ClassAMP: A Prediction Tool for Classification of Antimicrobial Peptides

Shaini Joseph, Shreyas Karnik, Pravin Nilawe, V.K. Jayaraman, and Susan Idicula-Thomas

Abstract—Antimicrobial peptides (AMPs) are gaining popularity as anti-infective agents. Information on sequence features that contribute to target specificity of AMPs will aid in accelerating drug discovery programs involving them. In this study, an algorithm called ClassAMP using Random Forests (RFs) and Support Vector Machines (SVMs) has been developed to predict the propensity of a protein sequence to have antibacterial, antifungal, or antiviral activity. ClassAMP is available at http://www.bicnirrh.res.in/classamp/.

Index Terms—Antibacterial, antifungal, antimicrobial, antiviral, prediction algorithm, random forests, SVM.

1 Introduction

A ntimicrobial peptides (AMPs) are an important component of the innate immune system and are present in all classes of life. These molecules inhibit growth of several bacteria, fungi, and viruses [1]. AMPs destroy microbes by disrupting their cell membrane, inhibiting extracellular polymer synthesis like peptidoglycan in case of bacteria and chitin in case of fungi or by acting on intracellular targets [2], [3]. Microbes do not easily develop resistance to AMPs since they interact with structural components of the microbial cell membrane and have multiple cellular targets. Differences in the membrane properties of microbes and mammalian cells enable AMPs to specifically act on microbes thereby reducing their toxicity [3]. Their broad spectrum of activity and reduced toxicity make them ideal substitutes to antibiotics as anti-infective agents, which are constantly rendered ineffective due to development of antibiotic resistance by microbes [4].

AMPs vary in their spectrum of activity and target specificity. Several studies have been carried out to understand the mechanism of action of AMPs. While some antibacterial peptides recognize and disrupt the bacterial cell membranes, others permeabilize the membrane and act on intracellular targets [2]. Studies on antifungal peptides indicate that they interact with the fungal cell membranes by binding to cell surface receptors (e.g., histatins) [2], [4], components of the fungal cell wall like chitin or form complexes with the lipid membrane. Some antifungal peptides like nikkomycins act by inhibiting chitin biosynthesis [4]. Antiviral peptides interact with specific viral receptors on host cells which influence viral attachment, entry or intracellular shuttling [4], [5], [6], [7]. Few antiviral peptides (e.g., retrocyclin 2) are also known to directly target the glycoproteins on the viral envelop leading to inactivation of the virus [4].

It is essential to understand the role of primary structure of AMPs in determining their target specificity. Although charge, hydrophobicity and amphipathicity have been identified as important determinants of antimicrobial activity, sequence features that contribute to target specificity have not been delineated [2], [3]. The AntiBP2 server predicts the antibacterial activity of peptides [8]. However there has not been an attempt to differentially predict the antibacterial, antifungal, or antiviral activity of peptides. In this study, a prediction algorithm has been developed to classify AMPs as antibacterial, antifungal, or antiviral peptides based on their sequence features. This tool will help aid in rational design of AMPs with specific activity against bacteria, fungi, and viruses.

- *S. Joseph and S. Idicula-Thomas are with the Biomedical Informatics Center of Indian Council of Medical Research, National Institute for Research in Reproductive Health, J.M. Street, Parel, Mumbai 400012, Maharashtra, India. E-mail: shainimarina31@gmail.com, thomass@nirrh.res.in.*
- *S. Karnik is with the School of Informatics, Indiana University, 1703 N Chestnut Avenue, Apartment 108, Marshfield, WI 54449. E-mail: sdkarnik@iupui.edu.*
- *P. Nilawe is with the Biomedical Informatics Center of Indian Council of Medical Research, National Institute for Research in Reproductive Health, Room no 201, Plot 17/8, Second Floor, Shivdarshan Building, Sector 4, Sanpada, Navi Mumbai 400705, Maharashtra, India. E-mail: pravinnilawe@gmail.com, 2pravinnilawe@gmail.com.*
- *V.K. Jayaraman is with the Center for Development of Advanced Computing, Pune University Campus, Pune 411007, India. E-mail: jayaramanv@cdac.in.*

Manuscript received 13 Sept. 2011; revised 21 May 2012; accepted 27 May 2012;
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2011-09-0234. Digital Object Identifier no. 10.1109/TCBB.2012.

