D1154–D1158 *Nucleic Acids Research, 2014, Vol. 42, Database issue* *Published online 21 November 2013*
doi:10.1093/nar/gkt1157

CAMP: Collection of sequences and structures of antimicrobial peptides

Faiza Hanif Waghu, Lijin Gopi, Ram Shankar Barai, Pranay Ramteke, Bilal Nizami and Susan Idicula-Thomas*

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

Received September 14, 2013; Revised and Accepted October 25, 2013

ABSTRACT

Antimicrobial peptides (AMPs) are gaining importance as anti-infective agents. Here we describe the updated Collection of Antimicrobial Peptide (CAMP) database, available online at http://www.camp.bicnirrh.res.in/. The 3D structures of peptides are known to influence antimicrobial activity. Although there exists databases of AMPs, information on structures of AMPs is limited in these databases. CAMP is manually curated and currently holds 6756 sequences and 682 3D structures of AMPs. Sequence and structure analysis tools have been incorporated to enhance the usefulness of the database.

INTRODUCTION

Antimicrobial peptides (AMPs) are widely studied as potential alternatives for antibiotics. Surge in research on AMPs has led to the development of several databases and prediction tools. Some of these are general databases such as APD2 (1), DAMPD (2) and LAMP (3), whereas others are specialized databases like—AMSdb (http://www.bbcm.units.it/~tossi/pag1.htm) that contains AMPs from only plant and animal sources; RAPD (4) provides information on recombinant methods to generate AMPs; PhytAMP (5) and BACTIBASE (6) are databases dedicated to AMPs from plant and bacterial sources, respectively; Defensins knowledgebase (7) and PenBase (8) are devoted to AMPs from defensin and penaeidin families, respectively; Peptaibol Database (9) is a database of peptaibols (unusual class of peptides); BAGEL (10) is a database of bacteriocins; and HIPdb (11) is a database of experimentally validated HIV-inhibiting peptides. The enormous amount of data on AMPs had motivated us to develop a general database, Collection of Antimicrobial Peptides (CAMP) (12), which included a sequence-based prediction tool for AMPs.

While all these databases provide comprehensive information on sequences of AMPs, information on structures of AMPs is limited. The topological features of peptides play a crucial role in dictating antimicrobial activity (13). Although many sequence-based prediction algorithms are available, the knowledge of 3D structural features of known AMPs has not been exploited to develop prediction algorithms. The lack of structural databases of AMPs is probably one of the main impediments in this direction. Presently, there are several AMPs whose structural information is available in the Protein Data Bank (PDB) (14). However, retrieving information on structures of AMPs from the structural databases such as PDB is not a trivial task; for example, the structures may have additional chains that are non-AMPs, and these have to be filtered out by manual curation. The structures may also not be easily retrieved from structure databases based on simple keyword searches such as 'antibacterial', 'antifungal', etc. To address these shortcomings, the current release of CAMP has been developed.

MATERIALS AND METHODS

Data collection and organization

Sequence and structural information of AMPs was retrieved from protein databases of NCBI, UniProtKB (15) and PDB using combination of keywords like 'antimicrobial', 'antibacterial', 'antifungal', 'antiviral' and 'antiparasitic'. Manually curated information related to 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, functional and structural classifications, protein family descriptions and links to external databases like UniProtKB, PDB, PubMed and other AMP databases is made available to the users.

Database architecture

The updated CAMP database is built on Apache HTTP server 2.0.59. MySQL Server 5.0 is used at the back-end,

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

whereas the front-end is built using PHP, HTML, JavaScript, Perl and Open Flash Chart 2.

