Title: Overall performance and taxonomic bias of antimicrobial peptide prediction tools in invertebrates

Authors: Zoltán Rádai 1*, Johanna Kiss 2, Nikoletta A. Nagy 2

1 – Lendület Seed Ecology Research Group, Institute of Ecology and Botany, Centre for Ecological Research, Vácrátót, Hungary

2 – MTA-DE Behavioural Ecology Research Group, Department of Evolutionary Zoology and Human Biology, University of Debrecen, Debrecen, Hungary

*author of correspondence; email: zozi.web@gmail.com

Abstract

Invertebrate antimicrobial peptides (AMPs) are at the forefront in the search for agents of therapeutic utility against multi-resistant microbial pathogens, and in recent years substantial advances took place in the *in silico* prediction of antimicrobial function of amino acid sequences. A yet neglected aspect is taxonomic bias in the performance of these tools. Owing to differences in the prediction algorithms and used training data sets between tools, and phylogenetic differences in sequence diversity, physicochemical properties and evolved biological functions of AMPs between taxa, notable discrepancies may exist in performance between the currently available prediction tools. Here we tested if there is a taxonomic bias in prediction power in 10 tools with a total of 20 prediction algorithms in 19 invertebrate taxa. We found that most of the tools exhibited considerable variation in performance between tested invertebrate groups, based on which we provide guidance in choosing the best-performing prediction tool for each taxa.

Keywords: antimicrobial peptides, AMP prediction, taxonomic bias, amino acids, prediction performance

Introduction

Antimicrobial peptides (AMPs) are low molecular weight components of the innate immunity present virtually in all living organisms. Their importance has long been recognized, especially in the search for efficient therapeutic agents against multi-resistant bacteria (Toke, 2005). In the last few decades, invertebrates (mainly arthropods) have become central model organisms to identify and classify novel AMPs with potential utility in medicine (Mylonakis et al., 2016; Vizioli & Salzet, 2002). With the rapid accumulation of available genomic and proteomic data from a wide range of taxa and novel methods for the prediction of structural, physicochemical, and biological properties of peptides, the search for therapeutic-potential AMPs is one of the most exciting and promising areas in the life sciences (Lee et al., 2018; Porto et al., 2017). In fact, by today a considerable number of AMP activity prediction tools has been built, mainly utilizing modern machine-learning algorithms to identify amino acid sequences with potential antimicrobial capacities, generally on the basis of their physicochemical parameters (Lee et al., 2018; Wu et al., 2018).

Beside to other challenges of AMP prediction using peptides from databases (see Porto et al., 2020 and references therein), neglected taxonomic (and, indeed, phylogenetic) information may also contribute to biases in the training of prediction algorithms. Although AMP prediction tools are quite available, and some articles made attempts of benchmarking how efficiently these tools can identify AMPs in general (e.g. Gabere & Noble, 2017), it still remains overlooked whether or not these prediction tools may provide the same performance across the taxa from which putative AMPs are extracted. This could be rather problematic, owing to that even among phylogenetically related AMPs substantial differences may exist in the evolutionary trajectories of the amino acids between species. For example this might be expected to lead to differences in taxon-specific gene clustering (i.e. sequence diversity, affecting scalar physicochemical properties), or in the evolved functionality of peptides belonging to the given AMP family (Patil et al., 2004; Patil et al., 2005). Indeed, the importance of evolutionary origin of sequence diversity in these molecules was emphasized recently in invertebrates (Tassanakajon et al., 2015). While there is little knowledge about how such

diversity may be associated with physicochemical properties and structural features of AMPs, it is not unreasonable to assume that unaccounted taxon-specific variation in them might affect the performance of prediction tools in recognizing antimicrobial function (e.g. due to taxon-based differences in the evolved physicochemical properties). Also, alternative tools using different methodological approaches and training sets further complicate the picture and may lead to discrepant performances, i.e. to differences between the prediction tools in how efficiently they recognize AMPs and non-AMPs in different taxa, which might depend on the taxonomic composition of the training datasets. In other words, the above mentioned factors may play a role in a low generalization-efficiency of the prediction algorithms that were trained on taxonomically (phylogenetically) limited training dataset. Consequently, neglecting this aspect may lead to an increased incidence of both type I and type II errors when using inadequate tools by falsely labelling non-antimicrobial peptides as (putative) AMPs, or not recognizing true AMPs to have antimicrobial activity, respectively.

