Bioinformatics, 31(15), 2015, 2553–2559
doi: 10.1093/bioinformatics/btv180
Advance Access Publication Date: 29 March 2015
Original Paper

Data and text mining

Overlap and diversity in antimicrobial peptide databases: compiling a non-redundant set of sequences

Longendri Aguilera-Mendoza[1], Yovani Marrero-Ponce[2,3,4,*], Roberto Tellez-Ibarra[1], Monica T. Llorente-Quesada[1], Jesús Salgado[4], Stephen J. Barigye[5] and Jun Liu[6]

[1]Grupo de Investigación de Bioinformática, Centro de Estudio de Matemática Computacional, Universidad de las Ciencias Informáticas, La Habana, Cuba, [2]Grupo de Investigación en Estudios Químicos y Biológicos, Facultad de Ciencias Básicas, Universidad Tecnológica de Bolívar, Cartagena de Indias, Bolívar, Colombia, [3]Facultad de Química Farmacéutica, Universidad de Cartagena, Cartagena de Indias, Bolívar, Colombia, [4]Instituto de Ciencia Molecular (ICMol), Universitat de València, C/ Catedrático José Beltrán, 2, 46980, Paterna (Valencia), Spain, [5]Departamento de Química, Universidade Federal de Lavras, UFLA Caixa Postal 3037, 37200-000 Lavras, MG, Brazil and [6]School of Computing and Mathematics, Faculty of Computing and Engineering, Ulster University, Jordanstown campus, Northern Ireland, UK

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on November 26, 2014; revised on February 23, 2015; accepted on March 24, 2015

Abstract

Motivation: The large variety of antimicrobial peptide (AMP) databases developed to date are characterized by a substantial overlap of data and similarity of sequences. Our goals are to analyze the levels of redundancy for all available AMP databases and use this information to build a new non-redundant sequence database. For this purpose, a new software tool is introduced.

Results: A comparative study of 25 AMP databases reveals the overlap and diversity among them and the internal diversity within each database. The overlap analysis shows that only one database (Peptaibol) contains exclusive data, not present in any other, whereas all sequences in the LAMP_Patent database are included in CAMP_Patent. However, the majority of databases have their own set of unique sequences, as well as some overlap with other databases. The complete set of non-duplicate sequences comprises 16 990 cases, which is almost half of the total number of reported peptides. On the other hand, the diversity analysis identifies the most and least diverse databases and proves that all databases exhibit some level of redundancy. Finally, we present a new parallel-free software, named Dover Analyzer, developed to compute the overlap and diversity between any number of databases and compile a set of non-redundant sequences. These results are useful for selecting or building a suitable representative set of AMPs, according to specific needs.

Availability and implementation: The regularly updated non-redundant sequence databases and the Dover Analyzer software to perform custom analysis are available at http://mobiosd-hub.com/doveranalyzer/.

Contact: ymarrero77@yahoo.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Antimicrobial peptides (AMPs) have emerged as an alternative solution against multi-drug-resistance infections (Engler *et al.*, 2012). These compounds are part of the innate host defense system in vertebrate and invertebrate organisms and exhibit a broad-spectrum activity against a variety of pathogens, including bacteria and fungi. Additionally, antitumor (Gaspar *et al.*, 2013), antiviral (Jenssen *et al.*, 2004) and antiparasitic (Mor, 2009) activities have been described for some of these peptides. This provides a valuable ground for the discovery of new drugs that mimic the antimicrobial activity or that possess in some cases anticancer, antiviral or antiparasitic functions (Fjell *et al.*, 2012).

A growing number of naturally occurring AMPs are being discovered thanks to huge research efforts. Additionally, synthetic new peptides are being designed to understand, imitate or improve the natural ones. For example, although one of the first published AMP databases (Wang and Wang, 2004) comprised 523 peptides, mainly from natural sources, a recent one (Zhao *et al.*, 2013) holds 3904 natural and 1643 synthetic cases. Today there are many AMP databases available, which depending on the origin of peptides can be classified as 'General' or 'Specific' (Waghu *et al.*, 2014). The General databases comprise entries from a large variety of sources, whereas the Specific ones are dedicated to store AMPs that come from a particular type of organism, like bacteria, fungi or plants, or belong to a particular family, like defensins or penaeidin. It is clear that the General databases must share some identical sequences, at least with the specialized ones. Similarly, there can be overlap between databases designed for different purposes due to the multiple functions of AMPs.

Non-identical but similar sequences may likewise be found in biological databases. Tolerating some redundancy of content can be useful to derive a set of sequence-based physicochemical properties, which would be essential for molecular functions (Torrent *et al.*, 2011). However, duplication and redundancy in a large collection of biological data can also introduce distortions in the analysis of sequence–structure or sequence–property relationships or affect the efficiency in a similarity search. To avoid these problems, the duplicate entries can be merged and closely similar sequences can be clustered, thus yielding a representative set that covers the sequence space. In this article, we aim to analyze the levels of overlap and diversity for the case of well-known AMP databases and to compile a new non-redundant sequence database. To fulfill this purpose, we have developed a software tool, which is also presented here.

