Int. J. Agricult. Stat. Sci. Vol. 10, Supplement 1, pp. 275-281, 2014 *ISSN : 0973-1903*

ORIGINAL ARTICLE

CLASSIFICATION AND IDENTIFICATION OF CATTLE ANTIMICROBIAL PEPTIDES USING ARTIFICIAL NEURAL NETWORK METHODOLOGY

M. A. Iquebal*, Sarika and Anil Rai

Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, Pusa, New Delhi - 110 012, India.
E-mail : jiqubal@gmail.com

Abstract : With the advent of machine learning techniques, a large number of biological problems have been given a solution. The interpretation of massive genomic data is a big challenge to the researchers, but literature shows many computational approaches to counter such problems. Out of many biological issues, one is regarding the Antimicrobial peptides (AMPs), which are the hosts' defence molecules gaining extensive research attention worldwide. Today, resistance to chemical antibiotics is an unsolved and growing problem. AMPs may be a natural alternative to chemical antibiotics and a potential area of research under applied biotechnology. The present work shows application of Artificial Neural Networks (ANN), a machine learning algorithm on bovine AMPs for prediction. Total of 99 AMPs related to cattle collected from various databases and published literature were taken into study. N-terminal residues, C-terminal residues and full sequences were used for model development and identification (prediction). For N-terminal residues, MultiLayer Perceptron (MLP 31-19-2) was found to be the best model with accuracy 94% while for C-terminal residues and full sequence, MLP 31-14-2 and MLP 31-16-2 were the best models with accuracies 94% and 92%, respectively for classification of bovine AMPs. The computational approach for AMPs identification from this study may be used to design potent peptides against microbial pathogens.

Key words : Antimicrobial peptides, Bovine, Genomics, Artificial Neural Network, Accuracy.

1. Introduction

Integration of domesticated animal with cropping system in agriculture decades ago has led to a revolution, which is emerging as an important growth leverage of the economy. Livestock renders essential food products *viz*., milk and meat, draught power, manure, employment, income and export earnings. Cattle domestication initiated sometime in the Neolithic (8,000-10,000 years ago) with subsequent spread of cattle throughout the world is intertwined with human migrations and trade [Willham (1986)]. At present, more than 1.5 billion cattle are reported, which is liable to expand to 2.6 billion by 2050, as per Food and Agriculture Organization [FAO (2012)]. India covers less than 3% of the world's total land area but sustains about 57% of the world's buffalo population, 16% of the cattle population and 20% of goat population. After the cattle genome sequencing, UMD 3.1 assembly is the third release of *Bos Taurus* assembly from CBCB, University of Maryland and enables to more understanding of mammalian evolution and accelerating livestock genetic improvement for milk and meat production.

In the era of genomics, the various machine learning approaches is applied for some relevant conclusion of the biological processes. The interpretation of massive genomic data is a big challenge to the researchers, but literature shows many computational approaches [Stanke and Waack (2003), Brusic *et al*. (1998), Peters *et al*. (2003), Saha *et al*. (2007), Ansari *et al*. (2010)] to counter such problems. In this regards, attention has been given to a peptidic group of bioactive molecules known as antimicrobial peptides (AMP). These are the hosts' defence molecules and an essential part of innate immunity in response to microbial challenges [Otvos (2000)]. 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)].

***Author for correspondence** **Received** June 24, 2014 **Revised** September 18, 2014 **Accepted** October 12, 2014*

These peptides are known to be alternative to chemical antibiotics to overcome the problem of resistance against pathogens, hence termed as "natural antibiotics". AMPs have their applicability in bioengineering and are used as a biotechnological tool for creating transgenic agricultural crops, biofuels etc [Bryksa *et al*. (2010)]. A good bioinformatics resource has been reported in relation to AMPs like 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.

