AmpGram: a novel tool for prediction of antimicrobial peptides

Jarosław Chilimoniuk[1], Michał Burdukiewicz[2], Katarzyna Sidorczuk[1], Stefan Rödiger[3] and Przemysław Gagat[1]*

*przemyslaw.gagat@uwr.edu.pl

[1]Department of Bioinformatics and Genomics, University of Wrocław, Wrocław, POLAND [2]Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, POLAND [3]Faculty of Natural Sciences, Brandenburg University of Technology Cottbus-Senftenberg, Senftenberg, Germany

Introduction

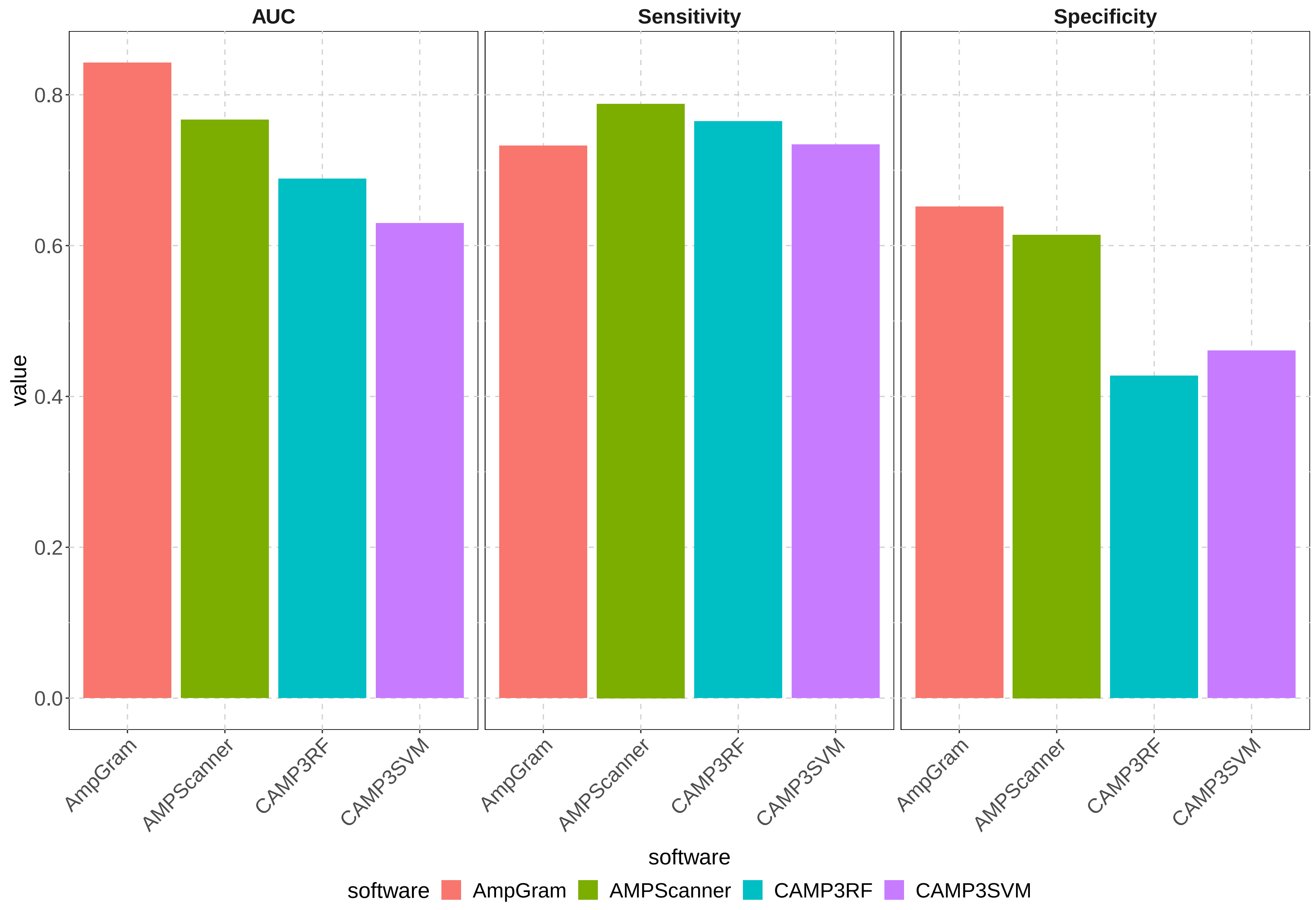
Question: Antimicrobial peptides (AMPs) are ancient and evolutionarily conserved molecules widespread in all living organisms that participate in host defense and/or microbial competition. Due to their positive charge, hydrophobicity and amphipathicity, they preferentially disrupt negatively-charged bacterial membranes. AMPs are considered an important alternative to traditional antibiotics, especially when the latter are drastically losing their effectiveness. Therefore, efficient computational tools for AMP prediction are essential to identify new AMP candidates without undertaking expensive experimental studies.