2 Materials and methods

2.1 Databases

There are 20 databases of AMPs freely available through the internet. Next we describe them briefly, in alphabetic order, mainly considering the sources of data and the particular purpose for their creation. Some of them were disaggregated into sub-databases, giving the total of 25 databases used for this work (listed in Table 1).

AMSDb (Tossi and Sandri, 2002) contains eukaryotic sequences produced by organisms ranging from fungi and protozoa to plants, insects, fish, amphibians and mammals. It was derived principally from the Swiss-Prot database.

AMPer (Fjell *et al.*, 2007) is a database compiled from AMSDb and UniProt. It encompasses all major AMP classes, including defensins, cathelicidins and granulin among others. This database was constructed to create hidden Markov models that enable recognition of individual classes of antimicrobials peptides.

APD (Wang *et al.*, 2009) provides peptides from all biological sources. They were collected from the literature *via* PubMed searches using keywords such as 'antimicrobial peptide', 'antibacterial peptide', 'antifungal peptide', 'anticancer peptide' and 'antitumor peptide'.

AVPdb (Qureshi *et al.*, 2014) contains AMPs that were experimentally verified for antiviral activity, excluding those targeting HIV because they were published in another database (see HIPdb in this section). The data were extracted mainly from Patent Lens and PubMed databases using text queries in the title/abstract fields, such as (((((virus OR viral) AND (peptide OR peptides) AND (inhibit * OR block*))))). The research articles returned by the above query were manually reviewed and filtered. Besides, the modified peptides were provided and, however, not considered in this study.

BACTIBASE (Hammami *et al.*, 2010) is specifically dedicated to Bacteriocins, which are AMPs produced by many bacteria and display growth-inhibition activity against other, closely related bacteria. All microbiological information was collected from the literature *via* PubMed search.

Bagel (de Jong *et al.*, 2010) is a web-based application that identifies the Open Reading Frames of putative bacteriocins, using knowledge-based bacteriocin and motif databases. The internal knowledge-based bacteriocin database was built from other databases such as ExPasY, NCBI and UniProt. Additionally, as not all known bacteriocins were present in these databases, a literature search was used to incorporate missing cases. Bagel also assigns each putative bacteriocin to a corresponding predicted Class. To accomplish this, the knowledge-based bacteriocin database has been split according to a classification scheme in Classes I, II and III. Class I contains three subclasses of lantibiotics according to Willey and van Der Donk (2007). Class II is made of four subclasses of bacteriocins according to (Cotter *et al.*, 2005) and Class III bacteriocins are relatively large proteins.

CAMP (Waghu *et al.*, 2014) contains sequences and structural information of AMPs retrieved from protein databases such as NCBI, UniProt and PDB, using combinations of keywords like 'antimicrobial', 'antibacterial', 'antifungal', 'antiviral' and 'antiparasitic'. The data in CAMP are sectioned into sequence, structure and patent databases. In turn, the sequence database is divided into two subdatabases, separating the experimentally validated peptides from the predicted ones. In any case, the subdatabase of predicted sequences has not been taken into account for this study.

DADP (Novković *et al.*, 2012) includes only anuran defense peptides obtained from research papers and UniProt. This database focuses on precursors and their specific regions (signal, acidic and bioactive). When the precursor structure is not reported, the peptide is only presented as the mature bioactive sequence.

DAMPD (Sundararajan *et al.*, 2012) is an update and a replacement of the ANTIMIC database (Brahmachary *et al.*, 2004). In this case, the AMPs were retrieved from UniProt and subsequently curated manually by selecting only the experimentally validated peptides.

DBAASP (Gogoladze *et al.*, 2014) contains information on AMPs of different origins (ribosomal, nonribosomal and synthetic) and level of complexity (monomers, dimers and two peptides) that have been collected from PubMed using the following keywords: 'antimicrobial', 'antibacterial', 'antifungal', 'antiviral', 'antitumour', 'anticancer' and 'antiparasitic peptides'. A two-peptide complex differs from a dimer in that the former consists of two different polypeptide chains without an interchain covalent bond, which act synergistically to yield a given antimicrobial activity, whereas the latter involves a covalent link between the peptides. In this study,

Table 1. List of databases used in this study, with summary of main characteristics

