Computers and Electronics in Agriculture 111 (2015) 55–61

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Species specific approach to the development of web-based antimicrobial peptides prediction tool for cattle

Sarika [*], M.A. Iquebal, Vasu Arora, Anil Rai, Dinesh Kumar

Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, PUSA, New Delhi 110012, India

ARTICLE INFO

Article history:
Received 28 July 2014
Received in revised form 24 November 2014
Accepted 11 December 2014

Keywords:
Antimicrobial peptides
Cattle
Prediction
Support vector machine
Web-server

ABSTRACT

Antimicrobial peptides (AMPs) are the defence molecules of the host gaining extensive attention worldwide as these are natural alternative to chemical antibiotics. Machine learning techniques have capabilities to analyse large biological data for detection of hidden pattern in understanding complex underlying biological problems. Presently, development of resistance to chemical antibiotics in cattle is unsolved and growing problem which needs immediate attention. In the present study, attempt was made to apply machine learning algorithms such as Artificial Neuron Network (ANN) and Support Vector Machine (SVM). It was found that performance of SVM based models for *in silico* prediction/identification of AMPs of cattle is superior than ANN. A total of 99 AMPs related to cattle collected from various databases and published literature were taken for this study. N-terminus residues, C-terminus residues and full sequences were used for model development and identification/prediction. It was found that best SVM models in this case for C-terminus residues, N-terminus residues and full sequence were with kernels Radial Basis Function (RBF), Sigmoid and RBF with accuracy as 95%, 99% and 97%, respectively. These SVM models were implemented on web server and made available to users at http://cabin.iasri.res.in/amp/ for classification/prediction of novel AMPs of cattle. This computational server can accelerate novel AMP discovery from whole genome proteins of a given cattle species for bulk discovery with very high accuracy. This is the first successful attempt for development of species specific approach for prediction/classification of AMPs, which may be used further as a model in other species as well.

1. Introduction

Animal domestication is the oldest economic activity associated with human civilization. Livestock products, not only contributes to human nutrition through milk, meat and other livestock foods but also provides draught power, manure, employment, income, and export earnings. Domestication of cattle since Neolithic (8,000–10,000 years ago) era with subsequent spread of cattle throughout the world was intertwined with human migrations and trade (Willham, 1986). Currently, global cattle population is 1.5 billion, which is likely to increase to 2.6 billion by 2050 (FAO, 2012).

One of the major factors contributing to low animal productivity in the country are biotic and abiotic stresses apart from genetic factors. Cattle suffers from wide range of infectious diseases. Therefore, preventing measures are important for animal health. The best means to achieve this is by vaccination. Antimicrobial peptides (AMPs) play important role in host defence and is known as an essential part of innate immunity in response to microbial challenges. Macrophages, neutrophils, epithelial cells, haemocytes, fat body, reproductive tract, etc. are the various sources of AMPs in animals. In case of animals, both immune systems, i.e. innate and adaptive provides protection against spreading infection due to pathogen. The major advantages of AMPs in clinical application include their potential for broad-spectrum activity, rapid microbicidal activity and low propensity for resistance development (Marr et al., 2006). They also offer advantage of 'peptide promiscuity' showing an enormous multiplicity of biological activities, including activities such as antimicrobial, cytotoxic, insecticidal, uterotonic, antivirus, neurotensin antagonism, hemolytic and anthelmintic (Franco, 2011). AMPs are much more versatile in therapeutic applications viz., as single anti-infective agents; in combination with conventional antibiotics or antivirals to promote any additive or synergistic effects; as immune-stimulatory agents that enhance natural innate immunity, and also as endotoxin-neutralizing agents to prevent the potentially fatal complications associated with bacterial virulence factors that cause septic shock (Gordon et al., 2005; Franco, 2011). The major challenges in use of AMP are higher cost, limited stability (especially when composed of L-amino acids), and

[*] Corresponding author.
E-mail address: sarika@iasri.res.in (Sarika).

http://dx.doi.org/10.1016/j.compag.2014.12.008
0168-1699/© 2014 Elsevier B.V. All rights reserved.

unknown toxicology and pharmacokinetics. Now success are coming up with the development of stable, more cost-effective and potent broad-spectrum synthetic peptides in industries (Marr et al., 2006; Franco, 2011).

The peptidic group of bioactive molecules i.e. AMPs have been gaining attention world wide in research. These are the host defence molecules for innate immunity in response to microbial challenges (Otvos, 2000). AMPs play vital role as a natural antibiotic alternative of their chemical counterpart for protecting animals against diseases. These are present in both type organisms i.e. prokaryotic and eukaryotic. These AMP molecules can be further classified as cationic or anionic depending on net charge on them. 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). Some of the bovine AMPs with commercial use like Lactoferricin, Lactoferrampin, Alpha and Beta lactoglobulin derived peptides, casein derived peptides, lysozyme derived peptide (Jabbari et al., 2012) are well documented. A good bioinformatics resource has been reported in relation to AMPs (Sarika et al., 2012) like AMSDb (Tossi and Sandri, 2002), ANTIMIC (Zheng and Zheng, 2002), AMPer (Fjell et al., 2007), APD2 (Wang et al., 2009) and CAMP (Thomas et al., 2010).

