D774–D780 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 18 November 2009*
doi:10.1093/nar/gkp1021

CAMP: a useful resource for research on antimicrobial peptides

Shaini Thomas[1], Shreyas Karnik[2], Ram Shankar Barai[1], V. K. Jayaraman[3,*] and Susan Idicula-Thomas[1,*]

[1]Biomedical Informatics Center of Indian Council of Medical Research, National Institute for Research in Reproductive Health, Mumbai, India, [2]School of Informatics, Indiana University, Indianapolis, IN 46202, USA and [3]SECG, Center for Development of Advanced Computing, Pune University Campus, Pune, India

Received August 13, 2009; Revised October 16, 2009; Accepted October 20, 2009

ABSTRACT

Antimicrobial peptides (AMPs) are gaining popularity as better substitute to antibiotics. These peptides are shown to be active against several bacteria, fungi, viruses, protozoa and cancerous cells. Understanding the role of primary structure of AMPs in their specificity and activity is essential for their rational design as drugs. Collection of Anti-Microbial Peptides (CAMP) is a free online database that has been developed for advancement of the present understanding on antimicrobial peptides. It is manually curated and currently holds 3782 antimicrobial sequences. These sequences are divided into experimentally validated (patents and non-patents: 2766) and predicted (1016) datasets based on their reference literature. Information like source organism, activity (MIC values), reference literature, target and non-target organisms of AMPs are captured in the database. The experimentally validated dataset has been further used to develop prediction tools for AMPs based on the machine learning algorithms like Random Forests (RF), Support Vector Machines (SVM) and Discriminant Analysis (DA). The prediction models gave accuracies of 93.2% (RF), 91.5% (SVM) and 87.5% (DA) on the test datasets. The prediction and sequence analysis tools, including BLAST, are integrated in the database. CAMP will be a useful database for study of sequence-activity and -specificity relationships in AMPs. CAMP is freely available at http://www.bicnirrh.res.in/antimicrobial.

INTRODUCTION

Microbial resistance to antibiotics is a rising concern among health care professionals, driving them to search for alternative therapies. In the past few years, antimicrobial peptides (AMPs) have attracted lot of attention as a substitute for conventional antibiotics (1). AMPs are naturally present in all organisms and play a vital role in their innate immunity. These peptides cause cell death either by disrupting the microbial cell membrane; inhibiting extracellular polymer synthesis or intracellular functions (2–4). They generally act on structural components of the cell wall and can have multiple cellular targets (2). The plausibility of microbial resistance to AMPs is highly reduced due to the considerable difficulty for microbes to modify their cell wall composition or alter each of their multiple targets. AMPs are also effective against multidrug resistant bacteria (5). The differences in membrane structure of microbes and higher eukaryotes help them in selectively targeting the microbial membranes and thereby making it less toxic for therapeutic use. Some AMPs also possess antitumor activity and can act as mitogens and signaling molecules (6).

AMPs vary in their spectrum of biological activity. Some peptides have a broad range of activity. For example, maximins from *Bombina maxima* are active against bacteria (Gram-positive and -negative), certain fungi and HIV-1 virus (7), while AMPs like formaecin-1 and 2 from *Myrmecia gulosa* are active against *Escherichia coli* but not active against Gram-positive bacteria and yeast *Candida albicans* (8). Experiments have revealed that small variations in the primary structure of peptides may lead to drastic changes in their specificity and activity (9). In case of carnobacteriocin B2, a single residue alteration renders the peptide inactive (10). Sequence changes

*To whom correspondence should be addressed. Tel.: 91-22-24192107/04 Fax: 91-22-24139412; Email: thomass@nirrh.res.in; susan@icmr.org.in
Correspondence may also be addressed to V. K. Jayaraman. Tel.: 91-20-25704228; Fax: 91-20-25694004; Email: jayaramanv@cdac.in; vkjayaram@yahoo.com

may not always render the peptide completely non-antimicrobial but may lead to changes in their minimum inhibitory concentration (MIC). Studies on HP 2–20 peptide and its analogs derived from *Helicobacter pylori* ribosomal protein L1, showed that a single L11S mutation almost doubles the MIC of the peptide whereas E16W and D18W mutations in HP analogs 1 and 2, respectively, decreases the MIC of the peptides almost by half (11). Several such experiments strongly indicate that primary structure of a peptide influences its antimicrobial activity.

