MACREL: antimicrobial peptide screening in genomes and metagenomes

Célio Dias Santos-Junior [1,2], **Shaojun Pan** [1,2], **Xing-Ming Zhao** [1,2], and **Luis Pedro Coelho** [1,2,*]

[1]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, 200433, China
[2]Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Ministry of Education, China
[*]to whom correspondence should be addressed: luispedro@big-data-biology.org

ABSTRACT

Motivation: Antimicrobial peptides (AMPs) have the potential to tackle multidrug-resistant pathogens in both clinical and non-clinical contexts. The recent growth in the availability of genomes and metagenomes provides an opportunity for *in silico* prediction of novel AMPs. However, due to the small size of these peptides, standard gene prospection methods cannot be applied in this domain and alternative approaches are necessary. In particular, standard gene prediction methods have low precision for short peptides, and functional classification by homology results have low recall.
Results: Here, we present a novel set of 22 peptide features. These were used to build classifiers which perform similarly to the state-of-the-art in the prediction of both antimicrobial and hemolytic activity of peptides, but with enhanced precision (using standard benchmarks). We use these classifiers to build MACREL—Meta(genomic) AMPs Classification and REtrievaL—an end-to-end tool which combines assembly, ORF prediction, and AMP classification to extract AMPs directly from genomes or metagenomes. We demonstrate that MACREL recovers high-quality AMP candidates from genomes and metagenomes using realistic simulations and real data.
Availability: MACREL is implemented in Python. It is available as open source at https://github.com/BigDataBiology/macrel and through bioconda. Classification of peptides or prediction of AMPs in contigs can also be performed on the webserver: http://big-data-biology.org/software/macrel.
Supplementary information: Supplementary data are available online.

1 Introduction

Antimicrobial peptides (AMPs) are short proteins (containing fewer than 100 amino acids) that can decrease or inhibit bacterial growth. They interact with microbial membranes or intracellular targets (Zhang and Gallo, 2016) and have remained potent for millions of years (Zasloff, 2002). Given the dearth of novel antibiotics in recent decades and the rise of antimicrobial resistance, prospecting naturally-occurring AMPs is a potentially valuable new source of antimicrobial molecules (Theuretzbacher et al., 2019). The increasing number of publicly available metagenomes and metatranscriptomes presents an opportunity to use them for finding novel AMP sequences. However, methods that have been successful in prospecting other microbial functionality, cannot be directly applied to small genes (Saghatelian and Couso, 2015), such as AMPs. In particular, there are two major computational challenges: the prediction of small genes in DNA sequences (either genomic or metagenomic contigs) and the prediction of AMP activity for small genes.

Current automated gene prediction methods typically exclude small open reading frames (smORFs) (Miravet-Verde et al., 2019), as the naïve use of the methods that work for larger sequences leads to unacceptably high rates of false positives when extended to short sequences (Hyatt et al., 2010). A few recent large-scale smORFs surveys have, nonetheless, shown that these methods can be employed if the results are subsequently filtered while revealing that prokaryotic smORFs are biologically active across a range of functions (Miravet-Verde et al., 2019; Sberro et al., 2019).

Similarly, the prediction of AMP activity requires different techniques than the homology-based methods that are applicable for longer proteins (Huerta-Cepas et al., 2017). In this context, several machine learning-based methods have demonstrated high accuracy in predicting antimicrobial activity in peptides, when tested on curated benchmarks (Xiao et al., 2013; Meher et al., 2017; Lata et al., 2010; Thakur et al., 2012; Sharma et al., 2016; Bhadra et al., 2018). However, to be applicable to the task of extracting AMPs from genomic data, an AMP classifier needs to be robust to gene mispredictions and needs to be benchmarked

in that context.

Here, we present MACREL (for *Meta(genomic) AMPs Classification and REtrievaL*, see Fig. 1), a simple, yet accurate, pipeline that processes either genomes (in the form of pre-assembled contigs) or metagenomes/metatranscriptomes (in the form of short reads) and predicts AMP sequences. We test MACREL with standard benchmarks in AMP prediction as well as both simulated and real sequencing data to show that, even in the presence of large numbers of (potentially artifactual) input smORFs, MACREL still outputs only a small number of high-quality candidates.

