Published online 12 November 2022 *Nucleic Acids Research, 2023, Vol. 51, Database issue* ***D377–D383***
https://doi.org/10.1093/nar/gkac933

CAMP$_{R4}$: a database of natural and synthetic antimicrobial peptides

Ulka Gawde[1,†], Shuvechha Chakraborty[1,†], Faiza Hanif Waghu[1,†], Ram Shankar Barai[1], Ashlesha Khanderkar[2], Rishikesh Indraguru[2], Tanmay Shirsat[1] and Susan Idicula-Thomas[1,*]

[1]Biomedical Informatics Centre, ICMR-National Institute for Research in Reproductive and Child Health, Mumbai 400012, Maharashtra, India and [2]Department of Bioinformatics, Guru Nanak Khalsa College, Nathalal Parekh Marg, Matunga, Mumbai 400019, Maharashtra, India

Received August 17, 2022; Revised September 25, 2022; Editorial Decision September 29, 2022; Accepted October 11, 2022

ABSTRACT

There has been an exponential increase in the design of synthetic antimicrobial peptides (AMPs) for its use as novel antibiotics. Synthetic AMPs are substantially enriched in residues with physicochemical properties known to be critical for antimicrobial activity; such as positive charge, hydrophobicity, and higher alpha helical propensity. The current prediction algorithms for AMPs have been developed using AMP sequences from natural sources and hence do not perform well for synthetic peptides. In this version of CAMP database, along with updating sequence information of AMPs, we have created separate prediction algorithms for natural and synthetic AMPs. CAMP$_{R4}$ holds 24243 AMP sequences, 933 structures, 2143 patents and 263 AMP family signatures. In addition to the data on sequences, source organisms, target organisms, minimum inhibitory and hemolytic concentrations, CAMP$_{R4}$ provides information on N and C terminal modifications and presence of unusual amino acids, as applicable. The database is integrated with tools for AMP prediction and rational design (natural and synthetic AMPs), sequence (BLAST and clustal omega), structure (VAST) and family analysis (PRATT, ScanProsite, CAMPSign). The data along with the algorithms of CAMP$_{R4}$ will aid to enhance AMP research. CAMP$_{R4}$ is accessible at http://camp.bicnirrh.res.in/.

INTRODUCTION

Antimicrobial resistance (AMR) is one of the major health crises affecting the health care system worldwide (1). The COVID-19 pandemic has further amplified AMR risk due to the rampant use of antibiotics; especially in low- and middle-income countries (2). The dearth of novel antimicrobials is a significant bottleneck to combat drug-resistant infections.

Over the years, antimicrobial peptides (AMPs) have gained attention as novel antibiotics. These are potent, broad-spectrum and quick-acting defence molecules produced by living organisms ranging from bacteria to mammals, as part of the innate immune response (3,4). Owing to the reduced risk of AMR of AMPs as compared to conventional antibiotics (5), there has been accelerated research on characterisation, discovery and rational design of AMPs. Consequentially, a large volume of data on AMPs is now accessible through various online databases (6–12).

We had first developed CAMP, a manually curated database on AMPs, in 2010 followed by updated versions in 2014 and 2016 (13–15). CAMP$_{R3}$ contained 10 247 sequences, 757 structures and 114 family-specific signatures of AMPs along with tools for AMP analysis (15). The data available in CAMP has been used by several research groups to create secondary AMP databases and prediction servers (16–31). The prediction algorithms in CAMP have been widely used to identify AMPs from natural sources and for rational design (32–45). In the present release, along with updating AMP sequences and associated data extracted from literature post 2015, we have dedicated a separate section for data and prediction algorithms pertaining to synthetic AMPs. Information related to the N and C terminal modifications, that are known to alter antimicrobial activity, has also been incorporated (Figure 1). CAMP$_{R4}$ presently contains 24243 sequences of which 11827 are of natural origin, 12416 are synthetic; 2143 patents; 933 3D structures and 263 family specific signatures.

*To whom correspondence should be addressed. Tel: +91 22 24192107; Fax: +91 22 24139412; Email: thomass@nirrch.res.in
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

