Bioinformatics, 2020, 1–3
doi: 10.1093/bioinformatics/btaa917
Advance Access Publication Date: 2 November 2020
Applications Note

Sequence analysis

amPEPpy 1.0: a portable and accurate antimicrobial peptide prediction tool

Travis J. Lawrence [1,*], Dana L. Carper[1], Margaret K. Spangler[1], Alyssa A. Carrell[1], Tomás A. Rush [1], Stephen J. Minter[2], David J. Weston[1] and Jessy L. Labbé[1,*]

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA and [2]Cryomagnetics, Inc., Oak Ridge, TN 37830, USA

*To whom correspondence should be addressed.

Associate Editor: Martelli Pier Luigi

Received on July 31, 2020; revised on October 7, 2020; editorial decision on October 13, 2020; accepted on October 16, 2020

Abstract

Summary: Antimicrobial peptides (AMPs) are promising alternative antimicrobial agents. Currently, however, portable, user-friendly and efficient methods for predicting AMP sequences from genome-scale data are not readily available. Here we present amPEPpy, an open-source, multi-threaded command-line application for predicting AMP sequences using a random forest classifier.

Availability and implementation: amPEPpy is implemented in Python 3 and is freely available through GitHub (https://github.com/tlawrence3/amPEPpy).

Contact: tlawrence3@ucmerced.edu or labbejj@ornl.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Multidrug-resistant microbial infections are among the most critical health crises plaguing the world today (Michael *et al.*, 2014), creating an urgent demand for novel therapies. Antimicrobial peptides (AMPs), an ancient, extremely diverse, innate defense mechanism against pathogens found in almost all forms of life (Jenssen *et al.*, 2006), have emerged as top contenders in the race against resistance (Zharkova *et al.*, 2019). AMPs are typically 5–100 amino acids long, amphipathic and positively charged, and thus able to disrupt negatively charged microbial membranes (Mishra and Wang, 2012). AMPs increase antibiotic activity against multidrug-resistant bacteria (Lázár *et al.*, 2018) and have a lower prevalence of horizontally transferred resistance relative to antibiotics (Kintses *et al.*, 2019). Based on these features, AMPs represent promising medicinal targets. Currently, a major challenge in AMP discovery is the lack of sequence conservation of these peptides, limiting the effectiveness of traditional sequence homology-based search tools such as BLAST (Mishra and Wang, 2012). This has led to the development of several machine learning approaches for AMP prediction (Cardoso *et al.*, 2020; Torres and de la Fuente-Nunez, 2019).

However, the current machine-learning methods are not easily portable, require proprietary software (e.g. MATLAB) (Bhadra *et al.*, 2018) or are accessible as web portals (Cardoso *et al.*, 2020) that are not amenable to processing of genome-scale data because of dataset size restrictions and limited computational resources. Furthermore, programming and machine-learning expertise are often required for training and optimizing current methods on novel training data. To address these issue, we developed amPEPpy, a Python 3 application that implements the amPEP (Bhadra *et al.*, 2018) classifier with improved portability, increased accuracy relative to similar methods (Cardoso *et al.*, 2020; Lata *et al.*, 2010; Lin *et al.*, 2019), utilities for easily training and optimizing random forest (RF) classifiers on novel data which promotes documentation and reproducibly, and a command-line user interface designed for the efficient processing of genome-scale data.

2 Materials and methods

2.1 Implementation

amPEPpy is written in Python 3 and has a command-line interface. The functionality of amPEPpy is contained in two modules, train and predict. The train module provides methods for training an RF classifier to identify AMP sequences, optimizing the number of decision trees within the classifier, and calculating feature importance using the drop-column method. The inputs to the train module are a set of AMP and non-AMP amino acid sequences partitioned into separate FASTA files, which are available on the amPEPpy GitHub repository. The predict module classifies amino acid sequences as AMPs or non-AMPs, and as input requires a trained RF classifier from the train module and a FASTA file of amino acid sequences to be classified.

2.2 Training data and encoding

The training data of Bhadra *et al.* (2018), downloaded from https://cbbio.cis.um.edu.mo/software/AmPEP/, included 3268 AMP sequences original retrieved from APD3 (Wang *et al.*, 2016), CAMPR3 (Waghu *et al.*, 2016) and LAMP databases (Zhao *et al.*,