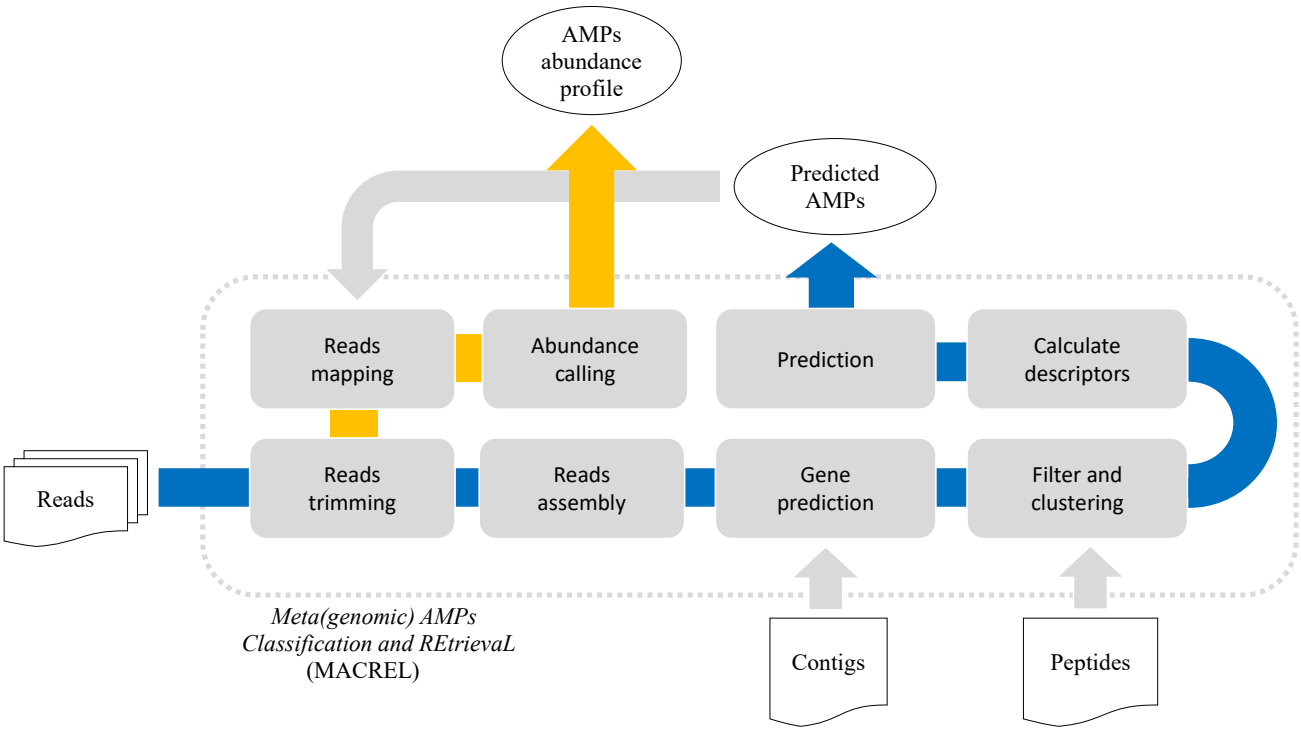

Figure 1. Meta(genomic) AMPs Classification and REtrievaL: MACREL pipeline. The blue arrows show the MACREL workflow from the processing of reads until AMP prediction. The user can also provide as input contigs or peptide sequences, if they are available. The yellow arrow shows the abundance profiling of AMPs using MACREL output and reads. Gray arrows show the alternative inputs accepted by MACREL.

2 System and Methods

2.1 MACREL Classifiers

Two binary classifiers are used in MACREL: one predicting AMP activity and other predicting hemolytic activity (which is invoked for putative AMPs). These are feature-based classifiers and use the same 22 descriptors.

2.1.1 Features

AMPs typically contain approximately 50% hydrophobic residues, usually positively charged and fold in a well-defined secondary structure (Zhang and Gallo, 2016). The peptide charge appears to be a key feature in the formation of amphiphilic ordered structures (Malmsten, 2014; Brogden, 2005; Pasupuleti et al., 2012; Hancock and Sahl, 2006; Shai, 2002; Strömstedt et al., 2009), which promote peptide-induced cell membrane disruption (Malmsten, 2014; Pasupuleti et al., 2012; Ringstad et al., 2006). These sequences can be predicted using local or global features (local features depend on the order of the amino-acids, while global ones do not). Local and global features are particularly valuable in different AMP-related prediction problems. Local features are more informative when predicting AMP activity, while global features are more informative when predicting the efficiency of an AMP (Bhadra et al., 2018; Fjell et al., 2009; Boone et al., 2018). Thus, MACREL combined both, and include 16 global and 6 local features (see Suppl. Table S1).

Bhadra et al. (2018) produced an accurate classifier of AMPs based on random forests using the composition-transition-distribution of amino acid residues (Dubchak et al., 1995, 1999) according to 7 physiochemical properties such as hydrophobicity, polarity, polarizability, and secondary structure. Therefore, the 6 local context descriptors adopted in the MACREL classifiers consisted of the amino acid distribution patterns of solvent accessibility and free energy to transfer from a relaxed coil in water to an ordered structure in the membrane, the last one being a novel descriptor proposed here and described in detail in Section 3.2.

Features used by MACREL classifiers (see Suppl. Table S1) also included the distribution at first residue of 3 amino acid classes organized according to the solvent accessibility (Bhadra et al., 2018) and extra variables, such as solubility. Solubility is a remarkable feature of AMPs (Fan et al., 2016; Wenzel et al., 2014), which was represented by an indirect measure—the isoelectric potential. AMPs usually have higher hydrophobicity, aliphatic index, and lower instability index when compared to typical proteins (Jhong et al., 2019). Thus, those variables were included in our set of descriptors.

MACREL models used other features related to the activity of AMPs, such as charge and percent composition of apolar residues (aromatic and aliphatic) (Nagarajan et al., 2019). The mechanism of anti-microbial activity was summarized in MACREL classifiers as the predisposition of a peptide to bing to membranes and its amphiphilicity (Boman index and hydrophobic moment, respectively). Additionally, MACREL classifiers used the percent composition of different amino acid groups (acidic, basic, charged, polar, non-polar, tiny and small) as AMPs have been shown to have a characteristic composition (Jhong et al., 2019; Nagarajan et al., 2019).

2.1.2 MACREL prediction models

For AMP prediction, our training set is adapted from the one presented by Bhadra et al. (2018) by eliminating redundant sequences. The resulting set contains 3,268 AMPs (from diverse databases, most bench-validated) and 166,182 non-AMPs (a ratio of approximately 1:50).

Tests comparing different AMPs classifiers showed that random forest (RF) classifiers achieved better performance than the alternatives (including Support vector machines and Bagged Foreststs, see Suppl. Table S2), as previously reported (Fernández-Delgado et al., 2014; Bhadra et al., 2018; Waghu et al., 2014, 2016). We tested this RF classifier built with a different number of trees (100, 200, or 500), and adopted 100 trees after a slight deterioration of accuracy with larger forests (see Suppl. Table S2).

The hemolytic activity classifier was built similarly to AMP classifier. For this, we used the training set HemoPI-1 from Chaudhary et al. (2016), which contains by 442 hemolytic and 442 non-hemolytic peptides.

2.1.3 Prediction in (meta)genomes

MACREL (see Fig. 1) accepts as inputs metagenomic paired-end or single-end reads in compressed FastQ format and performs quality-based trimming with NGLess (Coelho et al., 2019). After this initial stage, MACREL assembles contigs using MEGAHIT (Li et al., 2016) (a minimum length of 1,000 base pairs is used). Alternatively, if available, contigs can be passed directly to MACREL.

Genes are predicted on these contigs with a modified version of Prodigal (Hyatt et al., 2010), which predicts genes with a minimal length of 30 base pairs (compared to 90 base pairs in the standard Prodigal release). The original threshold was intended to minimize false positives (Hyatt et al., 2010), as gene prediction methods, in general, generate more false positives in shorter sequences (small ORFs, henceforth smORFs) (Höps et al., 2018). Sberro et al. (2019) showed that reducing the length threshold without further filtering could lead to as many as 61.2% of predicted smORFs being false positives. In MACREL, this filtering consists of outputting only those smORFs (10-100 amino acids) classified as AMPs.

