Published online 31 October 2018 *Nucleic Acids Research, 2019, Vol. 47, Database issue* ***D285–D297***
doi: 10.1093/nar/gky1030

dbAMP: an integrated resource for exploring antimicrobial peptides with functional activities and physicochemical properties on transcriptome and proteome data

Jhih-Hua Jhong[1], Yu-Hsiang Chi[1], Wen-Chi Li[2,3], Tsai-Hsuan Lin[1], Kai-Yao Huang[2,3] and Tzong-Yi Lee[1,2,3,*]

[1]Department of Computer Science and Engineering, Yuan Ze University, Taoyuan 320, Taiwan, [2]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518172, China and [3]Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen 518172, China

Received August 15, 2018; Revised October 15, 2018; Editorial Decision October 16, 2018; Accepted October 24, 2018

ABSTRACT

Antimicrobial peptides (AMPs), naturally encoded from genes and generally contained 10–100 amino acids, are crucial components of the innate immune system and can protect the host from various pathogenic bacteria, as well as viruses. In recent years, the widespread use of antibiotics has inspired the rapid growth of antibiotic-resistant microorganisms that usually induce critical infection and pathogenesis. An increasing interest therefore was motivated to explore natural AMPs that enable the development of new antibiotics. With the potential of AMPs being as new drugs for multidrug-resistant pathogens, we were thus motivated to develop a database (dbAMP, http://csb.cse.yzu.edu.tw/dbAMP/) by accumulating comprehensive AMPs from public domain and manually curating literature. Currently in dbAMP there are 12 389 unique entries, including 4271 experimentally verified AMPs and 8118 putative AMPs along with their functional activities, supported by 1924 research articles. The advent of high-throughput biotechnologies, such as mass spectrometry and next-generation sequencing, has led us to further expand dbAMP as a database-assisted platform for providing comprehensively functional and physicochemical analyses for AMPs based on the large-scale transcriptome and proteome data. Significant improvements available in dbAMP include the information of AMP–protein interactions, antimicrobial potency analysis for 'cryptic' region detection, annotations of AMP target species, as well as AMP detection on transcriptome and proteome datasets. Additionally, a Docker container has been developed as a downloadable package for discovering known and novel AMPs on high-throughput omics data. The user-friendly visualization interfaces have been created to facilitate peptide searching, browsing, and sequence alignment against dbAMP entries. All the facilities integrated into dbAMP can promote the functional analyses of AMPs and the discovery of new antimicrobial drugs.

INTRODUCTION

Antimicrobial peptides (AMPs), naturally encoded from genes and generally contained 10–100 amino acids, are produced by different organisms as a defense mechanism against microbial invasions (1–3). AMPs are a particularly functional group of protein molecules, and most of them are α-helical and amphipathic, which means they have hydrophilic residues on one side, and hydrophobic residues on the other side. In typical, AMPs contain more positively charged residues on hydrophilic side for attaching to the membrane surface of microbes. On the other hand, the hydrophobic side allows AMPs anchoring into the membrane lipid bilayer for leading to depolarization of the membrane and to cell death by worm-hole pore, barrel-stave pore, or carpet models (4–6). AMPs can play as the first line of innate immune systems of all living organisms, ranging from prokaryotes to humans, for enabling the cell death of microbes either by disrupting its cell membrane or its intracellular functions (7,8). Multicellular hosts can thus adapt to pathogenic microbes via their innate immunity through the rapid synthesis and release of diversely short peptides known as AMPs (9). AMPs are comprehensively distributed in nature and have been isolated from a variety of sources including bacteria, plants, vertebrates, and

*To whom correspondence should be addressed. Tel: +86 75523519551; Email: leetzongyi@cuhk.edu.cn

mammals (10–12). Recently, a couple of sophisticated approaches, such as AMP mimetics (13), AMP congeners, cyclotides and stabilized AMPs, AMP conjugates, and immobilised AMPs (14), were devoted to the discovery of AMPs.

AMPs have been discovered that they have a broad spectrum of antimicrobial activities against a variety of pathogens, including not only gram-positive and gram-negative bacterial, fungi, parasite, protozoans, viruses but also insects and several sorts of tumor (15,16). In addition to their antimicrobial properties, AMPs also can function as mediators of inflammation with impact on human epithelial and inflammatory cells such as cell proliferation, immune induction, and wound healing (17). AMPs have been validated to be associated with variously functional activities. A previous study showed that a little modification on the primary sequence of AMPs may influence its specificity and activity along with structural changes (18). Therefore, a full understanding of amino acid composition of AMPs is a key to manufacturing antimicrobial agents by reducing their cytotoxicities and increasing their activities. In the last decade, AMPs are becoming the attractive drug targets in clinical applications owing to the broad spectrum of antimicrobial activities and low propensity for drug-resistant development (19). AMPs were thus attracting the attention of many investigators as a substitute for conventional antibiotics. Hence, a further investigation into the structure–activity relationship of AMPs is an urgent need to the development of new antibiotics or drugs (20).

In recent year, the widespread use of antibiotics has inspired the rapid growth of antibiotic-resistant microorganisms that usually induce critical infection and pathogenesis. An increasing interest was therefore motivated to accumulate natural AMPs with an attempt to enable the development of new antibiotics. Several AMP databases have been developed for specific species, including PenBase for shrimp (21), PhytAMP for plants (22), DADP for anurans (23), as well as BACTIBASE (24), BAGEL4 (25) and YADAMP (26) for bacterial. Additionally, a couple of resources have integrated a broad spectrum of AMPs on multiple species, such as CAMPR3 (27), APD3 (28), AMPer (29), DAMPD (15), ADAM (20) and LAMP (30). A little number of databases dedicated to the function-specific AMPs, e.g. Antiviral Peptides (AVPdb (31)), Defensins Knowledgebase (32), synthetic peptides (SAPD (33) and DBAASP (34)), and recombinantly-produced AMPs (RAPD (16)). An increasing development of AMP databases has promoted many approaches dedicated to in silico prediction of AMPs based on amino acid composition (35). Recently, the advent of high-throughput technologies has led molecular biology into a data surge in both growth and scope (36). For instance, the next-generation sequencing (NGS) technology has been applied to generate large-scale DNA/RNA reads from foods (37), water (38), soil, air and specimen, for identifying microbiota and their functions based on metagenomics and metatranscriptomics, respectively. Additionally, mass spectrometry (MS) was also widely applied in proteomics studies for generating thousands of peptides in one experiment. Rapidly advancing technologies have offered us the opportunities to examine the genome, transcriptome, and proteome in comprehensive ways. Thus, we were motivated to design a database-assisted system (dbAMP: http://csb.cse.yzu.edu.tw/dbAMP) for exploring AMPs with functional activities and physicochemical properties on transcriptome and proteome data.

MATERIALS AND METHODS

The dbAMP is an open-access and manually curated database harboring diverse annotations of AMPs including sequence information, antimicrobial activities, post-translational modifications (PTMs), structural visualization, antimicrobial potency, target species with minimum inhibitory concentration (MIC), physicochemical properties, AMP–protein interactions, as well as the supporting references. In addition to the functional and physicochemical annotations, dbAMP provides an effective AMP prediction on proteome data from different species and a large-scale AMP detection on transcriptome data from NGS technologies. The flowchart of dbAMP construction is presented in Figure 1, including data integration and curation, functional and physicochemical analyses, characterization and identification of AMPs on different species, detection of AMPs on transcriptome data, and the development of dbAMP web interface. Moreover, a couple of external databases related to AMP functions are also integrated into the proposed resource.