Sequence studies on AMPs have revealed that they are 'mostly' cationic with length ranging from 6 to 100 amino acids (3,5). Few AMPs like maximin H5, dermcidin and enkelytin are known to be anionic in nature (3,12). Anionic peptides are generally more active when complexed with zinc or highly cationic peptides (3). AMPs also exhibit a high composition of hydrophobic residues (5). Most AMPs are amphipathic with hydrophilic domain on one side and hydrophobic domain on the other. It is proposed that the positive charge, hydrophobic nature and amphipathicity of these peptides help them to interact with the microbial cell membranes leading to cell permeation and lysis (2,13). A thorough understanding of the role of sequence of AMPs on their specificity and activity is essential to exploit them as antimicrobial drugs. A comprehensive database on AMPs with information on their activity is a pre-requisite to carry out sequence-specificity and sequence-activity studies. The existing databases on AMPs are APD (14), AMSdb (http://www.bbcm.units.it/~tossi/pag1.htm), RAPD (15), PhytAMP (16), BACTIBASE (17), Defensin knowledgebase (18), PenBase (19), Peptaibol Database (20), SAPD (21) and BAGEL (22). CAMP is created with an objective to provide a useful resource for sequence-specificity and sequence-activity studies on AMPs. A detailed comparison of the existing databases on AMPs with CAMP is shown in Supplementary Table S1.

CONSTRUCTION AND CONTENT

Data collection and organisation

Sequences of AMPs were collected from NCBI database using a combination of keywords like 'antimicrobial', 'antibacterial', 'antifungal', 'antiviral', 'antitumor', 'anticancer' and 'antiparasitic peptides'. Each of the obtained hits was validated with literature available for reference. Experimentally deduced sequences were included in the experimentally validated dataset while sequences patented to be antimicrobial were included under the patents dataset. Sequences predicted to be antimicrobial based on the similarity or with annotations in NCBI as 'antimicrobial regions' without experimental validation were included in a separate predicted dataset. The reference literature was used to ensure that the AMP sequence alone, excluding the signal and propeptide region, was included in CAMP. Information on accession numbers, protein definition, source, taxonomy, literature reference and target organisms with MIC values (if available) were extracted from NCBI and included in CAMP. Any other relevant information; for example, mention of peptides being inactive against certain organisms; were also included as comments for the database entries.

The data in CAMP is organized into 17 fields viz. CAMP ID, sequence, sequence length, source, taxonomy, activity, Gram nature, target organisms, hemolytic activity, PubMed ID, protein name, protein definition, GenInfo ID, Swiss-Prot, PDB accession numbers, comments and the dataset type (experimentally validated/patents/predicted). Based on their activity, peptides are classified as 'antibacterial', 'antifungal', 'antiviral' or 'antiparasitic'. The classification of AMPs in CAMP is similar to that of AMSdb (http://www.bbcm.units.it/~tossi/pag1.htm) and APD (14) databases. Peptides that have a wider range of activity are depicted as 'Antibacterial: Antifungal' or 'Antibacterial: Antifungal: Antiviral' etc. as the case may be. Links are provided to access further information on the peptides, if present in external databases like NCBI, NCBI Taxonomy Browser, Swiss-Prot and PDB.

Database architecture

CAMP is built on Apache HTTP Server 2.0.59 with MySQL Server 5.0 as the back-end and PHP 5.2.9, HTML and JavaScript as the front-end. Apache, MySQL and PHP technology were preferred as they are open-source softwares and platform independent. Besides these advantages, MySQL supports multithreading and multiuser environments.

Prediction algorithm

Creation of datasets. Datasets were created for antimicrobial (positive dataset) and non-antimicrobial (negative dataset) peptides. The positive dataset comprised of the experimentally validated (patents and non-patents) AMPs present in CAMP. Redundant sequences and sequences containing 'X' were eliminated to obtain the final positive dataset containing 2578 sequences. There are very few peptides that are experimentally proven to be non-antimicrobial. AMPs are generally secretory in nature (23). Hence, along with the experimentally proven non-antimicrobial peptides (25 sequences), non-secretory proteins randomly searched from the UniProt database without annotation as 'antimicrobial' (2413 sequences), arbitrary sequences generated using random numbers (1200 sequences) and proteins retrieved randomly without 'antimicrobial' annotation from the UniProt database (1200 sequences) were used to build the negative dataset. Since the length of peptides in the positive dataset ranged from 10–80 amino acids, the sequences in the negative dataset were truncated to be in the same range. The Cd-hit program (24) was used to eliminate sequences with >90% identity in the negative dataset. The entries in the negative dataset (4011 sequences) were restricted to ~1.5 times that of the positive dataset. These datasets were randomly split to generate the training (70%) and test (30%) datasets.