Methods: AmpGram is our novel tool for predicting AMPs based on the methodology that has already been used with success in our previous applications. It employes simplified alphabets to encode the information from highly variable AMPs into informative features suitable for machine learning, and n-gram analysis to reveal amino acid motifs associated with the presence or absence of antimicrobial properties.

Results: In order to test AmpGram performance, we benchmarked it against the state-of-the-art AMP classifiers, including AMPScanner 2.0, CAMPR3RF and CAMPR3SVM. AmpGram outperformed all of them in terms of AUC and specificity. Additionally, AmpGram provides a list of amino acid motifs associated with antimicrobial properties.

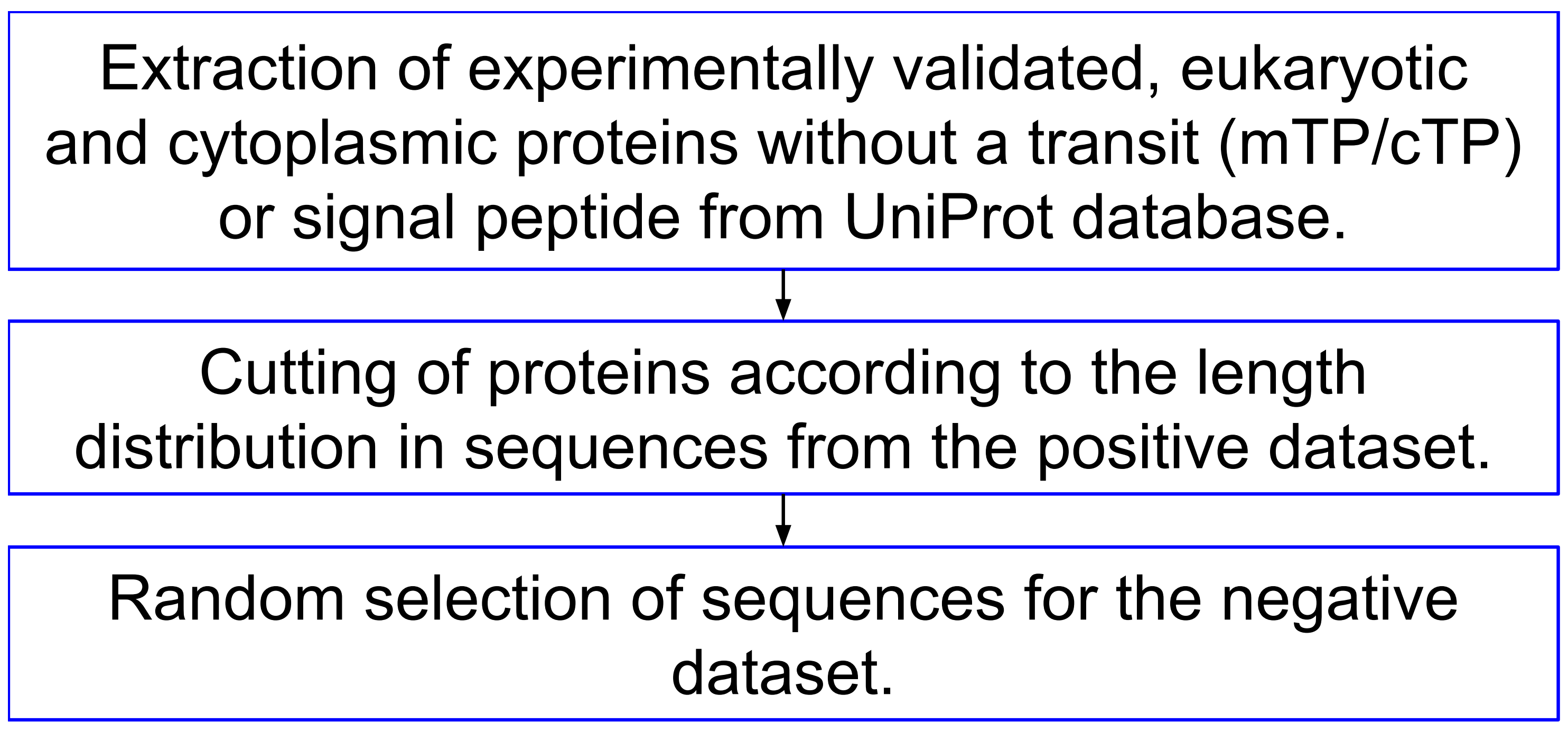
Conclusions: AmpGram is a novel tool for AMP prediction that outperforms all the existing tools. Moreover, it identifies amino acid motifs that can be used to study and design new AMPs.