This issue is anything but trivial, as a wide range of researchers rely on these AMP prediction tools and their outputs, from the search for invertebrate AMPs with therapeutical potential (Mylonakis et al., 2016), to ecophysiological and ecoimmunological studies (Shelomi et al., 2020), and to the studies on the evolutionary history of invertebrate AMPs (Tassanakajon et al., 2015). If “taxonomic bias” exists in the prediction performance of available tools (i.e. if there are differences between prediction tools in how reliably they can distinguish between AMPs and non-AMPs in different taxa), a practical resolution (in terms of the usage of available prediction tools) could be the identification of which tool performs best for specific taxa, and consistently using these tools in the given groups. As a first approach to address this problem, the taxon-specific assessment of the performance of currently available AMP prediction tools may help us to shed light on whether or not a taxon-bias is present at all. Accordingly, in the present study our aim is not to provide a comprehensive list of AMP prediction tools, or to introduce their algorithms and/or modern methods for AMP prediction, but merely to provide guidance in choosing the best-performing

prediction tool for specific taxa. In other words: our goal here is not to present the methodology and specific composition of training data sets in detail, but to assess whether currently available tools show considerable prediction bias between different invertebrate taxa, i.e. we don't offer a specific solution to this technical problem at the level of algorithms, but highlight a yet neglected aspect of AMP prediction practice.

To do so, we acquired 1525 unique AMP and 3050 non-AMP sequences from 19 invertebrate taxa, and used 10 freely available AMP prediction tools (with a total of 20 prediction algorithms), not only to assess the prediction tools' overall performance, but their performance per taxa.

Methods

All data handling and analyses were performed in the R software for statistical computing (ver. 4.0.1, Team R Core, 2020). AMP sequences were acquired from the protein database of the National Center for Biotechnology Information (*NCBI Protein Database*, 2020; NCBI Resource Coordinators, 2018) and the Antimicrobial Peptide Database APD3 (Wang et al., 2016) (specific keywords used in the search are listed in the Supplementary Material). Taxa for which we acquired AMP sequences were Acari (n = 76), Araneae (n = 45), Blattodea (n = 162), Coleoptera (n = 62), Decapoda (n = 64), Diptera (n = 447), Hemiptera (n = 63), Hymenoptera (n = 183), Lepidoptera (n = 136), Mytiloida (n = 31), Neogastropoda (n = 5), Orthoptera (n = 8), Ostreoida (n = 62), Pectinoida (n = 4), Scolopendromorpha (n = 10), Scorpiones (n = 103), Unionida (n = 4), Veneroida (n = 48), and Xiphosura (n = 12). Non-AMP sequences were downloaded from UniProt (The UniProt Consortium, 2019): for each taxon, a large batch of non-antimicrobial proteins were acquired, between the length of 10 and 200 amino acids (for specific search expression see Supplementary Material), from which 2n for each taxon (where "n" refers to the number of acquired AMP sequences in the given taxa) were randomly drawn, without duplicate sequences. Amino acid sequences and data tables used in our study have been made available at figshare (https://figshare.com/projects/Taxonomic_bias_in_AMP_prediction/114585).

After the acquisition, both AMPs and non-AMPs were used with 10 AMP prediction tools; as some of the tools included multiple algorithms, in total there were 20 algorithms for the prediction of AMP properties for each acquired sequence (for the detailed list of the used prediction tools see the Table 1).

Table 1. List of the used prediction tools (and algorithms within some tools).

Prediction tool	Algorithm	URL	Reference
ADAM	Hidden Markov models	http://bioinformatics.cs.ntou.edu.tw/ADAM/tool.html	(Lee et al., 2015)
	Support vector machine		
AmpGram*	N-gram encoding and random forest	http://biongram.biotech.uni.wroc.pl/AmpGram/	(Burdukiewicz et al., 2020)
AMP scan 2	Deep neural network	https://www.dveltri.com/ascan/v2/ascan.html	(Veltri et al., 2018)
$CAMP_{R3}$	Artificial neural network	http://www.camp.bicnirrh.res.in/predict/	(Waghu et al., 2016)
	Discriminant analysis		
	Random forest		
	Support vector machine		
Class AMP	Random forest	http://www.bicnirrh.res.in/classamp/predict.php	(Joseph et al., 2012)
	Support vector machine		
CS-AMPpred*	Support vector machine: linear kernel	https://sourceforge.net/projects/csamppred/	(Porto et al., 2012)
	Support vector machine: polynomial kernel		
	Support vector machine: radial kernel		
iAMP	Support vector machine	http://cabgrid.res.in:8080/amppred/server.php	(Meher et al., 2017)
iAMP-2L	Multi-label fuzzy K-nearest neighbour	http://www.jci-bioinfo.cn/bioinfo/iAMP-2L	(Xiao et al., 2013)
IAMPE	K-nearest neighbour	http://cbb1.ut.ac.ir/AMPClassifier/Index	(Kavousi et al., 2020)
	Support vector machine		
	Random forest		
	Extreme gradient boosting (XGBoost)		
StM	Hydrophobic momentum threshold	http://www.portoreports.com/stm	(Porto et al., 2020)

*Used on local computer, not on web-server.

