D586–D589 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh032

ANTIMIC: a database of antimicrobial sequences

M. Brahmachary[1,2], S. P. T. Krishnan[1], J. L. Y. Koh[1], A. M. Khan[1], S. H. Seah[1], T. W. Tan[2], V. Brusic[1,3] and V. B. Bajic[1,*]

[1]Institute of Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613, [2]Department of Biochemistry, National University of Singapore, Singapore and [3]Department of Microbiology, National University of Singapore, Singapore

Received August 15, 2003; Revised August 25, 2003; Accepted September 11, 2003

ABSTRACT

Antimicrobial peptides (AMPs) are important components of the innate immune system of many species. These peptides are found in eukaryotes, including mammals, amphibians, insects and plants, as well as in prokaryotes. Other than having pathogen-lytic properties, these peptides have other activities like antitumor activity, mitogen activity, or they may act as signaling molecules. Their short length, fast and efficient action against microbes and low toxicity to mammals have made them potential candidates as peptide drugs. In many cases they are effective against pathogens that are resistant to conventional antibiotics. They can serve as natural templates for the design of novel antimicrobial drugs. Although there are vast amounts of data on natural AMPs, they are not available through one central resource. We have developed a comprehensive database (ANTIMIC, http://research.i2r.a-star.edu.sg/Templar/DB/ANTIMIC/) of known and putative AMPs, which contains ~1700 of these peptides. The database is integrated with tools to facilitate efficient extraction of data and their analysis at molecular level, as well as search for new AMPs. These tools include BLAST, PDB structure viewer and the Antimic profile module.

INTRODUCTION

Antimicrobial peptides (AMPs) are important constituents of the innate immune defense for all species of life (1). The field of AMPs is vast and generic since many peptides may exhibit microbe-killing properties and thus can fall into the category of AMPs. For example, toxins from some organisms, such as lycotoxins from the spider *Lycosa carolinensis* (2) and the pardaxins (3), exert antibacterial activity. Many AMPs are gene encoded, while others are secondary metabolites. AMPs are found in eukaryotes, including mammals, amphibians, insects and plants, as well as in prokaryotes (4–11). As well as having pathogen-lytic properties, these peptides can have other activities like antitumor activity, mitogen activity and signaling molecule activity (12). AMPs are short, provide fast and efficient action against microbes, and in many cases have low toxicity to mammals (13). In many cases they are effective against pathogens, which are resistant to conventional antibiotics (14). AMPs have a broad target spectrum that includes Gram-positive and Gram-negative bacteria, yeasts and fungi, and even certain enveloped viruses and protozoa (15). Therefore, they could be considered as natural design templates for anti-infectious agents in humans (13).

The design of novel peptides with specific and enhanced antimicrobial activities requires the development of methods for narrowing down the candidate peptides so as to enable rational experimentation by wet-lab scientists. An indispensable supporting tool in this effort is a systematized and user-friendly organized repository of information on AMPs, their sequences, properties and effects, as well as the supporting tools to enhance the analysis of specific peptide properties and peptide class.

Currently, data related to AMPs can be found scattered across different repositories at different locations. For example, in addition to specialized AMP databases, such as AMSDb (http://www.bbcm.univ.trieste.it/~tossi/pag1.html) and Peptaibol (16), the data related to AMPs are contained in general purpose biological databases such as GenBank (17), EMBL (18), Swiss-Prot (19), etc. There is also a Synthetic Antibiotic Peptide Database (SAPD) (20), which contains both chemical and biological information on all published synthetic antibiotic peptides. To the best of our knowledge no common platform or resource exists that enables easy access to systematized information about this broad class of peptides and provide analytical tools that facilitate analysis of these peptides and search for new members of the AMP classes.

To fill this gap, we have made a systematic effort to collect, clean and analyze AMPs, and organize them into the ANTIMIC database. This database is the most comprehensive resource to date on the experimental and putative natural AMPs, and represents a public resource with several integrated bioinformatics tools for the analysis of these data.

BUILDING THE DATABASE

The ANTIMIC database contains an extensive collection of antimicrobial sequences from many families. The database has been created on an in-house-developed data-warehousing platform (BioWare, sdmc.i2r.a-star.edu.sg/Templar) that enables rapid building of specialized searchable biological

*To whom correspondence should be addressed. Tel: +65 6874 8800; Fax: +65 6774 8056; Email: bajicv@i2r.a-star.edu.sg

databases. BioWare comprises three program modules: BioWare Retreive Module, which retrieves raw data from diverse sources on the internet; BioWare-Prep Module for processing of retrieved data; and Templar Module for integration of this information into a central repository. The processing includes generation of a report summary for removal of redundant entries and renumbering of entries, and other sub-modules such as a module for the generation of multiple alignments and a module for viewing cysteine bridge patterns to help the database creator to manage the information more efficiently.