Integration of antimicrobial peptides from public domain and literature

Both experimentally validated and putative AMPs were retrieved from protein database of NCBI (39), UniProt (40), Protein Data Bank (41) and public AMP databases such as APD3 (28), CAMPR3 (27), ADAM (20), PhytAMP (22), AMPer (29), AntiBP2 (42), BACTIBASE (24) and LAMP (30). After the removal of redundant sequences by mapping all collected AMPs to UniProt protein entries, a total of 12,389 unique AMPs were integrated into the proposed resource which contains 4271 experimentally verified AMPs along with their functional activities obtained from 2048 organisms. In addition to databases integration, digging knowledge concerning AMPs in the related articles can enable a full understanding of the functional activities of AMPs and their targets. However, the surge in scope and scale of PubMed literature database has conducted a formidable challenge in manually curating AMPs from research articles. Thus, we designed a pipelined text extraction system for retrieving AMP-related articles by querying appropriate keywords such as ‘antimicrobial’, ‘antibacterial’, ‘anti-gram positive bacterial’, ‘anti-gram negative bacterial’, ‘antifungal’, ‘antiviral’, ‘antiparasitic’, ‘antibiofilm’, ‘antimalarial’, ‘antiprotozoal’, ‘antiyeast’, ‘anticancer’, ‘antitumor’, ‘wound healing’, ‘spermicidal’, ‘insecticidal’ and ‘surface immobilized’ against searchable fields ‘Title’, ‘Abstract’ and ‘Keywords’ of PubMed literature database. After obtaining the potentially AMP-related articles, an approach of name entity recognition (43) was adopted to detect name entities, which are summarized from validated AMPs existed in dbAMP, in the text and label them with appropriate tags. Then, a natural language processing algorithm (31) was employed to extract relations among those entities for obtaining functional activities and target species

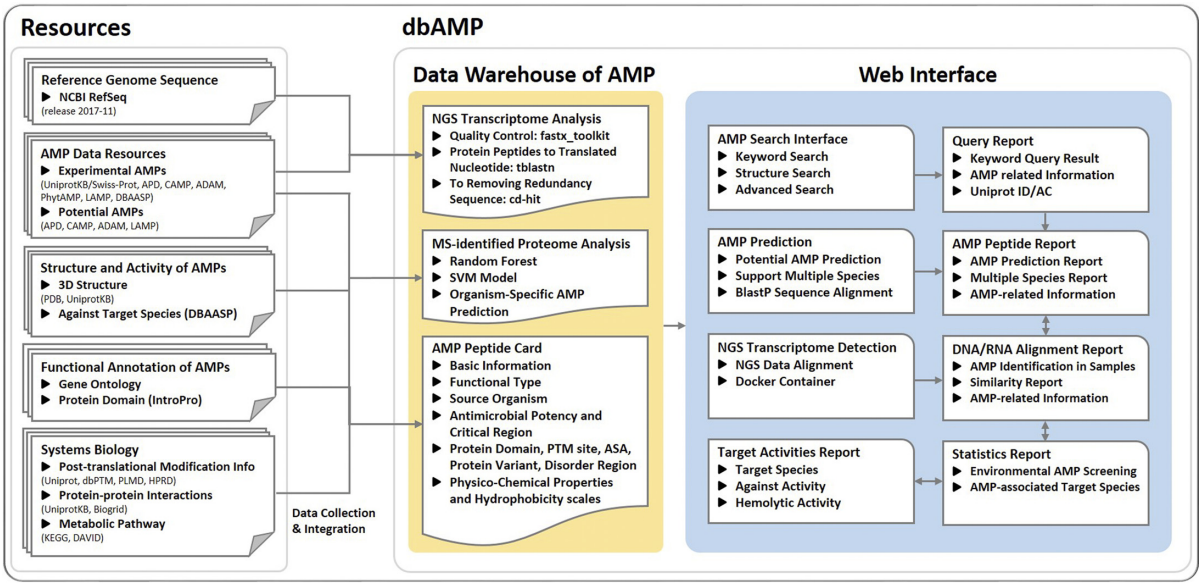

Figure 1. Schematic flowchart of dbAMP construction including data integration, functional analyses and web interface development.

of AMPs. After that, the articles labeled with specified name entities and functional contexts were further manually curated. Up to July 2018, a total of 1924 references were extracted and regarded as the evidence for supporting dbAMP data entries.

Comprehensive analyses for functional and physicochemical properties

An increasing interest in the functional and physicochemical investigation of AMPs motivated the mapping of all AMPs onto protein entries of UniProt and Protein Data Bank (PDB) based on sequence identity, which enables users to examine amino acid composition, post-translational modifications (PTMs), functional domains, solvent-accessible surface area, secondary structure, AMP–protein interactions, hydrophobicity, as well as the composition of positively and negatively charged residues. As for functional domains, the InterPro (44) is an integrated database for annotating 'signatures' like protein families, domains, and functional sites on proteins. It has been reported that a protein-interacting domain usually recognizes a short peptide motif of target protein but does not bind stably until the peptide has an appropriate PTM; this can create binding sites for specific protein-interaction domains that work together for cellular function (45). The information of PTMs on AMPs was obtained from dbPTM (46), which has accumulated most comprehensive data for validated substrate sites of various PTM types.

An increasing number of studies suggested that AMPs can play multiple roles not only in the interaction with membrane lipids and proteins but also in the intracellular targeting mechanisms, which include nucleic acids and protein biosynthesis, protein-folding, protease, cell division, cell wall biosynthesis and lipopolysaccharide inhibition (47,48). Thus, one of the aims in this resource is integrating the information of physical protein-protein interactions (PPIs) to explore the potentially intracellular targeting proteins of AMPs. For this purpose, the information of validated physical interactions was obtained from over ten PPI databases (as listed in Supplementary Table S1). Additionally, with an attempt to understand the potential target species and activity of AMPs, the BLAST program was used to query dbAMP peptides against the DBAASP database (34), with 100% identity and e-value $<1e10^{-5}$ as the threshold. In DBAASP data content, the information of AMP activity and cytotoxicity were extracted from published papers where the minimal inhibitory concentration (MIC, μM), based on microdilution assays, was reported for Escherichia coli, Staphylococcus aureus, or Pseudomonas aeruginosa. In order to increase the use of dbAMP for clinical treatment, cytotoxicity data with the measurement of hemolytic activity (HC) value for each AMP has been stored in our database. The HC50 value refers to peptide concentration, which is annotated as micromolar and is required for 50% haemolysis of red blood cells (RBCs). Each peptide has been annotated with different hemolytic potencies on different RBCs, if available.

Detection of critical region of antimicrobial potency

It has been reported that multicellular eukaryotes can secrete AMP-Releasing Proteins (AMP-RPs), which release active peptides after a partial proteolytic processing conducted by bacterial or host proteases (49–51). Computationally identifying the presences of putative antimicrobial regions inside a larger protein (AMP precursor) could be a useful scheme for the exploration of new cryptic AMPs naturally encoded from host genome (52). Hence, we have adopted an effective sliding window analysis to search for critical region of antimicrobial potency contained into the primary structure of larger proteins and precursors. This work focused on the identification of a critical antimicrobial region endowed with a significant antimicrobial potency for discovering cryptic AMP from host proteins. Pane *et al.* have demonstrated that the antimicrobial potency is

highly correlated to the formula: $C^m H^n L$ where C stands for the net charge of a peptide, H represents a measure of its hydrophobicity and L is peptide length (52). Coefficients m and n are conducted by accounting for the influence of bacterial strain features and environmental conditions on the interaction between the AMP and a cellular membrane. The Relative AntiMicrobial Score (RAMS) for a given peptide can be determined by:

$$\mathrm{RAMS} = C^m \ H^n / MaxScore \quad (1)$$

where C is the sum of net charge and H is the sum of hydrophobicity scores in a sliding window along the peptide. *MaxScore* is the maximal value of $C^m H^n$ score which can be obtained from a given peptide at optimal values of m and n, evaluated in a suited range. According to the evaluation of correlation coefficient between RAMS values and experimental antimicrobial potency values, the optimal values for m and n are 0.9 and 1.1, respectively. However, most AMPs are of diversity in sequence length. In order to formalize various antimicrobial scores owing to different peptide lengths, we provided a flexible scheme to analyze all possible peptide lengths ranging from 12 to 40 amino acid residues in a protein sequence with a sliding window analysis (52,53). The absolute RAMS (ARAMS) can thus be calculated for a given peptide length L, which is defined as (54):

$$\text{Absolute RAMS (ARAMS)} = \mathrm{RAMS} \times \mathrm{L} \quad (2)$$

The antimicrobial scoring function was carried out on all AMPs through the sliding window analysis. To present a clear investigation of antimicrobial potency on AMPs, an isometric plot was created to show the ARAMS values based on different window lengths, ranging from 12 to 40 amino acid residues, for each AMP sequence using parameters optimized for strain *S. aureus* C623. As presented in Supplementary Figure S1, the x axis stands for different values of window size and y axis labels the amino acid sequence of the peptide. A sliding window having ARAMS value higher than 3.0 in response to having MIC value lower than 200 μM is highlighted with blue color.