Databases[a]	Type[b]	Focused on	Single link to download all entries	Last update[c]	No. of entries		
					Total[d]	Non-duplicate dataset	Percentage[e]
AMPer	Specific	Eukaryotic AMPs	Present	February 2007	988	948	95.95
AMSDb	Specific	Eukaryotic AMPs	Absent	November 2004	893	858	96.08
APD	General	General AMPs	Present	January 2015	2495	2452	98.28
AVPdb	Specific	Antiviral peptides	Present	2013	2059	1817	88.25
Bactibase	Specific	Bacteriocins	Present	October 2014	227	215	94.71
Bagel_I	Specific	Bacteriocins	Absent	January 2013	158	154	97.47
Bagel_II	Specific	Bacteriocins	Absent	January 2013	228	217	95.18
Bagel_III	Specific	Bacteriocins	Absent	January 2013	93	71	76.34
CAMP_Patent	Specific	Patented AMPs	Absent	November 2013	1716	1675	97.61
CAMP_Structure	General	3D structures of AMPs	Absent	November 2013	682	496	72.73
CAMP_Validated	General	General AMPs	Absent	November 2013	2602	2498	96
DADP	Specific	Anuran defense peptides	Absent	March 2012	2571	2460	95.68
DAMPD	General	General AMPs	Absent	September 2011	1232	1199	97.32
DBAASP	General	General AMPs	Absent	2014	6427	5694	88.59
Defensins	Specific	Defensin family of AMPs	Absent	2007	536	507	94.59
HIPdb	Specific	HIV inhibiting peptides	Present	2013	981	887	90.42
LAMP_Experimental	General	General AMPs	Present	March 2013	3191	3190	99.97
LAMP_Patent	Specific	Patented AMPs	Present	March 2013	1491	1491	100
MilkAMP	Specific	AMPs of dairy origin	Present	2013	385	313	81.30
PenBase	Specific	Penaeidin family of AMPs	Absent	July 2008	28	28	100
Peptaibol	Specific	Peptaibol family	Absent	2004	316	198	62.66
PhytAMP	Specific	AMPs from plants	Present	January 2012	273	272	99.63
RAPD	Specific	Recombinantly produced AMPs	Absent	March 2010	179	122	68.16
UniProtKb	General	General AMPs	Present	January 2015	2788	2682	96.20
YADAMP	General	General AMPs	Absent	March 2013	2525	2525	100
Overall					35 064	16 990	48.45

[a]The citation corresponding to each database can be found in the description given in the text and the link to the website in Supplementary Information SI1.
[b]Type 'Specific' refers to data of a particular organism or to a particular class of AMP. Type 'General' refers to any organism or class of AMP.
[c]Date of last update of data deposited in the corresponding database, at the time when we downloaded such data.
[d]The total number of entries corresponds to those downloaded and processed successfully.
[e]Percentage of non-duplicate data with respect to the total number of entries.

only the polypeptide chains found as monomers were used, whereas the dimer and two-peptide entries were ignored.

Defensins (Seebah *et al.*, 2007) is focused on the defensin family (Ganz, 2003) of AMPs. It includes sequences from mammals, birds, invertebrates and plants. This information was gathered from bibliographic databases as a primary source and from sequence databases as a secondary source. For the primary source, a search query with the keyword 'defensin' was performed in SciFinder Scholar 2006. For the secondary source, UniProt and GenBank were used to collate the amino acid sequences of individual defensin peptides.

HIPdb (Qureshi *et al.*, 2013) provides experimentally verified HIV inhibiting peptides. They were collected from the PubMed database using the following search query ((((HIV) OR Human immunodeficiency virus)) AND ((peptide) OR peptides)) AND ((inhibit*) OR block*) in the title/abstract fields. The data so retrieved were manually curated.

LAMP (Zhao *et al.*, 2013) comprises natural and synthetic AMPs, which have been partitioned into three databases: experimental, predicted and patent. All the AMP sequences were collected manually from the scientific literature or from UniProt and other AMP-related databases, for which cross-links have been established. As for the case of CAMP, the predicted dataset was not included in this study.

MilkAMP (Théolier *et al.*, 2014) is dedicated to the AMPs of dairy origin that were found in PubMed and Google Scholar using a combination of keywords such as 'milk peptides', 'antimicrobial peptides' and common milk protein names, aliases and abbreviations.

PenBase (Gueguen *et al.*, 2006) has been developed to hold information about AMPs from the penaeid shrimp species. The database was built by collecting and analyzing all publicly accessible penaeidin sequence data.

Peptaibol (Whitmore and Wallace, 2004) was created to store the sequences and structure information of a class of peptides known as peptaibols (Chugh and Wallace, 2001). They are characterized by the presence of an unusual amino acid, α-aminoisobutyric acid and a C-terminal hydroxylated amino acid. Generally, these molecules have a fungal origin. To process the entries with the BioJava library (Holland *et al.*, 2008), the unusual amino acids in the sequences were replaced by 'X'.