2 Procedure

2.1 Creation of Data Sets and the Prediction Models

AMPs are known to be broad spectrum or specific in their activity. In this study, three different methods were employed to classify AMPs based on their acitivty. While Method I was designed to identify AMPs active against one or more classes, Methods II and III were implemented to find AMPs specifically active against a particular class of microbes.

The experimentally validated AMPs present in the Collection of Antimicrobial Peptides (CAMP) database [9] were retrieved to create the positive data set. Based on the activity of AMPs, this data set was further divided into three classes' viz., antibacterial, antifungal, and antiviral. The positive data sets for Method I, II, and III are identical. The negative data set in Method I comprised of experimentally proven nonantimicrobial sequences, sequences generated using random numbers, arbitrary sequences with no annotation as antimicrobial and nonsecretory protein sequences with no annotation as antimicrobial present in the UniProt database. The number of sequences in the negative data set of Method I was restricted to twice the number of sequences present in the positive data set. The negative data set in Method II comprised of sequences which were active against the other two classes. For e.g., in case of the antibacterial model, antibacterial peptides constitute the positive data set and the negative data set includes antifungal and antiviral peptides. In Method III, the multiclass classification algorithm was employed. The classifier was trained using antibacterial, antifungal, and antiviral positive training data sets. No separate negative training data set was used. For all the methods, Support Vectore Machines (SVMs) and Random Forests (RFs) were used to build the prediction models.

The positive and negative data sets were further split in 70:30 ratios to obtain the training and test data sets (Table 1). AMPs with activity against more than one class of microbes (e.g., sequences with antibacterial and antifungal activity) were excluded from the training data sets but retained in the test data set II.

2.2 Calculation of Sequence Features

Sequence-based features that are known to influence antimicrobial activity and in vivo stability were computed using in-house Perl scripts (Tables S1 and S2, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2012). The amino acid composition based on the frequencies of each of the 20 amino acids, physicochemical properties like charge, hydrophobicity [10], [11], etc., BLOSUM-50 matrix [12], conformational similarity [13], normalized van der Waals volume, polarity, polarizability, secondary structure propensity [11], [14], and frequencies of dipeptides and tripeptides

1545-5963/12/$31.00 © 2012 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

TABLE 1
Number of Sequences Present in the Various Data Sets (Method I)

Datasets	Training dataset		Test dataset I	Test dataset II
	Positive	Negative		
Antibacterial	454	908	582	387
Antifungal	61	122	78	370
Antiviral	54	108	69	38

were considered for classification. Transition and distribution of physicochemical properties along the sequences were also computed [14], [15], [16].

2.3 Prediction Method

Random Forests: RF is a collection of decision trees which in addition to classification and regression problems, can be used for variable selection, interaction detection, clustering, etc. It is widely used for classification of biological problems and in many cases has performed better than the other machine learning algorithms like Support Vector Machines, Hidden Markov Models, etc., [9], [17], [18]. The feature selection method enables identification of the most informative classification features, which can be further used for reduction of background noise and improvement of prediction accuracy. RF algorithm implemented in R statistical language package was used for developing the prediction functions [19], [20].

Support Vector Machine: SVM employs a linear hyperplane in a higher dimensional feature space for binary classification tasks. Further, to deal with intractability problem it employs appropriate kernel functions so that all computations can be performed in the original input space itself [21], [22], [23]. SVMs can be used to identify patterns which could help in classification of biological data [24], [25]. SVM implementation of e1071 package in R-language was employed in this study [26], [27]. Models were built with all three kernel functions, i.e., linear, radial, and polynomial for each class of AMPs. The kernel function which performed the best for each class of AMPs was retained.