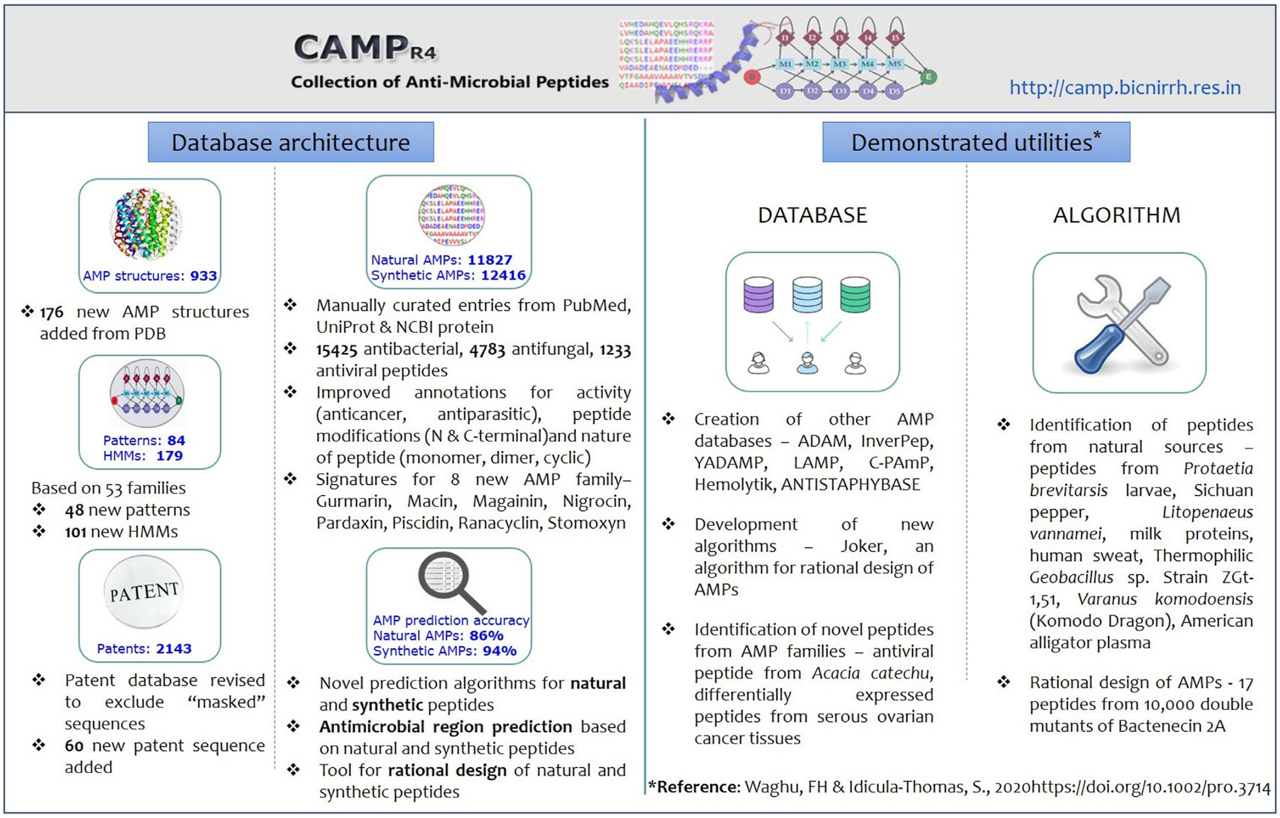

Figure 1. Overview of $CAMP_{R4}$ database update.

Table 1. Comparison of $CAMP_{R3}$ and $CAMP_{R4}$

Sr. No	Parameters of AMPs	$CAMP_{R3}$	$CAMP_{R4}$
1.	**Sequences**	8164	24243
	Experimentally validated	2766	16945
	Natural origin	2574	4839
	Synthetic origin	192	12106
	Predicted AMPs	5398	7298
2.	**Patents**	2083	2143
3.	**3D structures**	757	933
4.	**Signatures**	114	263
	Patterns	36	84
	HMMs	78	179
	Families	45	53
5.	**Prediction algorithms**	Single algorithm for prediction of natural and synthetic AMPs	Separate algorithms for prediction of natural and synthetic AMPs
6.	**Rational design tool**	Single tool for rational design of AMPs	Separate tools for rational design of natural and synthetic AMPs

MATERIALS AND METHODS

Data collection

$CAMP_{R4}$ database was updated using information available on AMPs from the NCBI protein (46), PDB (47), PubMed and Lens (patent) databases for the period post 2015. These databases were queried using keywords such as 'antimicrobial', 'antibacterial', 'antifungal' and 'antiviral'. The obtained hits were manually curated to extract information on sequence, structure, protein definition, accession numbers, reference literature, activity, taxonomy of the source organism, target organisms with minimum inhibitory concentration (MIC) values, hemolytic activity of the peptide and protein family description. Information on N and C terminal residues and other modifications including presence of alkyl groups or modified amino acids has been included in the comments section.

Database architecture

$CAMP_{R4}$ was developed using MySQL Server 5.1.33 as back-end and the front-end is built using PHP, HTML, JavaScript and Perl. The prediction server was developed using statistical software R version 4.0.5. The database interface consists of sections as described previously (15).