PhytAMP (Hammami *et al.*, 2009) is designated to store antimicrobial plant peptides, which were collected from the UniProt database and from the scientific literature using PubMed.

RAPD (Li and Chen, 2008) comprises data from the literature by searching PubMed with the keywords 'recombinant antimicrobial peptides'. This database facilitates extraction of relevant information on recombinant approaches to produce AMPs.

YADAMP (Piotto *et al.*, 2012) contains AMPs from all available biological sources, ranging from bacteria and plants to animals, including humans. The data have been collected from existing literature searches and other AMP databases.

UniProtKb (Magrane *et al.*, 2011) acts as a central hub of protein knowledge by providing a unified view of protein sequence and functional information. It consists of two sections, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. UniProtKB/Swiss-Prot is manually curated (reviewed by a curator), while the records in UniProtKB/TrEMBL are generated automatically. In this study, we used a subset of this database, gathered by the following query: (keyword: 'Antimicrobial [KW-0929]' AND reviewed: yes).

2.2 Data collection and integration

Most of the available databases (Table 1) do not provide a single link to download all entries. For those cases, we used the GNU Wget software (http://www.gnu.org/software/wget/) in a non-interactive way to obtain crude HTML pages of registries and to avoid manual downloading of a large number of entries.

During the collection of data, we encountered errors for only 52 peptide sequences. In the case of AMSDb, 2 out of 895 entries were rejected, one of them because of a broken link and the other due to absence of sequence information. Empty sequences were also a cause for rejection of 30 out of 566 registries of the Defensins database and 20 out of 405 of the MilkAMP.

All retrieved HTML files were processed with the Community Edition of Pentaho Data Integration, also known as Kettle (http://community.pentaho.com/), which enables extraction, transformation and loading of data (ETL process) to make it available for analysis, data mining or reporting. Kettle contains a number of tools and utilities. The Spoon allows us to implement the ETL process by the graphical construction of diagrams. The implemented diagrams read the HTML files and clean up all malformed or faulty HTML to convert it into eXtensible Hypertext Markup Language (XHTML), with the help of the jTidy Java API (http://jtidy.sourceforge.net/). Finally, by querying the well-formed XHTML files with the XML Path Language (XPath), the names and sequences of AMPs were integrated into a single file with either the FASTA or Comma Separated Value (CSV) format.

2.3 Analysis of the databases

The overlap and diversity across databases were calculated following an approach by Voigt *et al.* (2001). The overlap of identical or similar compounds found in two databases represents the percentage of peptides in the first, for which identical or similar sequences exist in the second. On the other hand, the diversity was estimated by applying a clustering technique (Holm and Sander, 1998) and computing the corresponding number of clusters divided by the total of entries reported in the clustered database. Before determining the overlap and diversity, the following elements were taken into account. First, all duplicate sequences were removed to form a set of unique sequences for each database. Second, the pairwise sequence identity, calculated as a ratio of identical residues between two aligned peptides, above a given threshold, was used as a measure of similarity. Third, the sequence alignments were calculated by the dynamic programming algorithm of Needleman and Wunsch (1970), implemented in BioJava with the Blosum 62 substitution matrix and simple gap penalty parameters.

The clustering technique used to calculate the diversity and to compile a non-redundant sequence database is as follows. First, the list of peptides is sorted by the length of their sequences in a decreasing order. Then, a set of cluster representatives is created and initialized with the longest sequence of the sorted list. This set is completed by aligning each remaining peptide in the sorted list with all existing representatives and selecting as new representatives those peptides with pairwise sequence identities smaller than a given threshold value. At the end of this process, the cardinality of the representative set is taken as the number of clusters, while the list of representatives made a new non-redundant sequence database at the given identity threshold.

3 Results and discussion

For the sake of a realistic analysis of overlap and diversity, we removed all duplicate sequences from the databases. Table 1 lists the total number of sequences as well as the number and ratio of unique sequences for each database. The complete size, considering all databases, is 35 064 registries. Of these, 16 990 entries are unique sequences, i.e. less than half (48.45%) of the total number of peptides included in the databases.

There are only three databases without duplicate sequences: LAMP_Patent, PenBase and YADAMP (Table 1). The databases with highest duplication rates are Peptaibol and RAPD. In Peptaibol, the ratio of non-duplicates in the original sequences is 89.24% (with 282 unique sequences). However, this ratio decreases to 62.66% after replacement of the unusual amino acid residues by a virtual X residue. This is because the remaining conventional amino acids tend to be conserved in some specific positions, thus causing the rise in the number of duplicate sequences. In the case of RAPD, whose percentage of non-duplicates is 68.16%, the registries with identical sequences differ mainly in the literature references from which they were retrieved. Other low values of non-duplicate sequences are 72.73% and 76.34%, for CAMP_Structure and Bagel_III, respectively. In CAMP_Structure, the repeated sequences correspond to peptides with more than one PDB code, and in Bagel_III, they are identical sequences associated to more than one source and organism. For the remaining databases, the non-identical sequences are above 80% and in most cases, the repeated sequences are due to slight differences in the corresponding entries.

