D1094–D1097 *Nucleic Acids Research, 2016, Vol. 44, Database issue* *Published online 13 October 2015*
doi: 10.1093/nar/gkv1051

$CAMP_{R3}$: a database on sequences, structures and signatures of antimicrobial peptides

Faiza Hanif Waghu, Ram Shankar Barai, Pratima Gurung and Susan Idicula-Thomas*

Biomedical Informatics Centre of Indian Council of Medical Research, National Institute for Research in Reproductive Health, Mumbai 400012, Maharashtra, India

Received September 15, 2015; Revised September 29, 2015; Accepted October 01, 2015

ABSTRACT

Antimicrobial peptides (AMPs) are known to have family-specific sequence composition, which can be mined for discovery and design of AMPs. Here, we present $CAMP_{R3}$; an update to the existing CAMP database available online at www.camp3.bicnirrh.res.in. It is a database of sequences, structures and family-specific signatures of prokaryotic and eukaryotic AMPs. Family-specific sequence signatures comprising of patterns and Hidden Markov Models were generated for 45 AMP families by analysing 1386 experimentally studied AMPs. These were further used to retrieve AMPs from online sequence databases. More than 4000 AMPs could be identified using these signatures. AMP family signatures provided in $CAMP_{R3}$ can thus be used to accelerate and expand the discovery of AMPs. $CAMP_{R3}$ presently holds 10247 sequences, 757 structures and 114 family-specific signatures of AMPs. Users can avail the sequence optimization algorithm for rational design of AMPs. The database integrated with tools for AMP sequence and structure analysis will be a valuable resource for family-based studies on AMPs.

INTRODUCTION

Antimicrobial peptides (AMPs) are host defense molecules produced by a wide range of organisms including bacteria or protozoa as well as animals, where they are produced by the innate immune system (1). AMPs kill microbes via various mechanisms, such as destruction of the microbial membrane, inhibition of macromolecule synthesis (2–4) etc. Due to these multiple mechanisms of action, it is difficult for microbes to gain resistance against AMPs as compared to conventional antibiotics. Few of the naturally occurring AMPs have also been observed to regulate various physiological functions such as anti-inflammatory properties, angiogenesis and wound healing besides their antimicrobial activity (5,6).

Development in sequencing technology has accelerated availability of genomic and proteomic data of various organisms in public sequence repositories. The annotations of AMPs in these large data sets using wet-lab methods are cost and resource-intensive. AMPs belong to various AMP families. These families exhibit distinctive sequence composition such as cysteine conservation in defensins (7), abundance of histidines in histatins (8), conservation of unusual amino acid such as aminoisobutyric acid in peptaibols (9) and lanthionine in bacteriocins (lantibiotics) (10) etc. This family-specific sequence conservation can be exploited to identify AMPs from a large pool of sequence data. Family-based signatures such as patterns and Hidden Markov Models (HMMs) can be powerful tools to retrieve and annotate sequences available in sequence databases.

Sequence signatures (patterns and HMMs) present in 1386 experimentally studied AMPs represented by 45 families were generated and used to fetch AMPs from sequence databases. This data has been collated and presented as an update to CAMP database. $CAMP_{R3}$ currently holds 10247 sequences, 757 structures and 114 signatures present in 45 AMP families.

MATERIALS AND METHODS

Data collection and organization

Sequences, structures and family information of AMPs. To update the existing CAMP database (11), protein data available in NCBI (12), UniProtKB (13) and PDB (14) databases post 2013 was queried using appropriate keywords such as 'antimicrobial', 'antibacterial', 'antifungal', 'antiviral' and 'antiparasitic'. The obtained hits were manually curated to extract information on sequence, structure, protein definition, accession numbers, reference literature, activity, taxonomy of the source organism, target organisms with minimum inhibitory concentration (MIC) values, hemolytic activity of the peptide and protein family descriptions. This information is made available in $CAMP_{R3}$. Links to UniProtKB, PDB, PubMed (12) and other databases dedicated to AMPs are also made available for the benefit of the users.

