Title: **amPEPpy 1.0: A portable and accurate antimicrobial peptide prediction tool**

Authors: 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,*]

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

Corresponding authors
Travis J. Lawrence and Jessy L. Labbé
Biosciences Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831-6301
Email: lawrencetj@ornl.gov or labbejj@ornl.gov

Target journal: Oxford University Press Bioinfomatics

Word count: 931 (excluding references, figure caption, and funding statement)
Abstract - 72
Introduction - 300
Methods - 241
Results - 236
Conclusion - 82
Number of references - 16
Figures: 1;Tables: 0; Supplemental Figures: 3

Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
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,[2] Cryomagnetics, Inc. Oak Ridge, TN 37830, USA

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

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

Contact: lawrencetj@ornl.gov 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, Dominey-Howes and Labbate 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, Hamill and Hancock 2006), have emerged as top contenders in the race against resistance (Zharkova *et al.* 2019). AMPs are typically 5 to 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).

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. Furthermore, programming and machine-learning expertise are often required for training and optimizing current methods on novel training data. To address these issues work, 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 (Lata, Mishra and Raghava 2010; Lin *et al.* 2019; Cardoso *et al.* 2020), utilities for easily training and optimizing random forest (RF) classifiers on novel data, and a command-line user interface designed for the efficient processing of genome-scale data.

2 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 3,268 AMP sequences and 166,791 non-AMP sequences from UniProt. We created two balanced training sets: 1) by randomly subsampling the non-AMP sequences (count balanced); and 2) 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). Subsampling was performed using DISCo-microbe (Carper *et al.* 2020) (Fig. S1). 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.

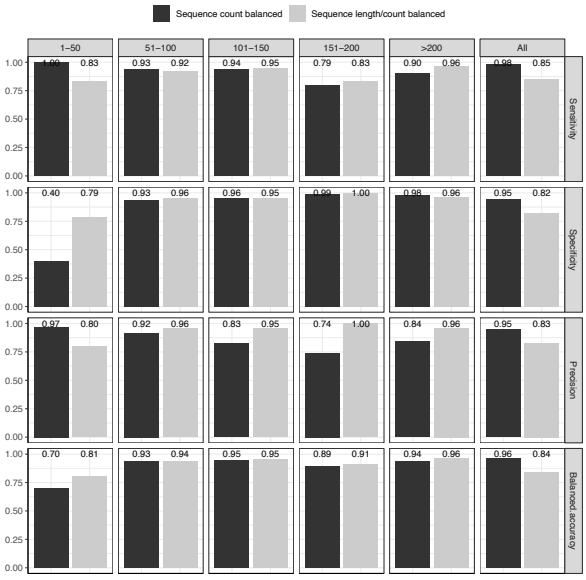

Figure 1. Sensitivity, specificity, precision, and balanced accuracy determined from OOB samples of the random forest classifiers trained on count-balanced or length/count-balanced training sets for each sequence length distribution and overall. These metrics are defined as sensitivity = tp/(tp+fn), specificity = tn/(tn+fp), precision = tp/(tp+fp), balanced accuracy = (sensitivity+specificity)/2, where tp = true positive, fp = false positive, tn = true negative, and fn = false negative.

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 to 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 (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 distribution (Fig 1). This is likely driven by the skewed distribution towards longer sequence lengths in the count-balanced dataset, limiting the number of short non-AMPs relative to short AMP sequences (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 (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, Soltani and Sardari 2009), and identify small secreted molecules responsible for host–microbe–community interactions (Plett *et al.* 2014).

Funding

DLC, AAC, TAR, and JLL 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. TJL and DJW 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, Yan J, Li J *et al.* AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. *Sci Rep* 2018;**8**:1–10.
Cardoso MH, Orozco RQ, Rezende SB *et al.* Computer-Aided Design of Antimicrobial Peptides: Are We Generating Effective Drug Candidates? . *Front Microbiol* 2020;**10**:3097.
Carper DL, Lawrence TJ, Carrell AA *et al.* DISCo-microbe: design of an identifiable synthetic community of microbes. *PeerJ* 2020;**8**:e8534–e8534.
Dubchak I, Muchnik I, Holbrook SR *et al.* Prediction of protein folding class using global description of amino acid sequence. *Proc Natl Acad Sci U S A* 1995;**92**:8700–4.
Jenssen H, Hamill P, Hancock REW. Peptide Antimicrobial Agents. *Clin Microbiol Rev* 2006;**19**:491–511.
Keymanesh K, Soltani S, Sardari S. Application of antimicrobial peptides in agriculture and food industry. *World J Microbiol Biotechnol* 2009;**25**:933–44.
Kintses B, Méhi O, Ari E *et al.* Phylogenetic barriers to horizontal transfer of antimicrobial peptide resistance genes in the human gut microbiota. *Nat Microbiol* 2019;**4**:447–58.
Lata S, Mishra NK, Raghava GPS. AntiBP2: Improved version of antibacterial peptide prediction. *BMC Bioinformatics* 2010;**11**:1–7.
Lázár V, Martins A, Spohn R *et al.* Antibiotic-resistant bacteria show widespread collateral sensitivity to antimicrobial peptides. *Nat Microbiol* 2018;**3**:718–31.
Lin Y, Cai Y, Liu J *et al.* An advanced approach to identify antimicrobial peptides and their function types for penaeus through machine learning strategies. *BMC Bioinformatics* 2019;**20**:1–10.
Mahlapuu M, Håkansson J, Ringstad L *et al.* Antimicrobial peptides: An emerging category of therapeutic agents. *Front Cell Infect Microbiol* 2016;**6**:1–12.
Michael CA, Dominey-Howes D, Labbate M. The antimicrobial resistance crisis: Causes, consequences, and management. *Front Public Heal* 2014;**2**:1–8.
Mishra B, Wang G. The importance of amino acid composition in natural amps: An evolutional, structural, and functional perspective. *Front Immunol* 2012;**3**:2010–3.
Oshiro KGN, Rodrigues G, Monges BED *et al.* Bioactive Peptides Against Fungal Biofilms. *Front Microbiol* 2019;**10**:1–17.
Plett JM, Daguerre Y, Wittulsky S *et al.* Effector MiSSP7 of the mutualistic fungus Laccaria bicolor stabilizes the Populus JAZ6 protein and represses jasmonic acid (JA) responsive genes. *Proc Natl Acad Sci* 2014;**111**:8299 LP – 8304.
Zharkova MS, Orlov DS, Golubeva OY *et al.* 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* 2019;**9**, DOI: 10.3389/fcimb.2019.00128.