3.1 Overlap analysis

An exhaustive study reveals the percentages of coincidences between any pair of the collected databases (Fig. 1a). Considering an ascending order of overlap, this analysis shows that the Peptaibol database does not present any sequence in common with the others. In the second place comes Bagel_III, whose large sequences are hardly incorporated in any other data source. For instance, only six records of Bagel_III (8.45%) are present in Bactibase, one record (1.41%) is in common with CAMP_Structure, three (4.23%) with DAMPD and seven (9.86%) with UniProtKb. Other cases like AVPdb contain sequences in common with a greater number of databases but show low percentage of overlap with most of them, reaching the highest values with CAMP_Patent (4.13%) and LAMP_Patent (3.96%). LAMP_Patent also displays low overlap levels, but this database is included completely as a subset of CAMP_Patent. Another group of uncommon sequences are stored in MilkAMP and HIPdb, but some percentages of these sequences (18.53% and 34.27%, respectively) are included in the most recent database DBAASP. The most promiscuous databases display the highest levels of overlap and, as expected, they are of the General type: APD, CAMP_Structure, DAMPD, UniProtKb and YADAMP overlap with other 22 databases; CAMP_Validated and LAMP_Experimental overlap with 21, whereas DBAASP overlaps with 20. With respect to CAMP_Structure, we have classified it within the General type, in spite of its specific purpose, because it holds the 3D structures of entries from most other databases, excluding Petaibol and Penbase.