Algorithm for prediction of natural and synthetic peptides

Dataset creation. *Positive class*: The positive class comprised of experimentally validated AMP sequences, from natural and synthetic origins, available in $CAMP_{R4}$. Synthetic AMPs, as defined in our study, are peptides that are rationally designed through single or multiple residue substitutions of natural AMPs or through de novo synthesis. The experimentally validated AMPs were further filtered to exclude sequences that had (i) non-standard amino

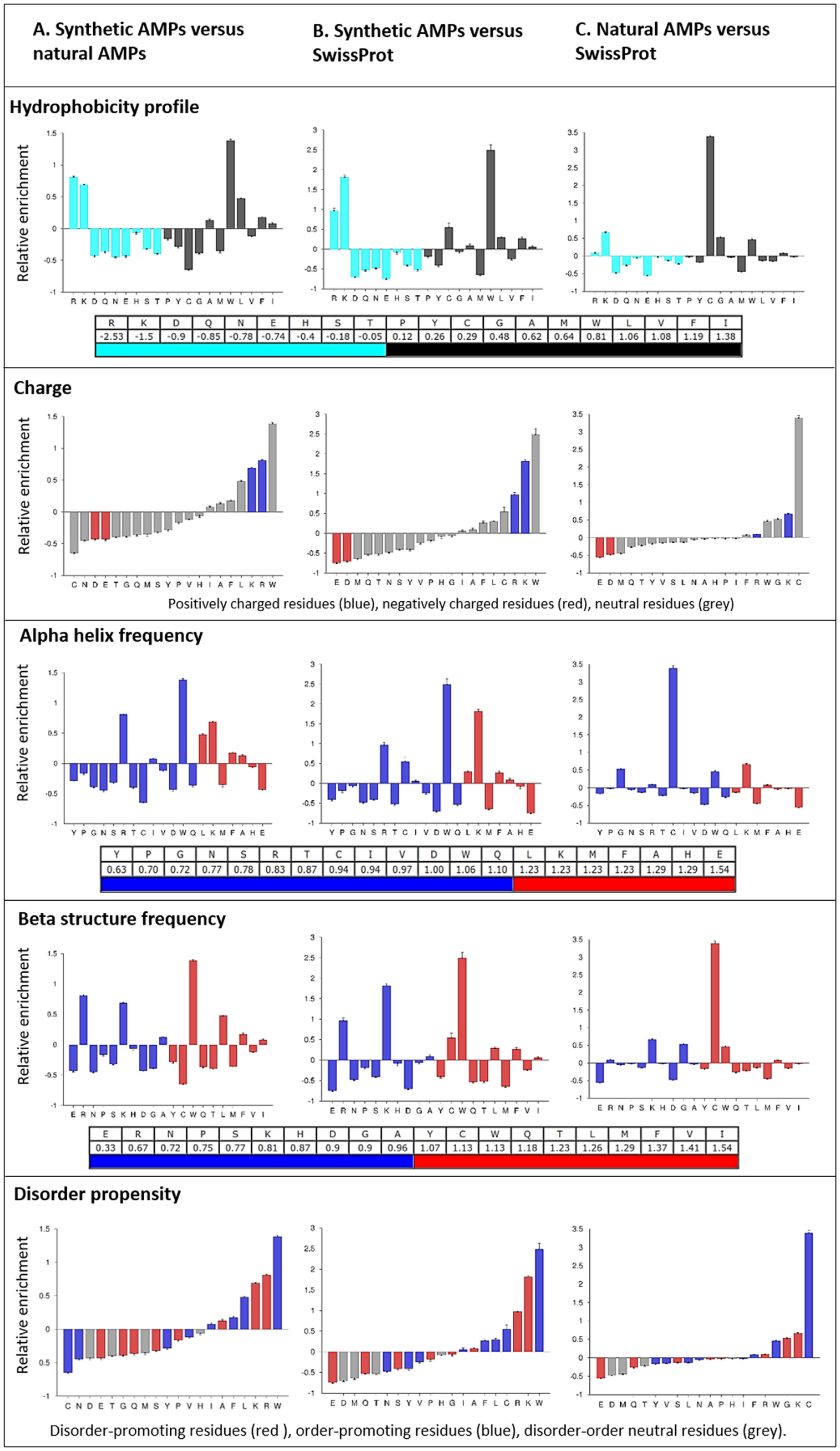

Figure 2. Enrichment and depletion analysis of amino acids of synthetic AMPs ($n = 955$) as compared to (**A**) natural AMPs ($n = 1397$) and (**B**) SwissProt 51 datasets (58). Amino acid composition of natural AMPs was compared with SwissProt 51 dataset in (**C**). The analysis was performed with Composition profiler online tool (59) using 10 000 bootstrap iterations; alpha value was set to 0.05 for statistical significance. Error bars represent standard deviations of observed relative frequencies of the bootstrap samples. Amino acids are arranged left to right in x-axis in order of increasing hydrophobicity (60), alpha helix and beta structure frequency (61) for each of the respective plots. Disorder propensity of amino acids are as defined by Dunker *et al.* (62).

acids (α,γ-diamino β-hydroxy butyric acid, D-ornithine, Z-α,β−dehydroarginine), stapled and circular peptides, and (ii) length >100 residues and ≤2. These sequences were further filtered through CD-HIT server (48), using a 90% sequence similarity cut-off, to generate non-redundant datasets comprising of 1592 synthetic and 2328 natural AMP sequences.