2. Materials and Methods

2.1 Extraction of AMPs in legumes

The antimicrobial peptide sequences were extracted from various specialized databases like AMSdb, SAPD, APD2, CAMP, ANTIMIC, AMPer etc. Approximately two hundred peptide sequences were considered for study for analysis purpose. These peptides belonged to two major classes of antimicrobial and non-antimicrobial peptides. Since, no experimentally validated non antimicrobial source exist, the peptide synthesized from mitochondria and other intracellular locations except the secretary proteins were considered as AMP are mostly secreted outside the cell [Kumar *et al*. (2006)].

2.2 Pre-processing of the sequences

Before using ANN algorithm for training and testing, the biological sequences need to be converted to format suitable for input to computer system. For this study, each instance was denoted by a vector, having 31 attributes (or *features*), 20 representing the 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 sulphur atoms, theoretical pI, estimated halflife, instability index, aliphatic index and grand average of hydropathicity (GRAVY)] are the physico-chemical parameters for that instance. These eleven features were computed using bioperl scripts (as in Annexure 1). *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,

Annexure 1 : Bioperl script for computing physic chemical parameters of peptides under study.

```
#!/bin/perl -w
use Bio::Seq;
use Bio::DB::GenBank;
use Bio::Tools::Protparam;
use Bio::SeqIO;
my $seqio_obj = Bio::SeqIO->new(-file => $ARGV[0], -format
    => "fasta" );
open (OUT,">$ARGV[0]-OUT");
open (OUT1,">res");
    "ID##",
    "Amino acid number##",
    "Number of negative amino acids##",
    "Number of positive amino acids##",
    "Molecular weight##",
    "Theoretical pI##",
    "Total number of atoms##",
    "Number of carbon atoms##",
    "Number of hydrogen atoms##",
    "Number of nitrogen atoms##",
    "Number of oxygen atoms##",
    "Number of sulphur atoms##",
    "Half life##",
    "Instability Index##",
    "Stability class##",
    "Aliphatic_index##",
    "Gravy##",
    print OUT1 "\n";

while( my $seq_obj = $seqio_obj->next_seq ) {
    my $pp=Bio::Tools::Protparam->new(seq=>$seq_obj->seq);
    $seq_obj->display_id,"##",
    $pp->amino_acid_number(),"##",
    $pp->num_neg(),"##",
    $pp->num_pos(),"##",
    $pp->molecular_weight(),"##",
    $pp->theoretical_pI(),"##",
    $pp->total_atoms(),"##",
    $pp->num_carbon(),"##",
    $pp->num_hydrogen(),"##",
    $pp->num_nitro(),"##",
    $pp->num_oxygen(),"##",
    $pp->num_sulphur(),"##",
    $pp->half_life(),"##",
    $pp->instability_index(),"##",
    $pp->stability(),"##",
    $pp->aliphatic_index(),"##",
    $pp->gravy(),"##",
print OUT "\n";
}
```

$$AAC_t = (N_t / N) \times 100$$

Where, AAC_t = ACC of i*th* amino acid residue.

N_t = Number of occurrences of i*th* 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 $N \times 31$ (here, N is 199) 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 Artificial Neural Network (ANN)

Artificial Neural Network (ANN) is a powerful machine learning technique commonly used in the field of bioinformatics. In this study, ANN was applied for prediction of antibacterial peptides. ANN have been developed as generalizations of mathematical models of biological nervous systems. The basic processing elements of neural networks are called artificial neurons, or simply neurons or nodes. In a simplified mathematical model of the neuron, the effects of the synapses are represented by connection weights that modulate the effect of the associated input signals and the nonlinear characteristic exhibited by neurons is represented by a transfer function. The neuron impulse is then computed as the weighted sum of the input signals, transformed by the transfer function. The learning capability of an artificial neuron is achieved by adjusting the weights in accordance with the chosen learning algorithm [Haykin (1994)].