Extensive literature is available related to antibacterial and antiviral peptides, describing their identification, characterization as well as mechanism of action. Unfortunately, antibacterial and antiviral peptides have no sequence homology, despite their common properties. Thus, it is difficult to develop techniques for predicting antibacterial and antiviral peptides based on homology. Moreover, experimental methods for identification and designing of antibacterial and antiviral peptides are resource intensive in terms of capital, time and manpower. Therefore, attempt was made to develop server for prediction of antimicrobial peptides. AntiBP2 (Snehlata et al., 2007) is the server that predicts antibacterial peptides. The prediction model of this server is very generic and developed considering all available antibacterial peptide sequences irrespective of organism. Since earlier approach are based on multispecies reference data, thus for any specific species, the prediction accuracy may not be accurate.

Since cattle has more than 30,000 genes but only 100 AMPs are reported in literature thus there is need to screen them *in silico* before evaluating them *in vitro* and *in vivo*. After excluding non-coding genes, *in vitro* or *in vivo* evaluation of more than 20,000 proteins is a great challenge in assay of AMP activity.

Though non-species specific AMP prediction servers are reported like ANTIBP (Snehlata et al., 2007), CAMP (Thomas et al., 2010) and CS-AMPred (Porto et al., 2012) but species specific approach has not been attempted so far. Thus, there is need to develop efficient computational tool for predicting antibacterial and antiviral peptides specific only for cattle, which could be used to design potent peptides against microbial pathogens. Therefore, in this study attempt has been made to develop prediction tool for antimicrobial peptides of cattle through *in silico* approach. Also estimated prediction/accuracy of the developed model has been obtained through cross validation technique. Therefore, this server will be quite useful in this process of protein evaluation and narrow down search of AMPs through lab experiments. Thus, *in silico* search for this server will be quite resource efficient.

2. Materials and methods

2.1. Data collection

The antimicrobial peptide sequences of bovidae family (cattle) were extracted from various specialized databases like AMSdb (Tossi and Sandri, 2002), SAPD (Wade and Englund, 2002), ANTIMIC (Brahmachary et al., 2004), AMPer (Fjell et al., 2007), APD2 (Wang et al., 2009) and CAMP (Thomas et al., 2010). Nearly two hundred peptide sequences were considered for this study. In order to build the SVM based model, we need to have non-antimicrobial peptides as control also. Since, no experimentally validated non antimicrobial source exists, thus peptide synthesized from mitochondria and other intracellular locations except the secretary proteins were considered as AMP which are mostly secreted outside the cell (Kumar et al., 2006). Eukaryotic mitochondrial organelle genome mimics prokaryotic genome features like common protein synthesis inhibitor and ribosome types. This is due to endosymbiont hypothesis endorsing prokaryotic origin of mitochondria during course of evolution (Martin and Mentel, 2010). Moreover, bovine AMP lactoferrin is known to have antimicrobial activities does not bind to mitochondrial proteins is well demonstrated in the species to be investigated (Richardson et al., 2009).

The extracted antimicrobial peptides were from different AMP family viz., Bactenecin, Lactoferricin, Defensin, Indolicidin, Seminalplasmin, Cathelicidin, Enkelytin, Casecidin, Vasostatin, Bactenecin, Cathelin, Melantropin, Aprotinin, Cascocidin, Lactoferricin, Proenkephalin, Casocidin and Apolipoprotein. The maximum number of data were extracted for "Defensin" family.

2.2. Pre-processing

In order to use these sequences for SVM based machine learning algorithm for training and testing, the biological sequences need to be converted in suitable feature for model building. In this study, each instance, i.e. biological sequence was denoted by a vector, having 31 attributes (or *features*), out of these, 20 representing Amino Acid Composition (AAC) for that instance and rest 11 features (*viz.* molecular weight, number of carbon atoms, number of hydrogen atoms, number of nitrogen atoms, number of oxygen atoms, number of sulfur atoms, theoretical pI, estimated halflife, instability index, aliphatic index, and grand average of hydropathicity (GRAVY) (Gasteiger et al., 2005) are the physico-chemical parameters for that sequence. These features were computed using bioperl scripts. AAC is a quantitative measure of the sequence that represents the sequence in terms of 20 values, one for each amino acid residue. For *i*th amino acid residue, AAC is defined as the percentage of *i*th residue in whole sequence. Mathematically,

$$\mathrm{AAC}_i = (N/N_i) * 100$$

where AAC_i is the Amino Acid Composition of *i*th amino acid residue, N_i is the Number of occurrences of *i*th amino acid residue in the sequence and N is the Total number of amino acid residue in the sequence.