Benchmark of AmpGram with cutting-edge classifiers



AmpGram uses n-grams (amino acid motifs typical of AMPs), reduced amino acid alphabets (containing information about relationships between amino acids) and random forests (a machine learning method) as an AMP classification algorithm. The positive dataset, after CD-HIT reduction (90%) (Fu et al., 2012), included 2463 experimentally validated AMP sequences retrieved from dbAMP database (Jhong et al., 2019). AmpGram applies a methodology that has already been used with success in our previous studies (Burdukiewicz et al., 2017, 2018).

Generation of the negative dataset for prediction



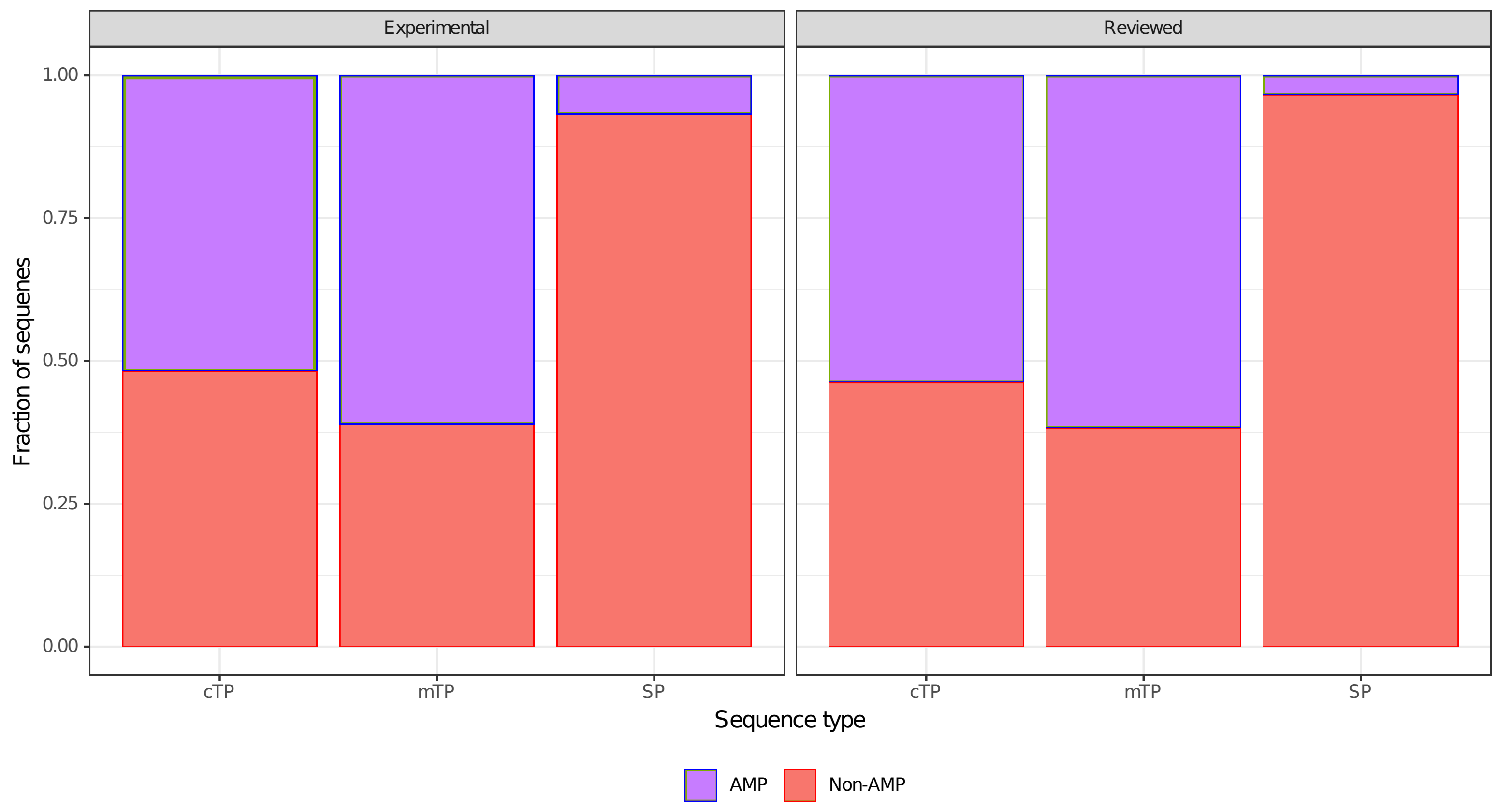
Extraction of experimentally validated, eukaryotic and cytoplasmic proteins without a transit (mTP/cTP) or signal peptide from UniProt database.