Activation function

Every neuron model consists of a processing element with synaptic input connections and a single output. The signal flow of neuron inputs, x_j is unidirectional. Fig. 1 illustrates the typical artificial neural network and the neuron output signal "O" given by the following relationship

$$O = f(net) = \left(\sum_{j=1}^{n} w_j x_j \right)$$

Where, w_j is the weight vector and the function *f(net)* is referred to as an activation (transfer) function.

The most important unit in neural network structure is their net inputs by using a scalar-to-scalar function called "the activation function or threshold function or transfer function", output a result value called the "unit's activation".

The variable *net* is defined as a scalar product of the weight and input vectors

$$net = w^T x = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n$$

Where, T is the transpose of a matrix and in the simplest case, the output value O is computed as

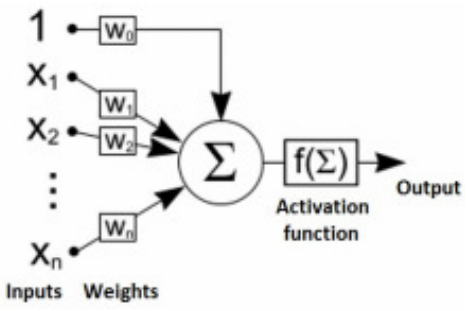

Fig. 1 : Architecture of an artificial neural network.

$$O = f(net) = \begin{cases} 1; & if\ w^T x \geq \theta \\ 0; & otherwise \end{cases}$$

Where, θ is called the threshold level and this type of node is called a linear threshold unit.

2.4 Five-fold cross validation

All models were evaluated using five-fold cross-validation technique in the study. 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 testing. Average of five sets is finally considered.

2.5 Assessment of the prediction accuracy

After model fitting, the performance needs to be assessed. Computational models that are valid, relevant, and properly assessed for accuracy is used for planning of complementary laboratory experiments. In the study, 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*), False Positive Rate (FPR), False Discovery Rate (FDR), Accuracy and Mathew's correlation coefficient (*MCC*) and F1 score.

These measures are defined as follows

$Sensitivity = TP / (TP + FN) * 100$

$Specificity = TN / (FP + TN) * 100$

$PPV = TP / (TP + FP) * 100$

$NPV = TN / (TN + FN) * 100$

$FPR = FP / (FP + TN) = 1 - Specificity$

$FDR = FP / (TP + FP) = 1 - PPV$

$$F_1 = 2TP/(2TP + FP + FN)$$

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

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

Where, *TP* = True Positive, *TN* = True Negative *FP* = False Positive and *FN* = False Negative

3. Results and Discussion

Prior *in silico* approaches help to get an idea of the AMP coding potentials of animal species under study, though it requires further biological validation. In our study, total of 98 antimicrobial peptides of cattle/ bovidae family belonging to following family of AMPs were extracted: Bactenecin, Lactoferricin, Defensin, Indolicidin, seminalplasmin, Cathelicidin, Enkelytin, casecidin, vasostatin, bactenecin, cathelin, melantropin, aprotinin, cascocidin, lactoferecin, proenlphlin, casocidin and apolipoprotein. The maximum number of data was extracted for "Defensin" family. Fig. 2 represents the main classes of collected antimicrobial peptide sequences from cattle along with their percentage contribution. *In silico* studies of these AMPs help to unravel the functional aspects of peptides.

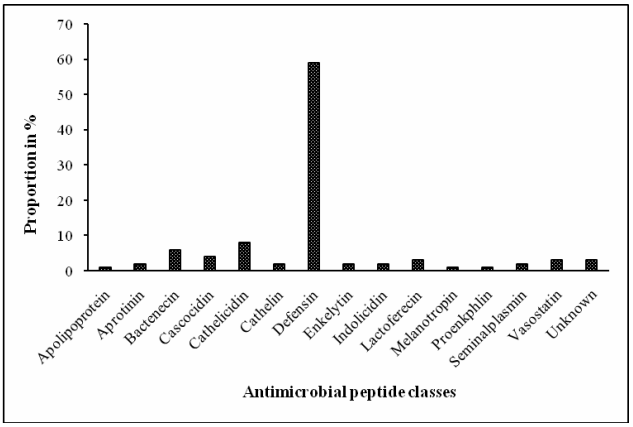

Fig. 2 : Class wise distribution of cattle AMPs.