In the case of prediction tools providing a probability for a given amino acid sequence to be an AMP, we used the conventional 0.5 threshold (i.e. sequences with >50% probability were considered to be classified as AMP). Results from the prediction tools were evaluated using the typical binary classification performance measures:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$
$$Presicion = \frac{TP}{TP + FP}$$
$$Sensitivity = \frac{TP}{TP + FN}$$
$$Specificity = \frac{TN}{TN + FP}$$

where TP, TN, FP, and FN stand for True Positive, True Negative, False Positive, and False Negative prediction, respectively. We also calculated the F_1 score:

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} = \frac{2}{Sensitivity^{-1} + Precision^{-1}}$$

Tool performances were also assessed separately for taxa. Additionally, taxonomic bias in prediction power of tools was calculated as the median distance between F_1 scores within taxa, separately for each prediction tool.

Results and Discussion

Overall, ADAM showed the best performance with both of its algorithms (hidden Markov model, and support vector machine). Also, the artificial neural network classifier of $CAMP_{R3}$ had relatively good performance, similarly to CS-AMPpred (all kernels, although the linear kernel did not perform as good as the polynomial and radial), and to iAMP-2L (Fig. 1). The weakest overall performance was observed in the case of StM.

AMP prediction tools showed substantial variation in how well they performed in different taxa (Fig. 1; also see Supplementary Material Fig. S1). Class AMP and IAMPE classifiers showed almost no between-taxa variation, which was likely due to their poor performance, as they almost unanimously predicted AMP activity in all (even non-AMP) amino acid sequences. The largest taxonomic bias (i.e. largest median difference in tool performance between taxa) was observed in iAMP-2L and in the linear kernel CS-AMPpred; in these tools the discrepancy between the minimum and maximum F_1 score was 0.46, and 0.32, respectively. Even those prediction tools showing best performances based on their F_1 scores had marked taxonomic biases: for example in the hidden Markov model ADAM the difference between minimum and maximum F_1 score was 0.10. Still, based on the F_1 values calculated per taxa, this classifier yielded the highest performance for most of the taxa (Table 2).

Table 2. Best preforming prediction tools for given taxa based on F_1.

Taxon	Best performing prediction tool	F_1
Acari	ADAM (hidden Markov model)	0.86
Araneae	ADAM (support vector machine)	0.93
Blattodea	ADAM (support vector machine)	0.86
Coleoptera	ADAM (hidden Markov model)	0.91
Decapoda	iAMP-2L	0.82
Diptera	ADAM (hidden Markov model)	0.95
Hemiptera	ADAM (hidden Markov model)	0.88
Hymenoptera	ADAM (hidden Markov model)	0.90
Lepidoptera	ADAM (hidden Markov model)	0.92
Mytiloida	iAMP-2L	0.82
Neogastropoda	ADAM (hidden Markov model)	0.73
Orthoptera	$CAMP_{R3}$ (artificial neural network)	0.89
Ostreoida	ADAM (hidden Markov model)	0.85
Pectinoida	ADAM (hidden Markov model)	0.80
Scolopendromorpha	CS-AMPpred (linear kernel)	0.80
Scorpiones	iAMP-2L	0.78
Unionida	$CAMP_{R3}$ (artificial neural network)	0.89
Veneroida	ADAM (hidden Markov model)	0.77
Xiphosura	ADAM (support vector machine)	0.88

Notably, in most cases taxonomic biases were tool specific, meaning that there were only a few taxa for which most of the prediction tools yielded consistently low or high performance (see Supplementary Material Fig. S2). Specifically, we found negative bias for Veneroida, Unionida, and Neogastropoda (meaning that for these groups F_1 scores were consistently low across prediction tools), and positive bias for Xiphosura and Araneae (meaning that for these groups F_1 scores were consistently high across prediction tools). In taxa for which sample sizes were small (such as Unionida and Neogastropoda) this bias might simply be an artefact of chance, as any given prediction affects the estimated performance score much stronger than in taxa of larger sample sizes. However, in Veneroida and Araneae the sample sizes were not quite small, and in other small-sample-sized taxa no such pattern was observed. At this point it is difficult to find a satisfactory explanation for this observation. Given that most of the tested tools rely on amino acid sequence-based physicochemical measures, one possible cause might relate to the similarity of amino acid sequences in a given taxon to those used in the training sets of prediction tools. For example, even if a given taxon is not represented in a tool's training set, similar physicochemical properties to the training sequences likely helps better recognition by the given tool. On the other hand, if physicochemical properties of a given taxon are significantly different from those taxa included in

the training set, it will likely lead to poorer performance for that given query taxon not represented in the set. This should be addressed in future studies to clarify what mechanisms can lead to such biases in prediction performance.