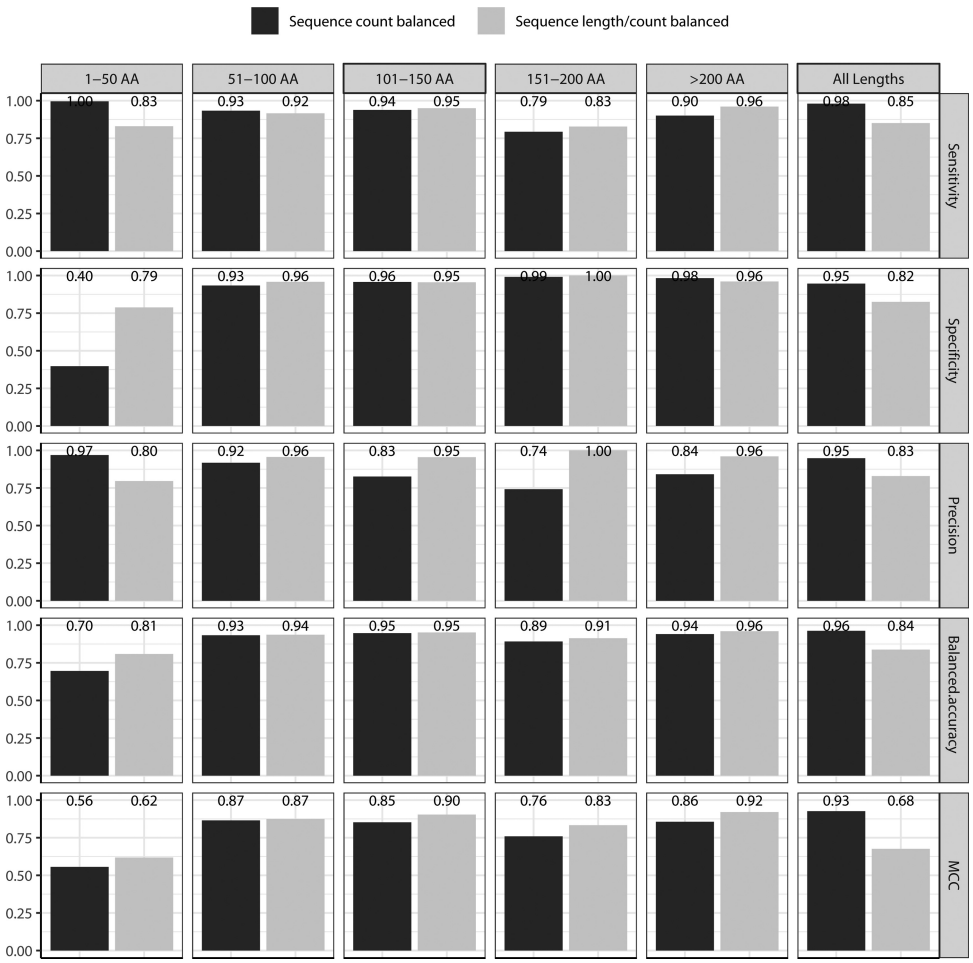

Fig. 1. Sensitivity, specificity, precision, balanced accuracy and the Matthews correlation coefficient (MCC) determined from OOB samples of the RF classifiers trained on count-balanced or length/count-balanced training sets for each sequence length distribution and overall. These metrics are defined as sensitivity=tp/(tp+f n), specificity=tn/(tn+fp), precision=tp/(tp+fp), balanced accuracy=(sensitivity+specificity)/2, MCC=$(\mathrm{tp} \times \mathrm{tn} - \mathrm{fp} \times \mathrm{fn})/\sqrt{(\mathrm{tp}+\mathrm{fp})(\mathrm{tp}+\mathrm{fn})(\mathrm{tn}+\mathrm{fp})(\mathrm{tn}+\mathrm{fn})}$, where tp = true positive, fp = false positive, tn = true negative and fn = false negative

2013) and 166 791 non-AMP sequences from UniProt. We created two balanced training sets containing the same number of AMP and non-AMP sequences: (i) by randomly subsampling the non-AMP sequences (count balanced); and (ii) by randomly subsampling sequences in five sequence length distributions of 1–50, 51–100, 101–150, 151–200, ≥200 amino acids to match the proportions of the AMP data (length/count balanced) (Supplementary Fig. S1; Supplementary Table S1). Subsampling was performed using DISCo-microbe (Carper *et al.*, 2020). Following Bhadra *et al.* (2018), we encoded 105 features for each sequence using the *Distribution* descriptor set from the Global Protein Sequence Descriptors (Dubchak *et al.*, 1995), which describes the distribution of physicochemical properties along the primary amino acid sequence.

2.3 Training and optimization of the RF AMP classifier

We implemented our RF classifier using scikit-learn v0.23.1 in Python v3.8.3. Using the Out-Of-Bag (OOB) error (1 - OOB accuracy) as the optimality criterion, we optimized the number of decision trees within our RF classifier, considering 25–175 decision trees. We optimized a classifier for the count-balanced and length/count-balanced training sets. To determine relative feature importance, we used the drop-column method, which drops a feature, retrains the classifier and calculates the difference in OOB accuracy relative to the full model.

