPs with accuracy and Mathews Correlation Coefficient of 96.4% and 0.931, respectively. The best performance was obtained at threshold 0.5, where the sensitivity, specificity were 1.000 and 0.929, respectively. The ROC curve was plotted and area under curve (AUC) was found to be 0.964 with standard error of 0.041, which indicated a good prediction performance. It is anticipated that the current prediction method would be a useful tool for the systematic analysis of genome data. AMPs identified from the studies may be used to confer disease resistance in other crops as transgenics, thus opening unsuspected alternative to provide agronomically relevant levels of disease control worldwide.

Key words : Antimicrobial peptides, Legumes, Genomics, Kernel function, ROC curve.

1. Introduction

Legume is one of the most agriculturally important family of crops known for their ability to fix atmospheric nitrogen and has symbiotic relationship with rhizobium found in root nodules. The enormous digital information of legume genomes has triggered the use of bioinformatics and other *in silico* approaches to retrieve important and useful information. Special attention has been given to a peptidic group of plant bioactive molecules known as antimicrobial peptides (AMP). These are usually small cysteine or glycine-rich peptides, antagonistic to several pathogens and component of plant innate defense. Main classes of AMPs comprise of classes like defensins, thionins, lipid-transfer proteins, cyclotides, snakins and hevein-like, according to amino acid sequence homology [Pestana-Calsa *et al*. (2010)]. AMPs are the hosts' defence molecules, identified as an essential part of innate immunity in

***Corresponding author.** *Received Jan. 11, 2013* *Revised June 19, 2013* *Accepted Aug. 18, 2013*

response to microbial challenges [Otvos (2000)]. The antimicrobial function of innate immunity is mediated by these potent, majorly cationic peptides having broad spectrum of antimicrobial activity against microbes such as gram positive and negative bacteria, viruses, fungi, parasites etc. These peptides are an innovative alternative to chemical antibiotics to overcome the problem of resistance against pathogens and hence termed as "natural antibiotics". AMPs have shown their presence in bioengineering and are used as a biotechnological tool for creating transgenic agricultural crops, biofuels etc. [Bryksa *et al*. (2010)]. Although, the source of AMPs vary from prokaryote to eukaryotes, but highest concentration of AMPs are found in animal tissues (~71%) as compared to plants (~14%) exposed to microbes or cell types that are involved in host defense [Wang and Wang (2009)]. In huge repository of digital biological data, a number of specialised databases like PhytAMP [Hammami *et al.* (2009)], AMSDb [Tossi and Sandri (2002)], APD2 [Wang and Wang (2009)], ANTIMIC [Zheng and Zheng (2002)], AMPer [Fjell *et al*. (2007)], CAMP [Thomas *et al*. (2010)] etc. have been created accounting for AMPs.

Computational methods complement laboratory experimentation for efficient identification of antimicrobial peptides. Intensive literature review revealed numerous approaches previously employed for solving classification problems in biology. To name a few approaches, binding motifs, quantitative matrices, hidden Markov models and network based prediction algorithms are widely used [Brusic *et al*. (2004)]. A non-parametric, *Generalized Portrait* algorithm, the Support Vector Machine (SVM) algorithm, developed in Russia in the 1960s by Vapnik and Chervonenkis (1974), is gaining popularity due to many attractive features and promising empirical performance. SVM is a powerful methodology for solving problems in classification, function estimation and density estimation with no prior assumptions about data and underlying distribution. Also, it does not necessitate large number of training data to avoid overfitting. It implements the Structural Risk Minimization (SRM) principle, which has been shown to be superior to traditional Empirical Risk Minimization (ERM) principle implemented in ANN models. Its solution is always unique and globally optimal.

In order to avail the benefits of the molecular data in terms of antimicrobial peptides from the specialized database and build the classification models, SVM was applied on AMPs especially derived from legumes for the present study. Here, attempts have been made to develop prediction programs in particular for legumes to have reasonably higher accuracy. These developed models may be treated as experimental analogous to standard laboratory procedures.

2. Materials and Methods

2.1 Extraction of AMPs in legumes

The antimicrobial peptide sequences were extracted from various specialized databases like PhytAMP, UniProt Knowledgebase, CAMP, APD2 etc. Around hundred peptide sequences were taken under study for analysis purpose. These peptides belonged to two major classes of antimicrobial and non-antimicrobial peptides. No check for similar sequences was applied in order to conserve the natural preference of certain patterns over others.