The data have been extracted from public databases. Specific keyword search terms like 'alpha defensin' and generic keyword terms like 'antibacterial', 'antifungal', etc. were used within the BioWare Retrieve Module to search the NCBI's GenBank and Swiss-Prot databases.

This preliminary data set was checked for duplicates and redundancies, such as entries that may have been earlier versions of another entry. This was facilitated by the BioWare-Prep Module, which generates a sequence comparison report summary based on pairwise alignment of entries in the data set. Entries that had 100% sequence identity were reported as duplicates. Duplicates having the same name and taxon (organism source) were compared. The entry judged to contain the most complete information was kept while the others were not considered. Duplicates originating from different source species were kept as separate entries. Sequences that shared fragment or partial identity, where one sequence was an identical fragment of another, were checked for their uniqueness by referring to both literature and the cross-references field from the public databases. Most of these entries were earlier versions of other entries in the data set and hence were deleted. All deleted entries have been added to the ANTIMIC file of rejected entries (FRE), which is used to avoid future retrieval of the same entry during database updates. The resultant data set will be referred to in this text as the preliminary cleaned data set.

Next, each of the entries were checked manually to ensure that they are AMP entries and not irrelevant entries, examples including 'Integrin' or 'Reticulon 4 receptor precursor', which may have been picked up by the keyword search. Records eliminated at this step have also been recorded in the FRE. The final cleaned data set was used as the input to the Templar Module, which generated the online version of the ANTIMIC database (http://research.i2r.a-star.edu.sg/Templar/DB/ANTIMIC/).

The ANTIMIC database has a structure viewer module, which contains the PDB structures of antimicrobial sequences. The structure viewer was populated by searching the PDB database for 3D structures of antimicrobial sequences present in the ANTIMIC database. PDB accession numbers present in the annotation of entries in the ANTIMIC database were linked to their corresponding 3D structures.

DATABASE ORGANIZATION

Each ANTIMIC entry includes a concise description of the sequence, the scientific name and taxonomy of the source organism, bibliographic references, a table of features listing areas of biological significance, coding regions, peptide regions, sites of mutation or modification and the protein translation.

The annotation of each entry in the database contains the following fields (see Supplementary Material). A unique accession number 'DBACC' that defines each record in the ANTIMIC database. The format is [D][six digit number], where D denotes an entry of an AMP and the six-digit number is a unique descriptor of the entry. Next, the field 'Date' identifies the date when the entry was made. The fields 'Locus name', 'Sequence length', and 'GenBank division' contain information on the locus, length of the sequence, and the division group to which the sequence belongs in GenBank. In some entries 'GenBank division' is also known as 'Molecular type'. The date when the entry was updated by public database is shown in the field 'GenBank modification date' or, in some entries, as 'Release date'. The 'Name' field contains the name of the AMP as used in the literature, and if available, its common names. The 'Accession' field provides hyperlinks to the corresponding entries in the relevant external databases, GenBank and Swiss-Prot. The organism source of the AMP can be found in the 'Source' field and its taxonomy is shown in the 'Species' field. The 'Reference' field contains the literature references, with author names and titles. Relevant comments or observations can be found in the 'Comment' field. Structural features of AMPs, such as residues forming disulfide bridges, helices or strands, are described in the field 'Features'. Putative structural information derived by similarity to known structures is indicated as 'BY SIMILARITY'. Many entries have the field 'Link', which links that entry to other databases: EMBL, Pfam, ProDom, etc. The field 'Translation' provides the amino acid sequence of an AMP entry. If the PDB structure is available, the field 'Structure' contains an internal hyperlink to the PDB structure stored in the ANTIMIC database for relevant records.

INTEGRATED TOOLS

The ANTIMIC database contains several integrated tools to help in data extraction and analysis of AMP sequences. Data extraction and sequence viewing tools include: (i) keyword search, (ii) BLAST search and (iii) structure viewer.

The keyword search feature allows users to search the database using keywords. BLAST (21) search enables users to perform sequence similarity searches against the antimicrobial sequences stored in the database. Structure viewer allows the 3D structures of individual AMPs to be viewed.

The analysis-based tools consist of the Antimic profile tool. This tool has multiple modules. The modules allow for building of new profiles, querying new sequences against the build profiles or against the predefined profile library, as well as against either ANTIMIC or nr databases.

Users can access ANTIMIC entries by using either a simple keyword search such as species name, type of antimicrobial activity, Swiss-Prot or GenBank accession numbers, or they can perform complex searches for more specific results by using more than one keyword with the support of Boolean operators. For example, a simple search would be to use a keyword like 'Protegrin' to retrieve entries of this family. A complex search would be 'mellitin and wasp', which will return mellitin family related entries that are specific to the wasp species. Therefore, any term that is present in the

annotation of the entries can be used in combination with others to retrieve more specific results. The results are displayed in tabular form as a list. The list displays accession numbers, the species from which the AMP originates and the antimicrobial sequence name. The accession number is hyperlinked within the database to the full data record.