Figure 1. User interfaces in CAMP.

Calculation of sequence features. Sequence features known to be important for antimicrobial activity and *in vivo* stability of AMPs were considered for classification. These features include composition, physicochemical properties and structural characteristics of amino acids. Amino acids were converted to the reduced alphabets based on BLOSUM-50 matrix (25), conformational similarity (26), hydrophobicity (27,28), normalized van der Waals volume, polarity, polarizability, charge, secondary structure and solvent accessibility (28,29; Supplementary Table S2). Along with composition, dipeptide and tripeptide frequencies of the reduced alphabets; the transition and distribution of some of the features along the sequence of peptides were also computed for classification (29–31). Thus, 257 features were used for classification (Supplementary Table S3).

Prediction methods

Random forest. RF uses an ensemble of trees for classification and regression problems (32). RF implementation in R statistical language based on the original FORTRAN code by Leo Breiman was used for this study (33,34).

Support vector machines. SVMs are a class of machine learning algorithms that can perform pattern recognition and regression (35,36). It can very effectively handle noise and large datasets and thus is increasingly used for classification of biological data. SVM non-linearly transforms the original input space into a higher dimensional feature space by means of kernel functions (37,38). Of the three SVM kernel functions viz., linear, polynomial and radial basis, polynomial function-based model performed the best and hence was adopted for this study. SVM implementation of Kernlab package in R-language was employed in this study (33,39).

Discriminant analysis. DA is a classification algorithm that uses linear combination of independent variables to predict the group membership for each of the dependent variables (40). Stepwise selection algorithm with backward elimination was used for variable selection. DA application in SPSS 16.0 package was used for this study.

Feature selection. Rigorous recursive feature elimination (RFE) method based on RF Gini score was adopted to handle the background noise and identify the most informative sequence-based features for classification (34). The features, based on their gini scores, were reduced to 50% at each step. Thus, starting with 257 features, models corresponding to 128, 64, 32, 16, 8 and 4 features were used to build the models for classification based on RF. These models were evaluated using 10-fold cross validation accuracy and Matthews Correlation Coefficient (MCC) on training and test datasets. MCC is considered to be a balanced measure for evaluating the performance of the algorithm, even in datasets of different sizes (41,42). The subset of 64 features gave the best performance with RF. These 64 features were further used to build SVM and DA models of classification.

Database interfaces

The interfaces in CAMP are designed in a manner to help users in easy navigation and use of the various tools integrated in the database (Figure 1). The database interfaces include: Home, Search, Tools, Prediction, BLAST (43), Submit Sequence, Feedback and Help. A brief description of the interfaces is given below.

Home. The CAMP database along with its various features is described in this section.

Search. The search features in CAMP are designed to accommodate all possible queries of the users. The different search options are:

(i) Simple search: this feature allows users to search the database with keywords in all fields or in a specific field by selecting the field of interest from the drop down menu.
(ii) Advanced search: this feature enables specific searches by using the advanced query form. Users can limit their search to a particular field without the use of field descriptors in the query. An example query for each search field is displayed when the field is in focus.
(iii) Search sequences with MIC values: users can search and download sequences active against a particular target organism.
(iv) Browse all sequences: user can browse all the sequences present in CAMP using this feature.
(v) Browse protein families: users can browse sequences present in CAMP belonging to a particular protein family.

Multiple records can be viewed at a time. Users can download the AMPs of interest and save in Excel format using the 'Export' feature.

Tools. Algorithms for calculation of some of the properties that are known to influence the *in vivo* stability and antimicrobial activity of peptides are included in CAMP. Properties that can be calculated are length, net charge, amino acid composition, aliphatic index (44), instability index (45), hydropathy (46) and secondary structure propensities (30). The users can either paste their peptide sequence/s in FASTA format or upload a text file with sequences of interest.

Prediction. The prediction interface in CAMP allows users to predict the antimicrobial activity of sequences. Users can input sequence/s in FASTA format and select an algorithm (RF, DA and SVM) for prediction. The predicted antimicrobial activity of the peptide/s along with the probability of the prediction being true is displayed. In case of DA, the discriminant scores for the peptide/s are displayed.