Below is a brief description of the user interface of CAMP:

(1) Home: The CAMP database along with its various features is described in this section.
(2) Databases: Data are sectioned into sequence, structure and patent databases.
(3) Tools: The following analysis tools are available to the users.
   (a) AMP prediction: Users can predict AMPs and/or scan for antimicrobial regions within the peptides using Support Vector Machine (SVM), Random Forests (RF) and Artificial Neural Network (ANN).
   (b) Feature calculator: Amino acid composition, secondary structural propensities and physicochemical properties such as net charge, hydrophobicity, etc of the peptides can be calculated.
   (c) BLAST: Users can use BLAST (16) tool against the sequence or structure database of CAMP to find homologous sequences or structures, respectively.
   (d) ClustalW: Multiple sequence alignment of the peptides can be obtained using ClustalW (17) tool from EMBL-EBI.
   (e) Vector Alignment Search Tool: Similar protein structures can be identified using this NCBI tool (18).
   (f) PRATT: This tool from ExPASy can be used to find patterns in a set of related AMPs (19,20).
   (g) Helical wheel: Alpha-helical AMPs can be studied using the helical wheel Java applet created by Edward K. O'Neil and Charles M. Grisham (University of Virginia in Charlottesville, Virginia).
   (h) PDB2PQR: This clone server can be used for converting PDB files into PQR file format, (PQR files are PDB files where B-factor and occupancy columns have been replaced by radius and per-atom charge, respectively) which could be used for further structural studies (21,22).
(4) Search: Users can search for sequences and/or structures of AMPs using basic and advanced search options.
(5) Links to other available AMP databases have been provided.
(6) Statistics: Coverage of the database based on the nature of data, taxonomy of source organism and activity has been depicted using pie charts and Venn diagram.
(7) Help: A detailed explanation about the features and tools available in the database has been provided in this section.

Prediction algorithm

Dataset creation

The positive dataset constituted of 3010 AMP sequences. These were obtained from the patent and experimentally validated datasets of CAMP, after removing sequences that (i) are redundant (100% similarity cut-off), (ii) have non-standard amino acids and (iii) have length >100. CD-HIT server was used for removing redundant sequences (23).

The negative dataset consists of 4011 sequences, generated in our previous work (12). It includes experimentally proven non-antimicrobial sequences, arbitrary sequences generated using random numbers and protein sequences retrieved from the UniProt database without annotation as 'antimicrobial'. The sequences had length approximately in the same range as the positive dataset. The CD-HIT program (23) was used to eliminate sequences with >90% identity. These datasets were randomly divided into training (70%) and test (30%) datasets.

Model generation

Sixty-four best peptide descriptors based on the RF Gini score were used for developing SVM-, RF- and ANN-based prediction models. All the models were evaluated using Matthews correlation coefficient (MCC), prediction accuracy and 10-fold cross-validation accuracy on training and test datasets. For developing the prediction models, implementation of SVM, RF and ANN in *R* (version 2.15.3) was used (24).

SVM

Kernlab package in *R* was used to train the SVM classifier (25). In this study, we have used polynomial kernel function. The values of the hyper parameters were set as follows: degree = 4, scale = 0.01 and offset = 1.

RF

'randomForest' package was used to train the RF classifier with a maximum of 1500 trees (26).

ANN

'nnet' package in *R* was used for building the ANN-based prediction model (27).

RESULTS AND DISCUSSION

The updated CAMP is a comprehensive database on sequences and structures of AMPs. It currently holds 6756 sequences of AMPs (experimentally validated (2602), predicted (2438) and patents (1716)), which include 2736 recently identified AMP sequences. The information on the sequence, AMP family, source, target organism and activity is captured in the database. As can be seen in Figure 1A–C, CAMP has a wide coverage on the above fields.

CAMP presently contains 682 AMP structures. Multiple structures of AMPs, if available in PDB, are also integrated in the database. Although structural information on AMPs is available in databases such as APD2, LAMP, etc, the structures can be directly viewed using Jmol viewer in CAMP. Direct viewing of structures is also available in Defensins knowledgebase, PhytAMP, HIPdb and BACTIBASE. However, these databases cater to specific class of AMPs.