Negative class: As it is difficult to get substantial number of experimentally validated non-AMPs from published literature, a dataset of 4011 peptides as previously described in $CAMP_{R1}$ was used as a negative class (13). The dataset comprised of experimentally proven non-AMPs (25 sequences), non-secretory proteins searched from the UniProt database (49) without annotation as 'antimicrobial' (2413 sequences), arbitrary sequences generated using random numbers (1200 sequences) and proteins retrieved randomly without 'antimicrobial' annotation from the UniProt database (1200 sequences). These sequences were then further filtered using CD-HIT server (48), for eliminating sequences with >90% similarity and the remaining 4011 sequences were used as a negative class.

An identical number of sequences were maintained for the positive and negative classes to create balanced datasets for model generation. The positive and negative classes were further randomly divided into training (60%), test (30%) and external validation (10%) datasets. We ensured that the positive class of external dataset did not contain AMPs that were part of $CAMP_{R3}$ and this dataset was used to compare the performance of $CAMP_{R4}$ with $CAMP_{R3}$ and other existing prediction algorithms.

Feature selection and model generation. 257 features, that represent sequence-based composition and physicochemical properties of AMPs, were used as descriptors for model building, as described previously (13). These 257 features were ranked using the Gini score based rigorous recursive feature elimination (RFE) method and RF models were generated by reducing 50% of the features at each step. Thus, classification models for synthetic and natural AMPs were generated using 257, 128, 64, 32, 16, 8 and 4 features. These models were evaluated using 10-fold cross validation accuracy and kappa values for selecting the optimum number of features. Kappa values compares observed accuracy and accuracy obtained by random chance. The models generated using subset of 64 and 32 features, respectively for natural and synthetic AMPs performed the best. These features were used for developing SVM, RF and ANN based prediction models.

All the models were generated by implementation of SVM, RF and ANN in R (version 4.0.5). Linear, polynomial and radial basis SVM kernel functions were evaluated using '*Kernlab*' (50) package. Polynomial and radial basis kernels were found to perform best and thus retained respectively for the natural and synthetic AMP final model generation. Hyper parameters such as degree, scale and offset were set to 3, 0.01 and 1 for natural and sigma and offset were set to 0.03 and 1 for synthetic AMP prediction. '*randomForest*' package (51) was used to train the RF classifier with a maximum of 500 trees. ANN-based prediction model for natural and synthetic AMPs were built using the '*nnet*' (52) package with parameters size and decay set as 1 and 0.1, respectively. The models were evaluated through 10-fold cross-validation using Matthews correlation coefficient (MCC) and prediction accuracy scores.

Rational design of natural and synthetic AMPs. Algorithms for generating single residue substitutions of user-defined sequence/s followed by their AMP prediction using developed models (RF, SVM and ANN) for natural and synthetic AMPs were created using in-house Perl scripts.

Generation and validation of family-specific signatures. Family-specific signatures, represented by patterns and hidden Markov models (HMMs), were generated for the updated experimentally validated natural AMPs and validated as explained in Waghu *et al.* (15,53). Clustal-omega 1.2.2 (54) was used for multiple sequence alignment; '*hmmbuild*' and '*hmmsearch*' commands (with default parameters) of HMMER downloadable version 3.3.2 (55) were used for generation and search using HMMs respectively. Patterns and HMMs that had precision and recall values of ≥0.5 were included in the database.

RESULTS AND DISCUSSION

The CAMP database has been updated to incorporate the large number of natural and synthetic AMPs that have been discovered and designed in the last five years after the release of $CAMP_{R3}$. Natural AMPs were majorly extracted from NCBI protein database (46). Synthetic AMPs were retrieved from published literature in PubMed database. A total of ∼65000 entries were retrieved from PubMed using keyword-based search. These entries were further filtered to 18355 PubMed articles using an in-house text mining code executed on the abstract of these publications. Subsequently, each of these articles was carefully reviewed to retrieve manually curated information on AMPs. A detailed description of the contents in updated $CAMP_{R4}$ can be viewed in Table 1. There has been a massive increase in the number of AMPs, especially in the discovery of synthetic AMPs as compared to the earlier years, 12170 of the 16079 new AMPs were of synthetic origin. This is expected as AMPs are being increasingly explored as new antibiotics. The databases on AMPs and subsequent sequence analysis have led to the identification of many sequence-related features of AMPs such as positive charge, hydrophobicity and helical propensity which could be exploited for rational design of AMPs (56,57).