AAC completely ignores the sequence order information and focuses only on the percentage amino acid residue content. Now, a matrix of order $N \times 31$ (here, N is 199) is obtained which is used as input for this analysis. The prediction target vector of two dimension comprises of binary class i.e. AMP or Non-AMP.

Separate models for N and C terminus were chosen because both termini contributes in AMP activity. C-terminus first interacts with the negatively charged membrane of the bacteria and penetrates (Park et al., 1998). The N-terminus also contributes in hampering crucial bacterial metabolic functions by interacting with intracellular components like DNA and RNA (Yonezawa et al., 1992). Due to this reason, the whole dataset was analyzed with three approaches, i.e. N-terminal residues, C-terminal residues and full sequence. For N-terminal and C-terminal, the available data were split with window size of 30 using PERL scripts, and redundancy was checked with CDHit (Li and Godzik, 2006) at 80%. We thoroughly checked each peptide in various available peptide prediction servers and selected total of 972 data for N-terminal and 970 data for C-terminal model fitting. The peptide size less than or equal to 30 was taken for training and testing since the average size of these sequences is generally around 30 (Wang et al., 2009). Among 31 features taken under study, only the significant features, selected on the basis of *p*-values at 5% level of significance were considered for further analysis. This was done based on Chi-Square values obtained using STATISTICA ver 6.0 software package (STATISTICA, 2001).

2.3. Support Vector Machine (SVM) technique

In order to build the prediction model, earlier, Artificial Neural Networks (ANNs) with back propagation algorithm have been used (Cheng and Titterington, 1994; Shukla et al., 2011), but it was found that it overfits and provides underestimation of actual prediction error specially in case of small sample size. Therefore, Support Vector Machine (SVM) technique developed by Vapnik (2000) was found to be quite reliable in case of small sample size due to it non-linear optimization property. It has attractive features and profound empirical performance for small sample, nonlinearity and high dimensional data application. It is based on Structural Risk Minimization (SRM) principle, which has been shown to be superior to traditional Empirical Risk Minimization (ERM) principle implemented in ANN models. Therefore, in this we used SVM for development of prediction models.

Support Vector Machine (SVM) is a nonparametric algorithm developed by Vapnik (2000). It is very promising and popular methodology for nonlinear classification in the field of supervised machine learning. It is proven to be very attractive to biological analysis due to their ability to handle noise and large input spaces (Brown et al., 2000; Ding and Dubchak, 2001).

Consider two-class classification problem and assume a set of samples, i.e. a series of input vectors $\boldsymbol{x}_i \in \Re^d, \quad (i = 1, 2, \ldots, N)$ with corresponding class levels $y_i \in \{+1, \ -1\}$, $(i = 1, \ 2, \ldots, N)$. For our study, +1 and −1 indicates two different classes and input vector dimension, i.e. N is 31. Main objective is to construct a binary classifier or derive a decision function from the available samples, which has a minimum probability of misclassifying a future sample. Further, Non-linear Support Vector Machine (NL-SVM) maps input vectors $\boldsymbol{x}_i \in \Re^d$ into a high dimensional feature, i.e. space $\phi(\boldsymbol{x}_i) \in H$ and constructs an Optimal Separating Hyperplane (OSH), which maximizes the margin, i.e., the distance between hyperplane and nearest data points of each class in the space H. Equation of a simple hyperplane is given by

$$y = \text{sign}[\boldsymbol{w}^T\boldsymbol{x} + \text{b}]$$

where $\boldsymbol{w}$ denotes a weight vector that can map the training data in the input space to the output space and b is the bias. In case, the data of the two classes are separable, it can be written as

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

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

$$y_i[\boldsymbol{w}^T\boldsymbol{x}_i + \text{b}] \geq 1 \quad i = 1, 2, \ldots, N$$

SVM formulations are done within a context of convex optimization theory. The primal form Quadratic Programming (QP) problem is obtained. After introducing Lagrangian with Lagrange multipliers $\alpha_i \geqslant 0$ for $i = 1, 2 \ldots, N$, the resulting linear classifier is

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

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 complementarily 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. Instead of using an arbitrary support vector $\boldsymbol{x}_i$, it is better to take an average over all the support vectors.