Investigation and identification of AMPs on different source species

Due to no existing resources dedicated to the prediction of AMPs on different source organisms, in addition to providing a user-friendly interface for browsing the collected AMPs in dbAMP, all the experimentally verified AMPs were utilized to generate AMP-prediction models on different host species, including bacteria, humans, amphibian, fish, plants, insects, and mammals, based on random forest (RF) algorithm. As depicted in Supplementary Figure S2, the positive (AMPs) and negative (non-AMPs) training datasets, used for the construction of predictive models, were constituted from dbAMP and UniProt entries, respectively. After the removal of homologous training sequences by using CD-HIT program (55) with 40% identity, the resulting numbers of positive training sequences for human, amphibia, fish, insect, plant, bacterial, and mammal are 232, 926, 118, 274, 454, 431 and 559, respectively. Because only 35 peptides are annotated with experimentally-verified no antimicrobial activity in UniProt, we followed the data preparation procedure conducted in other studies (56–59) to generate our negative dataset. The protein sequences, containing sequence length between 10 and 100 and annotated without the information of membrane, toxic, secretory, defensin, antimicrobial, antibiotic, anticancer, antiviral and antifungal, were obtained from UniProt and regarded as the non-AMPs for the seven studied species. Consistent with the processing criteria used in positive training dataset, the CD-HIT program was adopted again to remove homologous sequences among non-AMP sequences, based on sequence identity of 40%. Additionally, 20% of positive and negative training sequences were randomly selected to compose the positive and negative testing datasets, respectively. A summary of numbers of training dataset (80%) and testing dataset (20%) is given in Supplementary Table S2.

Both amino acid composition (60,61) and physicochemical properties (62) were considered as the attributes for characterization and identification of AMPs on seven species. In this analysis, the random forest (RF), a sort of ensemble model that involves the aggregation of multiple decision tree classifiers, was employed to generate predictive models. Based on the integration of multiple decision trees within a RF model, each tree was generated from a subset of k attributes randomly selecting from training dataset with a total of m attributes, where k is less than m. In this way, we can obtain multiple decision-making results. Typically, the majority voting method is adopted to integrate the results to make a final decision, based on the class label with the most votes. The performance of a RF decision-making system is associated with the dimension of random vector, which is the number of attributes (k) used in each decision tree. The value of k is typically defined as

$$k = \log_2 m + 1 \quad (3)$$

where m is the total number of attributes in training dataset (63). In this study, a package of random forest, which has been integrated into Weka toolkit (64), was utilized to construct RF classifiers based on various attribute sets. In the generation of RF models, the k-fold cross-validation was employed to evaluate their predictive performances. After the evaluation of k-fold cross-validation, the RF model reaching a best predictive performance was further evaluated by the testing dataset, which is totally independent to the training dataset.

Large-scale detection of AMPs on transcriptome data

An increasing interest of identifying natural AMPs enables the development of new antibiotics. The emerging NGS was widely utilized to obtain large-scale DNA or RNA sequencing datasets from various samples such as food, water, soil, air and specimen, for conducting genomic or transcriptomic analyses (65). The high-throughput RNA sequencing technology offered us an opportunity to discover cryptic AMPs in the transcriptome data obtained from environments or experiments. Thus, we were motivated to design a database-assisted system for identifying AMPs with their functional types based on the metatranscriptomic analysis of various transcriptome data. For efficiently searching out RNA reads that potentially encode

for AMPs, all the amino acid sequences of experimentally verified AMPs were transformed back to RNA sequences for constructing an AMP-encoded RNA database by using BLAST program with 100% identity, 80% sequence coverage, and e-value $<1e10^{-5}$ as the threshold. This resulted in a total of 2651 AMP-encoded RNA sequences with functional activities retrieved from their original annotations of AMPs. Then, the Bowtie2 program (66) was integrated to implement a downloadable pipeline for discovering AMPs from NGS sequencing data, based on a Docker virtualization software that can package applications and their dependencies in a virtual container running on Linux server (67,68). The developed container *csbyzu/ismap* has been uploaded onto the Docker open source, which enables flexibility and portability on where the package can run. As presented in Figure 2, users can submit a large-scale data of NGS reads or MS/MS-identified peptides to the Docker container *csbyzu/isamp* locally on your computer, and the package could identify known AMPs with their functional activities and predict novel AMPs by the constructed RF models. The analyzed results, including data statistics of total reads aligned to AMPs, summary table of functional activities, detailed alignment results of NGS reads, and comprehensive annotations of the mapped AMP in dbAMP, can be displayed by the web interface which has been integrated into the Docker container.

DATABASE CONTENT AND UTILITY

Data statistics of antimicrobial peptides in dbAMP

The dbAMP was created as a useful resource for accumulating natural and synthetic AMPs from scientific literature and published AMP databases. One of the aims of dbAMP is to provide most comprehensive information of AMPs as well as their antimicrobial activities and physicochemical characteristics. Currently in dbAMP there are 12 389 unique entries, including 4271 experimentally verified AMPs and 8118 putative AMPs along with their functional activities, supported by 1924 research articles. Each unique AMP sequence was assigned an identification number (i.e. dbAMP ID) beginning with the prefix 'dbAMP'. The basic information related to AMPs, including functional activities, physicochemical properties, taxonomy of the source organism, target species, PDB structures, and crosslinks to external databases, can be accessed via dbAMP ID. Table 1 summarizes the number of AMPs according to their functional activities in dbAMP as well as in other databases. Currently dbAMP has a repository of over 20 types of functional activities on AMPs. It is worth mentioning that 114 anti-MRSA peptides (69), a newly annotated type of antimicrobial activity, has been integrated into dbAMP.

Sequential properties of AMPs

With the infrastructure of integrating comprehensively validated data into dbAMP, we can conduct a full investigation of sequence-based features, such as sequence length and amino acid composition (AAC), on AMPs according to different source species. The distribution of AMP source organisms is illustrated in Supplementary Figure S3. Nearly 90% of the experimentally verified AMPs contain 80 amino acid residues or less with an average peptide length of 46. A detailed distribution of peptide length of validated AMPs has also been given in Supplementary Figure S4. Existing methods have demonstrated that the composition of amino acids is a potent characteristic for AMP identification (59,70); therefore, we compared amino acid composition profile between 4271 validated AMPs and a background set of reviewed proteins obtained from UniProt. Figure 3 shows the leucine (L), glycine (G) and lysine (K) are the top three abundant amino acids. The frequencies of Cysteine (C), G and K residues are significantly higher in AMPs. The cysteine-rich antimicrobial peptides are containing a couple of disulfide bonds for stabilizing its tertiary structures in plants and invertebrates (71–73). The glycine-rich AMPs, such as acanthoscurrin (74), hyastatin (75) and armadillidin H (76), induce a higher frequency of G residue in AMPs. The enrichment of K residue contributes positive net charge onto hydrophilic side of amphipathic AMPs for attaching to the membrane surface of microbes (77). A partial reason constitutes the positive net charge in AMPs is owing to the lack of aspartic (D) and glutamic (E) acids. In addition, Supplementary Figure S5 presents the amino acid composition of AMPs based on different species.