2.2 Pre-processing of the Sequences

Before using SVM algorithm for training and testing, the biological sequences need to be converted to format suitable for input to computer system. For the study, each instance was denoted by a vector, having 20 attributes (or *features*) representing the amino acid composition (*AAC*) for that instance. *AAC* is a quantitative measure of the sequence that represents the sequence in terms of 20 values, one for each amino acid residue. For ith amino acid residue, *AAC* is defined as the percentage of ith residue in whole sequence. Mathematically,

$AAC_i = (N_i / N) \times 100$

Where, ACC_i = AAC of ith amino acid residue.

N_i = Number of occurrences of ith amino acid residue in the sequence.

N = total number of amino acid residue in the sequence.

AAC completely omits the sequence order information and focuses only on the percentage amino acid residue content. The addressed problem is binary classification type. Hence, a matrix of order *N x 20* (here, *N* is 98) is obtained, which is used as input in further study. The target vector comprises of binary class *i.e.* AMP or Non-AMP.

2.3 Support Vector Machine (SVM)

Support vector machines are relatively new type of supervised machine-learning techniques, proven to be particularly attractive to biological analysis due to their ability to handle noise and large input spaces [Brown *et al.* (2000), Ding and Dubchak (2001)]. Following is the basic idea behind SVM for pattern recognition, mainly for two-class classification problem.

Considering two-class classification problem and assuming a set of samples, *i.e.* a series of input vectors $x_i \in \Re^d$ ($i = 1, 2, ..., N$) with corresponding levels $y_i \in \{+1, -1\}$ ($i = 1, 2,...,N$). Here, +1 and -1 indicate two classes. To predict antimicrobial property, the input vector dimension is 20 and each input vector unit stands for one amino acid. The objective now is to construct a binary classifier or derive a decision function from the available samples, which has a small probability of misclassifying a future sample. SVM maps input vectors $x_i \in \Re^d$ into a high dimensional feature space $\phi(x_i) \in H$ and constructs an optimal separating hyperplane (OSH), which maximises the margin, the distance between hyperplane and nearest data points of each class in the space H (Figure 1). The equation of a simple hyperplane is given by

$y = sign[w^T, x + b]$

Where, w denotes a weight vector that can map the training data in the input space to the output space and b is the bias.

When the data of the two classes are separable, it can be written as

$$\begin{cases} w^T x_i + b \geq +1, & if \quad y_i = +1 \\ w^T x_i + b \leq -1, & if \quad y_i = -1 \end{cases}$$

These two sets of inequalities can be combined into one single set as follows

$y_i [w^T x_i + b] \geq 1, \quad i = 1, 2, ..., N$

Support vector machine formulations are done within a context of convex optimization theory. The primal form Quadratic Programming (QP) problem is given by

$$\min_{w,b} \; J_P(w) = \frac{1}{2} w^T w$$

Such that

$y_i [w^T x_i - b] \geq 1, \quad i = 1, 2, ..., N$

This is called the primal optimization problem. In the present case, it will turn out that it is more convenient to deal with the dual. To derive it, the Lagrangianis introduced as

$$L(w,b;\alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i \left(y_i \left[w^T x_i + b \right] - 1 \right)$$

with Lagrange multipliers $\alpha_i \geq 0$ for $i = 1, 2, ..., N$. The solution can be obtained as

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 \Rightarrow \; w = \displaystyle\sum_{k=1}^{N} \alpha_i y_i x_i \\ \dfrac{\partial L}{\partial b} = 0 \Rightarrow \; \displaystyle\sum_{k=1}^{N} \alpha_i y_i = 0 \end{cases}$$

resulting linear classifier is

$$f(x) = sign\left[\sum_{k=1}^{N} \alpha_i y_i x_i^T x + b \right]$$

By replacing the expression for w in the Lagrangian, following Quadratic programming (QP) problem given as the Dual form problem in the Lagrange multipliers α_i is

$$\max_{\alpha} \; J_D(\alpha) = -\frac{1}{2} \sum_{i,j=1}^{N} y_i y_j x_i^T x_j \alpha_i \alpha_j + \sum_{i=1}^{N} \alpha_i$$

$$\text{such that } \sum_{k=1}^{N} \alpha_i y_i = 0$$

Note that this problem is solved in $\alpha = [\alpha_1, ..., \alpha_N]$, not in w. with resulting linear classifier.

$$f(x) = sign\left[\sum_{i=1}^{\#SV} \alpha_i y_i x_i^T x_i + b \right]$$

The index i run now over the number of support vectors, where training data points corresponding to non-zero α_i values are called support vectors.