AMP sequences are classified according to their hemolytic activity and classified into four different families by composition (cationic or anionic) and structure (linear or disulfide bond-forming). For convenience, duplicated sequences can be clustered and output as a single entity. For calculating AMP abundance profiles, MACREL uses Paladin (Westbrook et al., 2017) and NGLess (Coelho et al., 2019).

2.2 Benchmarking

2.2.1 Methods to be compared

We benchmark two AMP MACREL classifiers: the standard one (denoted as MACREL), built with the training set adapted from Bhadra et al. (2018) (see Section 2.1.2), and a second one (denoted MACREL^X), which was built using the same features and methods as MACREL, but using the training set from Xiao et al. (2013), which contains 770 AMPs and 2405 non-AMPs.

Both models were compared to the following state-of-art methods: AmPEP (Bhadra et al., 2018), CAMPR3 (including all algorithms) (Waghu et al., 2016), iAMP 2-L (Xiao et al., 2013), AMAP (Gull et al., 2019), iAMPpred (Meher et al., 2017) and Antimicrobial Peptides Scanner v2 (Veltri et al., 2018). For all these comparisons, we used the benchmark dataset from Xiao et al. (2013), which contains 920 AMPs and 920 non-AMPs, and the models publicly available from other methods (Meher et al., 2017; Bhadra et al., 2018; Gabere and Noble, 2017).

The datasets from (Xiao et al., 2013) do not overlap. However, the training set used in MACREL and the test set from Xiao et al. (2013) do overlap extensively. Therefore, for testing, we used the out-of-bag estimate for any sequences that were present in the training set.

The benchmarking of the hemolytic peptides classifier was performed using the HemoPI-1 benchmark dataset formed by 110 hemolytic proteins and 110 non-hemolytic proteins previously established by Chaudhary et al. (2016). MACREL model performance was compared against models created using different algorithms (Chaudhary et al., 2016): Support vector machines—SVM, K-Nearest Neighbor (IBK), Neural networks (Multilayer Perceptron), Logistic regression, Decision trees (J48) and RF. There is no overlap between the training set and the testing set for the benchmark of hemolytic peptides.

2.2.2 Simulated human gut metagenomes

To test the MACREL short reads pipeline, and the effect of sequencing depth on the discovery rate of AMPs, 6 metagenomes were simulated at 3 different sequencing depths (40, 60 and 80 million reads of 150 bp) with ART Illumina v2.5.8 (Huang et al., 2012) using the pre-built sequencing error profile for the HiSeq 2500 sequencer. To ensure realism, the simulated metagenomes contained species abundances estimated from real human gut microbial communities (Coelho et al., 2019).

We processed both the simulated metagenomes and the isolate genomes used to build the metagenomes with MACREL to verify whether the same AMP candidates could be retrieved and whether the metagenomic processing introduced false positive sequences not present in the original genomes.

The complete set of scripts used to benchmark MACREL is available at https://github.com/BigDataBiology/macrel2020benchmark and the newly simulated generated dataset of different sequencing depths is available at Zenodo (DOI:10.5281/zenodo.3529860).

2.2.3 AMP screening in real metagenomes

To evaluate MACREL in real data, we used 182 metagenomes and 36 metatranscriptomes generated with Illumina technology in a previous study of the human gut microbiome (Heintz-Buschart et al., 2016) (available from the European Nucleotide Archive, accession number PRJNA289586). MACREL was used to process metagenome reads (see Suppl. Table S3), and the genes encoding AMP candidates were mapped back to the metatranscriptomes using bwa-0.7.17 (r1188) (Li and Durbin, 2009) and samtools v1.9 (Li et al., 2009). Abundance profiles were calculated using NGLess (Coelho et al., 2019).

2.2.4 Detection of spurious sequences

To test whether spurious smORFs still appeared in MACREL results, we used Spurio (Höps et al., 2018) and considered a prediction spurious if the score was greater or equal to 0.8.

To identify putative gene fragments, the AMP sequences predicted with MACREL were validated through homology-searching against the non-redundant NCBI database (https://www.ncbi.nlm.nih.gov/). Predicted AMPs annotation was done by homology against the DRAMP database (Fan et al., 2016), which comprises circa 20k AMPs. The above-mentioned databases were used to perform a local search with the blastp algorithm (Camacho et al., 2009), using a maximum e-value of $1 \cdot 10^{-5}$ and a word size of 3. Hits with a minimum of 70% of identity and 95% query coverage were kept and parsed to the best-hits after ranking them by score, e-value, identity, and coverage.

To check whether the AMPs predicted by the MACREL pipeline were gene fragments, patented peptides or known AMPs, the alignments were manually evaluated.

2.3 Implementation

MACREL is implemented in Python (compatible with both versions 2 and 3), and R (Team, 2018). Descriptors are calculated using Peptides (Osorio et al., 2015), and the classification is performed with scikit-learn (Pedregosa et al., 2011). For ease of installation, we made available a bioconda package (Grüning et al., 2018). The source code of Macrel is archived at DOI:10.5281/zenodo.3608055 (with the specific version tested in this manuscript being available as DOI:hrefhttps://doi.org/10.5281/zenodo.360805610.5281/zenodo.3608056).

3 Results

3.1 MACREL: Meta(genomic) AMPs Classification and REtrievaL

As we aim to process both genomes and metagenomes, we built a consolidated pipeline (MACREL, for *Meta(genomic) AMPs Classification and REtrievaL*), which implements a full workflow from short-reads to the prediction and quantification of AMPs (see Fig. 1).

MACREL accepts as inputs metagenomes (in the form of short reads), (meta)genomic contigs, or peptides. If short reads are given as input, MACREL will preprocess and assemble them into larger contigs. Automated gene prediction then extracts smORFs from these contigs which are classified into AMPs or rejected from further processing (see Fig. 1 and Methods). Putative AMPs are further classified into hemolytic or non-hemolytic. Unlike other pipelines (Jhong et al., 2019), MACREL can not only quantify known sequences, but also discover novel AMPs.

MACREL is also available as a webserver (available at http://big-data-biology.org/software/macrel), which accepts both peptides and contig sequences.

3.2 Novel set of protein descriptors for AMP identification

MACREL classifiers use a set of 22 variables that capture the amphipathic nature of AMPs and their propensity to form transmembrane helices (see Methods and Suppl. Table S1). The same set of features is used in both classification steps (AMP and hemolytic activity predictions).