Some binary classification problems do not have simple hyperplane as a useful separating criterion. For those problems, there is a variant of mathematical approach that retains nearly all the simplicity of an SVM separating hyperplane. Let $\boldsymbol{x}$ be a vector in n dimensional input space and $\varphi(\cdot)$ 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 $\varphi(\cdot)$ is performed by kernel function $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, which defines an inner product in the space H. The decision function implemented by SVM is as follows:

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

where the coefficients α_i are obtained by solving the following convex quadratic programming problem:

$$\max_{\alpha} \quad -\frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i=1}^{N} \alpha_i$$

subject to $0 \leqslant \alpha_i \leqslant c$

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

Here, c is the regularization parameter that controls trade-off between margin and misclassification error. It is the learning parameter, where, larger values of c lead to few training errors and small values generate larger margin at the cost of more errors. The $\boldsymbol{x}_j$'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. $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T\boldsymbol{x}_j$ (Linear SVM)

b. $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\gamma\boldsymbol{x}_i^T\boldsymbol{x}_j + r)^d$ (Polynomial SVM of degree d)

c. $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left\{-\gamma\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right\}$ (Radial Basis function Kernel)

d. $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tan h(\gamma\boldsymbol{x}_i^T\boldsymbol{x}_j + r)$ (Sigmoid)

where r, d, $\gamma > 0$ 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.

Table 1
Feature selection with best predictors.

Features	Full sequence		C-terminal		N-terminal	
	Chi-square	*p*-value	Chi-square	*p*-value	Chi-square	*p*-value
Composition of Arginine	171.96	0.00	304.67	0.00	294.40	0.00
Gravy	149.46	0.00	489.58	0.00	483.96	0.00
Number of nitrogen atoms	142.90	0.00	295.45	0.00	266.63	0.00
Molecular weight	134.24	0.00	2.06	0.36	6.19	0.19
Number of hydrogen atoms	123.34	0.00	2.06	0.36	6.19	0.19
Theoretical pI	110.58	0.00	107.83	0.00	78.09	0.00
Number of carbon atoms	108.28	0.00	10.87	0.03	6.19	0.29
Composition of Tyrosine	105.30	0.00	47.79	0.00	42.91	0.00
Composition of Methionine	105.09	0.00	376.41	0.00	372.18	0.00
Composition of Cysteine	103.57	0.00	195.03	0.00	192.25	0.00
Number of sulphur atoms	98.99	0.00	71.46	0.00	70.59	0.00
Composition of Alanine	98.28	0.00	28.15	0.00	26.48	0.00
Composition of Lysine	93.07	0.00	218.96	0.00	228.65	0.00
Composition of Phenylalanine	92.98	0.00	83.16	0.00	83.68	0.00
Aliphatic_index	88.17	0.00	287.96	0.00	279.48	0.00
Composition of Asparagine	86.55	0.00	0.55	0.91	1.43	0.70
Composition of Glycine	85.97	0.00	23.33	0.00	19.07	0.00
Composition of Glutamine	85.46	0.00	117.85	0.00	115.66	0.00
Composition of Tryptophan	85.06	0.00	93.54	0.00	96.78	0.00
Composition of Leucine	84.36	0.00	171.93	0.00	158.12	0.00
Number of oxygen atoms	80.41	0.00	47.61	0.00	13.45	0.06
Composition of Serine	78.43	0.00	9.74	0.08	10.23	0.18
Composition of Isoleucine	75.80	0.00	207.81	0.00	218.53	0.00
Composition of Proline	68.96	0.00	39.25	0.00	39.87	0.00
Composition of Histidine	59.92	0.00	64.53	0.00	61.09	0.00
Composition of Threonine	59.44	0.00	114.47	0.00	131.94	0.00
Composition of Valine	52.59	0.00	14.73	0.01	25.56	0.00
Composition of Aspartic acid	31.34	0.00	121.00	0.00	116.36	0.00
Instability index	31.14	0.00	52.39	0.00	85.77	0.00
Half life	22.79	0.00	4.32	0.12	16.90	0.00
Composition of Glutamic acid	19.27	0.00	183.93	0.00	188.61	0.00

The variables having *p*-values >0.05 are not considered for model building.

2.4. Five-fold cross validation

All models obtained in this study were evaluated using five-fold cross-validation technique (Efron, 1983). In this case, dataset is randomly divided into five sets, each set containing around equal number of peptides. Four sets among five 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 test set. Average of five sets is finally considered.

2.5. Assessment of the prediction accuracy

The performance of each fitted model was assessed using test data. Several measures are available for the statistical estimation of the accuracy of these prediction models. The common statistical measures are Sensitivity, Specificity, Precision or Positive Predictive Value (PPV), Negative Predictive Value (NPV), False Positive Rate (FPR), False Discovery Rate (FDR), Accuracy and Mathew's correlation coefficient (MCC) and F1 score.

These measures are defined as follows:

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

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

$$\text{FPR} = \text{FP}/(\text{FP} + \text{TN}) \qquad \text{FDR} = \text{FP}/(\text{TP} + \text{FP}) = 1 - \text{PPV}$$

$$\text{F1} = 2\text{TP}/(2\text{TP} + \text{FP} + \text{FN})$$

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

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

where TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative.