The bias b determined from complementary conditions of the Karush-Kuhn-Tucker

(KKT) condition, which state that the product of the dual variable and the constraints should be zero at the optimal solution. Hence,

$y_i [w^T x + b] - 1 = 0$

Then,

$$b = y_i - \sum_{i=1}^{\#SV} \alpha_i y_i x_i^T x_i$$

Instead of using an arbitrary Support Vector x_i, it is better to take an average over all the Support Vectors.

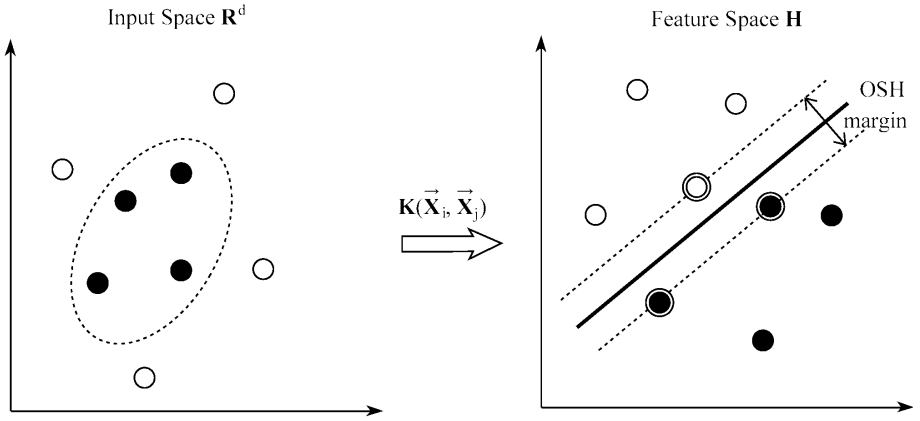

Fig. 1 : A separating hyperplane in the feature space corresponding to a non-linear boundary in the input space. Two classes denoted by circles and disks are linear non-separable in the input space $\Re^d$. SVMs constructs the OSH (the solid line), which maximises the margin between two classes by mapping the input spaceinto a high dimensional space, the feature space H. Mapping is determined by a kernel function $K(x_i, x_j)$. Support vectors are identified with double circle.

Some binary classification problems do not have a simple hyperplane as a useful separating criterion. For those problems, there is a variant of the mathematical approach that retains nearly all the simplicity of an SVM separating hyperplane. Let, x be a vector in the *n* dimensional input space and φ(.) be a nonlinear mapping function from the input space to the high dimensional feature space, which can be infinite dimension. Different mappings construct different SVMs. The mapping φ(.) is performed by kernel function $K(x_i, x_j)$, which defines an inner product in the space H. The decision function implemented by SVM is as follows:

$$f(x) = sign\left[\sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b\right]$$

Where, the coefficients α_i are obtained by solving the following convex quadratic programming problem

$$\max_{\alpha} \; -\frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) + \sum_{i=1}^{N} \alpha_i$$

subject to $0 \leq \alpha_i \leq C$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

Here, C is regularization parameter that controls the trade off between margin and misclassification error. These x_j's are called support vectors only, if corresponding $\alpha_i > 0$. The choice of the proper kernel function is an important issue for SVM training because the power of SVM comes from the kernel representation that allows the nonlinear mapping of input space to a higher dimensional feature space. Some typical choices of kernel function [Cristianini and Shawe-Taylor (2000)] are as follows:

a. $K(x_i, x_j) = x_i^T x_j$ (Linear SVM)

b. $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d$ (Polynomial SVM of degree d)

c. $K(x_i, x_j) = exp\{-\gamma \| x_i - x_j \|^2\}$ (Radial Basis function Kernel)

d. $K(x_i, x_j) = tanh(\gamma x_i^T x_j + r)$ (Sigmoid)

Where, d and $\gamma > 0$ are the kernel parameters.