One novel feature group (named *FET*) was designed to capture the fact that AMPs usually fold from random coils in the polar phase to well-organized structures in lipid membranes (Nagarajan et al., 2019). In particular, we clustered residues into three groups of increasing free-energy change (Bhadra et al., 2018) and used the composition-transition-distribution framework (Dubchak et al., 1995, 1999) to derive three features (see Fig. 2).

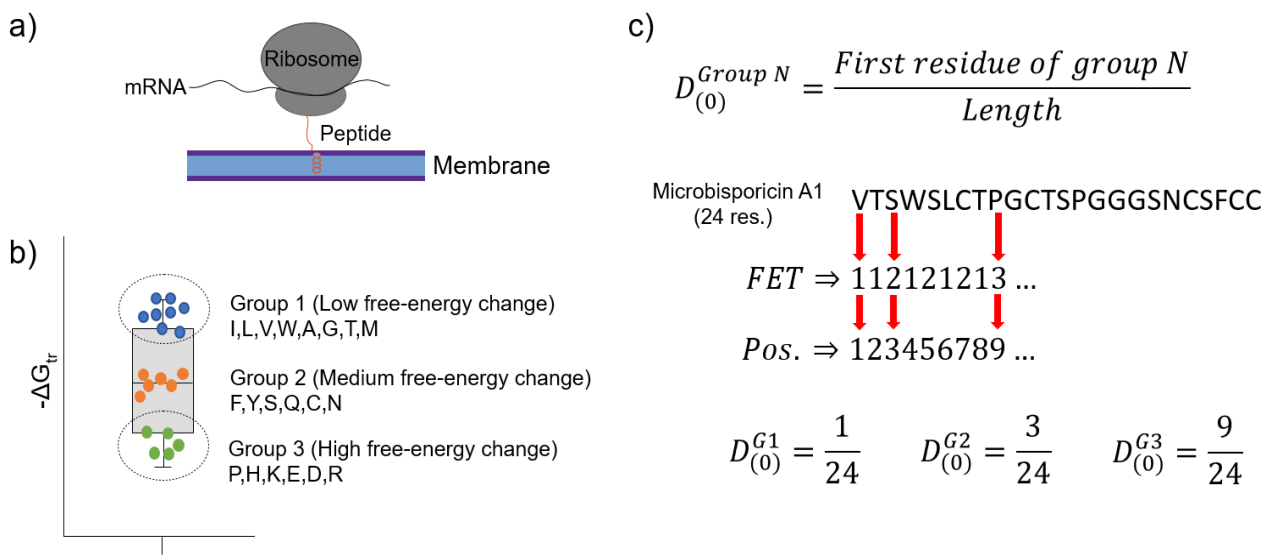

Figure 2. The FET measure estimates the propensity of peptides to fold when transferring from water to the membrane. The estimated change in the free-energy of the conformational change of an amino acid from random to organized structures in lipid membranes **(a)** was used to cluster the 20 amino acids into 3 groups **(b)**. These groups were used to encode peptide sequences as the relative position of the first amino acid in each group **(c)**.

All 22 descriptors used in MACREL models are important in classification (see Fig. 3). The fraction of acidic residues, electrical charge, and isoelectric point were the most important variables in the hemolytic peptides classifier. Those variables tend to capture the electrostatic interaction between peptides and membranes, a key step in hemolysis. For AMP prediction, charge and the distribution parameters using FET and solvent accessibility are the most important variables. This is consistent with reports that highly-charged peptides (typically glycine- or lysine-rich) show increased AMP activity (Bhadra et al., 2018; Jhong et al., 2019; Nagarajan et al., 2019).

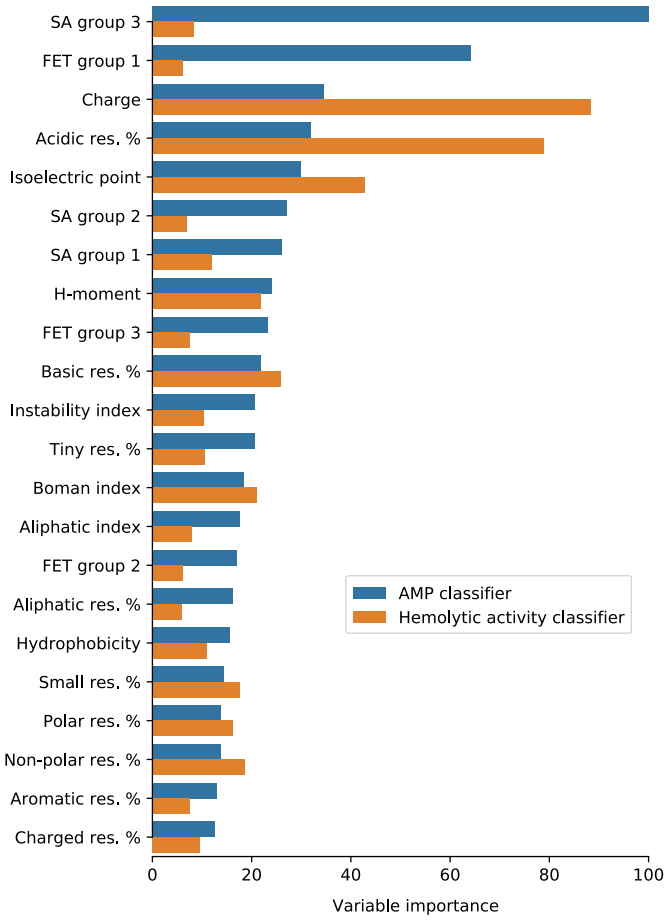

Figure 3. All 22 features are important in classification. Variable importance is measured as the percentage of times each variable is selected in the pruned models (MACREL classifiers of AMPs and hemolytic peptides). The solvent accessibility and free energy to transfer from water to lipophilic phase residues distribution at first position using 3 amino acid groups were summarized as SA and FET, respectively.

3.3 MACREL's performance is among the best methods in AMP classification

Benchmark results show that the AMP classifier trained with a more balanced dataset (MACREL^X, which was trained with an approximate ratio of 1:3, AMPs to non-AMPs, see Section 2.2.1) performs better than most of methods considered, with AmPEP (Bhadra et al., 2018) achieving the best results (see Table 1).