Physicochemical properties of AMPs

Antimicrobial peptides occur naturally as important innate immunity agents in a wide range of living organisms, and they are characterized by their positive net charge, modest length, and good solubility in water (78). As presented in Figure 4A, the comparison of net charge distribution between dbAMP and UniProt entries has revealed that a majority of AMPs had net charge values between +2 and +4 and less than 5% AMPs had a negative net charge value, which is consistent with a previous discovery (79). The average value of net charge of AMPs is around 3. The information of AMP tertiary structures was retrieved from PubMed or UniProt with a database ID crosslinking to PDB. Moreover, the EMBOSS pepinfo (80) has been integrated to obtain general information such as molecular weight, isoelectric point, charge, hydrophobicity values, as well as the composition of positively and negatively charged residues. Comparison of box plots of hydropathicity indices between dbAMP validated AMPs and UniProt reviewed proteins is depicted in Figure 4B. The information of hydrophobicity and net charge are considered as the crucial attributes for AMPs (81). Additionally, Figure 4C and D provides the comparisons of aliphatic and instability indices, respectively, for validated AMPs and UniProt reviewed proteins. Overall, these comparisons indicate that the validated AMPs contain more diverse values of hydropathicity, aliphatic and instability indices when comparing to UniProt reviewed proteins. Supplementary Figure S6 further shows that the net charge distributions of validated AMPs are noticeably diverse among different source species. In order to identify useful features for classifying between AMPs and non-AMPs on multiple species, the compositions of hydrophobic residues on different regions of AMPs, ranging from N-terminal end to C-terminal end, are provided in Supplementary Figure S7. Due to the variously

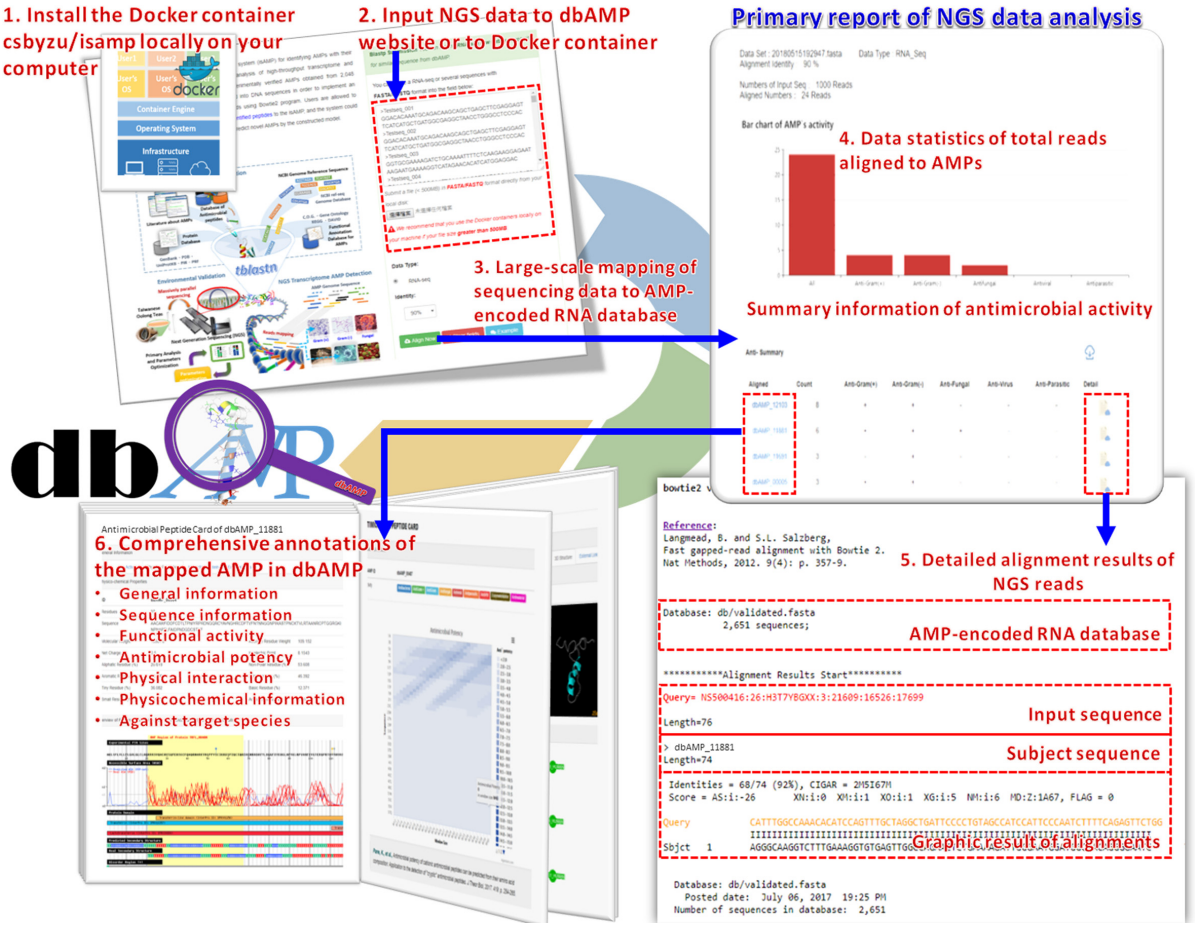

Figure 2. Flowchart of using the developed Docker container '*csbyzu/isamp*' to detect AMPs in NGS data.

Table 1. Comparison of data statistics between dbAMP and other AMP databases based on functional activities of AMPs

Functional activity	dbAMP	DBAASP	APD	CAMP	ADAM	PhytAMP	LAMP
Antibacterial	3006	2232	1536	2914	2350	52	-
Anti-Gram_positive	2726	2196	2124	1901	2080	-	-
Anti-Gram_negative	2323	2142	1722	1743	1913	-	-
Anti-fungal	1623	1433	1187	1150	1181	85	90
Anti-viral	300	85	186	117	181	15	78
Anti-parasitic	123	83	111	35	95	-	4
Anti-HIV	109	-	109	-	-	-	-
Wound Healing	19	-	18	-	-	-	-
Chemotactic	59	-	59	-	-	-	-
Cancer Cells	227	219	210	22	-	-	-
Anti-biofilm	40	-	31	-	-	-	-
Antimalarial	26	-	24	-	-	-	-
Antioxidant	22	-	22	-	-	-	-
Antiprotozoal	6	-	4	-	-	-	-
Spermicidal	13	-	13	-	-	-	-
Insecticidal	35	32	33	-	-	4	-
Antimicrobial	4816	-	494	4812	-	-	2840
Enzyme Inhibitor	26	-	26	-	-	-	-
Anti-tumour	9	-	4	8	-	-	-
Mammalian Cells	308	-	308	-	-	-	-
Surface Immobilized	19	-	19	-	-	-	-
Anti-oncogenic	1	-	-	-	-	-	-
Sodium Channel Blocker	2	-	-	-	-	2	-
Anti-yeast	4	-	-	-	-	4	-
Anti-inflammatory	2	-	-	-	-	-	-
Anti-MRSA	114	-	-	-	-	-	-

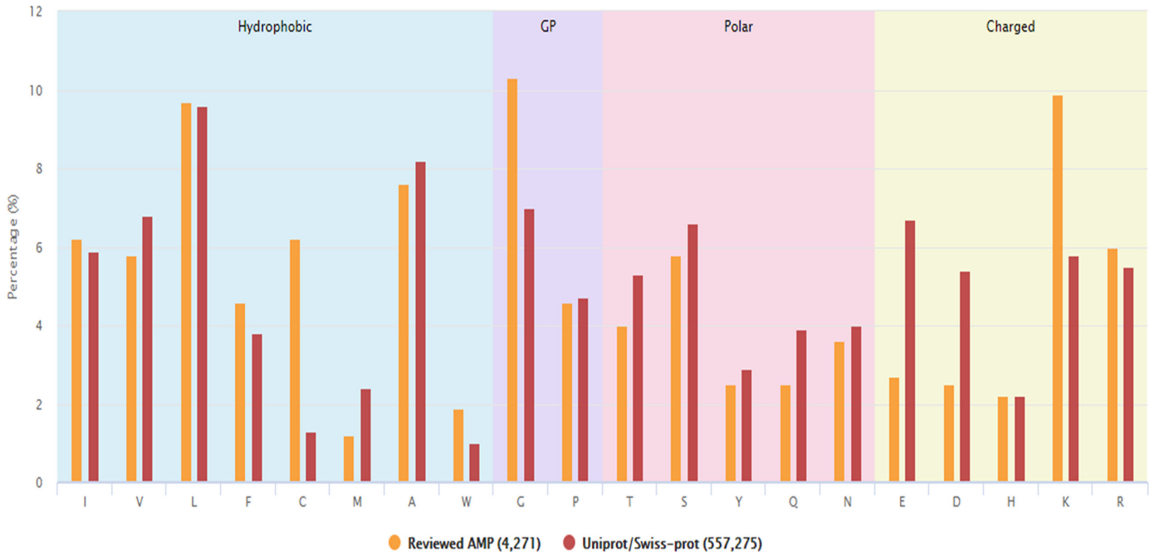

Figure 3. Comparison of amino acid composition between validated AMPs in dbAMP and a background set of reviewed proteins obtained from UniProt.