SVM can handle large feature spaces, effectively avoid overfitting by controlling the margin and automatically identify a small subset made up of informative points, *i.e.* support vectors, etc. The use of appropriate decision function can give better classification. For a given dataset, only the kernel function and regularization parameter C are selected to specify the model. SVM has many attractive features. For instance, the solution of the quadratic program (QP) problem is globally optimized while, with neural networks the gradient based training algorithms only guarantee finding a local minima. In addition, SVM can handle large feature spaces, effectively avoid overfitting by controlling the margin and automatically identify a small subset made up of informative points, *i.e.* the Support Vectors etc.

2.4 Five-fold cross Validation

In this study, all models were evaluated using five-fold cross-validation technique. Here, dataset is randomly divided into five sets, each set containing almost equal number of peptides. Among these, four sets are used for training and the remaining one set for testing. The process is repeated five times such that each set gets the opportunity to fall under testing. Average of five sets is calculated finally.

2.5 Assessment of the Prediction Accuracy

Computational models that are valid, relevant and properly assessed for accuracy can be used for planning of complementary laboratory experiments. The prediction quality was examined by testing the model, obtained after training the system, with test data set. Several measures are available for the statistical estimation of the accuracy of prediction models.

The common statistical measures are Sensitivity, Specificity, Precision or Positive predictive value (PPV), Negative predictive value (NPV), Accuracy and Mathew's correlation coefficient (MCC).

The sensitivity indicates the 'quantity' of predictions, *i.e.*, the proportion of real positives correctly predicted. The specificity indicates the 'quality' of predictions, *i.e.*, the proportion of true negatives correctly predicted. The PPV indicates the proportion of true positives in predicted positives- "the success rate" while NPV is the proportion of true negatives in predicted negatives.

These measures are defined as follows:

$$\text{Sensitivity} = TP/(TP+FN)*100 \qquad \text{Specificity} = TN/(FP+TN)*100$$

$$PPV = TP/(TP+FP)*100 \qquad NPV = TN/(TN+FN)*100$$

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

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

Where, *TP* = True Positive, *TN* = True Negative

FP = False Positive and *FN* = False Negative

ROC (Receiver Operator Characteristics) analysis is a visual and numerical method to assess the performance of classification algorithms. It is a graph obtained by selecting a series of threshold and representing a dependency of sensitivity versus specificity. ROC curve characterizes a probabilistic classifier, where each point on the curve corresponds to a discrete classifier. Area under the ROC Curve (AUC) is a widely used measure for predictive performance [Sonego *et al*. (2008)]. An AUC of 0.95 means that 95% of the pairs are correctly classified, whereas a test with an AUC of 0.50 is non-discriminative. By calculating the AUC, an approximated measure of probability is determined for each model, which is as follows :

AUC range	0.91-1.00	0.81-0.90	0.71-0.80	0.61-0.70	0.51-0.60
Model fitted	Excellent	Good	Fair	Poor	Failed

3. Results and Discussion

Initial *in silico* approaches lead to quickachievable AMP coding potentials of the plant species under study even, if it requires further biological validation. Plant AMPs represent almost 15% of deposited AMP sequences of which approximately 25% belong to legumes [Wang and Wang (2009)]. The main classes of collected antimicrobial peptide sequences from legume crops along with their percentage contribution reported in this study (Fig. 2) are

defensin, lipid-transfer protein (LTP), cicerarin, cicerin, cyclophilin, gymnin, thanumatin-like protein (TLP), antifungal lectin and arietin. It was observed that AMP class, defensinis abundantly reported in legumes followed by LTP. *In silico* studies of these AMP help to unravel the functional aspects of peptides.