In terms of overall accuracy, the AMP classifier implemented in MACREL (trained with an unbalanced dataset, at a ratio of approximately 1:50, see Section 2.1.2) is comparable to the best methods, with different trade-offs. In particular, MACREL achieves the highest precision and specificity at the cost of lower sensitivity. Although we do not possess good estimates of the proportion of AMPs in the smORFs predicted from real genomes (or metagenomes), we expect it to be much closer to 1:50 than to 1:3. Therefore, we chose to use the higher precision classifier in MACREL for AMP prediction from real data to minimize the number of false positives in the overall pipeline.

Table 1. The comparison of MACREL AMP classifier performance and state-of-art methods shows that MACREL is among the best methods across a range of metrics. The same test set (Xiao et al., 2013) was used to calculate the general performance statistics of the different classifiers, and the best value per column is in bold. MACREL refers to the MACREL classifier, while MACREL^X is the same system trained with the Xiao et al. (2013) training set. Legend: Accuracy (Acc), Sensitivity (Sn), Specificity (Sp), Precision (Pr), and Matthew's Correlation Coefficient (MCC).

Method	Acc.	Sp.	Sn.	Pr.	MCC	Refererence
AmPEP*	**0.981**	-	-	-	**0.92**	Bhadra et al. (2018)
MACREL^X	0.953	0.972	0.935	0.971	0.91	This study
iAMP-2L	0.947	0.92	0.974	0.924	0.90	Xiao et al. (2013)
MACREL	0.946	**0.998**	0.895	**0.998**	0.90	This study
AMAP	0.922	0.861	0.984	0.876	0.85	Gull et al. (2019)
CAMPR3-NN	0.799	0.709	0.89	0.753	0.61	Waghu et al. (2016)
APSv2	0.779	0.572	**0.987**	0.697	0.61	Veltri et al. (2018)
CAMPR3-DA	0.716	0.495	0.938	0.65	0.48	Waghu et al. (2016)
CAMPR3-SVM	0.676	0.398	0.954	0.613	0.42	Waghu et al. (2016)
CAMPR3-RF	0.65	0.335	0.965	0.592	0.39	Waghu et al. (2016)
iAMPpred	0.643	0.325	0.962	0.588	0.37	Meher et al. (2017)

* These data were retrieved from the original paper.

The hemolytic peptides prediction model implemented in MACREL is comparable to the state-of-the-art (Chaudhary et al., 2016). These models, using the same training and test sets, were built with different methods (composition-based or hybrid), and resulted in overall comparable performance (see Table 2).

Table 2. MACREL achieves accuracy comparable to the state-of-art in hemolytic peptides classification. Models implemented by Chaudhary et al. (2016) were generically called HemoPI-1 due to the datasets used in the training and benchmarking (the best values per column are in bold).

Methods	Sn.	Sp.	Acc.	MCC
HemoPI-1$^{C, SVM}$	0.957	**0.948**	**0.953**	**0.91**
HemoPI-1^{H}	**0.96**	0.946	**0.953**	**0.91**
HemoPI-1$^{C, IBK}$	0.955	0.937	0.946	0.89
HemoPI-1$^{C, RF}$	0.941	0.946	0.943	0.89
MACREL	0.918	0.964	0.941	0.88
HemoPI-1$^{C, Log}$	0.934	0.937	0.936	0.87
HemoPI-1$^{C, MP}$	0.939	0.928	0.933	0.87
HemoPI-1$^{C, JK48}$	0.896	0.885	0.89	0.78

3.4 MACREL recovers a small number of high-quality AMP candidates per meta(genome)

We ran MACREL on 484 reference genomes that had previously shown to be abundant in the human gut (Coelho et al., 2019). This resulted in 171,645 (redundant) smORFs. However, only 18 (17, after redundancy removal) of these were classified as potential AMPs. Neither Spurio (Höps et al., 2018) nor homology searches against the NCBI nr database provided any evidence that the 17 non-redundant AMP candidates are false positives.

Homology searches confirmed two AMP candidates as actual AMPs. One of them is a phenol-soluble modulin from the *Staphylococcus* genus and another is a stage V sporulation protein from multiple species in the *Clostridium* group. None of the AMPs found by MACREL in the reference genomes were present in the dataset used during model training.

To evaluate the impact of sequencing a mixed community using short reads, we simulated metagenomes composed by these same 484 reference genomes, using three different sequencing depths (40 million, 60 million, and 80 million reads) using abundance profiles estimated from real data (Coelho et al., 2019). In these simulations (see Fig. 4), it was clear that the number of smORFs increases with sequencing depth, with about 20k smORFs being predicted in the case of 80 million simulated reads.

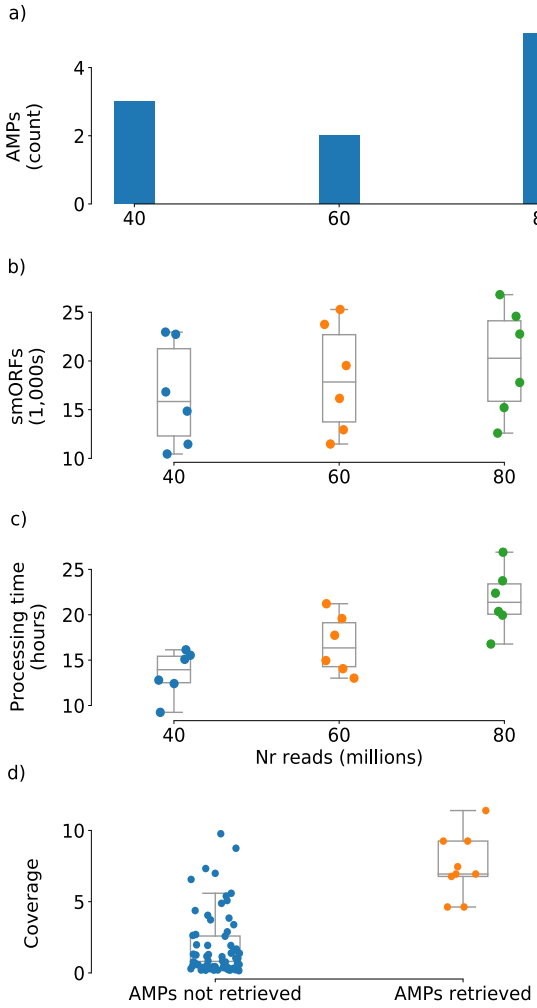

Figure 4. MACREL results in metagenome simulations involving a different number of reads (40-80 million). Communities with realistic species abundances were simulated with increasing sequencing depth (see Methods). MACREL recovers a small number of AMPs **(a)**, despite the large number of small ORFs (smORFs) detected in each metagenome **(b)**. Processing times increase with coverage, with the single largest sample taking 27 hours **(c)**. A minimal coverage of *circa* 5x is necessary (but not sufficient) for AMPs to be recovered **(d)** (for better visualization, outliers were removed, according to criteria defined by (Tukey, 1977)).

