SAMP: An Accurate Ensemble Model Based on Proportionalized Split Amino Acid Composition for Identifying Antimicrobial Peptides

Junxi Feng[1], Mengtao Sun[2], Weiwei Zhang[3], Guangshun Wang[3], Shibiao Wan[2]*
[1]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA 02115
[1]Department of Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, Omaha, Nebraska, 68198
[3]Department of Pathology and Microbiology, University of Nebraska Medical Center, Omaha, NE, United States 68198
*Correspondence: Dr. Shibiao Wan, swan@unmc.edu

University of Nebraska Medical Center

Abstract

Motivation:

1> Antimicrobial peptides (AMPs) have received significant attention for their capacity to combat a broad spectrum of pathogens, including viruses, bacteria, and fungi. Predicting AMPs has made it easy and efficient to find AMPs from large datasets with high accuracy.
2> Existing methods only use features including compositional, physiochemical, and structural properties of peptide sequences, which cannot fully capture information from AMPs.

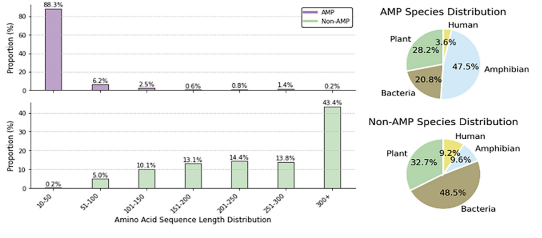
Proposal:

Here, we present SAMP, an ensemble random projection (RP) based computational model that leverages a new type of features called proportionalized split amino acid composition (PSAAC) in addition to conventional sequence-based features for AMP prediction. With this new feature set, SAMP captures the residue patterns like sorting signals at around both the N terminus and C terminus, while also retaining the sequence order information from the middle peptide fragments.

Findings:

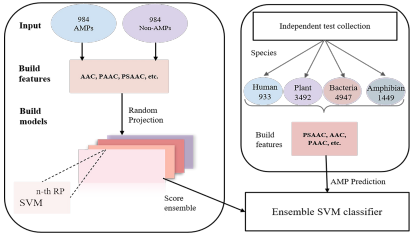
SAMP consistently outperforms existing state-of-the-art methods for identifying AMPs

SAMP Data

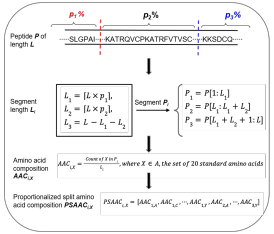


Methods

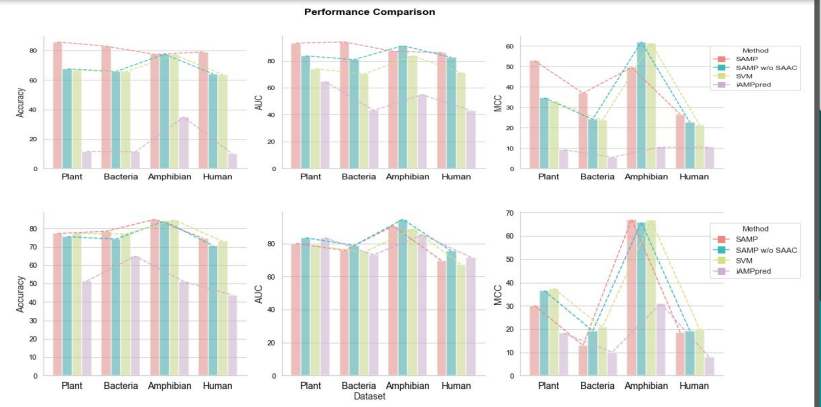
The flowchart of SAMP



Process of building features



Results



Conclusion

In this study, we introduce an ensemble model based on proportionalized split amino acid compositions, SAMP to accurately identify antimicrobial peptides based on peptide sequence data. The ensemble random projection architecture not only reduces the dimensionality of high-dimensional features, but also improves performance by incorporating prediction scores from individual layer of projected features.Our benchmarking tests on datasets from different species demonstrate that SAMP consistently outperforms existing state-of-the-art methods, such as iAMPpred and AMPScanner, in terms of accuracy, sensitivity, specificity and AUC. Future research may focus on the ensemble of different classification models and deep learning approaches. SAMP is freely and publicly available at https://github.com/wan-mlab/SAMP.

Funding

This work was supported by the Buffet Cancer Center, which is supported by the National Cancer Institute under award number CA036727, in collaboration with the UNMC/Children's Hospital & Medical Center Child Health Research Institute Pediatric Cancer Research Group. This work was also partly supported by the UNMC Alcohol Center of Research-Nebraska (ACORN) pilot grant.