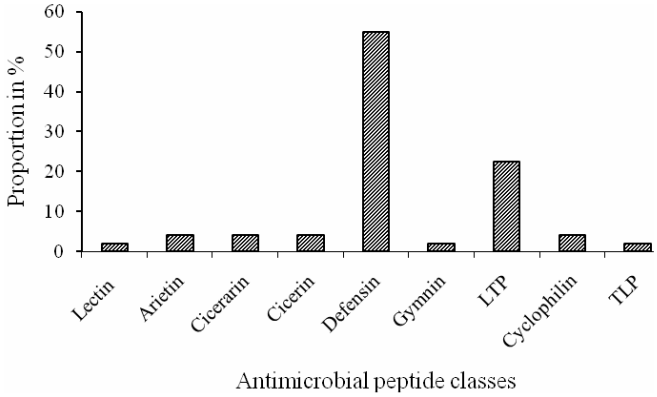

Fig. 2 : Class wise distribution of legume AMPs.

In this study, classification model has been developed using SVM. Total of 98 peptides sequences (49 from antimicrobial class and 49 from non-antimicrobial class) of legume crop were considered here. The peptide sequences were converted (*i.e.* pre-processing) to format suitable for input to SVM algorithm. Pre-processing of the sequences information and calculation of amino acid composition (*AAC*) was carried out by writing scripts in PERL. Each instance was denoted by a vector, having 20 attributes (or features) representing the amino acid composition (*AAC*) for that instance. This consists of series of input vectors $x_i \in \Re^d$ ($i = 1, 2, ..., N$). Hence, a matrix of order 98×20 is obtained, which is used as input in further study. The target vector comprises of binary class *i.e.* AMP and Non-AMP. Hence, this is a problem of binary classification type representing the vector y_i having +1 and -1 as values. About 70% of total data *i.e.* 70 was used for training purpose (model development) and remaining 28 for testing (model validation) purpose. All the relevant computer program for obtaining the classification models using SVM algorithms have been developed in Package 'e1071' under R software, which is free for non-commercial use and can be obtained from website (http://cran.r-project.org/). This program allows users to run SVM using various kernels and parameters.

Training of SVMs using all kernel functions, *viz.* Linear, Polynomial of degree 2, Polynomial of degree 3, Radial basis function (RBF) and Sigmoid function were performed to get best classification model. The performance measures (sensitivity, specificity, PPV, NPV, accuracy and MCC) were obtained for all kernels functions and results were presented in Table 1 for 5-fold cross validation. Number of support vectors was computed as 41. The highest accuracy (0.964) and MCC (0.931) were achieved for polynomial kernel function with degree 2. It was observed that both accuracy and MCC were decreased for other kernel functions (linear, polynomial of degree 3, RBF and sigmoid) used under this study. Hence, SVM with

polynomial of degree 2 kernel function was found to be best for classification of antimicrobial peptides from non-antimicrobial peptides. The distribution of observed and predicted classes was shown in Table 2. Table 3 represents the different evaluation measures for training and test sets as well as overall data. The graphical representation of performances of various kernels is shown in Fig. 3.

Table 1 : Prediction accuracy of AMPs with different kernel functions.

Kernel	Sensitivity	Specificity	PPV	NPV	Accuracy	MCC
Linear	0.714	1.000	1.000	0.778	0.857	0.745
Polynomial (Degree 3)	0.857	1.000	1.000	0.875	0.929	0.866
Polynomial (Degree 2)	**0.929**	**1.000**	**1.000**	**0.933**	**0.964**	**0.931**
Radial Basis Function	1.000	0.643	0.737	1.000	0.821	0.688
Sigmoid	0.857	0.857	0.857	0.857	0.857	0.714

Table 2 : Distribution of observed and predicted classes.

		Observed		Total
		Negative	Positive	
Predicted	Positive	46	0	46
	Negative	3	49	52
Total		49	49	98

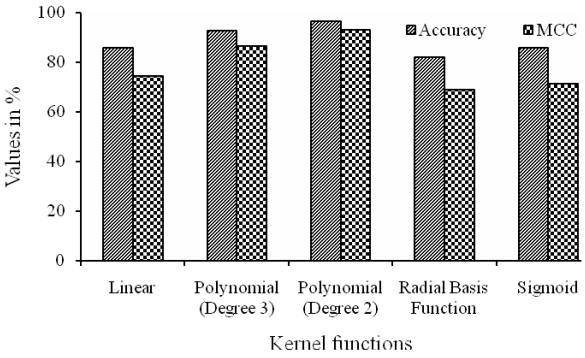

Fig. 3 : Performance of various kernel functions

Table 3 : Evaluation measures for best classifier (Polynomial with degree 2 kernel function).