Despite this large number of smORF candidates, only 6 non-redundant AMP candidates were predicted. Four of these are present in the underlying reference genomes, and were correctly recovered. *Post hoc*, we estimated that 5x coverage is required for AMP recovery (see Fig. 4), but that this coverage does not guarantee AMP recovery.

None of the AMP candidates recovered from simulated metagenomes were marked as spurious by Spurio, but we manually investigated the origin of the two extra peptides. One of them appears to be a gene fragment from a longer gene. The end of this gene contained one start codon in frame with the stop codon, which generated a mispredicted smORF as the gene was close to the contig end, so that the real start site was not present. The other smORF predicted by MACREL in the metagenomes was not present in the underlying reference genomes and appears to be an artifact of the simulated sequencing/assembly.

3.5 MACREL predicts putative AMPs in real human gut metagenomes

Of the 184 metagenomes in our dataset (Heintz-Buschart et al., 2016), 102 (56%) contain putative AMPs, resulting in a total of 75 non-redundant sequences (see Suppl. Table S3). Five of these were tagged as spurious (see Methods). Two of these appear to be fragments of larger proteins, suggesting that they are generated by fragmentary assemblies.

Our final dataset, after discarding smORFs identified as spurious (see Methods and Suppl. Table S3), consisted of 70 non-redundant AMPs, encoded by a total of 244 genes. None of them had hits to the DRAMP database and only 5 matched hypothetical proteins from NCBI.

To further strengthen MACREL predictions, we checked the co-prediction of the 70 non-redundant AMP candidates with alternative methods. In total, 96% of AMPs predicted with MACREL were also classified as such by at least one other classifier, and 64% of the times, half or more of the tested state-of-art methods agreed with MACREL results (see Suppl. Table S4). iAMPpred and iAMP-2L showed the highest agreement and co-predict 86% and 73% of the AMPs obtained with MACREL, respectively.

As this dataset contains metatranscriptomes produced from the same biological samples, we quantified the expression of the AMP genes. Over 70% of the predicted AMPs were being expressed (see Fig. 5). The gene expression of these AMPs could be detected in more than one metatranscriptome in 82% of the cases, thus bolstering the case that these are biologically active.

Taken together, we concluded that MACREL could find a set of high-quality AMPs candidates, whose genes show active transcription, and that can be co-predicted by other state-of-art methods.

3.6 MACREL requires only moderate computational resources

Tests reported here were carried out on a personal laptop (32 GB of RAM and 4 cores) to show that MACREL is a pipeline with modest computational requirements. The execution time, although dependent on the input size, was not greater than 27 h (recall that the largest simulated metagenomes contained 80 million reads). The reads trimming and assembly steps consumed 75-80% of the execution time, while gene prediction occupies another considerable part (10-15%) (see Fig. 4).

4 Conclusions

MACREL performs all operations from raw metagenomic reads assembly to the prediction of AMPs. Using a combination of local and global sequence encoding techniques, MACREL classifiers perform comparably to the state-of-the-art in benchmark datasets. These benchmarks are valuable for method development, but as they contain the same number of AMP and non-AMP sequences in the testing set, are not a good proxy for the setting in which we intend to use the classifiers. It is unlikely that half of peptide sequences predicted from (meta)genomes will have antimicrobial activity. Therefore, we chose a classifier that achieves a slightly lower accuracy on these benchmarks, but has very high precision and sensitivity.

The main challenge in computationally predicting smORFs (small ORFs, such as AMPs) with standard methods is the high rate of false-positives. However, after the filtering applied by MACREL classifiers, only a small number of candidate sequences remained. Supported by several lines of evidence (lack of obvious spurious origin, similar classification by other methods, and recovery of known sequences), we conclude that MACREL produces a small number of high-quality AMP candidates.

Here, we presented an initial analysis of publicly-available human gut metagenomes (Heintz-Buschart et al., 2016). Although only a small number of AMPs (70 non-redundant sequences) were predicted, AMPs predicted with MACREL were largely congruent (96%) with other state-of-art methods. This opens up the possibility of future work to understand the impact of these molecules on the microbial ecosystems or prospecting them for clinical or industrial applications.

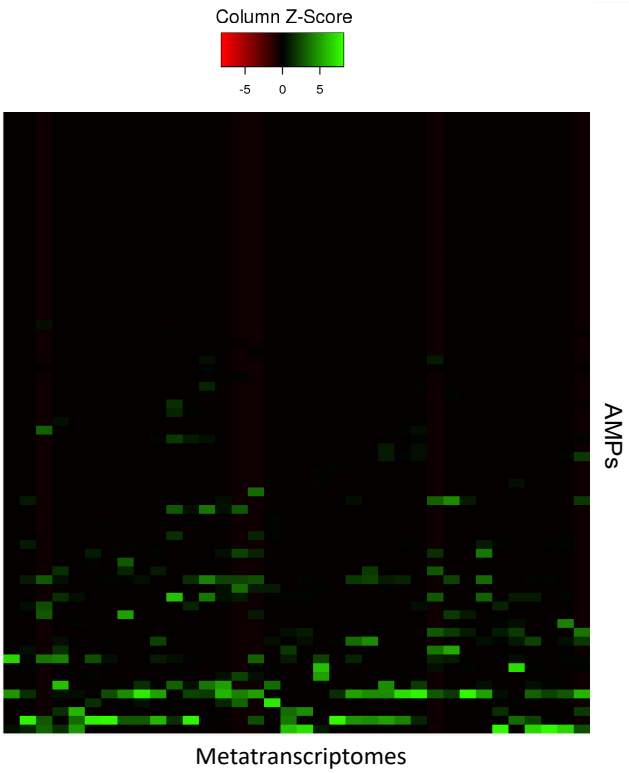

Figure 5. Gene expression of AMPs detected by MACREL in human gut metatranscriptomes. Non-redundant AMP gene variants from each cluster had their FKPM values expressed as an average. The heatmap shows values normalized by metatranscriptome using Z-score.