The prediction tool ADAM using the hidden Markov model provided the best performance in most assessed taxa. While it is out of scope of this study to give in-detail insights on the algorithms behind the prediction tools, it should be noted that this tool's strength likely comes from its set of homologues: it uses a probabilistic model based on which the already known AMPs are used to find the most likely homologue for the query peptide. This is a strength that few currently available prediction tools possess, namely that, instead of only using scalar physicochemical properties based on the amino acid sequence (which approach has its weaknesses, see Porto et al., 2020), sequences are classified based on their resemblance to known AMPs. Notably, such an approach may yield null-findings when presented a previously unknown AMP, but might potentially enable for accounting for evolutionary information. Although the hidden Markov model ADAM tool technically does not consider phylogeny for the taxa from which the query sequences originate, in future prediction tools homologue (and, indeed orthologue) finding may be a quite useful approach in the reliable *in silico* identification of AMPs. Surprisingly, tools accounting for (either species, or gene) evolution are still lacking, and studies addressing the question of functional typization of AMPs generally use amino acid composition (Porto et al., 2020), and/or quantitative structure–activity relationships (Juretić et al., 2011). Whilst indisputably understanding the association between sequence, structure, and biological function is the forefront in the study of AMPs with therapeutical potential (Ahmed & Hammami, 2019; Torres et al., 2018), evolutionary information may also help us in the identification (and, indeed, *de novo* design and synthesis) of candidate AMPs. Arguably, a taxonomically, or even phylogenetically informed perspective would be quite useful not only in the more robust and reliable identification of AMPs in invertebrates, but also in designing and engineering peptides with desirable properties (e.g. efficacy against antibiotic resistant bacteria, but low citotoxicity to eukaryotic cells), through a deeper understanding of the

evolution of the wide variety of AMPs evolved in invertebrates. Hence, comparative phylogenetic analyses of AMPs would be an important aspect helping to understand the evolution of these molecules.

Substantial variation exists in how reliably given tools can predict AMP function of amino acid sequences belonging to different taxa. From a practical perspective this gives reason for caution when selecting the prediction tool for identifying putative AMPs from amino acid sequences of a given invertebrate taxon *in silico*. The substantial taxonomic bias in the different prediction tools' performance may hint at important differences in the properties of amino acid sequences across taxa, but interpretation should be done with care, as this bias can have multiple origins. For example, various AMP families (e.g. defensins, cecropins) may have quite different physicochemical properties (e.g. Cytryńska et al., 2007), and there may be substantial between-taxa variation in amino acid sequences of same AMP family (Tassanakajon et al., 2015). Although these aspects should be taken into account when working with AMPs, it is outside of the scope of our present study to go into details of the background and their possible resolutions.

It should also be noted that our current approach is not without drawbacks. Firstly, the considerably uneven availability of AMPs for different taxa renders it difficult to consider our results as robust for all the assessed taxa. As noted before, in taxa with small sample sizes (e.g. Neogastropoda and Unionida) F_1 scores and prediction successes are more likely to be subjected to distortions by random chance than in taxa of large sample sizes. Still, we think that as a practical guide our results are still of value even for taxa with small sample sizes, although with more caution. Secondly, one could propose that overlap between our test data set and training sets of prediction tools might also bias our performance estimates. However, we found no evidence for that, since neither the number nor the proportion of common amino acid sequences in our and the used tools' data sets showed significant association with F_1 scores (see Supplementary Material Fig. S3-S4).

As previously mentioned, the different tools use different methodological approaches (e.g. CS-AMPpred uses support vector machine trained on 310 cysteine-stabilized peptides, utilizing scalar

physicochemical properties, whereas ADAM hidden Markov model is based on profile models of more than 7000 AMPs and 759 identified structures). Differences in the focus of AMP types undoubtedly restricts the potential range of peptides that can be recognised as to have antimicrobial function by a given prediction tool. It is also worth mentioning that taxonomic composition of the AMP and non-AMP sequence datasets used to train the different models might also shape the models' performance, in terms of whether or not a tested putative AMP sequence comes from a taxon included in the training set. Unexpectedly, there was a weak marginal trend that when a query taxon was present in the training set of a tool, then F_1 scores tended to be slightly smaller (see Supplementary Material Fig. S5). In general, however, there was no significant association of the number and evenness of taxonomic groups in the training data sets with the tools' F_1 scores (Supplementary Material Fig. S6).

Additionally, due to the variable availability of AMP sequences for the different taxa, AMP family could not have been included when testing performance, even though there might be tool-specific differences in it for the different AMP families (i.e. a given tool may differ in its capacity to recognize defensins *versus* cecropins as AMPs). Notably, we know about no studies explicitly assessing the effect of AMP family on the efficiency of AMP prediction, therefore this aspect might be a quite relevant and important one to address in future studies. Furthermore, owing to practical reasons, non-AMPs (used both for training prediction tools, or for the assessment of their performance) are not standardized, i.e. these peptides are generally drawn randomly from those peptides that are available for the given taxa, with no known functions in antimicrobial immunity. Non-overlapping biological functions between peptides from different taxa might introduce additional bias in the performance of prediction tools. For example, some membrane-active peptides might be expected to be falsely recognized as AMPs more often than others, due to their structural and physicochemical properties similar to that of AMPs (Lee et al., 2018). Hence, if the randomly drawn non-AMPs in a given taxon contain a higher proportion of such membrane peptides than in other taxa, this might confound the per-taxa-performance of the used prediction tool. Using

standardized sets of non-AMP protein families (preferably of several, distinct biological functions) that have available sequence data for a wide range of taxa might help improve the performance of new prediction tools.