Classification model was developed using ANN. Total of 199 peptide sequences comprising of 99 from antimicrobial class and 100 from non-antimicrobial class of cattle were considered here. These were pre-processed and converted quantitative required as input for ANN methodology for further analysis. Pre-processing of the sequences information and calculation of amino acid composition (*AAC*) was done in PERL scripts. Besides the amino acid composition, other

Table 1 : Artificial neural network models for identification of AMPsusing C-terminal residues.

Models	Sp/TNR	Sen/TPR	PPV	NPV	FPR	FDR	ACC	MCC	F1	Training Algorithm	Error Function	Activation Function	Output layer
MLP 31-14-2	**0.94**	**0.94**	**0.94**	**0.94**	**0.06**	**0.06**	**0.94**	**0.87**	**0.94**	**BFGS 38**	**Entropy**	**Exponential**	**Softmax**
MLP 31-6-2	0.94	0.92	0.93	0.92	0.06	0.07	0.93	0.86	0.93	BFGS 66	SOS	Tanh	Sine
MLP 31-5-2	0.94	0.91	0.93	0.91	0.06	0.07	0.92	0.85	0.92	BFGS 56	SOS	Tanh	Exponential
MLP 31-15-2	0.92	0.92	0.92	0.92	0.08	0.08	0.92	0.84	0.92	BFGS 51	SOS	Logistic	Sine
MLP 31-20-2	0.93	0.93	0.93	0.93	0.07	0.07	0.93	0.86	0.93	BFGS 55	SOS	Exponential	Identity
MLP 31-18-2	0.93	0.92	0.93	0.92	0.07	0.07	0.92	0.85	0.92	BFGS 69	SOS	Logistic	Sine
MLP 31-21-2	0.93	0.91	0.93	0.92	0.07	0.07	0.92	0.85	0.92	BFGS 52	SOS	Logistic	Tanh
MLP 31-17-2	0.92	0.93	0.92	0.93	0.08	0.08	0.93	0.85	0.93	BFGS 18	Entropy	Tanh	Softmax
MLP 31-17-2	0.93	0.93	0.93	0.93	0.07	0.07	0.93	0.86	0.93	BFGS 50	Entropy	Logistic	Softmax
MLP 31-13-2	0.94	0.93	0.94	0.93	0.06	0.06	0.94	0.87	0.93	BFGS 63	SOS	Logistic	Sine

Table 2 : Artificial neural network models for identification of AMPsusing N-terminal residues.

Models	Sp/TNR	Sen/TPR	PPV	NPV	FPR	FDR	ACC	MCC	F1	Training Algorithm	Error Function	Activation Function	Output layer
MLP 31-19-2	**0.94**	**0.94**	**0.94**	**0.94**	**0.06**	**0.06**	**0.94**	**0.88**	**0.94**	**BFGS 88**	**Entropy**	**Exponential**	**Softmax**
MLP 31-14-2	0.93	0.93	0.93	0.93	0.07	0.07	0.93	0.86	0.93	BFGS 33	Entropy	Exponential	Softmax
MLP 31-14-2	0.93	0.92	0.93	0.93	0.07	0.07	0.93	0.86	0.93	BFGS 103	SOS	Exponential	Identity
MLP 31-16-2	0.93	0.91	0.93	0.91	0.07	0.07	0.92	0.84	0.92	BFGS 25	Entropy	Logistic	Softmax
MLP 31-21-2	0.94	0.94	0.94	0.94	0.06	0.06	0.94	0.88	0.94	BFGS 45	Entropy	Tanh	Softmax
MLP 31-15-2	0.94	0.94	0.94	0.94	0.06	0.06	0.94	0.88	0.94	BFGS 63	Entropy	Tanh	Softmax
MLP 31-23-2	0.94	0.94	0.94	0.94	0.06	0.06	0.94	0.88	0.94	BFGS 34	Entropy	Tanh	Softmax
MLP 31-16-2	0.93	0.93	0.93	0.93	0.07	0.07	0.93	0.86	0.93	BFGS 30	Entropy	Exponential	Softmax
MLP 31-24-2	0.94	0.93	0.94	0.93	0.06	0.06	0.94	0.87	0.93	BFGS 57	Entropy	Tanh	Softmax
MLP 31-19-2	0.94	0.93	0.94	0.93	0.06	0.06	0.94	0.87	0.93	BFGS 45	Entropy	Logistic	Softmax