*To whom correspondence should be addressed. Tel: +91 22 24192107; Fax: +91 22 24139412; Email: thomass@nirrh.res.in

Signatures of AMPs. Experimentally validated AMPs, whose family information is available in CAMP (11) was used to generate family-based signatures. Families containing at least two members were considered for signature creation. 1386 sequences, representing 45 AMP families were used to generate patterns and HMMs. PRATT tool (15) was used for generation of patterns. Multiple sequence alignments of each AMP family were created using Clustal Omega (16,17) and these were used as input to build HMM models using 'hmmbuild' program of HMMER 3.1b1 package (18). A heuristically determined fitness value of 26 or above was used as a threshold for selecting patterns for retrieval of sequences. Since length is an important parameter for sequence alignment, length-based patterns and HMMs were also created. The generated patterns and HMMs were queried against the protein database of NCBI and UniProtKB using ScanProsite tool (19) and jackhmmer tool of HMMER web server (20), respectively, to retrieve hits. The HMMs were queried until convergence or stopped after three iterations. Sequences retrieved using HMMs, having a threshold e-value below 0.005 were considered for further screening. The retrieved hits were curated based on their AMP definitions. For each retrieved AMP; information related to sequence, protein definition, accession numbers, activity, source organism, target organisms, protein family descriptions and links to databases like UniProtKB and PubMed along with the generated signatures are provided in $CAMP_{R3}$.

Protein sequences, whose definition suggested antimicrobial activity and had at least one supporting literature reference in PubMed proving its antimicrobial activity by wet-lab methods, were included in the *Experimentally Validated* data set. 590 sequences were retrieved from APD2 (21). These sequences are integrated in the *Experimentally Validated* or *Predicted* data set based on the annotation provided by APD2.

AMPs that have annotations indicating their antimicrobial activity but do not have supporting PubMed reference literature were included in the *Predicted* data set. These sequences are predicted to be antimicrobial either based on their GO (22)/Pfam (23)/InterPro (24)/UniProtKB/NCBI annotations or they were retrieved based on the AMP family signatures.

Algorithm for rational design of AMPs

An in-house Perl script was created to generate all possible single residue substitutions of user defined sequence/s. These sequences are then run through the prediction models (Support Vector Machines (SVMs), Random Forests (RF) and Discriminant analysis (DA)) generated and available in the previous release of CAMP database (11).

Database architecture

The database is built using MySQL Server 5.1.33 as back-end and the front-end is built using PHP, HTML, JavaScript, Open Flash Chart 2 and Perl. The database is hosted on Apache web server 2.2.11. Statistical software R version 2.9.1 (25) was used for development of the prediction server. JSmol viewer (http://wiki.jmol.org/index.php/JSmol) has been integrated for AMP structure visualization.

A brief description of the user interface of $CAMP_{R3}$ is provided as follows.

Home: the home page provides information about various features of the database.

Databases: the data is divided into four databases which include sequence, structure, patents and the newly incorporated signature database.

Tools: the database includes the following tools for analysis. The AMP prediction tool has been developed in-house. Access to various tools relevant to sequence/structure analysis and available in public domain have also been provided in $CAMP_{R3}$ for the benefit of the users.

1. AMP prediction: users can (i) predict AMPs (ii) predict antimicrobial region within peptides and (iii) rationally design AMPs by generating an exhaustive combinatorial library of sequences for a user-defined sequence and predict effect of single residue substitutions on antimicrobial activity using SVMs, RF and DA.
2. BLAST: users can use BLAST tool (26) to query protein sequence/s against various data sets of $CAMP_{R3}$ which include the entire database, sequence, structure, patent, experimentally validated, predicted and predicted based on signature data sets to find homologous sequences, structures and other relevant information.
3. Clustal Omega: users can use Clustal Omega tool of EMBL-EBI to obtain multiple sequence alignment of peptides.
4. Vector Alignment Search Tool: users can identify similar protein structures and distant homologs that cannot be identified by sequence comparison using VAST of NCBI (27).
5. PRATT: users can generate AMP family-specific patterns using this tool from ExPASy.
6. ScanProsite: using this tool from Swiss Institute of Bioinformatics, users can (i) scan proteins against the PROSITE collection of PSSMs/patterns; (ii) scan patterns against protein sequence, structure or user defined database/s and (iii) scan user defined patterns against a set of protein sequences.
7. PHI-BLAST: users can use PHI-BLAST (28) to find AMPs similar to the query based on a family-specific pattern.
8. jackhmmer: users can iteratively search a protein sequence/structure database using a set of protein sequences/multiple sequence alignment/HMM as an input to find homologs using this tool from EMBL-EBI.

Search: basic and advanced search options are available for search of AMP families/sequences/structures and signatures.

Links: links to other online AMP databases are provided.

Statistics: information on $CAMP_{R3}$ statistics can be viewed.

Help: detailed description and use of the various features and tools incorporated in the database is provided for the benefit of the users.