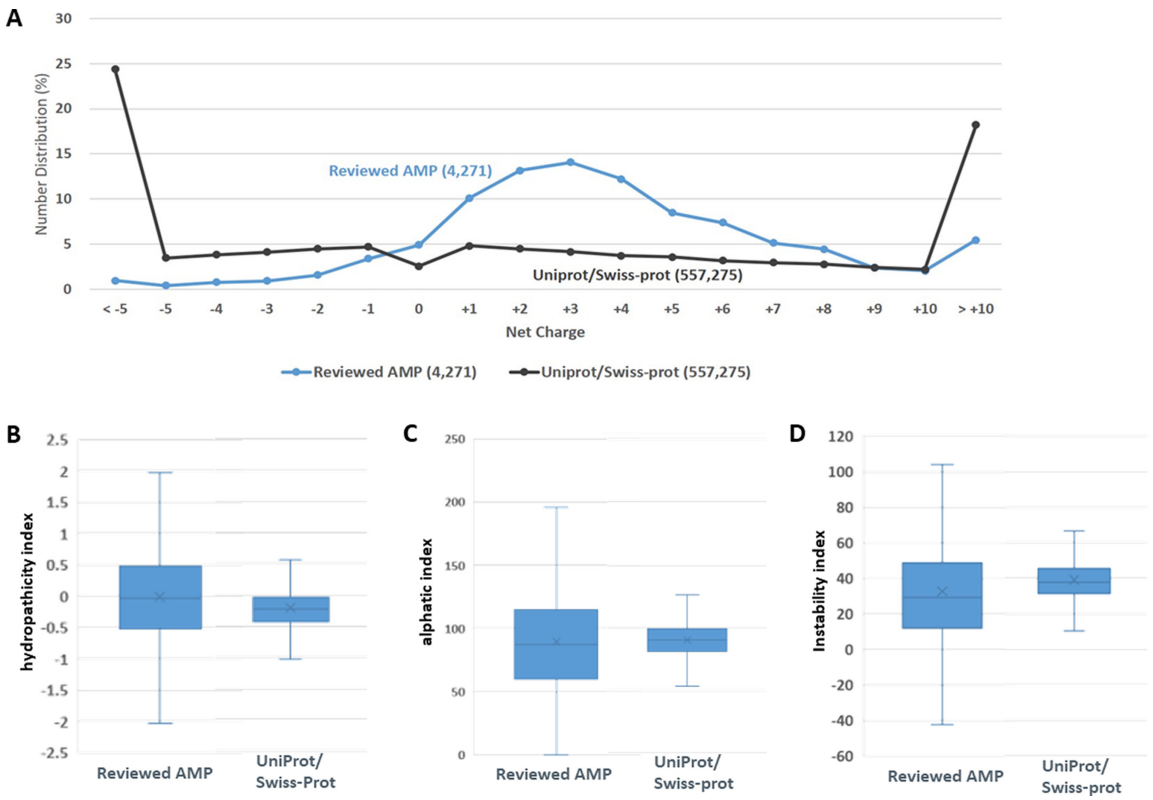

Figure 4. Physicochemical properties of validated AMPs. (**A**) Comparison of net charge distribution between dbAMP validated peptides and UniProt reviewed proteins. Comparisons of box plots of (**B**) hydropathicity, (**C**) alphatic and (**D**) instability indices between dbAMP validated peptides and UniProt reviewed proteins.

physicochemical characteristics on different source species, this work assigned an antimicrobial score to AMPs based on their net charge, hydrophobicity and length factors. This physicochemical investigation enables us to localize the position of cryptic peptides, for yielding an accurate map of the molecular determinants of their antimicrobial activity.

Functional analysis of AMPs

An increasing number of studies have suggested that AMPs can play the roles in the intracellular targeting mechanisms (47,48). For instance, the human antimicrobial peptide elafin domain can interact with other intracellular proteins for specifically participating into the apoptosis in-

Figure 5. Web interface of displaying comprehensive annotations for human antimicrobial peptide elafin. The comprehensively structural and functional analyses include (**A**) general information, (**B**) visualization of AMP tertiary structure, (**C**) solvent accessibility and functional domains, (**D**) amino acid composition, (**E**) antimicrobial potency, (**F**) antimicrobial target species, (**G**) histogram of hydrophobicity and composition of positively and negatively charged residues, as well as (**H**) AMP–protein interactions.

duction mechanism in melanoma cells (82,83). Targets of AMPs may be either bacterial membranes or diverse intracellular molecules; however, some peptides can operate through complex mechanisms that involve multiple targets (84). Hence, the dbAMP has integrated the information of functional domains and PPIs to explore the interactions between AMPs and potentially targeting proteins. Currently the dbAMP has accumulated 6,535 proteins that had the experimental evidence of physically interacting with AMPs, which provides a starting point for discovering potential targets of AMPs. Moreover, in order to explore the functions of potential AMP-targeting proteins, functional enrichment analysis (FEA) was performed by using the DAVID functional annotation tool (85). Next, gene ontology (GO) analysis was conducted to examine the dominant functions in biological processes, cellular components, and molecular functions. The distribution of top 20 GO terms for 6535 AMP-interacting proteins is displayed in Supplementary Figure S8. This functional analysis revealed that the potential AMP-targeting protein tends to be affected in terms of nucleotide metabolism and signaling pathways with response to regulations of signal transduction, cell adhesion, viral infection and apoptotic process.

Computational identification of AMPs on proteome data

It has been fully investigated that the dominant net charge and amino acids composition differ in various species. Machine learning (ML) method has been utilized to generate predictive models. Based on the evaluation of 5-fold cross-validation, four ML algorithms have been adopted to make the comparison of predictive performance in different species. Supplementary Table S3 shows that the proposed RF models could reach highest prediction accuracy on proteome data of different species. Accuracies for all the seven organisms are observed as higher than 93% with balanced sensitivities and specificities. This investigation has indicated that the RF model trained with the examined attributes could provide accurate predictions against multiple species. Moreover, Supplementary Table S4 also demonstrates that the proposed RF models could provide promising accuracies in different species, based on the independent testing dataset. To our knowledge, a variety of computational methods have been proposed for the prediction of antimicrobial peptides. However, there exists no online resource dedicated to characterizing and identifying AMPs on different source species, such as bacteria, humans, amphibian, fish, plants, insects and mammals. The dbAMP allows users to input a single sequence or multiple sequences with FASTA format or to upload a text file to perform the AMP prediction by specifying a species or life domain on the prediction page of website.

Detection of AMPs on NGS transcriptome data

Many methods have been developed to perform metagenomic and metatranscriptomic data analysis for NGS raw data. For instance, the UPARSE pipeline constructs a set of operational taxonomic unit from NGS amplicon reads used to understand the microbial community structure (86). The MG-RAST server is a SEED-based algorithm that characterizes the taxonomic composition, functional potential and diversity of the microbial assemblages (87). In contrast, we designed a database-assisted system specialized in identifying AMPs and their functional activities based on metatranscriptomic analysis of high-throughput transcriptome data. The dbAMP provides an intuitive graphical user interface (GUI) to execute a Docker container csbyzu/isamp on local machine. Metagenomics and metatranscriptomics analyses of diverse microscopic organisms in natural environments, including human body, have revolutionized the understanding of the relationship between microbes and their hosts. To showcase the new scheme of AMPs discovery, the RNA sequencing samples of Taiwanese oolong teas (Dayuling, Alishan, Jinxuan and Oriental Beauty teas), obtained from NCBI SRA with accession number SRP113601 (88), were subjected to the quality control and sequence alignment against dbAMP entries. As shown in Supplementary Table S5, among the reads mapped to plants, totally 1 077 775 (9.1%), 1 305 016 (6.8%), 1 152 178 (8.4%) and 366 454 (7.4%) RNA reads could be mapped to AMPs with sequence identity of 100% in Dayuling, Alishan, Jinxuan and Oriental Beauty teas, respectively. On the other hand, among the reads mapped to bacterial, a total of 8194 (6.5%), 26 220 (6.2%), 5753 (6.1%) and 106 683 (7.7%) RNA reads could be mapped to AMPs with sequence identity of 100% in Dayuling, Alishan, Jinxuan and Oriental Beauty teas, respectively. Furthermore, Supplementary Figures S9A presents the distribution of anti-Gram-positive and anti-Gram-negative AMPs of plants in four oolong teas. In addition, Supplementary Figures S9B also shows the distribution of anti-Gram-positive and anti-Gram-negative AMPs of bacterial in four oolong teas. As presented in Supplementary Figure S9C, the composition of anti-Gram-positive AMPs in Jinxuan and Oriental Beauty teas is more abundant than that in Dayuling and Alishan teas. Meanwhile, the composition of Gram-positive bacterial in Jinxuan and Oriental Beauty teas is less than that in Dayuling and Alishan teas.