Overall, the results presented here highlight the necessity of taking taxonomic information into account when predicting antimicrobial function of putative invertebrate AMPs. At this point, although arguably crucial, finding the specific background of taxonomic bias (e.g. composition of training sets, or characteristics of prediction algorithms) in prediction performance of the assessed tools is outside the scope of our paper, and we do not offer technical details or solutions. We merely aimed to highlight an overlooked phenomenon, and to provide a practical guide to use the currently available and most popular tools. We argue that our results will serve as a useful guide in choosing the appropriate AMP prediction tool for testing candidate peptides in invertebrates. This could be useful for the big-data / bioinformatic data style search for candidates in open databases and repositories, and for the identification of AMPs in invertebrates for ecophysiological studies. In the search for putative AMPs based on genomic and/or proteomic data, and immune profiling based on transcriptomics (i.e. screening produced peptides using transcriptome) in invertebrates may also benefit from our proposed approach. In addition, future studies to implement refined AMP prediction techniques, inclusion of taxonomic and/or phylogenetic information, utilization of explicit categorization of AMPs, and standardized sets of non-AMP sequences will undoubtedly help us to achieve more robust prediction of antimicrobial function in invertebrate models throughout our endeavour to identify peptides of potential therapeutical utility.

References

Ahmed, T. A. E., & Hammami, R. (2019). Recent insights into structure–function relationships of antimicrobial peptides. *Journal of Food Biochemistry*, *43*(1), e12546. https://doi.org/10.1111/jfbc.12546

Burdukiewicz, M., Sidorczuk, K., Rafacz, D., Pietluch, F., Chilimoniuk, J., Rödiger, S., & Gagat, P. (2020). Proteomic Screening for Prediction and Design of Antimicrobial Peptides with AmpGram. *International Journal of Molecular Sciences*, *21*(12), 4310. https://doi.org/10.3390/ijms21124310

Cytryńska, M., Mak, P., Zdybicka-Barabas, A., Suder, P., & Jakubowicz, T. (2007). Purification and characterization of eight peptides from Galleria mellonella immune hemolymph. *Peptides*, *28*(3), 533–546. https://doi.org/10.1016/j.peptides.2006.11.010

Gabere, M. N., & Noble, W. S. (2017). Empirical comparison of web-based antimicrobial peptide prediction tools. *Bioinformatics*, *33*(13), 1921–1929. https://doi.org/10.1093/bioinformatics/btx081

Joseph, S., Karnik, S., Nilawe, P., Jayaraman, V. K., & Idicula-Thomas, S. (2012). ClassAMP: A Prediction Tool for Classification of Antimicrobial Peptides. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *9*(5), 1535–1538. https://doi.org/10.1109/TCBB.2012.89

Juretić, D., Vukičević, D., Petrov, D., Novković, M., Bojović, V., Lučić, B., Ilić, N., & Tossi, A. (2011). Knowledge-based computational methods for identifyingor designing novel, non-homologous antimicrobial peptides. *European Biophysics Journal*, *40*, 371–385. https://doi.org/10.1007/s00249-011-0674-7

Kavousi, K., Bagheri, M., Behrouzi, S., Vafadar, S., Atanaki, F. F., Lotfabadi, B. T., Ariaeenejad, S., Shockravi, A., & Moosavi-Movahedi, A. A. (2020). IAMPE: NMR-Assisted Computational Prediction of Antimicrobial Peptides. *Journal of Chemical Information and Modeling*, *60*(10), 4691–4701. https://doi.org/10.1021/acs.jcim.0c00841

Lee, E. Y., Wong, G. C. L., & Ferguson, A. L. (2018). Machine learning-enabled discovery and design of membrane-active peptides. *Bioorganic & Medicinal Chemistry*, *26*(10), 2708–2718. https://doi.org/10.1016/j.bmc.2017.07.012

Lee, H.-T., Lee, C.-C., Yang, J.-R., Lai, J. Z. C., & Chang, K. Y. (2015, April 27). *A Large-Scale Structural Classification of Antimicrobial Peptides* [Resource Review]. BioMed Research International; Hindawi. https://doi.org/10.1155/2015/475062

Meher, P. K., Sahu, T. K., Saini, V., & Rao, A. R. (2017). Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Scientific Reports*, *7*(1), 42362. https://doi.org/10.1038/srep42362