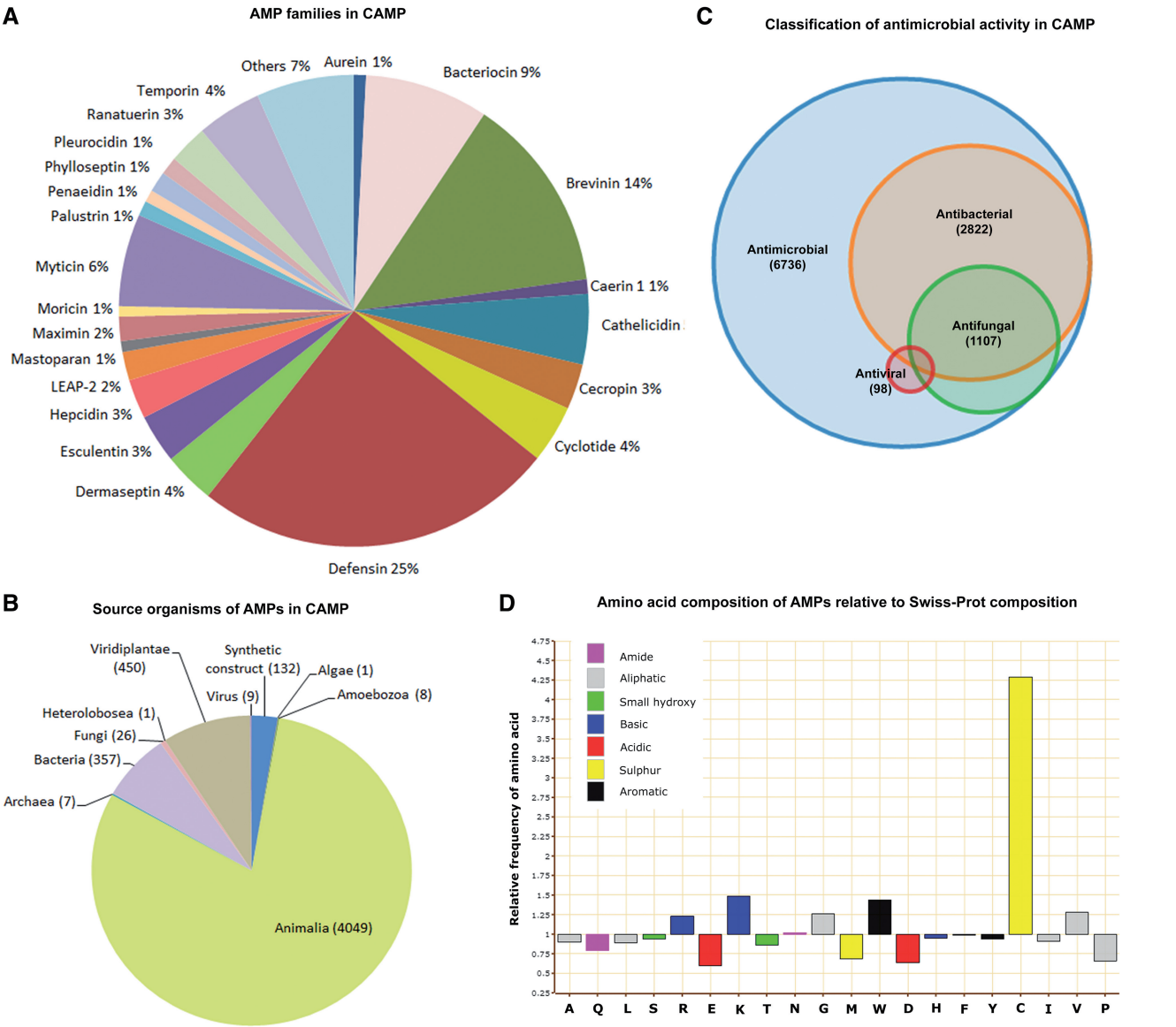

Figure 1. (**A**) Pie chart of AMP families in CAMP, (**B**) Pie chart of source organisms of AMPs in CAMP, (**C**) Venn diagram of classification of AMP activity in CAMP and (**D**) Relative amino acid composition of experimentally validated and predicted sequences of AMPs in CAMP as compared with Swiss-Prot composition.

Another interesting feature of the current release of CAMP is that users can selectively retrieve information on specific families of AMPs of their interest; e.g. cathelicidins, defensins and cecropins. The AMP family information for the peptides has been annotated manually using information from Pfam (28), InterPro (29) and associated literature. The distribution of the AMP families in the database can be seen in Figure 1A.

The prediction algorithm for AMPs has been modified using the updated sequence information. Supplementary Table S1 shows the prediction accuracy, MCC and cross-validation accuracy of the prediction models. Users can predict the antimicrobial activity of proteins and/or scan regions (with user-defined lengths) within proteins for antimicrobial activity.