Cutting of proteins according to the length distribution in sequences from the positive dataset.

Random selection of sequences for the negative dataset.

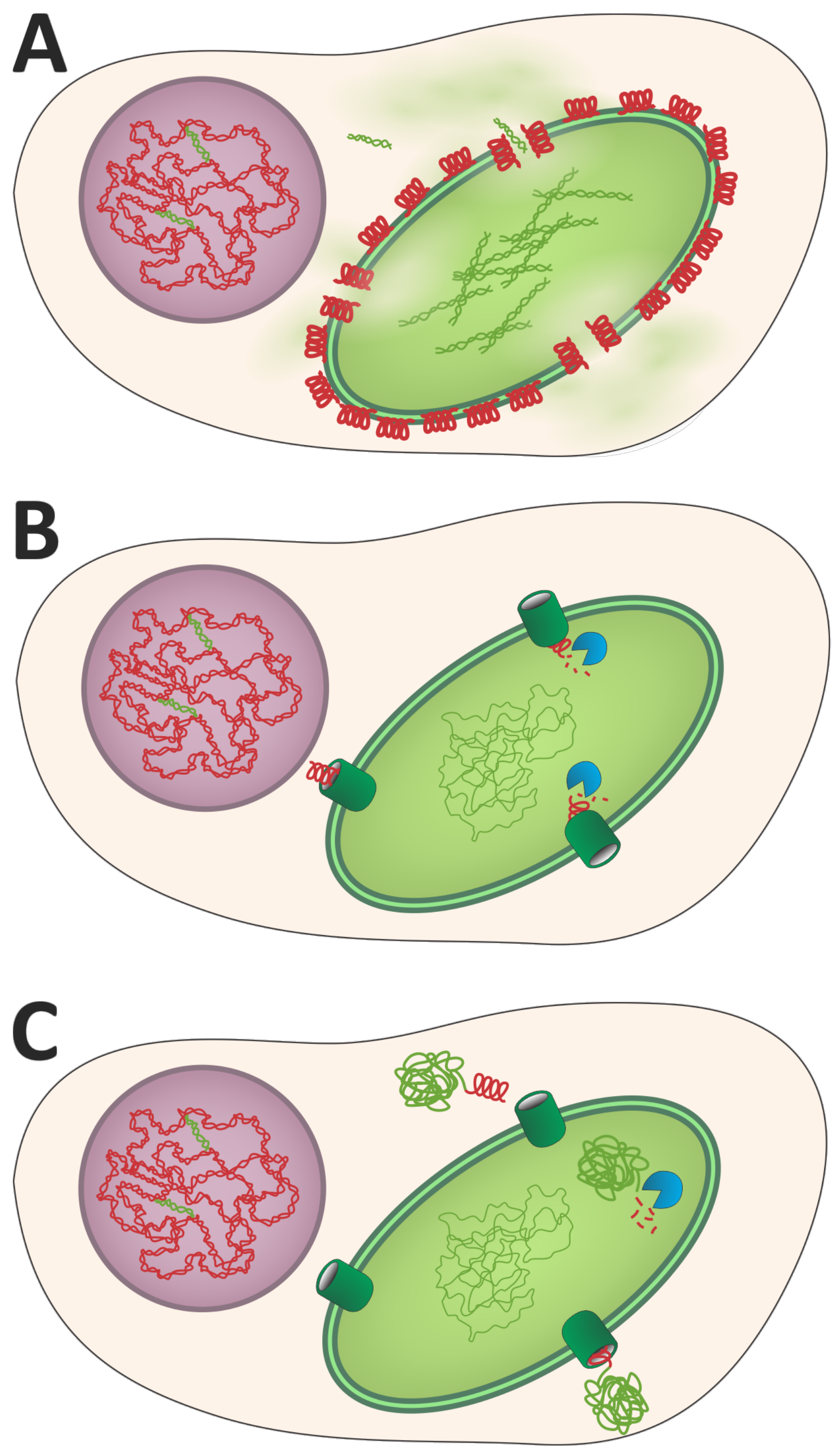
The negative dataset (2463 sequences) was build using peptides extracted from cytoplasmic proteins, which are not antimicrobial, antibacterial, antifungal and antiviral, similar to datasets presented elsewhere (Gabere and Noble, 2017) with the same length distribution as the sequences in the positive dataset.