In Method I and II, separate RF and SVM models were built for each class of AMPs whereas in Method III, a multiclass classification model to classify all three classes was built using RF and SVM.

2.4 Feature Selection

A total of 257 sequence features was used to build the classification models. Rigorous recursive Feature Elimination (RFE) method based on RF Gini score was used to identify the most significant sequence-based features that aid in classification [19], [28], [29]. Based on Gini scores, the features were reduced to 50 percent at each step. Thus starting with 257 features, models were built with the top-ranked 128, 64, 32, 16, 8, and 4 features. The performance of the prediction models were assessed based on 10-fold cross validation accuracy and Matthews Correlation Coefficient (MCC) on training data sets [30]. Independent test data sets were used to validate the prediction models.

3 Results and Discussion

ClassAMP, a prediction tool for identification of antibacterial, antifungal, and antiviral peptides have been developed. The classification functions were built using RF and SVM methods. Feature selection method in RF was used to identify the informative features for classification. Models were built with the most informative set of features. The antibacterial, antifungal, and antiviral models in Method I were built with 128, 8, and 64 features, respectively. In Methods II and III, feature selection was performed again. While 128 features were used for antibacterial and 64 features were used for antifungal, and antiviral models in Method II, the multiclass classifier in Method III was built using 32 features. The performance of the prediction models was evaluated based on MCC (Methods I and II) and prediction accuracy (Method III) on the training data sets (Tables 2, 3, and 4).

TABLE 2
Performance of RF and SVM Prediction Models Using Method I

Model	Method	MCC*		Prediction Accuracy (%)		
		Training dataset	Test dataset I	Training dataset	Test dataset I	Test dataset II
Antibacterial	RF	0.88	0.84	93.3	90.7	79.8
Antifungal	RF	0.97	1	98.4	100	95.1
Antiviral	RF	1	0.96	100	97.8	63.2
Antibacterial	SVM	0.91	0.83	92.9	93.1	80.0
Antifungal	SVM	0.85	0.91	92.6	96	65.20
Antiviral	SVM#	1	0.96	100	98.6	58.9

* Matthews Correlation Coefficient #In this case, the linear kernel function was used while for the other SVM-based models the RBF kernel function was used.

TABLE 3
Performance of RF and SVM Prediction Models Using Method II

Model	Method	MCC		Prediction Accuracy (%)	
		Training dataset	Test dataset I	Training dataset	Test dataset I
Antibacterial	RF	0.76	0.74	92.4	92.8
Antifungal	RF	0.63	0.68	94.2	94.7[#]
Antiviral	RF	0.93	0.87	98.6	97.9
Antibacterial	SVM[$]	0.92	0.63	92.3	91.8
Antifungal	SVM	0.83	0.65	93.5	94.2
Antiviral	SVM	0.96	0.87	99.3	96

[#] There are false positives since the size of the antifungal dataset is small [$]In this case, the linear kernel function was used while for the other SVM-based models the RBF kernel function was used.

In this work, models were built using three different methods as explained earlier. In Method II, three one-against-all classifier models were built with examples in one class taken as positive and other two as negative, test data sets were classified using each one of the models and the class obtaining the majority votes was assigned as its class. While the algorithm parameters were optimized separately for each of the models in Methods I and II, models in Method III, a conventional multiclass classifier were tuned with same set of parameters. The models built using Method II performed better than those built using Methods I and III and hence has been incorporated in ClassAMP.