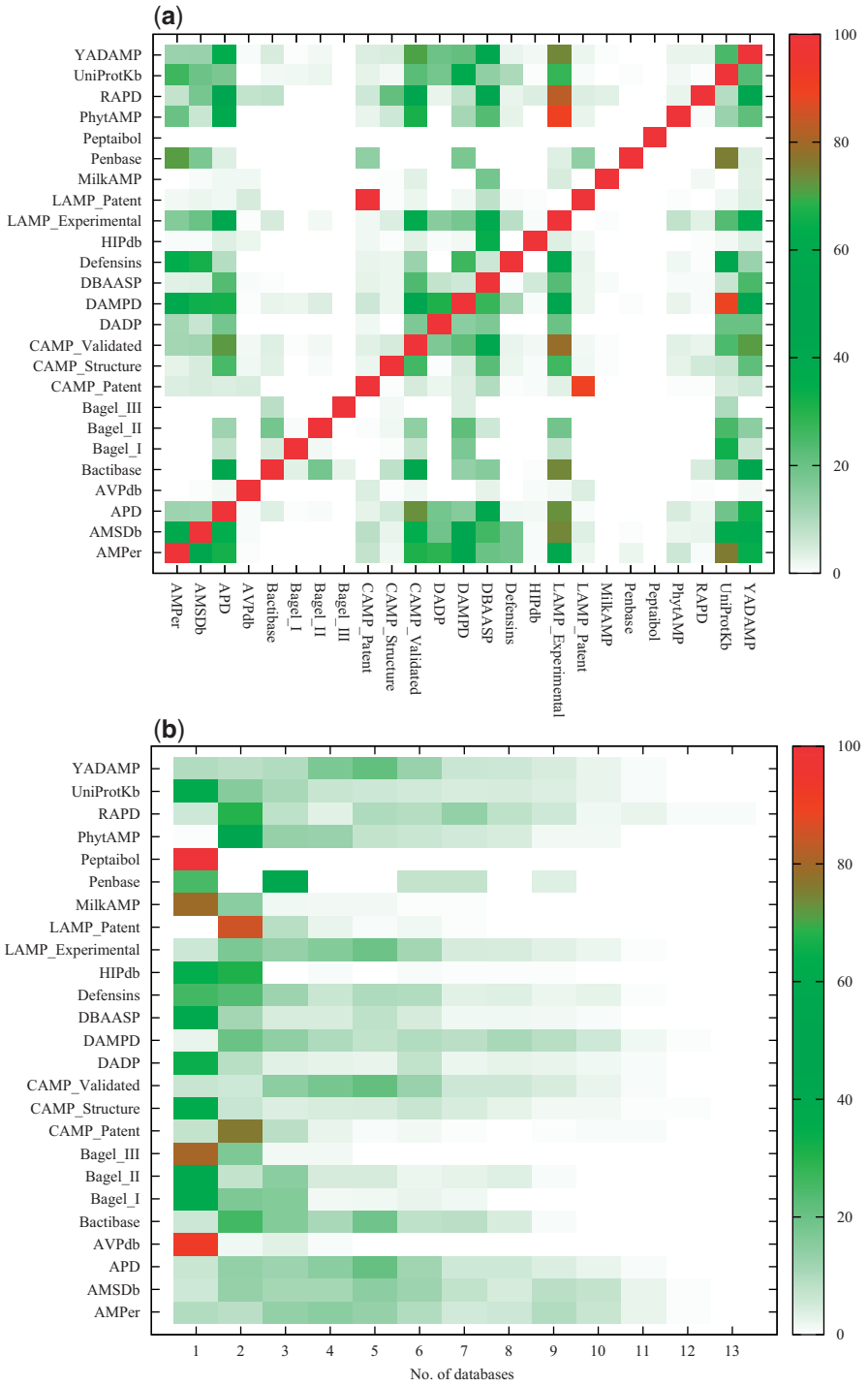

Fig. 1. Two heatmaps to visualize the percentages of overlap between databases. The percentage values are substituted by colored cells according to the scales given in the side bars, and they represent: (**a**) the fraction of the row database included in the column database and (**b**) the fraction of the row database stored exclusively in a given number of databases

Figure 1b illustrates another extension of the overlap analysis among databases. It consists of finding the peptide sequences shared exclusively by n databases. For $n = 1$, we found that all databases, except for LAMP_Patent, hold original sequences, i.e. not included in any other database. The highest levels of uniqueness are for Peptaibol (100%), AVPdb (92.52%), Bagel_III (80.28%) and MilkAMP (79.23%). For $n = 2$, we observed that LAMP_Patent and CAMP_Patent share the largest number of exclusive peptides. For $n = 3$, more than half of the peptides (57.14%) of Penbase are shared uniquely with two other databases. Nevertheless, in general, the percentage of overlap decreases with the increase of the variable n. Thus, there is only one sequence (KTCEHLAD TYRGVCFT NA SCDDHC KNKAHLIS GTCHNWKC FCTQNC) included in 13 databases ($n = 13$). This corresponds to a defense-related peptide found in pea seeds (*Pisum sativum*), which has antifungal activity and is included in AMPer, AMSDb, APD, CAMP_Structure, CAMP_Validated, DAMPD, DBAASP, Defensins, LAMP_Experimental, PhytAMP, RAPD, UniProtKb and YADAMP. From a

Table 2. Number of entries shared exclusively by a given number of databases

No. of databases	No. of entries
1	10 512
2	3140
3	982
4	726
5	688
6	406
7	177
8	151
9	112
10	67
11	23
12	5
13	1

global perspective, Table 2 lists the number of entries shared exclusively by n (1–13) databases. Note that, despite the overlap between databases, more than half of the entries in the complete dataset (10 512 out of 16 990) correspond to unique sequences deposited only in their own database.

3.2 Diversity analysis

Figure 2 shows the diversity ratio (number of clusters/number of entries) as a function of the identity threshold for each database and the complete dataset. According to the steepness of the curves, the databases may be categorized into three diversity groups: 'top', 'middle' and 'bottom'. The top group (characterized by the steepest slopes) comprised bacteriocins (Bagel_I, Bagel_II, Bagel_III, Bactibase) and RAPD. Note, however, that for identity threshold values greater than 0.5, the diversity ratio of Bagel_III decreases, showing as a result a trend similar to databases categorized as of middle diversity. As for the bottom diversity group, characterized by low gradient slopes for identity threshold values below 0.90, this contains Penbase, Peptaibol and DADP, where the low diversity is probably due to the abundance of conserved regions in the families of peptides in these particular databases.

Redundancy in the peptide databases is further evidenced by the fact that the number of clusters is less than the number of entries (number of clusters/number of entries < 1.0) for all identity thresholds less than 1; i.e. there is an identical or nearly identical fragment sequence shared by at least two entries (one cluster representative and, as a minimum, one cluster member) for all similarity cutoff values. The number of clusters is only equal to the number of entries (one entry per cluster) when the identity threshold is equal to 1, considering the exact match between sequences.

3.3 Software Dover Analyzer

To carry out most of the analysis described in this work, we have developed a software tool named Dover Analyzer. This software is a new wizard-like application that takes the collection of AMP databases as input and guides you through a few steps to compute the overlap, diversity and the corresponding non-redundant sequence databases (for table of the overlap, diversity ratio and sizes of the representative sets in the complete dataset, see Supplementary Information SI2 and SI3). Dover Analyzer is implemented in Java to achieve platform independence, and it is distributed free of charge along with a user manual. Although the set of AMP databases is embedded into the application, in the first step of the wizard, the user can keep or change the initial selection at will. In addition, new databases can be added using files in FASTA format.