Table 3 : Artificial neural network models for identification of AMPsusing full sequence.

Model	Sp/TNR	Sen/TPR	PPV	NPV	FPR	FDR	ACC	MCC	F1	Training Algorithm	Error Function	Activation Function	Output layer
MLP 31-16-2	**0.93**	**0.92**	**0.93**	**0.92**	**0.07**	**0.07**	**0.92**	**0.85**	**0.92**	**BFGS 6**	**Entropy**	**Tanh**	**Softmax**
MLP 31-12-2	0.92	0.90	0.92	0.90	0.08	0.08	0.91	0.82	0.91	BFGS 2	Entropy	Identity	Softmax
MLP 31-16-2	0.92	0.89	0.92	0.89	0.08	0.08	0.90	0.81	0.90	BFGS 7	SOS	Exponential	Exponential
MLP 31-11-2	0.92	0.92	0.92	0.92	0.08	0.08	0.92	0.84	0.92	BFGS 4	SOS	Identity	Logistic
MLP 31-23-2	0.92	0.92	0.92	0.92	0.08	0.08	0.92	0.84	0.92	BFGS 28	Entropy	Tanh	Softmax
RBF 31-23-2	0.92	0.76	0.90	0.79	0.08	0.10	0.84	0.69	0.82	RBFT	Entropy	Gaussian	Softmax
MLP 31-9-2	0.92	0.92	0.92	0.92	0.08	0.08	0.92	0.84	0.92	BFGS 5	SOS	Identity	Logistic
RBF 31-24-2	0.86	0.90	0.86	0.89	0.14	0.14	0.88	0.76	0.88	RBFT	Entropy	Gaussian	Softmax
MLP 31-17-2	0.92	0.92	0.92	0.92	0.08	0.08	0.92	0.84	0.92	BFGS 54	Entropy	Exponential	Softmax
MLP 31-18-2	0.92	0.90	0.92	0.90	0.08	0.08	0.91	0.82	0.91	BFGS 21	SOS	Tanh	Exponential

physico-chemical properties were considered like molecular weight, number of carbon atoms, number of hydrogen atoms, number of nitrogen atoms, number of oxygen atoms, number of sulphur atoms, theoretical pI, estimated halflife, instability index, aliphatic index, and grand average of hydropathicity (GRAVY). Each instance was denoted by a vector, having 31 attributes (or *features*) representing the amino acid composition (*AAC*) and other 11 physico-chemical parameters considered for that instance. This consists of series of input vectors $\mathbf{x}_i \in \Re^d$ (i = 1, 2, ..., N). Hence, a matrix of order 199 x 31 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 $\mathbf{y}_i$ having +1 and -1 as values. Approximately 70% of total data was used for training purpose (model development) and remaining 30% for testing (model validation) purpose. All the analyses for obtaining the classification models using ANN have been done in STATISTICA 8.0.

For AMPs, N-terminals play important role in bacteria-specific interaction process while C-terminus is responsible for membrane interaction and pore formation. For this reason, the whole dataset was analysed with three approaches, *i.e.* N-terminal residues, C-terminal residues and full sequence. Various ANN models were tried for N-terminal residues, C-terminal residues and full sequence. It was observed that for C-terminal residues MLP 31-14-2 was the best model with specificity, sensitivity, PPV, NPV, FPR, FDR, accuracy, MCC and F1 score as 0.94, 0.94, 0.94, 0.94, 0.06, 0.06, 0.94, 0.87 and 0.94 respectively. The training algorithm was Broyden-Fletcher-Goldfarb-Shanno (BGFS) 38 with entropy error function, activation function as exponential and softmax output layer.