The sequence features important for classification of AMPs into antibacterial, antifungal, and antiviral peptides were analyzed (TableS3, available online). Hydrophobicity, charge [31] and composition of glutamic acid, aspargine, cysteine, and serine residues were features which were necessary for classification of all three classes. While antifungal peptides were found to be more hydrophobic as compared to the other two classes, antibacterial peptides were observed to be more positively charged. Hydrophobicity and charge of AMPs are critical for the interaction and permeabilization of microbial cell membranes eventually leading to cell death either by causing cell lysis or acting on intracellular targets [2], [3]. Antibacterial and antifungal peptides were rich in neutrally charged residues while antiviral peptides were rich in polar residues.

Comparison with existing prediction algorithms. Several web-based tools are available for prediction of antimicrobial activity like APD [32], CAMP [9], etc. However these algorithms do not classify AMPs based on their target-specificity. Studies have been undertaken to investigate the sequence and structural features critical for antibacterial activity of peptides. These studies were mainly devoted to a subset of antibacterial peptides like cecropin-melittin derivatives [33] and small cationic amphipathic antibacterial peptides (SCAAPs) [34]. As these models are trained on

TABLE 4
Performance of RF and SVM Prediction Models Using Method III

Activity	Method	Prediction Accuracy (%)	
		Training dataset	Test dataset I
Antibacterial	RF	98.7	98.5
Antifungal		49.2	61.5
Antiviral		92.6	82.6
Antibacterial	SVM*	100	97
Antifungal		100	57
Antiviral		100	87

*The polynomial kernel function was used

a limited subset of antibacterial peptides, the sensitivity of these algorithms when extended to all classes of antibacterial peptides cannot be predicted. Absence of a web interface for these algorithms makes it further difficult for the users. AntiBP2 [8] is the only prediction server available for antibacterial peptides. The prediction models in AntiBP2 are built using SVM. The antibacterial prediction model of ClassAMP and AntiBP gave a prediction accuracy of ~95 percent with the test sequences (TableS4). An ANN-based model using sequence-based features has been developed for prediction of antifungal activity of peptides [35]. However, absence of web-based interface makes it difficult to access and review the model.

While independent studies have been undertaken to identify sequence features important for antibacterial and antifungal activity, there has not been an attempt for classification of peptides based on their target-specificity. ClassAMP enables classification of peptides as antibacterial, antifungal, or antiviral peptides.

4 Conclusions

Study of AMPs is currently an active area of research. Efforts are being made to develop AMPs as potent drugs. Knowledge of sequence features important for antibacterial, antifungal, and antiviral activity will help researchers to design specific drugs against bacteria, fungi, and viruses. The ClassAMP server is designed to predict the propensity of a given sequence to have antibacterial, antifungal, or antiviral activity. The performances of the RF- and SVM-based models are dependent on the available information on the target specificities of various AMPs. The performance can be enhanced with increase in the availability of information on AMPs in the public domain.

5 Availability

The ClassAMP server is freely available at http://www.bicnirrh.res.in/classamp/.

Acknowledgments

The authors are grateful to Dr. S.D. Mahale for all the help and support. They also acknowledge the assistance provided by Mr. Ram Shankar Barai in designing the web interface. We are also grateful to the anonymous reviewers for their comments which have helped in improving the manuscript. This work has been funded by Department of Science and Technology, Government of India (SR/S3/CE/52/2007), NIRRH (NIRRH/A/15/11), Indian Council of Medical Research (63/128/2001-BMS), and CSIR Emeritus Scientist Grant.

References

[1] B.M. Peters, M.E. Shirtiff, and M.A. Jabra-Rizk, "Antimicrobial Peptides: Primeval Molecules or Future Drugs," *PLoS Pathogens*, vol. 6, no. 10, p. e1001067, Oct. 2010. doi: 10.1371/journal.ppat.1001067.

[2] K.A. Brogden, "Antimicrobial Peptides: Pore Formers or Metabolic Inhibitors in Bacteria?" *Nature Rev. Microbiology*, vol. 3, no. 3, pp. 238-250, Mar. 2005.