MACREL is available as open-source software at https://github.com/BigDataBiology/macrel and the functionality is also available as a webserver: http://big-data-biology.org/software/macrel.

Acknowledgments

We thank Hiram He, Fudan University, who helped set up the MACREL website and kindly offered coding support as well as members of the Coelho group for helpful comments in previous versions of the manuscript.

Funding

This work was partly supported by National Natural Science Foundation of China (61932008, 61772368, 61572363), National Key R&D Program of China (2018YFC0910500), Natural Science Foundation of Shanghai (17ZR1445600), Shanghai Municipal Science and Technology Major Project (2018SHZDZX01) and ZJLab.

References

P. Bhadra, J. Yan, J. Li, S. Fong, and S. W. I. Siu. AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. *Scientific Reports*, 8(1):1–10, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-19752-w.

K. Boone, K. Camarda, P. Spencer, and C. Tamerler. Antimicrobial peptide similarity and classification through rough set theory using physicochemical boundaries. *BMC Bioinformatics*, 19(1):469, 2018. ISSN 1471-2105. doi: 10.1186/s12859-018-2514-6.

K. A. Brogden. Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria? *Nature Reviews. Microbiology*, 3(3): 238–250, 2005. ISSN 1740-1526. doi: 10.1038/nrmicro1098.

C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. BLAST+: architecture and applications. *BMC bioinformatics*, 10:421, 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-421.

K. Chaudhary, R. Kumar, S. Singh, A. Tuknait, A. Gautam, D. Mathur, P. Anand, G. C. Varshney, and G. P. S. Raghava. A web server and mobile app for computing hemolytic potency of peptides. *Scientific Reports*, 6:22843, 2016. ISSN 2045-2322. doi: 10.1038/srep22843.

L. P. Coelho, R. Alves, P. Monteiro, J. Huerta-Cepas, A. T. Freitas, and P. Bork. NG-meta-profiler: fast processing of metagenomes using NGLess, a domain-specific language. *Microbiome*, 7(1):84, 2019. ISSN 2049-2618. doi: 10.1186/s40168-019-0684-8.

I. Dubchak, I. Muchnik, S. R. Holbrook, and S. H. Kim. Prediction of protein folding class using global description of amino acid sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 92(19):8700–8704, 1995. ISSN 0027-8424.

I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk, and S. H. Kim. Recognition of a protein fold in the context of the structural classification of proteins (SCOP) classification. *Proteins*, 35(4):401–407, 1999. ISSN 0887-3585.

L. Fan, J. Sun, M. Zhou, J. Zhou, X. Lao, H. Zheng, and H. Xu. DRAMP: a comprehensive data repository of antimicrobial peptides. *Scientific Reports*, 6:24482, 2016. ISSN 2045-2322. doi: 10.1038/srep24482.

M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15:3133–3181, 2014.

C. D. Fjell, H. Jenssen, K. Hilpert, W. A. Cheung, N. Panté, R. E. W. Hancock, and A. Cherkasov. Identification of novel antibacterial peptides by chemoinformatics and machine learning. *Journal of Medicinal Chemistry*, 52(7):2006–2015, 2009. ISSN 0022-2623. doi: 10.1021/jm8015365.

M. N. Gabere and W. S. Noble. Empirical comparison of web-based antimicrobial peptide prediction tools. *Bioinformatics (Oxford, England)*, 33(13):1921–1929, 2017. ISSN 1367-4811. doi: 10.1093/bioinformatics/btx081.

B. Grüning, R. Dale, A. Sjödin, B. A. Chapman, J. Rowe, C. H. Tomkins-Tinch, R. Valieris, J. Köster, and Bioconda Team. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, 15(7):475–476, 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0046-7.

S. Gull, N. Shamim, and F. Minhas. AMAP: Hierarchical multi-label prediction of biologically active and antimicrobial peptides. *Computers in Biology and Medicine*, 107:172–181, 2019. ISSN 1879-0534. doi: 10.1016/j.compbiomed.2019.02.018.

R. E. W. Hancock and H.-G. Sahl. Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. *Nature Biotechnology*, 24(12):1551–1557, 2006. ISSN 1087-0156. doi: 10.1038/nbt1267.

A. Heintz-Buschart, P. May, C. C. Laczny, L. A. Lebrun, C. Bellora, A. Krishna, L. Wampach, J. G. Schneider, A. Hogan, C. d. Beaufort, and P. Wilmes. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nature Microbiology*, 2(1):1–13, 2016. ISSN 2058-5276. doi: 10.1038/nmicrobiol.2016.180.

W. Huang, L. Li, J. R. Myers, and G. T. Marth. ART: a next-generation sequencing read simulator. *Bioinformatics (Oxford, England)*, 28(4):593–594, 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/btr708.

J. Huerta-Cepas, K. Forslund, L. P. Coelho, D. Szklarczyk, L. J. Jensen, C. von Mering, and P. Bork. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular Biology and Evolution*, 34(8):2115–2122, 2017. ISSN 1537-1719. doi: 10.1093/molbev/msx148.

D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-119.

W. Höps, M. Jeffryes, and A. Bateman. Gene unprediction with spurio: A tool to identify spurious protein sequences. *F1000Research*, 7:261, 2018. ISSN 2046-1402. doi: 10.12688/f1000research.14050.1.

J.-H. Jhong, Y.-H. Chi, W.-C. Li, T.-H. Lin, K.-Y. Huang, and T.-Y. Lee. dbAMP: an integrated resource for exploring antimicrobial peptides with functional activities and physicochemical properties on transcriptome and proteome data. *Nucleic Acids Research*, 47:D285–D297, 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1030.

S. Lata, N. K. Mishra, and G. P. S. Raghava. AntiBP2: improved version of antibacterial peptide prediction. *BMC bioinformatics*, 11 Suppl 1:S19, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-S1-S19.

D. Li, R. Luo, C. M. Liu, C. M. Leung, H. F. Ting, K. Sadakane, H. Yamashita, and T. W. Lam. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, 102:3–11, 2016. doi: 10.1016/j.ymeth.2016.02.020.

H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp324.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp352.

M. Malmsten. Antimicrobial peptides. *Upsala Journal of Medical Sciences*, 119(2):199–204, 2014. ISSN 0300-9734. doi: 10.3109/03009734.2014.899278.