Similarly for N-terminal residues, MLP 31-19-2 was the best model with specificity, sensitivity, PPV, NPV, FPR, FDR, accuracy, MCC and F1 score as 0.94, 0.94, 0.94, 0.94, 0.06, 0.06, 0.94, 0.88 and 0.94 respectively. The training algorithm was BFGS 88 with eror function as entropy, activation function as exponential and softmax output layer.

For the full sequence, best model was found to be MLP 31-16-2 with specificity, sensitivity, PPV, NPV, FPR, FDR, accuracy, MCC and F1 score as 0.93, 0.92, 0.93, 0.92, 0.07, 0.0.7, 0.92, 0.85 and 0.92 respectively. The training algorithm was BFGS 6 with error function as entropy, activation function as Tanh and softmax output layer.

The performance measures (sensitivity, specificity, PPV, NPV, FPR, FDR, accuracy, MCC and F1 score) for N-terminal, C-terminal and full-sequence were obtained and results were presented in Tables 1, 2 and 3, respectively for 5-fold cross validation.

4. Conclusion

Computational prediction is an important immunoinformatic technology supporting the determination of AMPs. For N-terminal residues, MLP 31-19-2 was found to be the best model, while for C-terminal residues and full sequence, MLP 31-14-2 and MLP 31-16-2 were the best models respectively for classification of bovine AMPs. The parameters were also further fine-tuned to achieve the best performance in terms of misclassification error. This developed model may further be used for identification of antimicrobial peptides from candidate peptides. The current prediction method can be a useful tool for the systematic analysis of bovine AMP 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 domestic animals.

Acknowledgements

Authors are highly thankful to the Editor and learned reviewers for their constructive comments to improve the earlier draft of this manuscript.

References

Ansari, H. R. and G. P. S. Raghava (2010). Identification of conformational B-cell Epitopes in an antigen from its primary sequence. *Immunome Research*, **6**, 6.

Brusic, V., G. Rudy and M. Honeyman (1998). Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics*, **14**, 121–130.

Bryksa, B. C., Y. Horimoto and R. Y. Yada (2010). Rational redesign of porcine pepsinogen containing an antimicrobial peptide. Protein engineering design selection. *PEDS*, **23(9)**, 711-719.

FAO (2012). The state of the world's animal genetics resources for food and agriculture. http://wwwfaoorg/docrep/010/a1250e/a1250e00htm.

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.

Haykin, S. (1994). *Neural Networks : A Comprehensive Foundation*, MacMillan College Publishing Co., New York.

Kumar, M., R. Verma and G. P. S. Raghava (2006). Prediction of mitochondrial proteins using support vector machine and hidden Markov model. *Journal of Biological Chemistry*, **281(9)**, 5357-5363.

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.

Peters, B. (2003). Modeling the MHC-I pathway. *Thesis (PhD)* Berlin, Germany, Humboldt University.

Saha, S. and G. P. S. Raghava (2007). Prediction methods for B-cell epitopes. *Methods Molecular Biology*, **409**, 387-394.

Stanke, M. and S. Waack (2003). Gene prediction with a Hidden Markov Model and a new intron submodel. *Bioinformatics*, **19**, 215–225.

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** (Database issue), D774–D780.

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

Wang, G., X. Li and Z. Wang (2009). APD2 : the updated antimicrobial peptide database and its application in peptide design. *Nucleic Acids Research*, **37**, D933-7.

Willham, R. (1986). From husbandry to science : A highly significant facet of our livestock heritage. *Journal of Animal Science*, **62**, 1742–1758.

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.