Mylonakis, E., Podsiadlowski, L., Muhammed, M., & Vilcinskas, A. (2016). Diversity, evolution and medical applications of insect antimicrobial peptides. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1695), 20150290. https://doi.org/10.1098/rstb.2015.0290

NCBI Protein database. (2020). Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/protein/

NCBI Resource Coordinators. (2018). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *46*(D1), D8–D13. https://doi.org/10.1093/nar/gkx1095

Patil, A. A., Cai, Y., Sang, Y., Blecha, F., & Zhang, G. (2005). Cross-species analysis of the mammalian β-defensin gene family: Presence of syntenic gene clusters and preferential expression in the male reproductive tract. *Physiological Genomics*, *23*(1), 5–17. https://doi.org/10.1152/physiolgenomics.00104.2005

Patil, A., Hughes, A. L., & Zhang, G. (2004). Rapid evolution and diversification of mammalian α-defensins as revealed by comparative analysis of rodent and primate genes. *Physiological Genomics*, *20*(1), 1–11. https://doi.org/10.1152/physiolgenomics.00150.2004

Porto, W. F., Pires, A. S., & Franco, O. L. (2017). Computational tools for exploring sequence databases as a resource for antimicrobial peptides. *Biotechnology Advances*, *35*(3), 337–349. https://doi.org/10.1016/j.biotechadv.2017.02.001

Porto, W., Ferreira, K. C. V., Ribeiro, S. M., & Franco, O. L. (2020). Sense the Moment: A highly sensitive antimicrobial activity predictor based on hydrophobic moment. *BioRxiv*. https://doi.org/10.1101/2020.07.15.205419

Porto, William F., Pires, Á. S., & Franco, O. L. (2012). CS-AMPPred: An Updated SVM Model for Antimicrobial Activity Prediction in Cysteine-Stabilized Peptides. *PLOS ONE*, *7*(12), e51444. https://doi.org/10.1371/journal.pone.0051444

Shelomi, M., Jacobs, C., Vilcinskas, A., & Vogel, H. (2020). The unique antimicrobial peptide repertoire of stick insects. *Developmental & Comparative Immunology*, *103*, 103471. https://doi.org/10.1016/j.dci.2019.103471

Tassanakajon, A., Somboonwiwat, K., & Amparyup, P. (2015). Sequence diversity and evolution of antimicrobial peptides in invertebrates. *Developmental & Comparative Immunology*, *48*(2), 324–341. https://doi.org/10.1016/j.dci.2014.05.020

Team R Core. (2020). *R: A language and environment for statistical computing* (3.6.3) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

The UniProt Consortium. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(D1), D506–D515. https://doi.org/10.1093/nar/gky1049

Toke, O. (2005). Antimicrobial peptides: New candidates in the fight against bacterial infections. *Peptide Science*, *80*(6), 717–735. https://doi.org/10.1002/bip.20286

Torres, M. D. T., Pedron, C. N., Higashikuni, Y., Kramer, R. M., Cardoso, M. H., Oshiro, K. G. N., Franco, O. L., Silva Junior, P. I., Silva, F. D., Oliveira Junior, V. X., Lu, T. K., & de la Fuente-Nunez, C. (2018). Structure-function-guided exploration of the antimicrobial peptide polybia-CP identifies activity determinants and generates synthetic therapeutic candidates. *Communications Biology*, *1*(1), 1–16. https://doi.org/10.1038/s42003-018-0224-2

Veltri, D., Kamath, U., & Shehu, A. (2018). Deep learning improves antimicrobial peptide recognition. *Bioinformatics*, *34*(16), 2740–2747. https://doi.org/10.1093/bioinformatics/bty179

Vizioli, J., & Salzet, M. (2002). Antimicrobial peptides from animals: Focus on invertebrates. *Trends in Pharmacological Sciences*, *23*(11), 494–496. https://doi.org/10.1016/S0165-6147(02)02105-3

Waghu, F. H., Barai, R. S., Gurung, P., & Idicula-Thomas, S. (2016). CAMPR3: A database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Research*, *44*(D1), D1094-1097. https://doi.org/10.1093/nar/gkv1051

Wang, G., Li, X., & Wang, Z. (2016). APD3: The antimicrobial peptide database as a tool for research and education. *Nucleic Acids Research*, *44*(D1), D1087–D1093. https://doi.org/10.1093/nar/gkv1278

Wu, Q., Ke, H., Li, D., Wang, Q., & Zhou*, J. F. and J. (2018, December 31). *Recent Progress in Machine Learning-based Prediction of Peptide Activity for Drug Discovery*. Current Topics in Medicinal Chemistry. https://www.eurekaselect.com/169247/article

Xiao, X., Wang, P., Lin, W.-Z., Jia, J.-H., & Chou, K.-C. (2013). iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Analytical Biochemistry*, *436*(2), 168–177. https://doi.org/10.1016/j.ab.2013.01.019