BLAST. CAMP has two modules:

(i) BLAST AGAINST NCBI-PROT: this module allows users to compare peptide sequence/s with the non-redundant protein dataset of NCBI. The user-defined parameters are database for comparison, *E*-value, alignment type (gapped/ungapped), matrix for alignment and proteome of the organism.
(ii) BLAST AGAINST CAMP: this module allows users to compare sequences with the CAMP database. The user-defined parameters available here are *E*-value, alignment type (gapped/ungapped) and matrix for alignment.

Submit sequence. Researchers can submit new antimicrobial sequences using this feature.

Database statistics. This link can be used to understand the composition and access the entries of the CAMP database based on validation, source and activity of AMPs.

Feedback. Users can submit their suggestions/comments/queries using this feature.

Help. A detailed description on the use of the various features incorporated in CAMP is provided in this section for the benefit of users.

RESULTS AND DISCUSSION

CAMP is a comprehensive database on AMPs. It has 3782 peptides, which include experimentally validated peptides, patents and sequences predicted to be antimicrobial based on similarity. These peptides are separated into three different datasets (experimentally validated, patents and predicted) and thus users can restrict study on the dataset of their choice. The information on sequence, activity (MIC), target and non-target organisms are essential to delineate the sequence features important for specificity and activity of AMPs. Literature references, taxonomy and structural information of AMPs aid in understanding the nature of these peptides and identifying sequence patterns conserved across species. For this purpose, links to external databases like NCBI, NCBI Taxonomy Browser, Swiss-Prot and PDB are provided in CAMP. The browse and search features in the database facilitate easy retrieval of information present in CAMP. Hydrophobicity, net charge and secondary structure of the peptides are known to influence their antimicrobial activity (3). Tools for calculation of these parameters and the stability of AMPs like aliphatic and instability index are integrated in CAMP. Aliphatic index is a measure of the thermostability of the peptides (44) and instability index predicts the *in vivo* half-life of the peptide (45).

Research on AMPs is currently focused on rational design of peptides that would act as substitutes to antibiotics. Often, researchers would also want to enhance the antimicrobial activity of the peptide/s of their interest. Comparison of the user-defined sequence with that of existing AMPs will help in design of mutations that could enhance the antimicrobial activity of the query sequence. For this purpose, BLAST module has been incorporated in the database, which would

Table 1. Performance of prediction algorithms

Algorithm	MCC		Prediction accuracy for test dataset (%)		
	Training dataset	Test dataset	Overall	Positive dataset	Negative dataset
DA	0.75	0.74	87.5	87.8	87.4
RF	0.86	0.86	93.2	89.9	95.4
SVM	0.88	0.82	91.5	88.0	93.8

identify similar sequences in CAMP and NCBI-protein database. A prediction algorithm for antimicrobial activity, based on machine learning algorithms like RF, SVM and DA, is integrated in CAMP to help in estimating the antimicrobial potential of the peptides.

The performance of the algorithms can be understood from the MCC of test datasets, which are 0.86 (RF), 0.82 (SVM) and 0.74 (DA). All models have good sensitivity and specificity (Table 1). The prediction algorithm developed using RF performed the best with an MCC of 0.86 on the training dataset and 0.86 on the test dataset. SVM and DA also gave good prediction accuracies. It was observed that SVM and/or DA correctly predicted the antimicrobial potential of a few peptides that were wrongly predicted by RF. Hence, to increase the stringency of prediction, all three-prediction models are incorporated into CAMP and the probability values are included in the output. It is to be noted that not all of the experimentally validated AMP sequences have MIC value information and the use of different methods for the evaluation of MIC could be a deterrent in getting a good prediction model.

The results of the feature selection algorithm indicated that composition and distribution of charged and hydrophobic residues of peptides are the major determinants of their antimicrobial activity. AMPs are known to be mostly cationic and hydrophobic in nature. They elicit antimicrobial activity by attraction, attachment and permeation of the microbial cell membrane (3). The positive charge on AMPs brings about their interaction with the negatively charged microbial cell membrane leading to cell permeation and lysis. The hydrophobic domains in AMPs help in effective membrane permeabilization by partitioning the lipid bilayer (2,3).