3. Results and discussions

The entire data analysis is carried out using STATISTICA, Ver. 6.0 software package (STATISTICA, 2001). Table 1 shows the features selected for model development. For full-sequence model, all the features were selected for model building. For C-terminus residues, molecular weight, number of hydrogen atoms, composition of asparagine and half-life were discarded. For N-terminus residues, molecular weight, number of hydrogen atoms, carbon atoms, oxygen atoms, composition of asparagine and serine were not included in model development. In this study, we have developed classification models using ANN and SVM. While comparing, these two techniques, models based on SVM was found to be superior (Tables 2–4) which has also been reported in various other studies (Snehlata et al., 2007; Bhasin and Raghava, 2004). In case of ANN, before training, available observations are divided into two subsets: (i) first sub-set is training set comprising 70% observations, which is used for computing and updating the network weight and biases, and (ii) test set comprise the remaining 30% observations. The two most popular and widely used networks namely, MultiLayer Perceptron (MLP) and Radial Basis Function (RBF) are trained using all the three learning algorithms, *viz.* Gradient Descent Algorithm (GDA), *Broyden–Fletcher–Goldfarb–Shanno* (BFGS), and Conjugate Gradient Descent Algorithm (CGDA) with a view to minimizing sum of the squared error function of the network output. Several learning rates (Cheng and Titterington, 1994) are considered for training the networks as well as for adjusting the weights. A higher learning rate may converge more quickly, but may also exhibit greater instability. The present ANN model is having typically three-layer feed forward network viz., input, hidden and output layer. For our data, best result is obtained for 0.01 learning rate. For hidden units and output units, several activation

Table 2
Performance of the models for C-terminus residues using SVM and ANN methodology.

Models	C	γ	No. SVs	Sp	Sen	PPV	NPV	FPR	FDR	ACC	MCC	F1
SVM methodology												
Linear	2.00	–	140	0.95	0.93	0.95	0.94	0.05	0.05	0.94	0.88	0.94
Polynomial-2	9.75	0.05	330	0.95	0.93	0.94	0.93	0.05	0.06	0.94	0.87	0.94
Polynomial-3	6.25	0.05	167	0.91	0.94	0.91	0.94	0.09	0.09	0.93	0.85	0.93
RBF	**9.00**	**0.11**	**180**	**0.96**	**0.94**	**0.96**	**0.94**	**0.04**	**0.04**	**0.95**	**0.90**	**0.95**
Sigmoid	19.50	0.22	177	0.96	0.93	0.95	0.93	0.04	0.05	0.94	0.89	0.94
ANN methodology												
MLP 31-14-2	–	–	–	0.94	0.94	0.94	0.94	0.06	0.06	0.94	0.87	0.94

The model in bold is the best model for C-terminus residues and implemented at the back-end.

Table 3
Performance of the models for N-terminus residues using SVM and ANN methodology.

Models	*c*	γ	No. of SVs	Sp	Sen	PPV	NPV	FPR	FDR	ACC	MCC	F1
SVM methodology												
Linear	17.50	–	123	0.95	0.95	0.95	0.95	0.05	0.05	0.95	0.90	0.95
Polynomial-2	9.25	0.08	176	0.95	0.94	0.95	0.95	0.05	0.05	0.95	0.90	0.95
Polynomial-3	19.25	0.08	417	0.93	0.94	0.93	0.94	0.07	0.07	0.93	0.87	0.93
RBF	12.00	0.2	172	0.97	0.98	0.97	0.98	0.03	0.03	0.97	0.95	0.97
Sigmoid	**18.00**	**0.11**	**240**	**0.98**	**0.99**	**0.98**	**0.99**	**0.02**	**0.02**	**0.99**	**0.98**	**0.99**
ANN methodology												
MLP 31-19-2	–	–	–	0.94	0.94	0.94	0.94	0.06	0.06	0.94	0.88	0.94

The model in bold is the best model for N-terminus residues and implemented at the back-end.

Table 4
Performance of the models for full sequence using SVM and ANN methodology.

Models	*c*	γ	No. of SVs	Sp	Sen	PPV	NPV	FPR	FDR	ACC	MCC	F1
SVM methodology												
Linear	2.00	–	17	0.95	0.95	0.95	0.95	0.05	0.05	0.95	0.90	0.95
Polynomial-2	2.50	0.05	133	0.95	0.96	0.95	0.96	0.05	0.05	0.95	0.91	0.95
Polynomial-3	17.75	0.05	129	0.96	0.96	0.96	0.96	0.04	0.04	0.96	0.92	0.96
RBF	**20.00**	**0.05**	**20**	**0.97**	**0.97**	**0.97**	**0.97**	**0.03**	**0.03**	**0.97**	**0.94**	**0.97**
Sigmoid	1.50	0.31	81	0.97	0.96	0.97	0.96	0.03	0.03	0.96	0.93	0.96
ANN methodology												
MLP 31-16-2	–	–	–	0.93	0.92	0.93	0.92	0.07	0.07	0.92	0.85	0.92

The model in bold is the best model for full sequence and implemented at the back-end.

functions, *viz.* Identity, tanh, logistic, exponential and sine are tried. Performance of the trained network is assessed by computing different measures as mentioned in Section 2.5 on the training and test sets. As described in Section 2.2, the whole analysis is done for three types *viz.* C-terminus residues, N-terminus residues and full sequence.