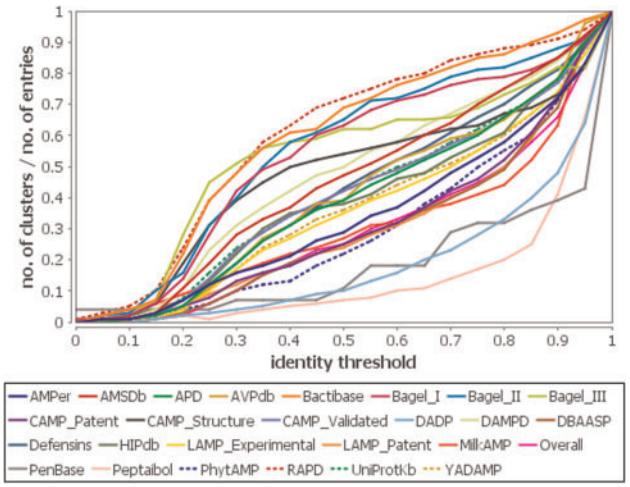

Fig. 2. Plot of the diversity ratio (number of clusters/number of entries) in each database and the complete dataset (overall), indicated in the legend, as a function of the identity threshold

In the second step of the wizard, the user specifies an output directory and a criterion for comparison. There are two criteria to compare sequence entries in the AMP databases: (i) string comparison of sequences or (ii) pairwise sequence identity above a given threshold. If the first option is chosen, the identical overlap and non-redundant sequence databases (without duplication) are computed. Upon choosing the second option, one can compute the similarity overlap, the diversity ratio and the non-redundant sequence databases (without the neighbors of the cluster representatives). Additionally, the user can choose the type of sequence alignment to be used (Needleman–Wunsch or Smith–Waterman), the scoring matrix (BLOSUM or PAM) and the sequence identity definition (number of identical residues/number of columns or number of identical residues/length of shorter sequence). From the third step onwards, additional settings which affect the output depending on the case study may be configured.

The computations are launched in the last step of the wizard. Most of the calculation time is spent on building the pairwise identity matrix, which stores the sequence identity between all pairs of peptides (if this type of comparison was specified). For n sequences, there are $n * (n-1)/2$ pairs of sequences that must be aligned. This is a time-consuming task and a parallel implementation has been developed to take advantage of current multi-core processors. The parallelization strategy to construct the matrix is as follows: because this matrix is symmetric, only the lower triangular part is used. This part is divided into squared and triangular sections, which are in turn subdivided until the portions of the big task are small enough to be executed in parallel by the multi-core processors.

4 Conclusion and future work

This study assesses the overlap and diversity across 25 AMP databases that have been created during the last decade. The overlap analysis reveals three main facts: (i) all databases, except for LAMP_Patent (which is included completely in CAMP_Patent), have some percentage of uniqueness. (ii) All databases, except for Peptaibol (which contains exclusive data), share some sequences with at least one database, with the largest mutual overlap between

the 13 databases being single peptide commonalities. (iii) Presently, there is no general database that collects the universe of reported AMP peptides. On the other hand, a comparison of the diversity ratios of all databases has revealed that bacteriocins are the most diverse sequences. Conversely, the AMPs corresponding to anuran species, peptaibol or penaeidin families are the least diverse. Moreover, there is some degree of redundancy in each database. To address this issue, we have developed a parallel-free software, named Dover Analyzer, which allows easy calculation of the overlap and diversity of any set of databases (existent, updated or new), as well as building (from them) a new non-redundant sequence database. This should allow potential users to compute their own representative set of AMP peptides using reported sequences. In the future, an integrated database and a search tool will be developed and published. It will require a meticulous effort to reconcile data from different sources, like the biological activity and other annotations, which were not considered for this study.

Funding

This work was supported by the Spanish MINECO (BFU2010-19118 to J.S., supported in part by the European Regional Development Fund).

Conflict of Interest: none declared.

References

Brahmachary,M. *et al*. (2004) Antimic: a database of antimicrobial sequences. *Nucleic Acids Res.*, **32**(Suppl. 1), D586–D589.

Chugh,J. and Wallace,B. (2001) Peptaibols: models for ion channels. *Biochem. Soc. Trans.*, **29**, 565–570.

Cotter,P.D. *et al*. (2005) Bacteriocins: developing innate immunity for food. *Nat. Rev. Microbiol.*, **3**, 777–788.

de Jong,A. *et al*. (2010) Bagel2: mining for bacteriocins in genomic data. *Nucleic Acids Res.*, **38**(Suppl. 2), W647–W651.

Engler,A.C. *et al*. (2012) Emerging trends in macromolecular antimicrobials to fight multi-drug-resistant infections. *Nano Today*, **7**, 201–222.