Comparison with existing databases and prediction tools

Presently, there exist few databases of AMPs. Most of these databases are dedicated to specific classes of AMPs. For example, AMSdb (http://www.bbcm.units.it/~tossi/pag1.htm), PhytAMP (16), BACTIBASE (17) and PenBase (19) are databases of AMPs from eukaryotes, plants, bacteria and shrimps, respectively.

While RAPD (15) deals with recombinant AMPs, SAPD (21) contains information on synthetic antimicrobial peptides. The Defensin knowledgebase (18) and Peptaibol Database (20) deal with defensins and peptaibols respectively. These databases are very useful while searching for AMPs belonging to specific classes. Although Antimic (47), AMPer (48) and APD (14) are databases of AMPs, which include information of all classes of AMPs, APD (45) is the lone database that is currently available. APD also includes a tool for design of AMPs based on some known principles. CAMP contains nearly thrice the number of sequences as compared to APD (14) with additional information on taxonomy and activity (MIC values). The data in CAMP is divided into three datasets (experimentally validated, patents and predicted). The search features present in CAMP allow search against all or each of these datasets. The advanced search features in CAMP are similar to that of APD (14). The database can be queried with a combination of keywords using the Boolean operators 'and/or/not'. Records can be viewed and exported to Excel format for analysis using the 'Export' feature of CAMP. A detailed comparison of the existing databases on AMPs with CAMP is shown in Supplementary Table S1.

Some of the existing prediction servers for AMPs are APD (14), BACTIBASE (17), PhytAMP (16) and AntiBP (49). AMP prediction of APD is based on similarity and some known principles of AMPs. BACTIBASE and PhytAMP have an HMM-based model that predicts the AMP family of the query sequence/s. While these algorithms help in prediction of AMPs, they are trained and tested only with their specific source organisms. AntiBP predicts antibacterial peptides based on Quantitative Matrices (QM), Artificial Neural Network (ANN) and SVM. The training datasets are limited to N and/or C termini residues of antibacterial peptides (positive dataset) and non-secretory proteins (negative dataset). The users are constrained to predict the antibacterial activity for peptides that have to be at least 15 residues long. In contrast, the prediction algorithms in CAMP are trained on all classes of AMPs (antibacterial, antifungal and antiviral) and different classes of negative datasets (UniProt, non-secretory, non-antimicrobial and random sequences). The algorithms are trained on complete sequences and the antimicrobial activity can be predicted for sequences of variable length.

CONCLUSION

The role of AMPs in therapeutics is well-known. However, not much progress has been made in exploiting them as potent drugs. The understanding of the role of sequence of AMPs in their activity is important for their rational design as drugs. However, the precise sequence features important for antimicrobial activity are still not clear. The accuracy of prediction algorithms for AMPs heavily depends on the correctness and extent of information available in the training datasets used for the study. Hence, CAMP has been created with an objective to help researchers understand the role of primary structure of AMPs in their antimicrobial activity. The prediction algorithm available in CAMP will be useful in identifying potential AMPs based on their primary structure. In addition, tools to calculate some of the properties known to influence antimicrobial activity and *in vivo* stability of the peptides are also included. CAMP will be

updated monthly and should provide a valuable resource for research on AMPs.

AVAILABILITY AND REQUIREMENTS

CAMP is freely available at http://www.bicnirrh.res.in/antimicrobial/.

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. The authors also acknowledge the assistance provided by Mr Pravin Nilawe in database design and Ms Archana Sonawani in data collection. The authors are thankful to Dr P.V. Balaji for the resources provided by him. The authors are grateful to the anonymous reviewers whose comments have helped in improving the quality of the database.

FUNDING

Department of Science and Technology, Government of India (SR/S3/CE/52/2007); Indian Council of Medical Research (63/128/2001-BMS).

Conflict of interest statement. None declared.