Probably for the same reason, synthetic AMPs were found to be significantly enriched with residues that are known to be critical for antimicrobial activity such as positively charged (K, R), hydrophobic (W, L), higher alpha helical propensity (L, K) and flexibility (W, L) as compared to natural AMPs (Figure 2). This observation prompted us to investigate the effectiveness of the current AMP prediction algorithms, that are trained on natural AMP sequences, for predicting synthetic AMPs. A dataset of 159 synthetic AMPs and 159 sequences from negative dataset that were not part of the training models (external validation dataset; see Methods) was predicted with an accuracy of 92.5% using $CAMP_{R3}$ and 71.7% using DBAASP (Table 2). In this update, taking cognizance of the difference in the sequence

Table 2. Comparison of prediction accuracy of $CAMP_{R4}$ algorithms with other AMP prediction algorithms

External dataset*	Algorithms	$CAMP_{R4}$	$CAMP_{R3}$	DBAASP
Natural source	RF	86.5%	85.0%	68.5%
	SVM	84.1%	82.4%	
	ANN	82.2%	79.8%	
Synthetic source	RF	94.3%	92.5%	71.7%
	SVM	90.3%	89.3%	
	ANN	90.6%	85.5%	

*Number of natural peptides used as *external* dataset each for *Positive* and *Negative* class is 233. Number of synthetic peptides used as *external* dataset each for *Positive and Negative* class is 159.

composition of these two classes of AMPs, we created independent prediction algorithms for natural and synthetic AMPs which have also been applied for rational design of natural and synthetic AMPs. The performance metrics and the top features used for these algorithms are provided in Supplementary Tables S1 and S2.

Conclusion

$CAMP_{R4}$ contains updated information on sequences (natural and synthetic), structures and families of AMPs. The database hosts algorithms for predicting natural and synthetic AMPs. Comparison of $CAMP_{R4}$ with presently available manually curated AMP databases is provided in Supplementary Table S3.

The highlights of this update are as follows:

Comprehensive update on AMP-related data: The updated version has information on 24243 sequences (of which 11827 are natural and 12416 are synthetic), 2143 patents, 933 structures and 263 AMP family signatures.

Prediction algorithms for natural and synthetic AMPs: Independent algorithms for prediction of synthetic and natural AMPs based on physicochemical properties and sequence composition have been developed. These algorithms have better prediction accuracy for natural (86.5%) and synthetic AMPs (94.3%) as compared to the currently available online algorithms (Table 2).

Tool for rational design of natural and synthetic AMPs: The tool allows the rational design of AMPs by generating single residue mutant sequences for a user-defined sequence and predicts the effect of single residue substitutions on antimicrobial activity using separate models generated for predicting natural and synthetic AMPs.

Updated family information and signatures: The database now contains information on 53 AMP families (8 new families included) and has 263 AMP family-specific signatures that can promote AMP family-based studies and novel AMP discovery. Signatures for 8 AMP families namely gurmarin, macin, magainin, nigrocin, pardaxin, piscidin, ranacyclin and stomoxyn have been included in this update.

Improved annotations: Information on features such as N and C terminal modification of amino acids, presence of unusual amino acids, cyclic nature of peptides that are important determinants of antimicrobial activity; have been included in this update. Information relating to other functions of AMPs such as anticancer, antiviral activity has also been added.

DATA AVAILABILITY

$CAMP_{R4}$ is freely accessible at http://camp.bicnirrh.res.in/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to Dr Geetanjali Sachdeva, Director, ICMR-NIRRCH for support. The authors thank Ms. Indra Kundu, Ms. Karishma Desai, Dr Chandan Kumar, Dr Prayagraj Fandilolu, Ms. Anam Arshi, and Ms. Komal Chaudhari at ICMR-NIRRCH, Ms. Rekha Chataule, Ms. Rohini Balmiki, Mr. Rakesh Poojari, Ms. Priya Pandey, Ms. Shabari Prakashan at Guru Nanak Khalsa College, Mumbai for assistance with data curation.

FUNDING

This work [RA/1121/09-2021] was supported by research funds from Department of Biotechnology (DBT), India [BT/PR40165/BTIS/137/12/2021] and Indian Council of Medical Research. The open access publication charge for this paper has been waived by Oxford University Press - *NAR*.

Conflict of interest statement. None declared.