The database has integrated the BLAST program, which consists of a set of similarity search programs for protein or DNA sequences. The BLAST feature allows users to perform sequence comparison using the BLAST algorithm. A query amino acid sequence can be compared against all sequences in the ANTIMIC database. Users can choose to return the results either in standard BLAST output or as a color-coded multiple sequence alignment generated by Mview program (22). Mview highlights the positions of conserved and homologous amino acids in the multiple sequence alignment returned by BLAST.

For antimicrobial sequences that have an entry in the PDB the corresponding peptide structures can be seen through the structure feature using the Chime (23) or Swiss PDB viewer (24). The PDB files can also be downloaded. The latest version of Chime 2.6 SP4 is functional with Netscape (version 4.x) (Supplementary Material).

The Antimic profile tool aims to facilitate tentative classification of query sequences into different antimicrobial families. It uses a predefined antimicrobial-specific library of profiles, although users can generate profiles from their specific sequences. The profile library has been created from mature peptide regions of AMPs of different families. The Antimic profile tool suggests positions that represent the signature for the selected family and potentially may be crucial for antimicrobial activity, as well as those that are 'non-critical' in the functional domain of a family of sequences. The profiles used by this module can serve as templates to suggest to which family of antimicrobial sequence a query sequence may belong. The use of profiles enables the capture from public databases of homologs that have a high likelihood of belonging to a particular family.

The Antimic profile tool is based on HMMER (25) (a program that uses hidden Markov models for motif description). The Antimic profile tool has multiple features. It consists of a profile-building module known as 'Build profiles' that enables the creation of profiles from the sequences submitted by the user. The input sequences in this module can be in any format that is accepted by the program readseq. The module generates a ClustalW alignment of the sequences, which is used to generate the profile. The user can view the ClustalW alignment in the web browser. The results page gives the user the option to view the profile that has been generated or to use the profile for querying. If the option use profile is selected the user is directed to the 'Query profile' module. Using this module the user can input query sequences for query against the profile. The 'Query profile' module stores the profiles built by the user with an ID tag and also stores a permanent profile library 'antimicrobial.hmm'. The antimicrobial.hmm consists of HMM profiles of several families of AMPs. The families currently included are mellitin, magainin, bacteriocin, cecropin and protegrin. HMM profiles of individual families are also provided separately. We plan to gradually cover all those families of AMPs included in the ANTIMIC database for which the hmm profile models can create acceptable results.

The 'Query profile' module helps a user to predict to which family a query sequence most likely belongs and whether it is likely to share the same mode of action as the matched family of sequences. The results contain three sections: a ranked list of the best scoring HMMs, a list of the best scoring domains in order of their occurrence in the sequence, and alignments for the highest scoring domains. The matches are shown with scores (bits) and E-values. The bits score indicates how well the sequences match an HMM profile. The E-value, which is calculated from the bits score, shows the number of false positives that is expected to be seen at or above this bit score. Therefore an E-value of 0.1 indicates that there is only a 10% chance that the hit is false or has come up by chance. Hence, a low E-value is best. The best hits appear at the top of the results list. The critical residues (highly conserved residues) for both the query sequence and the consensus pattern for a family are shown in capital letters (Supplementary Material).

The second module is known as 'Query db'. Query db allows users to search for sequences in the GenBank 'nr' and ANTIMIC databases, which match specific profiles. These AMP profiles are predefined (for five AMP families) and could be used either as single profiles or as a library. Additionally, users may employ their own generated profiles.

COMPARISON WITH OTHERS

An attempt has been made in Italy to consolidate information about AMPs and store it in a database called AMSDb (http://www.bbcm.univ.trieste.it/~tossi/pag1.html). This database contains annotated AMP sequence data and enables a keyword search for categories such as ID, date, family, category, activity, organism source and generic keywords. The AMSdb database consists of a total of 804 entries (on 05 August 2003) of eukaryotic origin. This database does not provide any tools for the analysis of data.

Another database (http://public-1.cryst.bbk.ac.uk/peptaibol/home.shtml), the Peptaibol database, is a highly specialized one that contains over 300 entries of antibiotic peptides known as Peptaibols (16), which originate from fungal organisms like *Trichoderma* and *Emericellopsis*. This database enables users to search for information about Peptaibols by name or Peptaibol group. It also allows for searching of entries using motifs specific for Peptaibols (which are known to have non-standard amino acid residues in them). The Peptaibol database has a structure-viewing tool. The database has stored Peptaibol entries with PDB entries and enables users to view the structure from the database. The authors of this database have classified the Peptaibols into subfamilies based on the alignments of these sequences with common sequence features thought to be important for channel formation (16).