Tools that aid in sequence and structure analysis such as feature calculator, PRATT, ClustalW, Vector Alignment Search Tool, BLAST and PDB2PQR have also been incorporated in CAMP. Effect of mutations on the structure of AMPs and/or their analogs can be visualized using the Jmol visualizer integrated in the database. Helicity is known to influence antimicrobial activity (30) and therefore, tool for helical wheel projection is also available. AMPs are known to be rich in hydrophobic and cationic amino acids. The ratio of the percentage frequency of amino acids in CAMP to the percentage frequency of amino acids in UniProtKB/Swiss-Prot protein knowledgebase (Release 2013_08 of 24 July 2013) is plotted in Figure 1D. As expected, AMPs were observed to be enriched in positively charged and hydrophobic residues such as Arg, Lys, Gly, Cys, Trp and Val residues.

Table 1. Comparison of CAMP with existing AMP databases

Features	Database											
	RAPD	PhytAMP	BACTIBASE second release	Defensins knowledg-ebase	PenBase	Peptaibol database	AMSDb	HIPdb	APD2	DAMPD	LAMP	CAMP
Type	Specific (Recombinantly produced AMPs only)	Specific (Plant AMPs only)	Specific (Bacteriocins only)	Specific (Defensin family AMPs only)	Specific (Penaeidin family AMPs only)	Specific (Peptaibols only)	Specific (Eukaryotic AMPs only)	Specific (HIV inhibiting peptides only)	General	General	General	General
Total number of entries	179	273	220	566	28	317	895	1068	2307	1232	5547	7438
Prediction algorithm	Absent	Present	Present	Absent	Absent	Absent	Absent	Absent	Present	Present	Absent	Present
Structural information	Absent	Present	Present	Present	Absent	Present[a]	Present[a]	Present	Present[a]	Present[a]	Present[a]	Present
Search based on AMP family	Present	Present	Absent	Present	Absent	Absent	Present	Present	Absent	Present	Absent	Present
MIC values	Absent	Present	Present	Present	Absent	Absent	Present	Present	Present	Present	Present	Present
Separate searches for experimental and predicted datasets	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Absent	Present	Present
Tools	DNA translator, peptide calculator, DNA sequence convertor	BLAST, FASTA, Smith-Waterman search, ClustalW, muscle, physiochemical profile	BLAST, FASTA, Smith-Waterman search, ClustalW, Muscle, T-coffee, physiochemical profile, MODELLER	BLAST and ClustalW	BLAST and ClustalW	Absent	HydroMCalc and HydroPlot	HIPdb map, HIPdb BLAST	AMP designer	BLAST, ClustalW, NJPLOT, HMMER, hydrocalculator, signalp, graphical views.	BLAST	ClustalW, PRATT, helical wheel, vector alignment search tool , BLAST, PDB2PQR, Feature calculator

[a]The PDB IDs are available. Structures cannot be directly viewed.

CONCLUSIONS

CAMP holds a massive update on AMP sequences and incorporates several tools relevant to design of AMPs. The 3D conformations of peptides are known to be critical determinants of antimicrobial activity. The prominent feature of the current release of CAMP is the addition of experimentally derived structures of AMPs, which can be directly viewed using the Jmol viewer. The update also facilitates family-based study on AMPs. A detailed comparison of CAMP with the existing databases on AMPs is presented in Table 1. The information, present in an easily searchable and downloadable form, is envisaged to accelerate sequence–structure–activity studies on AMPs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to Dr Smita D. Mahale (PI of Biomedical Informatics Centre) for all the help and support. They also acknowledge the assistance provided by Ms Shaini Joseph and Ms Pratima Gurung in data collection.