	Training Set	Test Set	Overall
Sensitivity	0.943	0.929	0.939
Specificity	1.000	1.000	1.000
PPV	0.946	0.933	0.942
NPV	1.000	1.000	1.000
Accuracy	0.971	0.964	0.969
MCC	0.944	0.931	0.941

The choice of parameters for training purpose with polynomial of degree 2 was made after running a grid analysis on the input data set. After fine tuning of the model, it was found that optimal results were obtained at regularity parameter $C = 0.3$ and kernel parameters, $\gamma = 0.1$ and $\gamma = 0.5$ performance (error) is 0.10. Fig. 4 (A and B) shows the relationship of error with kernel parameter gamma and regularity parameter C.

The performance of SVM based method with 'Polynomial kernel function of degree 2' was measured at various threshold values (Table 4).

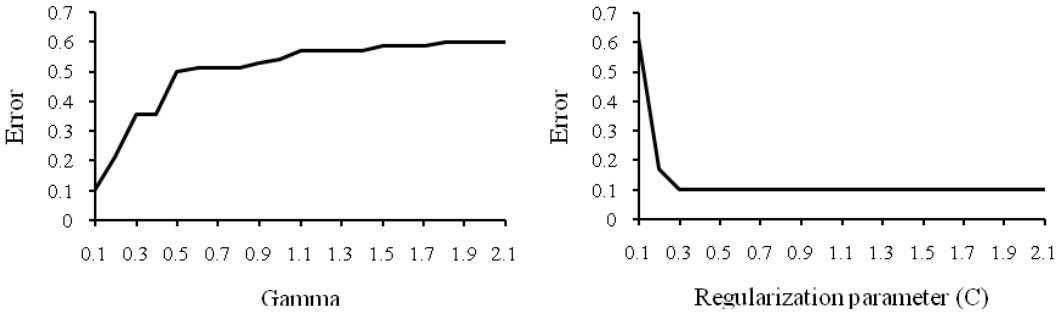

Fig. 4 : Relation of Gamma (A) and Regularization parameter, C (B) with error.

Table 4 : Performance of SVM based method with 'Polynomial kernel function of degree 2' at various threshold values.

Threshold	Sensitivity	Specificity	PPV	NPV	Accuracy	MCC
0.00	1.000	0.429	0.636	1.000	0.714	0.522
0.25	1.000	0.643	0.737	1.000	0.821	0.688
0.50	**1.000**	**0.929**	**0.933**	**1.000**	**0.964**	**0.931**
0.75	0.929	0.929	0.929	0.929	0.929	0.857
1.00	0.143	1.000	1.000	0.538	0.571	0.277

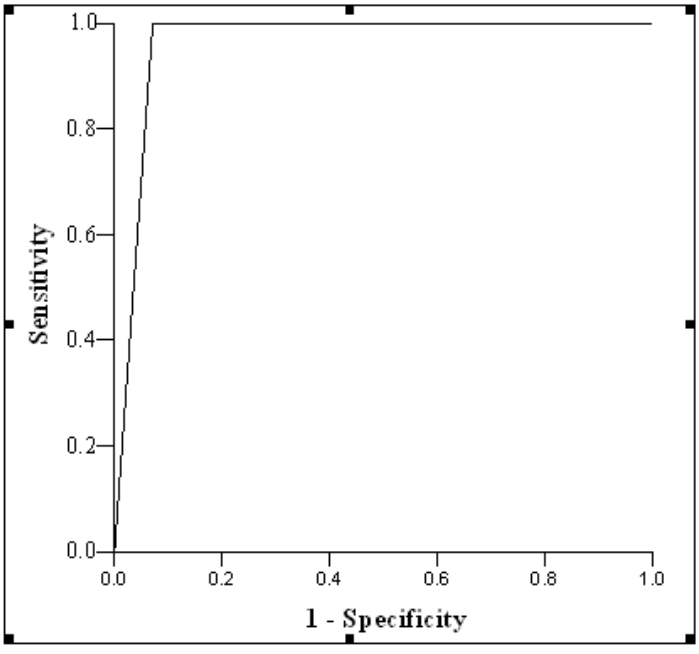

Fig. 5 : Area under ROC at threshold value 0.5.