REFERENCES

1. Aslam,B., Wang,W., Arshad,M.I., Khurshid,M., Muzammil,S., Rasool,M.H., Nisar,M.A., Alvi,R.F., Aslam,M.A., Qamar,M.U. *et al.* (2018) Antibiotic resistance: a rundown of a global crisis. *Infect. Drug Resist.*, **11**, 1645–1658.
2. Lucien,M., Canarie,M.F., Kilgore,P.E., Jean-Denis,G., Fénélon,N., Pierre,M., Cerpa,M., Joseph,G.A., Maki,G., Zervos,M.J. *et al.* (2021) Antibiotics and antimicrobial resistance in the COVID-19 era: perspective from resource-limited settings. *Int. J. Infect. Dis.*, **104**, 250–254.
3. Zasloff,M. (2002) Antimicrobial peptides of multicellular organisms. *Nature*, **415**, 389–395.
4. Waghu,F.H., Joseph,S., Ghawali,S., Martis,E.A., Madan,T., Venkatesh,K.V. and Idicula-Thomas,S. (2018) Designing antibacterial peptides with enhanced killing kinetics. *Front. Microbiol.*, **9**, 325.
5. Spohn,R., Daruka,L., Lázár,V., Martins,A., Vidovics,F., Grézal,G., Méhi,O., Kintses,B., Számel,M., Jangir,P.K. *et al.* (2019) Integrated evolutionary analysis reveals antimicrobial peptides with limited resistance. *Nat. Commun.*, **10**, 4538.
6. Qureshi,A., Thakur,N. and Kumar,M. (2013) HIPdb: a database of experimentally validated HIV inhibiting peptides. *PLoS One*, **8**, e54908.
7. Pirtskhalava,M., Amstrong,A.A., Grigolava,M., Chubinidze,M., Alimbarashvili,E., Vishnepolsky,B., Gabrielian,A., Rosenthal,A., Hurt,D.E. and Tartakovsky,M. (2021) DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Res.*, **49**, D288–D297.
8. Wang,G., Li,X. and Wang,Z. (2016) APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.*, **44**, D1087–D1093.
9. Jhong,J.H., Yao,L., Pang,Y., Li,Z., Chung,C.R., Wang,R., Li,S., Li,W., Luo,M., Ma,R. *et al.* (2022) dbAMP 2.0: updated resource for antimicrobial peptides with an enhanced scanning method for genomic and proteomic data. *Nucleic Acids Res.*, **50**, D460–D470.
10. Ye,G., Wu,H., Huang,J., Wang,W., Ge,K., Li,G., Zhong,J. and Huang,Q. (2020) LAMP2: a major update of the database linking antimicrobial peptides. *Database*, **2020**, baaa061.