[3] M.R. Yeaman and N.Y. Yount, "Mechanisms of Antimicrobial Peptide Action and Resistance," *Pharmacological Rev.*, vol. 55, no. 1, pp. 27-55, Mar. 2003.

[4] H. Jenssen, P. Hamill, and R.E. Hancock, "Peptide Antimicrobial Agents," *Clinical Microbiology Rev.*, vol. 19, no. 3, pp. 491-511, July 2006.

[5] Y. Aboudy, E. Mendelson, I. Shalit, R. Bessalle, and M. Fridkin, "Activity of Two Synthetic Amphiphilic Peptides and Magainin-2 against Herpes Simplex Virus Types 1 and 2," *Int'l J. Peptide and Protein Research*, vol. 43, no. 6, pp. 573-582, June 1994.

[6] A. Belaid, M. Aouni, R. Khelifa, A. Trabelsi, M. Jemmali, and K. Hani, "In Vitro Antiviral Activity of Dermaseptins against Herpes Simplex Virus Type 1," *J. Medical Virology*, vol. 66, no. 2, pp. 229-234, Feb. 2002.

[7] W.E. Robinson Jr., B. McDougall, D. Tran, and M.E. Selsted, "Anti-HIV-1 Activity of Indolicidin, an Antimicrobial Peptide from Neutrophils," *J. Leukocyte Biology*, vol. 63, no. 1, pp. 94-100, Jan. 1998.

[8] S. Lata, N.K. Mishra, and G.P. Raghava, "AntiBP2: Improved Version of Antibacterial Peptide Prediction," *BMC Bioinformatics*, vol. 11, pp. S1-S19, Jan. 2010.

[9] S. Thomas, S. Karnik, R.S. Barai, V.K Jayaraman, and S. Idicula-Thomas, "CAMP: A Useful Resource for Research on Antimicrobial Peptides," *Nucleic Acids Research*, vol. 38, pp. D774- D780, Jan. 2010.

[10] G.D. Rose, A.R. Geselowitz, G.J. Lesser, R.H. Lee, and M.H. Zehfus, "Hydrophobicity of Amino Acid Residues in Globular Proteins," *Science*, vol. 229, no. 4716, pp. 834-838, Aug. 1985.

[11] K. Tomii and M. Kanehisa, "Analysis of Amino Acid Indices and Mutation Matrices for Sequence Comparison and Structure Prediction of Proteins," *Protein Eng.*, vol. 9, no. 1, pp. 27-36, Jan. 1996.

[12] L.R. Murphy, A. Wallqvist, and R.M. Levy, "Simplified Amino Acid Alphabets for Protein Fold Recognition and Implications for Folding," *Protein Eng.*, vol. 13, no. 3, pp. 149-152, Mar. 2000.

[13] P. Chakrabarti and D. Pal, "The Interrelationships of Side-Chain and Main-Chain Conformations in Proteins," *Progress in Biophysics and Molecular Biology*, vol. 76, nos. 1/2, pp. 1-102, 2001.

[14] Z.R. Li, H.H Lin, L.Y. Han, L. Jiang, X. Chen, and Y.Z. Chen, "PROFEAT: A Web Server for Computing Structural and Physicochemical Features of Proteins and Peptides from Amino Acid Sequence," *Nucleic Acids Research*, vol. 34, pp. W32-37, July 2006.

[15] I. Dubchak, I. Muchnik, S.R. Holbrook, and S.H. Kim, "Prediction of Protein Folding Class Using Global Description of Amino Acid Sequence," *Proc. Nat'l Academy of Sciences of USA*, vol. 92, no. 19, pp. 8700-8704, Sept. 1995.

[16] I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk, and S.H. Kim, "Recognition of a Protein Fold in the Context of the SCOP Classification," *Proteins*, vol. 35, no. 4, pp. 401-407, June 1999.