FUNDING

This work [RA/18-09/2013] was supported by grants from Department of Science and Technology, Government of India [SB/S3/CE/028/2013]; and Indian Council of Medical Research. Funding for open access charge: Waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Seshadri Sundararajan,V., Gabere,M.N., Pretorius,A., Adam,S., Christoffels,A., Lehväslaiho,M., Archer,J.A. and Bajic,V.B. (2012) DAMPD: a manually curated antimicrobial peptide database. *Nucleic Acids Res.*, **40**, D1108–D1112.
3. Zhao,X., Wu,H., Lu,H., Li,G. and Huang,Q. (2013) LAMP: a database linking antimicrobial peptides. *PLoS One.*, **8**, e66557.
4. Li,Y. and Chen,Z. (2008) RAPD: a database of recombinantly produced antimicrobial peptides. *FEMS Microbiol. Lett.*, **289**, 126–129.
5. 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.
6. 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.
7. Seebah,S., Anita,S., Zhuo,S.W., Yong,H.C., Chua,H., Chuon,D., Beuerman,R. and Verma,C.S. (2006) Defensins knowledgebase: a manually curated database and information source focused on the defensins family of antimicrobial peptides. *Nucleic Acids Res.*, **35**, D265–D268.
8. 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.* (2005) PenBase, the shrimp antimicrobial peptide penaeidin database: sequence-based classification and recommended nomenclature. *Dev. Comp. Immunol.*, **30**, 283–288.
9. Whitmore,L. and Wallace,B.A. (2004) The Peptaibol database: a database for sequences and structures of naturally occurring peptaibols. *Nucleic Acids Res.*, **32**, D593–D594.
10. de Jong,A., van Heel,A.J., Kok,J. and Kuipers,O.P. (2010) BAGEL2: mining for bacteriocins in genomic data. *Nucleic Acids Res.*, **38**, W647–W651.
11. Qureshi,A., Thakur,N. and Kumar,M. (2013) HIPdb: a database of experimentally validated HIV inhibiting peptides. *PLoS One.*, **8**, e54908.
12. Thomas,S., Karnik,S., Barai,R.S., Jayaraman,V.K. and Idicula-Thomas,S. (2010) CAMP: a useful resource for research on antimicrobial peptides. *Nucleic Acids Res.*, **38**, D774–D780.
13. Sitaram,N. and Nagaraj,R. (2002) Host-defense antimicrobial peptides: importance of structure for activity. *Curr. Pharm. Des.*, **8**, 727–742.
14. Bernstein,F.C., Koetzle,T.F., Williams,G.J., Meyer,E.F. Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1978) The Protein data bank: a computer-based archival file for macromolecular structures. *Arch. Biochem. Biophys.*, **185**, 584–589.
15. The UniProt Consortium. (2013) Update on activities at the universal protein resource (UniProt) in 2013. *Nucleic Acids Res.*, **41**, D43–D47.
16. 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.
17. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
18. Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.
19. Jonassen,I., Collins,J.F. and Higgins,D. (1995) Finding flexible patterns in unaligned protein sequences. *Protein Sci.*, **4**, 1587–1595.
20. Jonassen,I. (1997) Efficient discovery of conserved patterns using a pattern graph. *Comput. Appl. Biosci.*, **13**, 509–522.
21. Dolinsky,T.J., Nielsen,J.E., McCammon,J.A. and Baker,N.A. (2004) PDB2PQR: an automated pipeline for the setup, execution, and analysis of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.*, **32**, W665–W667.
22. Dolinsky,T.J., Czodrowski,P., Li,H., Nielsen,J.E., Jensen,J.H., Klebe,G. and Baker,N.A. (2007) PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.*, **35**, W522–W525.
23. Huang,Y., Niu,B., Gao,Y., Fu,L. and Li,W. (2010) CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.
24. R Development Core Team. (2009) R: A Language and Environment for Statistical Computing, Vienna, Austria.
25. Karatzoglou,A., Smola,A., Hornik,K. and Zeileis,A. (2004) Kernlab - an S4 package for Kernel methods. *R. J. Stat. Softw.*, **11**, 1–20.
26. Liaw,A. and Wiener,M. (2002) Classification and regression by random forest. *R News*, **2**, 18–22.
27. Venables,W.N. and Ripley,B.D. (2002) *Modern Applied Statistics with S*, 4th edn. Springer, New York, ISBN 0-387-95457-0.
28. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
29. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.
30. Chen,H.C., Brown,J.H., Morell,J.L. and Huang,C.M. (1988) Synthetic magainin analogues with improved antimicrobial activity. *FEBS Lett.*, **236**, 462–426.