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/

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