[17] V.L. Ravich, M. Masso, and I.I. Vaisman, "A Combined Sequence-Structure Approach for Predicting Resistance to the Non-Nucleoside HIV-1 Reverse Transcriptase Inhibitor Nevirapine," *Biophysical Chemistry*, vol. 153, nos. 2/3, pp. 168-172, Jan. 2011.

[18] G. Riddick, H. Song, S. Ahn, J. Walling, D. Borges-Rivera, W. Zhang, and H.A. Fine, "Predicting in Vitro Drug Sensitivity Using Random Forests," *Bioinformatics*, vol. 27, no. 2, pp. 220-224, Jan. 2011.

[19] L. Breiman, "Random Forests," *Machine Learning*, vol. 35, no. 1, pp. 5-32, Oct. 2001, doi: 10.1023/A:1010933404324.

[20] A. Liaw and M. Wiener, "Classification and Regression by Random Forest," *R News*, vol. 2, pp. 18-22, Dec. 2002.

[21] V. Vapnik, *Statistical Learning Theory*. Wiley and Sons, 1998.

[22] S.R. Gunn, "Support Vector Machines for Classification and Regression," http://www.svms.org/tutorials/, 2012.

[23] S. Idicula-Thomas, A.J. Kulkarni, B.D. Kulkarni, V.K. Jayaraman, and P.V. Balaji, "A Support Vector Machine-Based Method for Predicting the Propensity of a Protein to be Soluble or to form Inclusion Body on Overexpression in Escherichia Coli," *Bioinformatics*, vol. 22, no. 3, pp. 278-284, Feb. 2006.

[24] V. Vapnik, *The Nature of Statistical Learning Theory*, first ed. Springer, 1995.

[25] K.R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An Introduction to Kernel-Based Learning Algorithms," *IEEE Trans. Neural Networks*, vol. 12, pp. 181-201, Mar. 2001.

[26] R Development Core Team, "R: A Language and Environment for Statistical Computing," *R Foundation for Statistical Computing*, 2009.

[27] A. Kulkarni, B.D. Kulkarni, and V.K. Jayaraman, "Support Vector Classification with Parameter Tuning Assisted by Agent-Based Technique," *Computers and Chemical Eng.*, vol. 28, pp. 311-318, Mar. 2004.

[28] http://en.wikipedia.org/wiki/Decision_tree_learning, 2012.

[29] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," *Machine Learning*, vol. 46, no. 1, pp. 389-422, 2002.

[30] P. Baldi, S. Brunak, Y. Chauvin, C.A. Andersen, and H. Nielsen, "Assessing the Accuracy of Prediction Algorithms for Classification: An Overview," *Bioinformatics*, vol. 16, no. 5, pp. 412-424, May 2000.

[31] M. Charton and B.I. Charton, "The Dependence of the Chou-Fasman Parameters on Amino Acid Side Chain Structure," *J. Theoretical Biology*, vol. 102, no. 1, pp. 121-134, May 1983.

[32] Z. Wang and G. Wang, "APD: The Antimicrobial Peptide Database," *Nucleic Acids Research*, vol. 32, pp. D590-D592, Jan. 2004.

[33] A. Cherkasov and B. Jankovic, "Application of 'Inductive' QSAR Descriptors for Quantification of Antibacterial Activity of Cationic Polypeptides," *Molecules*, vol. 9, no. 12, pp. 1034-1052, Dec. 2004.

[34] C. Polanco and J.L. Samaniego, "Detection of Selective Cationic Amphipatic Antibacterial Peptides by Hidden Markov Models," *Acta Biochimica Polonica*, vol. 56, no. 1, pp. 167-176, Mar. 2009.

[35] S. Soltani, K. Keymanesh, and S. Sardari, "Evaluation of Structural Features of Membrane Acting Antifungal Peptides by Artificial Neural Network," *J. Biological Sciences*, vol. 8, no. 5, pp. 834-845, 2008, doi:10.3923/jbs.2008.834.845.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.