11. Di Luca,M., Maccari,G., Maisetta,G. and Batoni,G. (2015) BaAMPs: the database of biofilm-active antimicrobial peptides. *Biofouling*, **31**, 193–199.
12. Qureshi,A., Thakur,N., Tandon,H. and Kumar,M. (2014) AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic Acids Res.*, **42**, D1147–D1153.
13. Thomas,S., Karnik,S., Barai,R.S., Jayaraman,V.K. and Idicula-Thomas,S. (2010) CAMP: a useful resource for research on antimicrobial peptides. *Nucleic Acids Res.*, **38**, D774–D780.
14. Waghu,F.H., Gopi,L., Barai,R.S., Ramteke,P., Nizami,B. and Idicula-Thomas,S. (2014) CAMP: collection of sequences and structures of antimicrobial peptides. *Nucleic Acids Res.*, **42**, D1154–D1158.
15. Waghu,F.H., Barai,R.S., Gurung,P. and Idicula-Thomas,S. (2016) $Camp_{R3}$: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Res.*, **44**, D1094–D1097.
16. Lee,H.T., Lee,C.C., Yang,J.R., Lai,J.Z. and Chang,K.Y. (2015) A large-scale structural classification of antimicrobial peptides. *Biomed. Res. Int.*, **2015**, 475062.
17. Gómez,E.A., Giraldo,P. and Orduz,S. (2017) InverPep: a database of invertebrate antimicrobial peptides. *J. Global Antimicrob. Resist.*, **8**, 13–17.
18. Piotto,S.P., Sessa,L., Concilio,S. and Iannelli,P. (2012) YADAMP: yet another database of antimicrobial peptides. *Int. J. Antimicrob. Agents*, **39**, 346–351.
19. Zhao,X., Wu,H., Lu,H., Li,G. and Huang,Q. (2013) LAMP: a database linking antimicrobial peptides. *PLoS One*, **8**, e66557.
20. Niarchou,A., Alexandridou,A., Athanasiadis,E. and Spyrou,G. (2013) C-PAmP: large scale analysis and database construction containing high scoring computationally predicted antimicrobial peptides for all the available plant species. *PLoS One*, **8**, e79728.
21. Gautam,A., Chaudhary,K., Singh,S., Joshi,A., Anand,P., Tuknait,A., Mathur,D., Varshney,G.C. and Raghava,G.P. (2014) Hemolytik: a database of experimentally determined hemolytic and non-hemolytic peptides. *Nucleic Acids Res.*, **42**, D444–D449.
22. 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 Res.*, **47**, D285–D297.
23. Zouhir,A., Taieb,M., Lamine,M.A., Cherif,A., Jridi,T., Mahjoubi,B., Mbarek,S., Fliss,I., Nefzi,A., Sebei,K. *et al.* (2017) ANTISTAPHYBASE: database of antimicrobial peptides (AMPs) and essential oils (Eos) against methicillin-resistant staphylococcus aureus (MRSA) and staphylococcus aureus. *Arch. Microbiol.*, **199**, 215–222.
24. Gautam,A., Sharma,A., Jaiswal,S., Fatma,S., Arora,V., Iquebal,M.A., Nandi,S., Sundaray,J.K., Jayasankar,P., Rai,A. *et al.* (2016) Development of antimicrobial peptide prediction tool for aquaculture industries. *Probiotics Antimicrob. Proteins*, **8**, 141–149.
25. Sarika,I.M.A., Arora,V., Rai,A. and Kumar,D. (2015) Species specific approach to the development of web-based antimicrobial peptides prediction tool for cattle. *Comput. Electron. Agric.*, **111**:55–61. .
26. Meher,P.K., Sahu,T.K., Saini,V. and Rao,A.R. (2017. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into chou's general PseAAC. *Sci. Rep.*, **7**, 42362.
27. Vishnepolsky,B. and Pirtskhalava,M. (2014) Prediction of linear cationic antimicrobial peptides based on characteristics responsible for their interaction with the membranes. *J. Chem. Inf. Model.*, **54**, 1512–1523.
28. Romani,A.A., Baroni,M.C., Taddei,S., Ghidini,F., Sansoni,P., Cavirani,S. and Cabassi,C.S. (2013) In vitro activity of novel in silico-developed antimicrobial peptides against a panel of bacterial pathogens. *J. Peptide Sci.*, **19**, 554–565.
29. Ng,X.Y., Rosdi,B.A. and Shahrudin,S. (2015) Prediction of antimicrobial peptides based on sequence alignment and support vector machine-pairwise algorithm utilizing LZ-complexity. *Biomed. Res. Int.*, **2015**, 212715.
30. Holton,T.A., Pollastri,G., Shields,D.C. and Mooney,C. (2013) CPPpred: prediction of cell penetrating peptides. *Bioinformatics*, **29**, 3094–3096.
31. Tyagi,A., Kapoor,P., Kumar,R., Chaudhary,K., Gautam,A. and Raghava,G.P. (2013) In silico models for designing and discovering novel anticancer peptides. *Sci. Rep.*, **3**, 2984.
32. Li,Z., Meng,M., Li,S. and Deng,B. (2019) The transcriptome analysis of protaetia brevitarsis lewis larvae. *PLoS One*, **14**, e0214001.
33. Hou,X., Li,S., Luo,Q., Shen,G., Wu,H., Li,M., Liu,X., Chen,A., Ye,M. and Zhang,Z. (2019) Discovery and identification of antimicrobial peptides in sichuan pepper (Zanthoxylum bungeanum maxim) seeds by peptidomics and bioinformatics. *Appl. Microbiol. Biotechnol.*, **103**, 2217–2228.
34. Yang,S., Huang,H., Wang,F., Aweya,J.J., Zheng,Z. and Zhang,Y. (2018) Prediction and characterization of a novel hemocyanin-derived antimicrobial peptide from shrimp litopenaeus vannamei. *Amino Acids*, **50**, 995–1005.
35. Dziuba,B. and Dziuba,M. (2014) New milk protein-derived peptides with potential antimicrobial activity: an approach based on bioinformatic studies. *Int. J. Mol. Sci.*, **15**, 14531–14545.
36. Yu,Y., Prassas,I., Muytjens,C.M. and Diamandis,E.P. (2017) Proteomic and peptidomic analysis of human sweat with emphasis on proteolysis. *J. Proteomics*, **155**, 40–48.
37. Alkhalili,R.N., Bernfur,K., Dishisha,T., Mamo,G., Schelin,J., Canbäck,B., Emanuelsson,C. and Hatti-Kaul,R. (2016) Antimicrobial protein candidates from the thermophilic geobacillus sp. Strain ZGt-1: production, proteomics, and bioinformatics analysis. *Int. J. Mol. Sci.*, **17**, 1363.
38. Bishop,B.M., Juba,M.L., Russo,P.S., Devine,M., Barksdale,S.M., Scott,S., Settlage,R., Michalak,P., Gupta,K., Vliet,K. *et al.* (2017) Discovery of novel antimicrobial peptides from varanus komodoensis (Komodo dragon) by large-scale analyses and de-novo-assisted sequencing using electron-transfer dissociation mass spectrometry. *J. Proteome Res.*, **16**, 1470–1482.
39. Juba,M.L., Russo,P.S., Devine,M., Barksdale,S., Rodriguez,C., Vliet,K.A., Schnur,J.M., van Hoek,M.L. and Bishop,B.M. (2015) Large scale discovery and de novo-assisted sequencing of cationic antimicrobial peptides (CAMPs) by microparticle capture and electron-transfer dissociation (ETD) mass spectrometry. *J. Proteome Res.*, **14**, 4282–4295.
40. Azkargorta,M., Soria,J., Ojeda,C., Guzmán,F., Acera,A., Iloro,I., Suárez,T. and Elortza,F. (2015) Human basal tear peptidome characterization by CID, HCD, and ETD followed by in silico and in vitro analyses for antimicrobial peptide identification. *J. Proteome Res.*, **14**, 2649–2658.
41. Kim,I.W., Markkandan,K., Lee,J.H., Subramaniyam,S., Yoo,S., Park,J. and Hwang,J.S. (2016) Transcriptome profiling and in silico analysis of the antimicrobial peptides of the grasshopper oxya chinensis sinuosa. *J. Microbiol. Biotechnol.*, **26**, 1863–1870.
42. Lin,C.H., Chang,M.W. and Chen,C.Y. (2014). A potent antimicrobial peptide derived from the protein lsgrp1 of lilium. *Phytopathology*, **104**, 340–346.
43. Hovde,B.T., Deodato,C.R., Hunsperger,H.M., Ryken,S.A., Yost,W., Jha,R.K., Patterson,J., Monnat,R.J., Jr,Barlow, S.,B. *et al.* (2015) Genome sequence and transcriptome analyses of chrysochromulina tobin: metabolic tools for enhanced algal fitness in the prominent order prymnesiales (Haptophyceae). *PLoS Genet.*, **11**, e1005469.
44. Leoni,G., De Poli,A., Mardirossian,M., Gambato,S., Florian,F., Venier,P., Wilson,D.N., Tossi,A., Pallavicini,A. and Gerdol,M. (2017) Myticalins: a novel multigenic family of linear, cationic antimicrobial peptides from marine mussels (Mytilus spp.). *Mar. Drugs*, **15**, 261.
45. Porto,W.F., Fensterseifer,I., Ribeiro,S.M. and Franco,O.L. (2018) Joker: an algorithm to insert patterns into sequences for designing antimicrobial peptides. *Biochim. Biophys. Acta. Gen. Subj.*, **1862**, 2043–2052.
46. Sayers,E.W., Bolton,E.E., Brister,J.R., Canese,K., Chan,J., Comeau,D.C., Connor,R., Funk,K., Kelly,C., Kim,S. *et al.* (2022) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **50**, D20–D26.
47. Burley,S.K., Bhikadiya,C., Bi,C., Bittrich,S., Chen,L., Crichlow,G.V., Christie,C.H., Dalenberg,K., Di Costanzo,L., Duarte,J.M. *et al.* (2021) RCSB protein data bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.*, **49**, D437–D451.