P. K. Meher, T. K. Sahu, V. Saini, and A. R. Rao. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into chou's general PseAAC. *Scientific Reports*, 7:42362, 2017. ISSN 2045-2322. doi: 10.1038/srep42362.

S. Miravet-Verde, T. Ferrar, G. Espadas-García, R. Mazzolini, A. Gharrab, E. Sabido, L. Serrano, and M. Lluch-Senar. Unraveling the hidden universe of small proteins in bacterial genomes. *Molecular Systems Biology*, 15(2):e8290, 2019. ISSN 1744-4292. doi: 10.15252/msb.20188290.

D. Nagarajan, T. Nagarajan, N. Nanajkar, and N. Chandra. A uniform in vitro efficacy dataset to guide antimicrobial peptide design. *Data*, 4(1):27, 2019. doi: 10.3390/data4010027.

D. Osorio, P. Rondon-Villarreal, and R. Torres. Peptides: A package for data mining of antimicrobial peptides. *The R Journal*, 7(1):4–14, 2015.

M. Pasupuleti, A. Schmidtchen, and M. Malmsten. Antimicrobial peptides: key components of the innate immune system. *Critical Reviews in Biotechnology*, 32(2):143–171, 2012. ISSN 1549-7801. doi: 10.3109/07388551.2011.594423.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

L. Ringstad, A. Schmidtchen, and M. Malmsten. Effect of peptide length on the interaction between consensus peptides and DOPC/DOPA bilayers. *Langmuir: the ACS journal of surfaces and colloids*, 22(11):5042–5050, 2006. ISSN 0743-7463. doi: 10.1021/la060317y.

A. Saghatelian and J. P. Couso. Discovery and characterization of smORF-encoded bioactive polypeptides. *Nature Chemical Biology*, 11(12):909–916, 2015. ISSN 1552-4469. doi: 10.1038/nchembio.1964.

H. Sberro, B. J. Fremin, S. Zlitni, F. Edfors, N. Greenfield, M. P. Snyder, G. A. Pavlopoulos, N. C. Kyrpides, and A. S. Bhatt. Large-scale analyses of human microbiomes reveal thousands of small, novel genes. *Cell*, 178(5):1245–1259.e14, 2019. ISSN 0092-8674. doi: 10.1016/j.cell.2019.07.016.

Y. Shai. Mode of action of membrane active antimicrobial peptides. *Biopolymers*, 66(4):236–248, 2002. ISSN 0006-3525. doi: 10.1002/bip.10260.

A. Sharma, P. Gupta, R. Kumar, and A. Bhardwaj. dPABBs: A novel in silico approach for predicting and designing anti-biofilm peptides. *Scientific Reports*, 6:21839, 2016. ISSN 2045-2322. doi: 10.1038/srep21839.

A. A. Strömstedt, M. Pasupuleti, A. Schmidtchen, and M. Malmsten. Evaluation of strategies for improving proteolytic resistance of antimicrobial peptides by using variants of EFK17, an internal segment of LL-37. *Antimicrobial Agents and Chemotherapy*, 53(2):593–602, 2009. ISSN 1098-6596. doi: 10.1128/AAC.00477-08.

R. C. Team. R: The r project for statistical computing. 2018. URL https://www.r-project.org/.

N. Thakur, A. Qureshi, and M. Kumar. AVPpred: collection and prediction of highly effective antiviral peptides. *Nucleic Acids Research*, 40:W199–204, 2012. ISSN 1362-4962. doi: 10.1093/nar/gks450.

U. Theuretzbacher, K. Outterson, A. Engel, and A. Karlén. The global preclinical antibacterial pipeline. *Nature Reviews Microbiology*, pages 1–11, 2019. ISSN 1740-1534. doi: 10.1038/s41579-019-0288-0.

J. W. Tukey. *Exploratory data analysis*. Addison-Wesley, 1977.

D. Veltri, U. Kamath, and A. Shehu. Deep learning improves antimicrobial peptide recognition. *Bioinformatics (Oxford, England)*, 34(16):2740–2747, 2018. ISSN 1367-4811. doi: 10.1093/bioinformatics/bty179.

F. H. Waghu, L. Gopi, R. S. Barai, P. Ramteke, B. Nizami, and S. Idicula-Thomas. CAMP: Collection of sequences and structures of antimicrobial peptides. *Nucleic Acids Research*, 42:D1154–1158, 2014. ISSN 1362-4962. doi: 10.1093/nar/gkt1157.

F. H. Waghu, R. S. Barai, P. Gurung, and S. Idicula-Thomas. CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Research*, 44:D1094–1097, 2016. ISSN 1362-4962. doi: 10.1093/nar/gkv1051.

M. Wenzel, A. I. Chiriac, A. Otto, D. Zweytick, C. May, C. Schumacher, R. Gust, H. B. Albada, M. Penkova, U. Krämer, R. Erdmann, N. Metzler-Nolte, S. K. Straus, E. Bremer, D. Becher, H. Brötz-Oesterhelt, H.-G. Sahl, and J. E. Bandow. Small cationic antimicrobial peptides delocalize peripheral membrane proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 111(14):E1409–1418, 2014. ISSN 1091-6490. doi: 10.1073/pnas.1319900111.

A. Westbrook, J. Ramsdell, T. Schuelke, L. Normington, R. D. Bergeron, W. K. Thomas, and M. D. MacManes. PALADIN: protein alignment for functional profiling whole metagenome shotgun data. *Bioinformatics (Oxford, England)*, 33(10): 1473–1478, 2017. ISSN 1367-4811. doi: 10.1093/bioinformatics/btx021.

X. Xiao, P. Wang, W.-Z. Lin, J.-H. Jia, and K.-C. Chou. iAMP-2l: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Analytical Biochemistry*, 436(2):168–177, 2013. ISSN 1096-0309. doi: 10.1016/j.ab.2013.01.019.

M. Zasloff. Antimicrobial peptides of multicellular organisms. *Nature*, 415(6870):389–395, 2002. ISSN 1476-4687. doi: 10.1038/415389a.

L.-J. Zhang and R. L. Gallo. Antimicrobial peptides. *Current biology: CB*, 26(1):R14–19, 2016. ISSN 1879-0445. doi: 10.1016/j.cub.2015.11.017.