Web interface of dbAMP

To enable the comprehensive analyses of AMPs, the dbAMP has provided users the web interface with enhanced designs. The dbAMP was built on the Centos (CentOS release 6.9 Final) Linux operating system using the freeware Apache (Apache/2.2.15) web server, PHP programming language, and MySQL database system (5.5.56 MySQL Community Server). All the collected AMPs have been manually curated and systematically stored on the database management system of MySQL. In order to visualize AMP tertiary structures obtained from both the PDB and molecular dynamics (MD) simulations, the Jmol program (89) has been integrated into the PHP web program. The web interface also constructs a couple of links connecting to external resources, including UniProtKB, PubMed, PDB, Pfam (90), as well as other AMP databases, for providing additionally functional information. Users are allowed to browse all the AMPs and submit RNA sequencing reads or MS/MS-identified peptides to the dbAMP, and the system could identify known AMPs with their functional activities and discover novel AMPs by the predictive models. Figure

Table 2. Comparison of system functionality between dbAMP and other AMP databases

Database	$CAMP_{R3}$	APD3	ADAM	LAMP	AMPer	AntiBP2	BACTIBASE	PhytAMP	dbAMP
PubMed ID	26467475	26602694	26000295	23825543	17341497	20122190	20105292	18836196	-
Number of AMP sequences	10 247	2889	7007	5548	1298	999	228	273	12 389
Number of organisms	773	1396	794	Multiple	Eukaryotes	Multiple	Bacteriocins	Plant	2048
Number of AMP tertiary structures	757	388	759	189	N/A	N/A	72	39	1169
AMP prediction model	Random forest	-	SVM/HMM	-	HMM	SVM	-	HMM	Random forest
AMP detection on multiple species	N/A	-	N/A	-	N/A	N/A	-	N/A	Seven life domains
Number of AMP targets	-	-	-	-	-	-	-	-	2172
Number of AMP-interacting proteins	-	-	-	-	-	-	-	-	6338
Antimicrobial potency Analysis	-	-	-	-	-	-	-	-	Antimicrobial potency of cationic AMPs can be determined from their amino acid composition
Detection of cryptic region in AMPs	-	-	-	-	-	-	-	-	Determining the cryptic region by referring to net charge and hydrophobicity
Against target species									Providing the activity of AMPs against target species based on sequence analysis
NGS data analysis	-	-	-	-	-	-	-	-	Using the developed Docker container to explore AMPs for transcriptome data

5 showcases the web interface providing comprehensively functional and physicochemical analyses.

CONCLUSION

In recent years, many studies have been conducted to dig out new drugs for multidrug-resistant bacteria. The results of these studies have concluded that antimicrobial peptides have a high potential to be considered as an alternative drug for conventional antibiotics and become a computational model for the development of new antimicrobial drugs that can solve the problem with regarding to an increasing number of multidrug resistances of pathogenic microorganisms (91). However, it is labor-intensive and time-consuming to design experimental methods to discover natural AMPs. We are thus inspired to design a database-assisted platform (dbAMP) for providing comprehensively functional and physicochemical analyses for AMPs based on the large-scale transcriptome and proteome data. Table 2 shows the comparison of system functionalities between dbAMP and other AMP databases. Significant improvements available in dbAMP comprise the information of AMP–protein interactions, antimicrobial potency analysis for 'cryptic' region detection, annotations of AMP against target species and AMP detection on transcriptome and proteome datasets. To our knowledge, the dbAMP is the first resource for providing a downloadable package to discover known and novel AMPs on high-throughput omics data. Additionally, user-friendly visualization interfaces have been designed to facilitate peptide searching, browsing and sequence alignment against dbAMP entries. All in all, the dbAMP can promote functional analyses of antimicrobial peptides and become a valuable resource for the discovery of new antimicrobial drugs.

DATA AVAILABILITY

The data content in dbAMP will be maintained and updated quarterly by continuously surveying the public resources and research articles. The database-assistant system is now freely accessed online at http://csb.cse.yzu.edu.tw/dbAMP/. All of the experimentally verified AMPs as well as the putative AMP dataset could be downloaded in the text format. Additionally, the Supplementary Figures S1–S9 and Tables S1–S5 are available at NAR online.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors sincerely appreciate the Warshel Institute for Computational Biology. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

FUNDING

The Chinese University of Hong Kong (ShenZhen). Funding for open access charge: Start-up fund from The Chinese University of Hong Kong (ShenZhen).
Conflict of interest statement. None declared.