48. Huang,Y., Niu,B., Gao,Y., Fu,L. and Li,W. (2010) CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.
49. UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
50. Karatzoglou,A., Smola,A., Hornik,K. and Zeileis,A. (2004) kernlab - An S4 package for kernel methods in R. *J. Stat. Softw.*, **11**, 1–20.
51. Liaw,A and Wiener,M. (2002) Classification and regression by random forest. *R News*. **2**, 18–22.
52. Venables,W.N. and Ripley,B.D. (2002) In: *Modern Applied Statistics with S. Fourth Edition*. Springer, NY.
53. Waghu,F.H., Barai,R.S. and Idicula-Thomas,S. (2016) Leveraging family-specific signatures for AMP discovery and high-throughput annotation. *Sci. Rep.*, **6**, 24684.
54. Sievers,F, Barton,GJ and Higgins,DG (2020) Multiple sequence alignment. *Bioinformatics*, **227**, 227–250.
55. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
56. Jiang,Z., Vasil,A.I., Hale,J.D., Hancock,R.E., Vasil,M.L. and Hodges,R.S. (2008) Effects of net charge and the number of positively charged residues on the biological activity of amphipathic alpha-helical cationic antimicrobial peptides. *Biopolymers*, **90**, 369–383.
57. Huang,Y., He,L., Li,G., Zhai,N., Jiang,H. and Chen,Y. (2014. Role of helicity of α-helical antimicrobial peptides to improve specificity. *Protein Cell*, **5**, 631–642.
58. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The universal protein resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
59. Vacic,V., Uversky,V.N., Dunker,A.K. and Lonardi,S. (2007) Composition profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinf.*, **8**, 211.
60. Eisenberg,D., Schwarz,E., Komaromy,M. and Wall,R. (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.*, **179**, 125–142.
61. Nagano,K. (1973) Logical analysis of the mechanism of protein folding. I. Predictions of helices, loops and beta-structures from primary structure. *J. Mol. Biol.*, **75**, 401–420.
62. Dunker,A.K., Lawson,J.D., Brown,C.J., Williams,R.M., Romero,P., Oh,J.S., Oldfield,C.J., Campen,A.M., Ratliff,C.M., Hipps,K.W. *et al.* (2001) Intrinsically disordered protein. *J. Mol. Graph. Model.*, **19**, 26–59.