The models MLP 31-14-2, MLP 31-19-2 and MLP 31-16-2 with accuracy 0.94, 0.94 and 0.92 were found to be best for C-terminus residues, N-terminus residues and full sequence, respectively. In case of C-terminus residues, training algorithm BFGS, entropy error function, exponential activation function for hidden layer and for output layer softmax function were found to be best. For N-terminus residues, activation function was exponential, while for full sequence, it was Tanh. Training algorithm and error functions for all the three were BFGS and entropy respectively whereas output layer was Softmax. The best models were selected on the basis of measures discussed in Section 2.5 and the same are reported in Tables 2–4 respectively for C-terminus residues, N-terminus residues and full sequence.

Subsequently, SVM models using all kernel functions, viz. Linear, Polynomial of degree 2, Polynomial of degree 3, RBF, and Sigmoid function have been developed with a stopping criteria as maximum number of iterations as or stop if error is less than 0.001. Further, 5-fold cross validation is applied here. For C-terminus residues, SVM model with RBF kernel function was found to be best with c parameter as 9.00 with 180 number of support vectors. Sensitivity, specificity, PPV, NPV, FPR, FDR, accuracy, MCC and F1 score were found to be 0.96, 0.94, 0.96, 0.94, 0.04, 0.04, 0.95, 0.90 and 0.95 respectively (Table 2). Table 3 shows the results tabulated for N-terminus residues. For N-terminus residues, SVM model with sigmoid kernel function was found to be best with c parameter as 18.00 and 240 number of support vectors. Sensitivity, Specificity, PPV, NPV, FPR, FDR, accuracy, MCC and F1 score were found to be 0.98, 0.99, 0.98, 0.99, 0.02, 0.02, 0.99, 0.98 and 0.99, respectively. Table 4 shows SVM models based on full sequence. Here, again, SVM with kernel function RBF was found to be best with c parameter as 20.00 and 20 number of support vectors. Sensitivity, Specificity, PPV, NPV, FPR, FDR, accuracy, MCC and F1 score were found to be 0.97, 0.97, 0.97, 0.97, 0.03, 0.03, 0.97, 0.94 and 0.97, respectively. The receiver-operating characteristic (ROC) curves obtained for each kernel function for C-terminal, N-terminal and full sequence models are represented in Fig. 1.

The compromised accuracy of ABP2 (92.14% by SVM) (Snehlata et al., 2007) and CAMP (93.2% by Random Forest approach and 91.5% by SVM) (Thomas et al., 2010) as compared to our server, which gave the predicted accuracy of 97% was seen. This speculates that during course of evolution, the AMP diversification might be on specialized species specific parameters which might have led to higher accuracy in our finding.

The performances of different peptide prediction models developed using machine learning techniques were compared. It was concluded that for classification and prediction of AMP of cattle using SVM methodology performs superior as compared to ANN methodology. Therefore, the SVM model was implemented in the webserver.