REFERENCES

1. Vizioli,J. and Salzet,M. (2002) Antimicrobial peptides from animals: focus on invertebrates. *Trends Pharmacol. Sci.*, **23**, 494–496.
2. Brogden,K.A., Ackermann,M., McCray,P.B. Jr. and Tack,B.F. (2003) Antimicrobial peptides in animals and their role in host defences. *Int. J. Antimicrob. Agents*, **22**, 465–478.
3. Maroti,G., Kereszt,A., Kondorosi,E. and Mergaert,P. (2011) Natural roles of antimicrobial peptides in microbes, plants and animals. *Res. Microbiol.*, **162**, 363–374.
4. Papagianni,M. (2003) Ribosomally synthesized peptides with antimicrobial properties: biosynthesis, structure, function, and applications. *Biotechnol.. Adv*, **21**, 465–499.
5. Sitaram,N. and Nagaraj,R. (2002) Host-defense antimicrobial peptides: importance of structure for activity. *Curr. Pharm. Des.*, **8**, 727–742.
6. Durr,U.H., Sudheendra,U.S. and Ramamoorthy,A. (2006) LL-37, the only human member of the cathelicidin family of antimicrobial peptides. *Biochim. Biophys. Acta*, **1758**, 1408–1425.
7. Yeaman,M.R. and Yount,N.Y. (2003) Mechanisms of antimicrobial peptide action and resistance. *Pharmacol. Rev.*, **55**, 27–55.
8. Brogden,K.A. (2005) Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria? *Nat. Rev. Microbiol.*, **3**, 238–250.
9. Kim,I.W., Lee,J.H., Subramaniyam,S., Yun,E.Y., Kim,I., Park,J. and Hwang,J.S. (2016) De novo transcriptome analysis and detection of antimicrobial peptides of the American Cockroach Periplaneta americana (Linnaeus). *PLoS One*, **11**, e0155304.
10. Scott,M.G. and Hancock,R.E. (2000) Cationic antimicrobial peptides and their multifunctional role in the immune system. *Crit. Rev. Immunol.*, **20**, 407–431.
11. Bradshaw,J. (2003) Cationic antimicrobial peptides: issues for potential clinical use. *BioDrugs*, **17**, 233–240.
12. Wang,Z. and Wang,G. (2004) APD: the antimicrobial peptide database. *Nucleic Acids Res.*, **32**, D590–D592.
13. Giuliani,A. and Rinaldi,A.C. (2011) Beyond natural antimicrobial peptides: multimeric peptides and other peptidomimetic approaches. *Cell. Mol. Life Sci.*, **68**, 2255–2266.
14. Brogden,N.K. and Brogden,K.A. (2011) Will new generations of modified antimicrobial peptides improve their potential as pharmaceuticals? *Int. J. Antimicrob. Agents*, **38**, 217–225.
15. Gaspar,D., Veiga,A.S. and Castanho,M.A. (2013) From antimicrobial to anticancer peptides. A review. *Front. Microbiol.*, **4**, 294.
16. Chu,H.L., Yip,B.S., Chen,K.H., Yu,H.Y., Chih,Y.H., Cheng,H.T., Chou,Y.T. and Cheng,J.W. (2015) Novel antimicrobial peptides with high anticancer activity and selectivity. *PLoS One*, **10**, e0126390.
17. Koczulla,A.R. and Bals,R. (2003) Antimicrobial peptides: current status and therapeutic potential. *Drugs*, **63**, 389–406.
18. 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.
19. Marr,A.K., Gooderham,W.J. and Hancock,R.E. (2006) Antibacterial peptides for therapeutic use: obstacles and realistic outlook. *Curr. Opin. Pharmacol.*, **6**, 468–472.
20. 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.
21. Gueguen,Y., Garnier,J., Robert,L., Lefranc,M.P., Mougenot,I., de Lorgeril,J., Janech,M., Gross,P.S., Warr,G.W., Cuthbertson,B. *et al.* (2006) PenBase, the shrimp antimicrobial peptide penaeidin database: sequence-based classification and recommended nomenclature. *Dev. Comp. Immunol.*, **30**, 283–288.
22. Hammami,R., Ben Hamida,J., Vergoten,G. and Fliss,I. (2009) PhytAMP: a database dedicated to antimicrobial plant peptides. *Nucleic Acids Res.*, **37**, D963–D968.
23. Novkovic,M., Simunic,J., Bojovic,V., Tossi,A. and Juretic,D. (2012) DADP: the database of anuran defense peptides. *Bioinformatics*, **28**, 1406–1407.
24. Hammami,R., Zouhir,A., Le Lay,C., Ben Hamida,J. and Fliss,I. (2010) BACTIBASE second release: a database and tool platform for bacteriocin characterization. *BMC Microbiol.*, **10**, 22.
25. van Heel,A.J., de Jong,A., Song,C., Viel,J.H., Kok,J. and Kuipers,O.P. (2018) BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. *Nucleic Acids Res.*, **46**, W278–W281.
26. 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.
27. Waghu,F.H., Barai,R.S., Gurung,P. and Idicula-Thomas,S. (2016) CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Res.*, **44**, D1094–D1097.
28. 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.
29. Fjell,C.D., Hancock,R.E. and Cherkasov,A. (2007) AMPer: a database and an automated discovery tool for antimicrobial peptides. *Bioinformatics*, **23**, 1148–1155.
30. Zhao,X., Wu,H., Lu,H., Li,G. and Huang,Q. (2013) LAMP: A database linking antimicrobial peptides. *PLoS One*, **8**, e66557.
31. Wi,C.I., Sohn,S., Rolfes,M.C., Seabright,A., Ryu,E., Voge,G., Bachman,K.A., Park,M.A., Kita,H., Croghan,I.T. *et al.* (2017) Application of a natural language processing algorithm to asthma ascertainment. An automated chart review. *Am. J. Respir. Crit. Care Med.*, **196**, 430–437.
32. Seebah,S., Suresh,A., Zhuo,S., Choong,Y.H., Chua,H., Chuon,D., Beuerman,R. and Verma,C. (2007) Defensins knowledgebase: a manually curated database and information source focused on the defensins family of antimicrobial peptides. *Nucleic Acids Res.*, **35**, D265–D268.
33. Baltzer,S.A. and Brown,M.H. (2011) Antimicrobial peptides: promising alternatives to conventional antibiotics. *J. Mol. Microbiol. Biotechnol.*, **20**, 228–235.
34. Pirtskhalava,M., Gabrielian,A., Cruz,P., Griggs,H.L., Squires,R.B., Hurt,D.E., Grigolava,M., Chubinidze,M., Gogoladze,G., Vishnepolsky,B. *et al.* (2016) DBAASP v.2: an enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides. *Nucleic Acids Res.*, **44**, D1104–D1112.
35. Liu,S., Fan,L., Sun,J., Lao,X. and Zheng,H. (2017) Computational resources and tools for antimicrobial peptides. *J. Pept. Sci.*, **23**, 4–12.
36. MacLean,D., Jones,J.D. and Studholme,D.J. (2009) Application of'next-generation'sequencing technologies to microbial genetics. *Nat. Rev. Microbiol.*, **7**, 287.
37. Pompanon,F., Deagle,B.E., Symondson,W.O., Brown,D.S., Jarman,S.N. and Taberlet,P. (2012) Who is eating what: diet assessment using next generation sequencing. *Mol. Ecol.*, **21**, 1931–1950.
38. Tan,B., Ng,C., Nshimyimana,J.P., Loh,L.L., Gin,K.Y. and Thompson,J.R. (2015) Next-generation sequencing (NGS) for assessment of microbial water quality: current progress, challenges, and future opportunities. *Front. Microbiol.*, **6**, 1027.
39. Coordinators,N.R. (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **44**, D7–D19.
40. UniProt Consortium, T. (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **46**, 2699.
41. Rose,P.W., Prlic,A., Bi,C., Bluhm,W.F., Christie,C.H., Dutta,S., Green,R.K., Goodsell,D.S., Westbrook,J.D., Woo,J. *et al.* (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res.*, **43**, D345–D356.
42. Lata,S., Mishra,N.K. and Raghava,G.P. (2010) AntiBP2: improved version of antibacterial peptide prediction. *BMC Bioinformatics*, **11**, S19.
43. Wang,H. and Li,Y. (2015) Co-decision matrix framework for name entity recognition in biomedical text. *Int. J. Data Mining Bioinformatics*, **11**, 412–423.
44. Finn,R.D., Attwood,T.K., Babbitt,P.C., Bateman,A., Bork,P., Bridge,A.J., Chang,H.Y., Dosztanyi,Z., El-Gebali,S., Fraser,M. *et al.* (2017) InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.*, **45**, D190–D199.
45. Seet,B.T., Dikic,I., Zhou,M.M. and Pawson,T. (2006) Reading protein modifications with interaction domains. *Nat. Rev. Mol. Cell. Biol.*, **7**, 473–483.
46. Huang,K.Y., Su,M.G., Kao,H.J., Hsieh,Y.C., Jhong,J.H., Cheng,K.H., Huang,H.D. and Lee,T.Y. (2016) dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins. *Nucleic Acids Res.*, **44**, D435–D446.