Figures

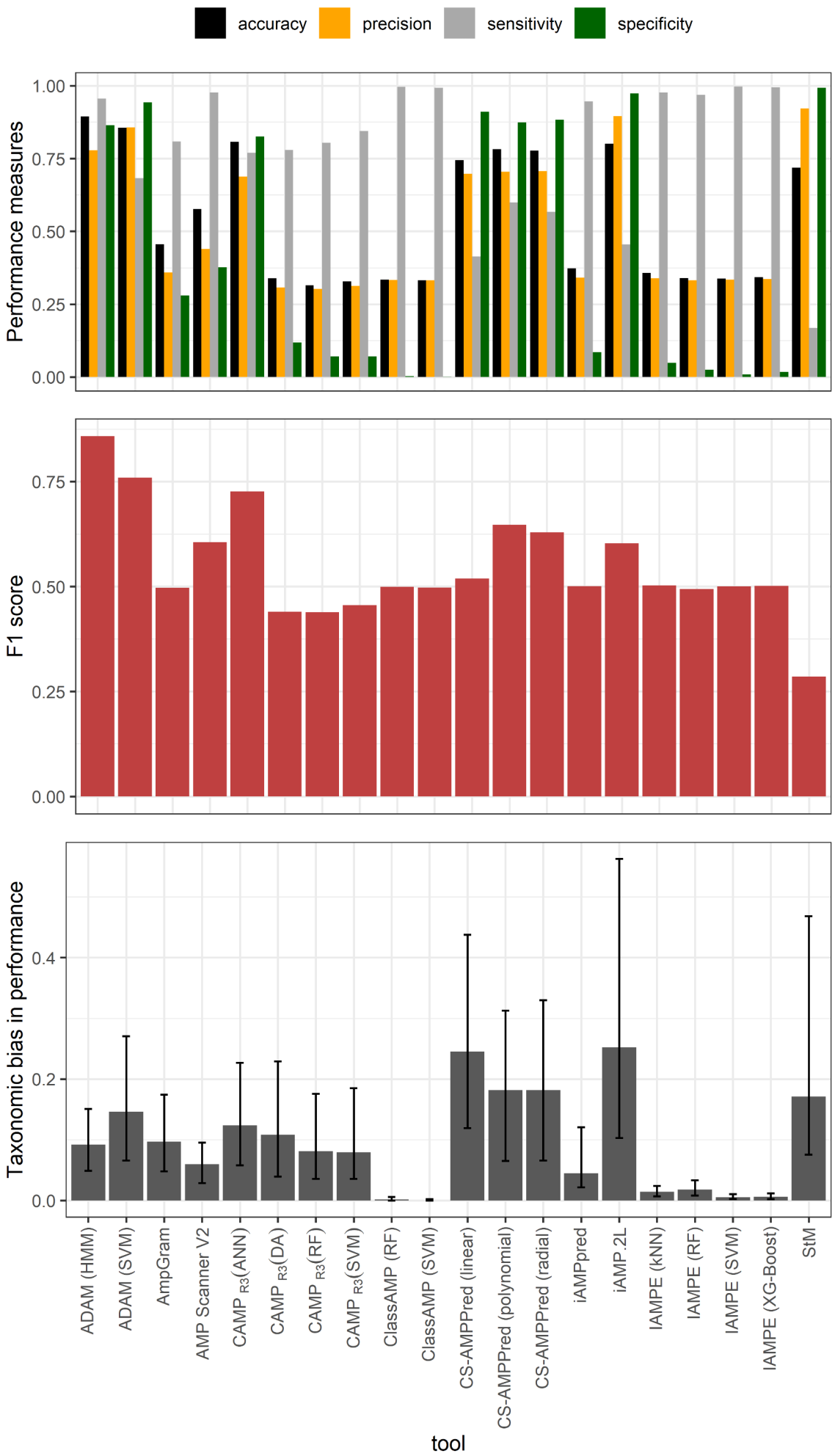
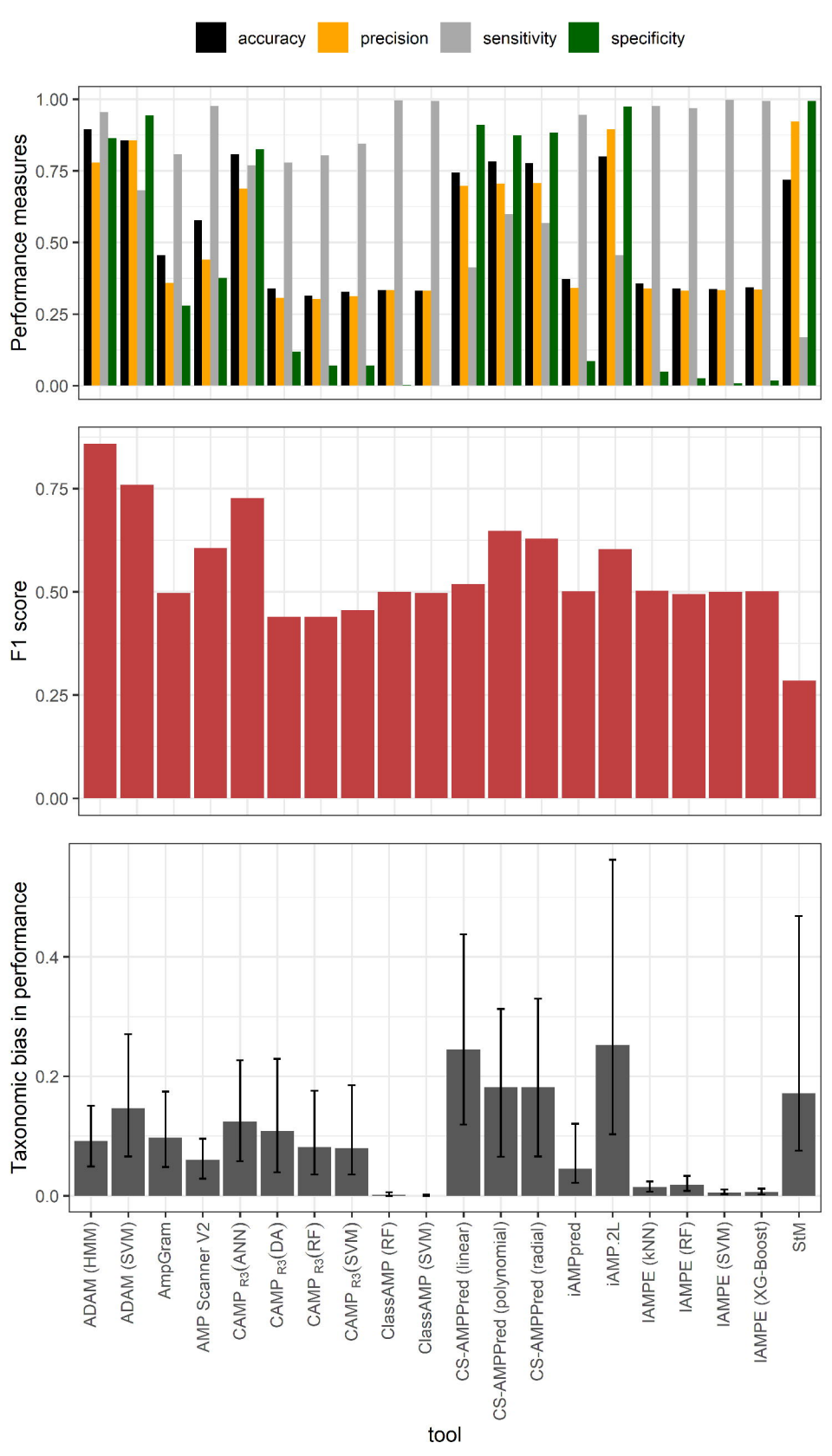