3 Results

Our RF classifiers achieved an optimized OOB error of 0.036 with 128 decision trees on the count-balanced data and an OOB error of 0.163 with 160 decision trees on the length/count-balanced training data (Supplementary Fig. S2). To investigate biases of our classifiers, we calculated sensitivity, specificity, precision and balanced accuracy separately for each sequence length distribution (Fig. 1). Notably, the classifier trained on the length/count-balanced data had a higher or equal balanced accuracy for every sequence length distribution; however, the classifier trained on count-balanced data had a higher balanced accuracy on the dataset overall (Fig. 1). The difference in balanced accuracy between our classifiers was greatest for the sequence length distribution of 1–50 amino acids, primarily because the classifier trained on length/count-balanced data had a greater ability to distinguish non-AMP sequences in this length the number of short non-AMPs relative to short AMP sequences (Supplementary Fig. S1). This suggests that the RF is capable of learning sequence length indirectly and highlights the importance of balancing sequence length distribution in AMP training data. A notable result from this work is that ordinal ranking of feature importance was not consistent among classifiers trained on the two datasets (Supplementary Fig. S3) or among those estimated by Bhadra *et al.* (2018), suggesting that feature importance is highly dependent on the training data.

4 Conclusion

Here, we introduced amPEPpy, an easily installable and publicly available Python application for predicting AMP protein sequences using a RF classifier. amPEPpy should prove useful for the identification and study of AMP sequences that could be used to combat invasive microbial pathogens (Oshiro *et al.*, 2019), develop novel pharmacological agents (Mahlapuu *et al.*, 2016) and products in the agricultural and food industries (Keymanesh *et al.*, 2009), and identify small secreted molecules responsible for host–microbe–community interactions (Plett *et al.*, 2014).

Funding

D.L.C., A.A.C., T.A.R. and J.L.L. were supported by the Genomic Science Program, U.S. Department of Energy (DOE), Office of Science, Biological and Environmental Research as part of the Plant Microbe Interfaces Scientific Focus Area. T.J.L. and D.J.W. were supported by the Laboratory Directed Research and Development Program at Oak Ridge National Laboratory (ORNL). ORNL is managed by UT-Battelle, LLC, for the DOE under contract DE-AC05-00OR22725.

Conflict of Interest: none declared.

References

Bhadra,P. *et al.* (2018) AmPEP: sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. *Sci. Rep.*, **8**, 10.

Cardoso,M.H. *et al.* (2020) Computer-aided design of antimicrobial peptides: are we generating effective drug candidates? *Front. Microbiol.*, **10**.

Carper,D.L. *et al.* (2020) DISCo-microbe: design of an identifiable synthetic community of microbes. *PeerJ*, **8**, e8534–e8534.

Dubchak,I. *et al.* (1995) Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. USA*, **92**, 8700–8704.

Jenssen,H. *et al.* (2006) Peptide antimicrobial agents. *Clin. Microbiol. Rev.*, **19**, 491–511.

Keymanesh,K. *et al.* (2009) Application of antimicrobial peptides in agriculture and food industry. *World J. Microbiol. Biotechnol.*, **25**, 933–944.

Kintses,B. *et al.* (2019) Phylogenetic barriers to horizontal transfer of antimicrobial peptide resistance genes in the human gut microbiota. *Nat. Microbiol.*, **4**, 447–458.

Lata,S. *et al.* (2010) AntiBP2: improved version of antibacterial peptide prediction. *BMC Bioinformatics*, **11**, S19–7.

Lázár,V. *et al.* (2018) Antibiotic-resistant bacteria show widespread collateral sensitivity to antimicrobial peptides. *Nat. Microbiol.*, **3**, 718–731.

Lin,Y. *et al.* (2019) An advanced approach to identify antimicrobial peptides and their function types for penaeus through machine learning strategies. *BMC Bioinformatics*, **20**, 1–10.

Mahlapuu,M. *et al.* (2016) Antimicrobial peptides: an emerging category of therapeutic agents. *Front. Cell. Infect. Microbiol.*, **6**, 1–12.

Michael,C.A. *et al.* (2014) The antimicrobial resistance crisis: causes, consequences, and management. *Front. Public Health*, **2**, 1–8.

Mishra,B. and Wang,G. (2012) The importance of amino acid composition in natural amps: an evolutional, structural, and functional perspective. *Front. Immunol.*, **3**, 2010–2013.

Oshiro,K.G.N. *et al.* (2019) Bioactive peptides against fungal biofilms. *Front. Microbiol.*, **10**, 1–17.

Plett,J.M. *et al.* (2014) Effector MiSSP7 of the mutualistic fungus Laccaria bicolor stabilizes the Populus JAZ6 protein and represses jasmonic acid (JA) responsive genes. *Proc. Natl. Acad. Sci. USA*, **111**, 8299–8304.

Torres,M.D.T. and de la Fuente-Nunez,C. (2019) Toward computer-made artificial antibiotics. *Curr. Opin. Microbiol.*, **51**, 30–38.

Waghu,F.H. *et al.* (2016) CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Res.*, **44**, D1094–D1097.

Wang,G. *et al.* (2016) APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.*, **44**, D1087–D1093.

Zhao,X. *et al.* (2013) LAMP: a database linking antimicrobial peptides. *PLoS One*, **8**, e66557.

Zharkova,M.S. *et al.* (2019) Application of antimicrobial peptides of the innate immune system in combination with conventional antibiotics—a novel way to combat antibiotic resistance? *Front. Cell. Infect. Microbiol.*, **9**. doi: 10.3389/fcimb.2019.00128.