47. Le,C.F., Fang,C.M. and Sekaran,S.D. (2017) Intracellular targeting mechanisms by antimicrobial peptides. *Antimicrob. Agents Chemother.*, **61**, e02340-16.
48. Nicolas,P. (2009) Multifunctional host defense peptides: intracellular-targeting antimicrobial peptides. *FEBS J.*, **276**, 6483–6496.
49. Warfield,R., Bardelang,P., Saunders,H., Chan,W.C., Penfold,C., James,R. and Thomas,N.R. (2006) Internally quenched peptides for the study of lysostaphin: An antimicrobial protease that kills Staphylococcus aureus. *Org. Biomol. Chem.*, **4**, 3626–3638.
50. Ellermeier,C.D. and Losick,R. (2006) Evidence for a novel protease governing regulated intramembrane proteolysis and resistance to antimicrobial peptides in Bacillus subtilis. *Genes Dev.*, **20**, 1911–1922.
51. Shinnar,A.E., Butler,K.L. and Park,H.J. (2003) Cathelicidin family of antimicrobial peptides: proteolytic processing and protease resistance. *Bioorg. Chem.*, **31**, 425–436.
52. Pane,K., Durante,L., Crescenzi,O., Cafaro,V., Pizzo,E., Varcamonti,M., Zanfardino,A., Izzo,V., Di Donato,A. and Notomista,E. (2017) Antimicrobial potency of cationic antimicrobial peptides can be predicted from their amino acid composition: application to the detection of "cryptic" antimicrobial peptides. *J. Theor. Biol.*, **419**, 254–265.
53. Notomista,E., Falanga,A., Fusco,S., Pirone,L., Zanfardino,A., Galdiero,S., Varcamonti,M., Pedone,E. and Contursi,P. (2015) The identification of a novel Sulfolobus islandicus CAMP-like peptide points to archaeal microorganisms as cell factories for the production of antimicrobial molecules. *Microb. Cell Fact.*, **14**, 126.
54. Wiradharma,N., Khoe,U., Hauser,C.A., Seow,S.V., Zhang,S. and Yang,Y.Y. (2011) Synthetic cationic amphiphilic alpha-helical peptides as antimicrobial agents. *Biomaterials*, **32**, 2204–2212.
55. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
56. Xiao,X., Wang,P., Lin,W.Z., Jia,J.H. and Chou,K.C. (2013) iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal. Biochem.*, **436**, 168–177.
57. Wang,P., Hu,L., Liu,G., Jiang,N., Chen,X., Xu,J., Zheng,W., Li,L., Tan,M., Chen,Z. *et al.* (2011) Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. *PLoS One*, **6**, e18476.
58. 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.
59. Bhadra,P., Yan,J., Li,J., Fong,S. and Siu,S.W.I. (2018) AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. *Sci. Rep.*, **8**, 1697.
60. Weng,S.L., Huang,K.Y., Kaunang,F.J., Huang,C.H., Kao,H.J., Chang,T.H., Wang,H.Y., Lu,J.J. and Lee,T.Y. (2017) Investigation and identification of protein carbonylation sites based on position-specific amino acid composition and physicochemical features. *BMC Bioinformatics*, **18**, 66.
61. Wong,Y.H., Lee,T.Y., Liang,H.K., Huang,C.M., Wang,T.Y., Yang,Y.H., Chu,C.H., Huang,H.D., Ko,M.T. and Hwang,J.K. (2007) KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res.*, **35**, W588–W594.
62. Bui,V.M., Weng,S.L., Lu,C.T., Chang,T.H., Weng,J.T. and Lee,T.Y. (2016) SOHSite: incorporating evolutionary information and physicochemical properties to identify protein S-sulfenylation sites. *BMC Genomics*, **17**, 9.
63. Liang,S.Y., Wu,S.W., Pu,T.H., Chang,F.Y. and Khoo,K.H. (2014) An adaptive workflow coupled with Random Forest algorithm to identify intact N-glycopeptides detected from mass spectrometry. *Bioinformatics*, **30**, 1908–1916.
64. Hall,M., Frank,E., Holmes,G., Pfahringer,B., Reutemann,P. and Witten,I.H. (2009) The WEKA data mining software: an update. *ACM SIGKDD Explor. Newslett.*, **11**, 10–18.
65. Weng,J.T., Wu,L.C., Chang,W.C., Chang,T.H., Akutsu,T. and Lee,T.Y. (2014) Novel bioinformatics approaches for analysis of high-throughput biological data. *Biomed. Res. Int.*, **2014**, 814092.
66. Langdon,W.B. (2015) Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Mining*, **8**, 1.
67. Menegidio,F.B., Jabes,D.L., Costa de Oliveira,R. and Nunes,L.R. (2018) Dugong: a Docker image, based on Ubuntu Linux, focused on reproducibility and replicability for bioinformatics analyses. *Bioinformatics*, **34**, 514–515.
68. Bordin,N. and Devos,D.P. (2018) ICBdocker: a Docker image for proteome annotation and visualization. *Bioinformatics*, doi:10.1093/bioinformatics/bty493.
69. Zouhir,A., Jridi,T., Nefzi,A., Ben Hamida,J. and Sebei,K. (2016) Inhibition of methicillin-resistant Staphylococcus aureus (MRSA) by antimicrobial peptides (AMPs) and plant essential oils. *Pharm Biol*, **54**, 3136–3150.
70. Mishra,B. and Wang,G. (2012) The importance of amino acid composition in natural AMPs: An evolutional, structural, and functional perspective. *Front Immunol*, **3**, 221.
71. Dimarcq,J.L., Bulet,P., Hetru,C. and Hoffmann,J. (1998) Cysteine-rich antimicrobial peptides in invertebrates. *Biopolymers*, **47**, 465–477.
72. Kumar,P., Kizhakkedathu,J.N. and Straus,S.K. (2018) Antimicrobial peptides: diversity, mechanism of action and strategies to improve the activity and biocompatibility in vivo. *Biomolecules*, **8**, E4.
73. Tam,J.P., Wang,S., Wong,K.H. and Tan,W.L. (2015) Antimicrobial Peptides from Plants. *Pharmaceuticals (Basel)*, **8**, 711–757.
74. Lorenzini,D.M., da Silva,P.I. Jr., Fogaca,A.C., Bulet,P. and Daffre,S. (2003) Acanthoscurrin: a novel glycine-rich antimicrobial peptide constitutively expressed in the hemocytes of the spider Acanthoscurria gomesiana. *Dev. Comp. Immunol.*, **27**, 781–791.
75. Sperstad,S.V., Haug,T., Vasskog,T. and Stensvag,K. (2009) Hyastatin, a glycine-rich multi-domain antimicrobial peptide isolated from the spider crab (Hyas araneus) hemocytes. *Mol. Immunol.*, **46**, 2604–2612.
76. Verdon,J., Coutos-Thevenot,P., Rodier,M.H., Landon,C., Depayras,S., Noel,C., La Camera,S., Moumen,B., Greve,P., Bouchon,D. *et al.* (2016) Armadillidin H, a Glycine-Rich peptide from the terrestrial crustacean armadillidium vulgare, displays an unexpected wide antimicrobial spectrum with membranolytic activity. *Front. Microbiol.*, **7**, 1484.
77. Chang,K.Y., Lin,T.P., Shih,L.Y. and Wang,C.K. (2015) Analysis and prediction of the critical regions of antimicrobial peptides based on conditional random fields. *PLoS One*, **10**, e0119490.
78. Yin,L.M., Edwards,M.A., Li,J., Yip,C.M. and Deber,C.M. (2012) Roles of hydrophobicity and charge distribution of cationic antimicrobial peptides in peptide-membrane interactions. *J. Biol. Chem.*, **287**, 7738–7745.
79. 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.
80. Chojnacki,S., Cowley,A., Lee,J., Foix,A. and Lopez,R. (2017) Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. *Nucleic Acids Res.*, **45**, W550–W553.
81. Chen,Y., Guarnieri,M.T., Vasil,A.I., Vasil,M.L., Mant,C.T. and Hodges,R.S. (2007) Role of peptide hydrophobicity in the mechanism of action of alpha-helical antimicrobial peptides. *Antimicrob. Agents Chemother.*, **51**, 1398–1406.
82. Steinert,P.M. and Marekov,L.N. (1997) Direct evidence that involucrin is a major early isopeptide cross-linked component of the keratinocyte cornified cell envelope. *J. Biol. Chem.*, **272**, 2021–2030.
83. Yu,K.S., Lee,Y., Kim,C.M., Park,E.C., Choi,J., Lim,D.S., Chung,Y.H. and Koh,S.S. (2010) The protease inhibitor, elafin, induces p53-dependent apoptosis in human melanoma cells. *Int. J. Cancer*, **127**, 1308–1320.
84. Anunthawan,T., de la Fuente-Nunez,C., Hancock,R.E. and Klaynongsruang,S. (2015) Cationic amphipathic peptides KT2 and RT2 are taken up into bacterial cells and kill planktonic and biofilm bacteria. *Biochim. Biophys. Acta*, **1848**, 1352–1358.
85. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
86. Edgar,R.C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods*, **10**, 996–998.

87. Keegan,K.P., Glass,E.M. and Meyer,F. (2016) MG-RAST, a metagenomics service for analysis of microbial community structure and function. *Methods Mol. Biol.*, **1399**, 207–233.
88. Huang,K.Y., Chang,T.H., Jhong,J.H., Chi,Y.H., Li,W.C., Chan,C.L., Robert Lai,K. and Lee,T.Y. (2017) Identification of natural antimicrobial peptides from bacteria through metagenomic and metatranscriptomic analysis of high-throughput transcriptome data of Taiwanese oolong teas. *BMC Syst. Biol.*, **11**, 131.
89. Herraez,A. (2006) Biomolecules in the computer: Jmol to the rescue. *Biochem. Mol. Biol. Educ.*, **34**, 255–261.
90. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
91. Park,S.C., Park,Y. and Hahm,K.S. (2011) The role of antimicrobial peptides in preventing multidrug-resistant bacterial infections and biofilm formation. *Int. J. Mol. Sci.*, **12**, 5971–5992.