Prediction accuracy varied from 57.1 to 96.4% while, *MCC* ranged from 0.522 to 0.931. The best performance was obtained at threshold 0.5, where the sensitivity, specificity were 1.000 and 0.929, respectively. At threshold value 0.5, the ROC was plotted using SPSS 17.0 version and area under curve (AUC) was found to be 0.964 with standard error of 0.041 (Fig. 5), which indicated an excellent prediction performance of the classifier.

4. Conclusion

Computational prediction is an important immuno-informatic technology supporting the determination of AMPs. The SVM based method with polynomial kernel function with degree 2 was found to be best model for classification of legume AMPs and the kernel parameters, gamma and regularization parameter were also further fine tuned to achieve best performance in terms of misclassification error. This developed model may further be used for identification of antimicrobial peptides from candidate peptides. It is anticipated that the current prediction method would be a useful tool for the systematic analysis of genome data. Although, computational analyses and predictions may complement, but cannot exactly replace laboratory experiments. However, this analysis may help to minimize number of required laboratory experiments. AMPs identified from the studies may be used to confer disease resistance in other crops as transgenics, thus opening unsuspected alternative to provide agronomically relevant levels of disease control worldwide.

Acknowledgement

The authors would like to thank Editor and Reviewer for constructive comments and suggestions to improve the quality of the last version of this article.

References

Brown, M. P. S., W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares and D. Haussler (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of the Sciences of the United States of America*, **97**, 262–267.

Brusic, V., V. B. Bajic and N. Petrovsky N. (2004). Computational approach for prediction of Antimicrobial peptides in legumes-a framework for modelling, testing and applications. *Methods*, **34**, 436-443.

Bryksa, B. C., Y. Horimoto and R. Y. Yada (2010). Rational redesign of porcine pepsinogen containing an antimicrobial peptide. *Protein Engineering Design & Selection*, **23(9)**, 711-719.

Cristianini, N. and J. Shawe-Taylor (2000). *An Introduction to Support Vector Machines and other Kernel-based Learning methods*. Cambridge University Press.

Ding, C. H. Q. and I. Dubchak (2001). Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics,* **17**, 349–358.

Fjell, C. D., R. E. W. Hancock and A. Cherkasov (2007). AMPer : a database and an automated discovery tool for antimicrobial peptides. *Bioinformatics*, **23**, 1148-1155.

Hammami, R., J. B. Hamida, G. Vergoten and I. Fliss (2009). PhytAMP : a database dedicated to antimicrobial plant peptides. *Nucleic Acids Research*, **37**, D963–8.

Otvos, L. J. (2000). Antibacterial peptides isolated from insects. *Journal of Peptide Science*, **6**, 497-511.

Pestana-Calsa, M. C., I. L. Ribeiro and T. Calsa Jr. (2010). Bioinformatics-coupled molecular approaches for unravelling potential antimicrobial peptides coding genes in Brazilian native and crop plant species. *Current Protein and Peptide Science*, **11(3)**, 199-209.

Sonego, P., A. Kocsor and S. Pongor (2008). ROC analysis : application to the classification of biological sequences and 3D structures. *Briefings in Bioinformatics*, **9(3)**, 198-209.

Thomas, S., S. Karnik, R. S. Barai, V. K. Jayaraman and S. I. Thomas (2010). CAMP : a useful resource for research on antimicrobial peptides. *Nucleic Acids Research*, **38**, D774–D780.

Tossi, A. and L. Sandri (2002). Molecular diversity in Gene-Encoded, Cationic Antimicrobial Polypeptides. *Current Pharmaceutical Design*, **8**, 742-761.

Vapnik, V. N. and A. Y. Chervonenkis (1974). Theory of pattern recognition: Statistical problems of learning. Moscow: Nauka.

Wang, G., X. Li and Z. Wang (2009). APD2 : the updated antimicrobial peptide database and its application in peptide design. *Nucleic Acids Research*, **37**, D933-D937.

Zheng, X. L. and A. L. Zheng (2002). Genomic organization and regulation of three cecropin genes in *Anopheles gambiae*. *Insect Molecular Biology*, **11**, 517-525.