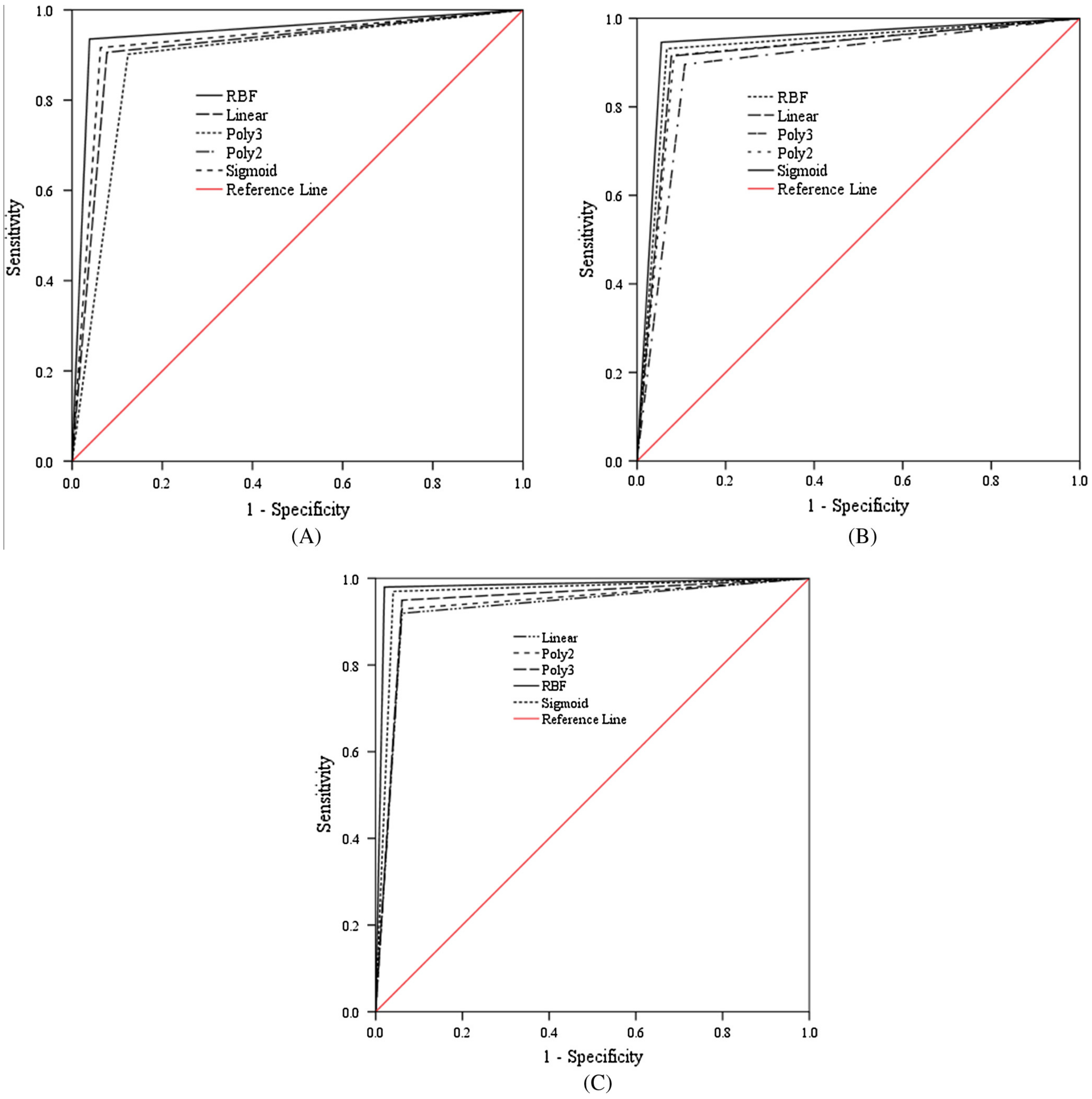

Fig. 1. ROC curves for (A) C-terminal, (B) N-terminal and (C) full sequence models.

3.1. Web implementation

The best models were implemented for N-terminus residues, C-terminus residues and full sequence and is made available at http://cabin.iasri.res.in/amp/. The server is developed using CGI-Perl script, Hyper Text Markup Language (HTML) and Java Scripts to make it more user-friendly. This is launched using open source web server software program, Apache. The user needs to submit the peptide sequence either by pasting in the text box or uploading through "upload" button. The user can select from the three options *viz.* N-terminus residues, C-terminus residues and full sequence using radio-button. The webserver has six tabs, viz. Home, Algorithm, Submission, Links, Tutorial and Team. The "Tutorial" has the full guide to use the server.

4. Conclusion

This is the first successful attempt to develop species specific approach for AMP prediction. In this study, SVM methodology which is used for nonlinear classification is described and implemented through web (http://cabin.iasri.res.in/amp/) for users. It is freely available tool, which is very cost and time effective for prediction of unknown peptides with prediction accuracy up to 97%. Computational prediction is an important immunoinformatic technology supporting the determination of AMPs. Antimicrobial peptides are potent and effective molecules responsible for innate immune response of eukaryotes. They are widely distributed in both plants and animals, but most dominantly present in animals. The application of AMPs have shown very promising results in production of animal/plant and agricultural produce and therefore the latest tools and technology for production of AMPs with specific activity and wide microbe range of action can be effectively utilized to develop genetically modified disease resistance varieties/breeds through biotechnology and genetic engineering. This first report on species specific tool can be used to decipher large number of putative AMPs over available genomes of cattle having more than 20,000 proteins. The future bulk discovery of AMP by this approach needs further wet lab validation before practicing/applying them for therapeutics and industrial application.

Availability: This application can be freely accessible to non-commercial users at http://cabin.iasri.res.in/amp/.

Acknowledgements

We are thankful to Director, IASRI, Indian Council of Agricultural Research, Ministry of Agriculture, Government of India, India for providing financial and infrastructural support to carry out this study under project (Number SIX1121).

References

Bhasin, M., Raghava, G.P.S., 2004. Prediction of CTL epitopes using QM, SVM and ANN techniques. Vaccine 22 (23–24), 3195–3204.

Brahmachary, M., Krishnan, S.P., Koh, J.L., Khan, A.M., Seah, S.H., Tan, T.W., et al., 2004. ANTIMIC: a database of antimicrobial sequences. Nucl. Acids Res. 32 (1), D586–D589 (Database issue).

Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., et al., 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl. Acad. Sci. 97, 262–267.

Cheng, B., Titterington, D.M., 1994. Neural networks: a review from a statistical perspective. Stat. Sci. 9, 2–54.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and other Kernel-based Learning Methods. Cambridge University Press, U.K.