accuracy
precision
sensitivity
specificity
Performance measures
1.00
0.75
0.50
0.25
0.00
F1 score
0.75
0.50
0.25
0.00
Taxonomic bias in performance
0.4
0.2
0.0
ADAM (HMM)
ADAM (SVM)
AmpGram
AMP Scanner V2
CAMP R3(ANN)
CAMP R3(DA)
CAMP R3(RF)
CAMP R3(SVM)
ClassAMP (RF)
ClassAMP (SVM)
CS-AMPPred (linear)
CS-AMPPred (polynomial)
CS-AMPPred (radial)
iAMPpred
iAMP.2L
IAMPE (kNN)
IAMPE (RF)
IAMPE (SVM)
IAMPE (XG-Boost)
StM
tool

Figure 1. Performance measures (top panel), F_1 scores (middle panel), and taxonomic bias in performance assessed as median of the differences in F_1 scores between taxa, within a given prediction tool (see Methods for details; bottom panel); black segments on the bottom panel represent the value range between the first and third quartiles.


accuracy
precision
sensitivity
specificity
Performance measures
1.00
0.75
0.50
0.25
0.00
F1 score
0.75
0.50
0.25
0.00
Taxonomic bias in performance
0.4
0.2
0.0
ADAM (HMM)
ADAM (SVM)
AmpGram
AMP Scanner V2
CAMP R3(ANN)
CAMP R3(DA)
CAMP R3(RF)
CAMP R3(SVM)
ClassAMP (RF)
ClassAMP (SVM)
CS-AMPPred (linear)
CS-AMPPred (polynomial)
CS-AMPPred (radial)
iAMPpred
iAMP 2L
IAMPE (kNN)
IAMPE (RF)
IAMPE (SVM)
IAMPE (XG-Boost)
StM
tool