Our ANTIMIC database is the most comprehensive source of natural AMPs to date, which has been manually curated. It contains over 1700 antimicrobial sequences that have entries extracted from GenBank and Swiss-Prot. The entries come from both eukaryotic and prokaryotic organisms. The database is created with the intention of aiding molecular analysis of AMPs. In addition to comprehensive peptide information and AMP specifics, the ANTIMIC database has integrated data extraction tools, sequence similarity search tools, BLAST, a

peptide structure viewer tool and analytical tools like the Antimic profile module, all of which facilitate analysis and classification of AMPs.

CONCLUSION

ANTIMIC is a specialized database that has been built with the aim at making a comprehensive repository of natural AMPs complemented by data extraction and analysis tools to help further analysis of AMPs. One of the integrated tools, the Antimic profile module, enables users to assign a new putative antimicrobial sequence to a family and functional domain. It also enables the capture of new peptide homologs from other public databases.

SUPPLEMENTARY MATERIAL

The following figures are available as Supplementary Material at NAR Online: example of a full data record; example of an image produced by Structure viewer; example of the output of Query profile.

REFERENCES

1. Hancock,R.E. and Diamond,G. (2000) The role of cationic antimicrobial peptides in innate host defences. *Trends Microbiol.*, **8**, 402–410.
2. Yan,L. and Adams,M.E. (1998) Lycotoxins, antimicrobial peptides from venom of the wolf spider *Lycosa carolinensis*. *J. Biol. Chem.*, **273**, 2059–2066.
3. Bloch-Shilderman,E., Jiang,H. and Lazarovici,P. (2002) Pardaxin, an ionophore neurotoxin, induces PC12 cell death: activation of stress kinases and production of reactive oxygen species. *J. Nat. Toxins*, **11**, 71–85.
4. Garcia-Olmedo,F., Molina,A., Alamillo,J.M. and Rodriguez-Palenzuela,P. (1998) Plant defense peptides. *Biopolymers*, **47**, 479–491.
5. Rinaldi,A.C. (2002) Antimicrobial peptides from amphibian skin: an expanding scenario. *Curr. Opin. Chem. Biol.*, **6**, 799–804.
6. Lehrer,R.I. and Ganz,T. (2002) Defensins of vertebrate animals. *Curr. Opin. Immunol.*, **14**, 96–102.
7. Cole,A.M. and Ganz,T. (2000) Human antimicrobial peptides: analysis and application. *Biotechniques*, **29**, 822–831.
8. Hoffmann,J.A., Reichhart,J.M. and Hetru,C. (1996) Innate immunity in higher insects. *Curr. Opin. Immunol.*, **8**, 8–13.
9. Hoffmann,J.A. (1995) Innate immunity of insects. *Curr. Opin. Immunol.*, **7**, 4–10.
10. Hoffmann,J.A. and Hetru,C. (1992) Insect defensins: inducible antibacterial peptides. *Immunol. Today*, **13**, 411–415.
11. Zhai,Y. and Saier,M.H.,Jr (2000) The amoebapore superfamily. *Biochim. Biophys. Acta*, **1469**, 87–99.
12. Kamysz,W. Okrüj,M. and Lukasiak,J. (2003) Novel properties of antimicrobial peptides. *Acta Biochim. Pol.*, **50**, 461–469.
13. van't Hof,W., Veerman,E.C., Helmerhorst,E.J. and Amerongen,A.V. (2001) Antimicrobial peptides: properties and applicability. *Biol. Chem.*, **382**, 597–619.
14. Ganz,T. and Lehrer,R.I. (1999) Antibiotic peptides from higher eukaryotes: biology and applications. *Mol. Med. Today*, **5**, 292–297.
15. Hancock,R.E. and Lehrer,R. (1998) Cationic peptides: a new source of antibiotics. *Trends Biotechnol.*, **16**, 82–88.
16. Chugh,J.K. and Wallace,B.A. (2001) Peptaibols: models for ion channels. *Biochem. Soc. Trans.*, **29**, 565–570.
17. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27
18. Stoesser,G., Baker,W., van den Broek,A., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V., Lopez,R. *et al.* (2003) The EMBL Nucleotide Sequence Database: major new developments. *Nucleic Acids Res.*, **31**, 17–22.
19. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
20. Wade,D. and Englund,J. (2002) Synthetic antibiotic peptides database. *Protein Pept. Lett.*, **9**, 53–57.
21. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
22. Brown,N.P., Leroy,C. and Sander,C. (1998) MView: A web-compatible database search or multiple alignment viewer. *Bioinformatics*, **14**, 380–381.
23. Horton,R.M. (1999) Scripting Wizards for Chime and RasMol. *Biotechniques*, **26**, 874–876.
24. Guex,N. and Peitsch,M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.
25. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.