Fjell,C.D. *et al*. (2007) AMPer: a database and an automated discovery tool for antimicrobial peptides. *Bioinformatics*, **23**, 1148–1155.

Fjell,C.D. *et al*. (2012) Designing antimicrobial peptides: form follows function. *Nat. Rev. Drug Discov.*, **11**, 37–51.

Ganz,T. (2003) Defensins: antimicrobial peptides of innate immunity. *Nat. Rev. Immunol.*, **3**, 710–720.

Gaspar,D. *et al*. (2013) From antimicrobial to anticancer peptides. A review. *Front. Microbiol.*, **4**, 294.

Gogoladze,G. *et al*. (2014) DBAASP: database of antimicrobial activity and structure of peptides. *FEMS Microbiol. Lett.*, **357**, 63–68.

Gueguen,Y. *et al*. (2006) Penbase, the shrimp antimicrobial peptide penaeidin database: sequence-based classification and recommended nomenclature. *Dev. Comp. Immunol.*, **30**, 283–288.

Hammami,R. *et al*. (2009) Phytamp: a database dedicated to antimicrobial plant peptides. *Nucleic Acids Res.*, **37**(Suppl. 1), D963–D968.

Hammami,R. *et al*. (2010) Bactibase second release: a database and tool platform for bacteriocin characterization. *BMC Microbiol.*, **10**, 22.

Holland,R.C. *et al.* (2008) Biojava: an open-source framework for bioinformatics. *Bioinformatics*, **24**, 2096–2097.

Holm,L. and Sander,C. (1998) Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*, **14**, 423–429.

Jenssen,H. *et al*. (2004) A wide range of medium-sized, highly cationic, α-helical peptides show antiviral activity against herpes simplex virus. *Antiviral Res.*, **64**, 119–126.

Li,Y. and Chen,Z. (2008) RAPD: a database of recombinantly-produced antimicrobial peptides. *FEMS Microbiol. Lett*, **289**, 126–129.

Magrane,M. *et al*. (2011) Uniprot knowledgebase: a hub of integrated protein data. *Database*, **2011**, doi: 10.1093/database/bar009.

Mor,A. (2009) Multifunctional host defense peptides: antiparasitic activities. *FEBS J.*, **276**, 6474–6482.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Novković,M. *et al*. (2012) DADP: the database of anuran defense peptides. *Bioinformatics*, **28**, 1406–1407.

Piotto,S.P. *et al*. (2012) Yadamp: yet another database of antimicrobial peptides. *Int. J Antimicrob. Agents*, **39**, 346–351.

Qureshi,A. *et al*. (2013) Hipdb: a database of experimentally validated HIV inhibiting peptides. *PloS One*, **8**, e54908.

Qureshi,A. *et al*. (2014) Avpdb: a database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic Acids Res.*, **42**, D1147–D1153.

Seebah,S. *et al*. (2007) Defensins knowledgebase: a manually curated database and information source focused on the defensins family of antimicrobial peptides. *Nucleic Acids Res.*, **35**(Suppl. 1), D265–D268.

Sundararajan,V.S. *et al*. (2012) DAMPD: a manually curated antimicrobial peptide database. *Nucleic Acids Res.*, **40**, D1108–D1112.

Théolier,J. *et al*. (2014) Milkamp: a comprehensive database of antimicrobial peptides of dairy origin. *Dairy Sci. Technol.*, **94**, 181–193.

Torrent,M. *et al*. (2011) Connecting peptide physicochemical and antimicrobial properties by a rational prediction model. *PloS One*, **6**, e16968.

Tossi,A. and Sandri,L. (2002) Molecular diversity in gene-encoded, cationic antimicrobial polypeptides. *Curr. Pharm. Des.*, **8**, 743–761.

Voigt,J.H. *et al*. (2001) Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci*, **41**, 702–712.

Waghu,F.H. *et al*. (2014) Camp: collection of sequences and structures of antimicrobial peptides. *Nucleic Acids Res.*, **42**, D1154–D1158.

Wang,G. *et al*. (2009) APD2: the updated antimicrobial peptide database and its application in peptide design. *Nucleic Acids Res.*, **37**(Suppl. 1), D933–D937.

Wang,Z. and Wang,G. (2004) APD: the antimicrobial peptide database. *Nucleic Acids Res.*, **32**(Suppl. 1), D590–D592.

Whitmore,L. and Wallace,B. (2004) The peptaibol database: a database for sequences and structures of naturally occurring peptaibols. *Nucleic Acids Res.*, **32**(Suppl. 1), D593–D594.

Willey,J.M. and van Der Donk,W.A. (2007) Lantibiotics: peptides of diverse structure and function. *Annu. Rev. Microbiol.*, **61**, 477–501.

Zhao,X. *et al*. (2013) Lamp: a database linking antimicrobial peptides. *PloS One*, **8**, e66557.