Table 1. Comparison of $CAMP_{R3}$ with few of the existing AMP databases

Database	Sequences	Structures	Signatures	Nature of data	Reference
$CAMP_{R3}$	10247	757	114 (36 Patterns and 78 HMMs)	General	-
APD2	2604	350	Absent	General	(21)
AMPer	1298	Absent	186 HMMs	Eukaryotic AMPs	(30)
LAMP	5548	Present[a]	Absent	General	(31)
BACTIBASE	228	72	Present[a]	Bacteriocins	(32)
YADAMP	2525	Present[a]	Absent	General	(33)
PhytAMP	273	39	Present[a]	Plant AMPs	(34)
Peptaibiotics database	1344	Absent	Absent	Peptaibols	(35)
Defensins Knowledgebase	566	Present[a]	Absent	Defensins	(36)

[a]Difficult to retrieve total count.

RESULTS AND DISCUSSION

$CAMP_{R3}$ provides comprehensive information on AMPs and their families as represented by their sequences, structures, activity, signatures, source and target organisms. The unique feature of $CAMP_{R3}$ as compared to other AMP databases is that information of family-specific signatures has been provided for a large set of both eukaryotic as well as prokaryotic AMPs. It presently contains 114 AMP family-specific sequence signatures (36 patterns and 78 HMMs). Using these signatures, a total of 4222 AMPs were identified, out of which 2739 were absent in the previous CAMP database.

Use of signatures is particularly significant for retrieving sequences that have to be queried specifically by their definitions. For example, AMPs such as thionin-2.1 (UniProt ID: Q42596), varv peptide A/kalata-B1 (UniProt ID: Q5USN7) etc. could not be retrieved from UniProtKB database using search keywords such as 'antimicrobial' but could be retrieved using their family signatures.

$CAMP_{R3}$ currently holds 10247 AMP sequences, of which 4857 are experimentally validated, and 5390 are predicted. Of these, 3491 have been recently identified. The structure database has also been updated to include 757 antimicrobial structures.

Sequence composition is an important determinant of antimicrobial activity. It has been well demonstrated by antimicrobial assays of AMPs and their analogues that minor variations in peptide sequence can drastically alter its antimicrobial activity (29). The prediction algorithm for AMPs, available in $CAMP_{R3}$ now includes an additional feature for rational design of AMPs. This feature can be used to predict the effect of single residue substitutions on antimicrobial activity.

The features incorporated in $CAMP_{R3}$ will significantly promote AMP family-based studies. AMPs belonging to a particular AMP family can be effortlessly obtained using the family-based search. This feature, along with the family signatures and tools available in $CAMP_{R3}$ for sequence and structure analysis, will allow users to study the various AMP families independently and effectively.

CONCLUSION

The database is available for retrieval of sequences/structures/patents/signatures and families of AMPs. Comparison of $CAMP_{R3}$ with the existing databases dedicated to AMPs is presented in Table 1. AMPs that are not easily retrievable using simple keyword search have been identified/retrieved from public sequence databases using family signatures.

The highlights of this updated database are as follows.

1. Massive update on AMP sequences and structures (10247 AMP sequences and 757 AMP structures).
2. Family-specific signatures of eukaryotic and prokaryotic AMPs.
3. Sequence optimisation prediction algorithm for antimicrobial activity.

$CAMP_{R3}$ has been developed with an objective to expand and accelerate research on AMPs.

ACKNOWLEDGEMENTS

The authors are grateful to Dr Smita D. Mahale (PI of Biomedical Informatics Centre) for all the assistance and support. They also acknowledge the help provided by Ms Shaini Joseph and Mr Lijin Gopi in data collection and design of the web interface, respectively.

FUNDING

This work [RA/296/09-2015] was supported by grants from Department of Science and Technology, Government of India [SB/S3/CE/028/2013] and Indian Council of Medical Research. The open access publication charge for this paper has been waived by Oxford University Press - *NAR*.

Conflict of interest statement. None declared.