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

Efron, B., 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. J. Am. Stat. Assoc. 78, 316–331.

FAO, 2012. The state of the world's animal genetics resources for food and agriculture, Rome, Italy. <http://wwwfaoorg/docrep/010/a1250e/a1250e00htm>.

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

Franco, O.L., 2011. Peptide promiscuity: an evolutionary concept for plant defense. FEBS Lett. 585 (7), 995–1000.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., et al., 2005. Protein identification and analysis tools on the ExPASy server. In: John, M., Walker, J. (Eds.), The Proteomics Protocols Handbook. Humana Press, pp. 571–607.

Gordon, Y.J., Romanowski, E.G., McDermott, A.M., 2005. A review of antimicrobial peptides and their therapeutic potential as anti-infective drugs. Curr. Eye Res. 30 (7), 505–515.

Jabbari, S., Hasani, R., Kafilzadeh, F., Janfeshan, S., 2012. Antimicrobial peptides from milk proteins: a prospectus. Annal. Biol. Res. 3 (11), 5313–5318.

Kumar, M., Verma, R., Raghava, G.P.S., 2006. Prediction of mitochondrial proteins using support vector machine and hidden Markov model. J. Biol. Chem. 281 (9), 5357–5363.

Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659.

Marr, A.K., Gooderham, W.J., Hancock, R.E.W., 2006. Antibacterial peptides for therapeutic use: obstacles and realistic outlook. Curr. Opin. Pharmacol. 6, 468–472.

Martin, W., Mentel, M., 2010. The origin of mitochondria. Nat. Ed. 3 (9), 58.

Otvos, L.J., 2000. Antibacterial peptides isolated from insects. J. Pept. Sci. 6, 497–511.

Park, C.B., Kim, H.S., Kim, S.C., 1998. Mechanism of action of the antimicrobial peptide buforin II: buforin II kills microorganisms by penetrating the cell membrane and inhibiting cellular functions. Biochem. Biophys. Res. Commun. 244, 253–257.

Pestana-Calsa, M.C., Ribeiro, I.L., Calsa, J.T., 2010. Bioinformatics-coupled molecular approaches for unravelling potential antimicrobial peptides coding genes in Brazilian native and crop plant species. Curr. Protein. Pept. Sci. 11, 199–209.

Porto, W.F., Pires, A.S., Franco, O.L., 2012. CS-AMPPred: an updated SVM model for antimicrobial activity prediction in cysteine-stabilized peptides. PLoS ONE 7 (12), e51444.

Richardson, A., deAntueno, R., Duncan, R., Hoskin, D.W., 2009. Intracellular delivery of bovine lactoferricin's antimicrobial core (RRWQWR) kills T-leukemia cells. Biochem. Biophys. Res. Commun. 388, 736–741.

Sarika, Iquebal, M.A., Rai, A., 2012. Biotic stress resistance in agriculture through antimicrobial peptides. Peptides 36, 322–330.

Shukla, R.P., Tripathi, K.C., Pandey, A.C., Das, I.M.L., 2011. Prediction of Indian summer monsoon rainfall using Niño indices: a neural network approach. Atmos. Res. 102, 99–109.

Snehlata, Sharma, B.K., Raghava, G.P.S., 2007. Analysis and prediction of antibacterial peptides. BMC Bioinform. 8, 263.

StatSoft, Inc. 2001. STATISTICA (Data Analysis Software System). Version 6.0. <www.statsoft.com>.

Thomas, S., Karnik, S., Barai, R.S., Jayaraman, V.K., Thomas, S.I., 2010. CAMP: a useful resource for research on antimicrobial peptides. Nucl. Acids Res. 38, D774–D780 (Database issue).

Tossi, A., Sandri, L., 2002. Molecular diversity in gene-encoded, cationic antimicrobial polypeptides. Curr. Pharm. Des. 8, 742–761.

Vapnik, V. (Ed.), 2000. The Nature of Statistical Learning Theory. Springer-Verlag Press, New York.

Wade, D., Englund, J., 2002. Synthetic antibiotic peptides database. Protein Pept. Lett. 9, 53–57.

Wang, G., Li, X., Wang, Z., 2009. APD2: the updated antimicrobial peptide database and its application in peptide design. Nucl. Acids Res. 37, D933–D937.

Willham, R., 1986. From husbandry to science: a highly significant facet of our livestock heritage. J. Anim. Sci. 62, 1742–1758.

Yonezawa, A., Kuwahara, J., Fujii, N., Sugiura, Y., 1992. Binding of tachyplesin I to DNA revealed by footprinting analysis: significant contribution of secondary structure to DNA binding and implication for biological action. Biochemistry 31, 2998–3004.

Zheng, X.L., Zheng, A.L., 2002. Genomic organization and regulation of three cecropin genes in Anopheles gambiae. Insect Mol. Biol. 11, 517–525.