AmpGram verification of AMP signal in mTPs, cTPs and SPs



In order to check the AMP signal in mTPs and cTPs, Viridiplantae protein sequences were retrieved from the UniProt database (UniProt Consortium, 2018) that were either experimentally verified (Experimental) or manually curated (Reviewed) to bear the appropriated presequences and localize to mitochondria and plastids, respectively. The mTPs and cTPs were next cut off from mature proteins. The final dataset included 134 cTPs. We used 295 Experimental and 4180 Reviewed SPs as a control also retrieved from the UniProt database (UniProt Consortium, 2018).

Model for evolution of protein import into bacteria-derived organellas



AMPs greatly contributed to the establishment of mitochondria and plastids by facilitating endosymbiont gene transfer as a result of bacterial cell lysis (A). At some point, ingested bacteria resisted the host AMPs by their uptake using a specific transporter and proteolytic degradation (B). In the next evolutionary stage, the host genome underwent rearrangements that led to the fusion of AMP genes with other nuclear genes. As a result, nuclear-encoded proteins acquired cleavable TPs capable of directing them into bacteria-derived organelles along with mitochondrial and plastid translocase complexes that evolved from the specific transporters (C). The cell organelles are not in scale, the nucleus is marked in purple and a bacteria-derived organelle in green.

Funding

National Science Center 2017/26/D/NZ8/00444
National Science Center 2018/31/N/NZ2/01338

Bibliography

Burdukiewicz, M., Sobczyk, P., Chilimoniuk, J., Gagat, P., and Mackiewicz, P. (2018). Prediction of Signal Peptides in Proteins from Malaria Parasites. *International Journal of Molecular Sciences*, 19(12).

Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. *Scientific Reports*, 7(1):12961.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152.

Gabere, M. N. and Noble, W. S. (2017). Empirical comparison of web-based antimicrobial peptide prediction tools. *Bioinformatics (Oxford, England)*, 33(13):1921–1929.

Jhong, J.-H., Chi, Y.-H., Li, W.-C., Lin, T.-H., Huang, K.-Y., and Lee, T.-Y. (2019). dbAMP: An integrated resource for exploring antimicrobial peptides with functional activities and physicochemical properties on transcriptome and proteome data. *Nucleic Acids Research*, 47(Database issue):D285–D297.

UniProt Consortium, T. (2018). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, 46(5):2699–2699.