REFERENCES

1. Cruz,J., Ortiz,C., Guzmán,F., Fernández-Lafuente,R. and Torres,R. (2014) Antimicrobial peptides: promising compounds against pathogenic microorganisms. *Curr. Med. Chem.*, **21**, 2299–2321.
2. Haney,E.F., Petersen,A.P., Lau,C.K., Jing,W., Storey,D.G. and Vogel,H.J. (2013) Mechanism of action of puroindoline derived tryptophan-rich antimicrobial peptides. *Biochim. Biophys. Acta.*, **1828**, 1802–1813.
3. Roy,R.N., Lomakin,I.B., Gagnon,M.G. and Steitz,T.A. (2015) The mechanism of inhibition of protein synthesis by the proline-rich peptide oncocin. *Nat. Struct. Mol. Biol.*, **22**, 466–469.
4. Wang,S., Thacker,P.A., Watford,M. and Qiao,S. (2015) Functions of Antimicrobial Peptides in Gut Homeostasis. *Curr. Protein Pept. Sci.*,**16**, 582–591.
5. Frasca,L. and Lande,R. (2012) Role of defensins and cathelicidin LL37 in auto-immune and auto-inflammatory diseases. *Curr. Pharm. Biotechnol.*, **13**, 1882–1897.

6. Guilhelmelli,F., Vilela,N., Albuquerque,P., Derengowski,Lda, S., Silva-Pereira,I. and Kyaw,C.M. (2013) Antibiotic development challenges: the various mechanisms of action of antimicrobial peptides and of bacterial resistance. *Front. Microbiol.*, **4**, 1–12.
7. Ganz,T. (2013) Defensins: antimicrobial peptides of innate immunity. *Nat. Rev. Immunol.*, **3**, 710–720.
8. van Dijk,I.A., Nazmi,K., Bolscher,J.G., Veerman,E.C. and Stap,J. (2015) Histatin-1, a histidine-rich peptide in human saliva, promotes cell-substrate and cell-cell adhesion. *FASEB J.*, **29**, 3124–3132.
9. Duclohier,H. (2010) Antimicrobial peptides and peptaibols, substitutes for conventional antibiotics. *Curr. Pharm. Des.*, **16**, 3212–3223.
10. Lohans,C.T. and Vederas,J.C. (2014) Structural characterization of thioether-bridged bacteriocins. *J. Antibiot. (Tokyo).*, **67**, 23–30.
11. 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.
12. NCBI Resource Coordinators. (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **43**, D6–D17.
13. UniProt Consortium. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
14. Rose,P.W., Prlić,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.
15. Jonassen,I., Collins,J.F. and Higgins,D.G. (1995) Finding flexible patterns in unaligned protein sequences. *Protein Sci.*, **4**, 1587–1595.
16. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Söding,J., Thompson,J.D. and Higgins,D.G. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 1–6.
17. McWilliam,H., Li,W., Uludag,M., Squizzato,S., Park,Y.M., Buso,N., Cowley,A.P. and Lopez,R. (2013) Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res.*, **41**, W597–W600.
18. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics.*, **14**, 755–763.
19. de Castro,E., Sigrist,C.J., Gattiker,A., Bulliard,V., Langendijk-Genevaux,P.S., Gasteiger,E., Bairoch,A. and Hulo,N. (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.*, **34**, W362–W365.
20. Finn,R.D., Clements,J. and Eddy,S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.
21. Wang,G., Li,X. and Wang,Z. (2009) APD2: the updated antimicrobial peptide database and its application in peptide design. *Nucleic Acids Res.*, **37**, D933–D937.
22. Gene Ontology Consortium. (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
23. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
24. Mitchell,A., Chang,H.Y., Daugherty,L., Fraser,M., Hunter,S., Lopez,R., McAnulla,C., McMenamin,C., Nuka,G., Pesseat,S. *et al.* (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, **43**, D213–D221.
25. R Development Core Team. (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
26. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
27. Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.
28. Zhang,Z., Schäffer,A.A., Miller,W., Madden,T.L., Lipman,D.J., Koonin,E.V. and Altschul,S.F. (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.*, **26**, 3986–3990.
29. Vila-Perelló,M., Sánchez-Vallet,A., García-Olmedo,F., Molina,A and Andreu,D. (2003) Synthetic and structural studies on Pyrularia pubera thionin: a single-residue mutation enhances activity against Gram-negative bacteria. *FEBS Lett.*, **536**, 215–219.
30. 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.
31. Zhao,X., Wu,H., Lu,H., Li,G. and Huang,Q. (2013) LAMP: A Database Linking Antimicrobial Peptides. *PLoS One.*, **8**, e66557.
32. 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**, 1–5.
33. 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.
34. 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.
35. Neumann,N.K., Stoppacher,N., Zeilinger,S., Degenkolb,T., Brückner,H. and Schuhmacher,R. (2015) The peptaibiotics database–a comprehensive online resource. *Chem. Biodivers.*, **12**, 743–751.
36. 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.