REFERENCES

1. Jenssen,H., Hamill,P. and Hancock,R. (2006) Peptide antimicrobial agents. *Clin. Microbiol. Rev.*, **19**, 491–511.
2. Yeaman,M.R. and Yount,N.Y. (2003) Mechanisms of antimicrobial peptide action and resistance. *Pharmacol. Rev.*, **55**, 27–55.
3. Brogden,K.A. (2005) Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria? *Nat. Rev. Microbiol.*, **3**, 238–250.
4. Ong,P.Y., Ohtake,T., Brandt,C., Strickland,I., Boguniewicz,M., Ganz,T., Gallo,R.L. and Leung,D.Y.M. (2002) Endogenous antimicrobial peptides and skin infections in atopic dermatitis. *N. Engl. J. Med.*, **347**, 1151–1160.
5. Giuliani,A., Pirri,G. and Nicoletto,S.F. (2007) Antimicrobial peptides: an overview of a promising class of therapeutics. *Cent. Eur. J. Biol.*, **2**, 1–33.
6. Kamysz,W., Okroj,M. and Lukasiak,J. (2003) Novel properties of antimicrobial peptides. *Acta Biochim. Pol.*, **50**, 461–469.
7. Lai,R., Zheng,Y.T., Shen,J.H., Liu,G.J., Liu,H., Lee,W.H., Tang,S.Z. and Zhang,Y. (2002) Antimicrobial peptides from skin secretions of Chinese red belly toad Bombina maxima. *Peptides*, **23**, 427–435.
8. Mackintosh,J.A., Veal,D.A., Beattie,A.J. and Gooley,A.A. (1998) Isolation from an ant Myrmecia gulosa of two inducible O-glycosylated proline-rich antibacterial peptides. *J. Biol. Chem.*, **273**, 6139–6143.
9. Ganz,T. (2003) The role of antimicrobial peptides in innate immunity. *Integr. Comp. Biol.*, **43**, 300–304.
10. Sprules,T., Kawulka,K.E., Gibbs,A.C., Wishart,D.S. and Vederas,J.C. (2004) NMR solution structure of the precursor for Carnobacteriocin B2, an antimicrobial peptide from Carnobacterium piscicola. *Eur. J. Biochem.*, **271**, 1748–1756.
11. Lee,K.H., Lee,D.G., Park,Y., Kang,D., Shin,S.Y., Hahm,K.S. and Kim,Y. (2006) Interactions between the plasma membrane and the antimicrobial peptide HP (2-20) and its analogues derived from Helicobacter pylori. *Biochem. J.*, **15**, 105–114.
12. Tasiemski,A., Salzet,M., Benson,H., Fricchione,G.L., Bilfinger,T.V., Goumon,Y., Metz-Boutigue,M.H., Aunis,D. and Stefano,G.B. (2000) The presence of antibacterial and opioid peptides in human plasma during coronary artery bypass surgery. *J. Neuroimmunol.*, **109**, 228–235.
13. Zasloff,M. (2002) Antimicrobial peptides of multicellular organisms. *Nature*, **415**, 389–395.
14. Wang,Z. and Wang,G. (2004) APD: the antimicrobial peptide database. *Nucleic Acids Res.*, **32**, D590–D592.
15. Li,Y. and Chen,Z. (2008) RAPD: a database of recombinantly-produced antimicrobial peptides. *FEMS Microbiol. Lett.*, **289**, 126–129.
16. 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.
17. Hammami,R., Zouhir,A., Ben Hamida,J. and Fliss,I. (2007) BACTIBASE: a web-accessible database for bacteriocin characterization. *BMC Microbiology*, **7**, 89.
18. 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.
19. 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.
20. 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.
21. Wade,D. and Englund,J. (2002) Synthetic antibiotic peptides database. *Protein Pept. Lett.*, **9**, 53–57.
22. de Jong,A., van Hijum,S.A., Bijlsma,J.J., Kok,J. and Kuipers,O.P. (2006) BAGEL: a web-based bacteriocin genome mining tool. *Nucleic Acids. Res.*, **34**, W273–W279.
23. Bals,R. (2000) Epithelial antimicrobial peptides in host defense against infection. *Respir. Res.*, **1**, 141–150.
24. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
25. Murphy,L.R., Wallqvist,A. and Levy,R.M. (2000) Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Eng.*, **13**, 149–152.
26. Chakrabarti,P. and Pal,D. (2001) The interrelationships of side-chain and main-chain conformations in proteins. *Prog. Biophys. Mol. Biol.*, **76**, 1–102.
27. Rose,G.D., Geselowitz,A.R., Lesser,G.J., Lee,R.H. and Zehfus,M.H. (1985) Hydrophobicity of amino acid residues in globular proteins. *Science*, **229**, 834–838.
28. Tomii,K. and Kanehisa,M. (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.*, **9**, 27–36.
29. Li,Z.R., Lin,H.H., Han,L.Y., Jiang,L., Chen,X. and Chen,Y.Z. (2006) PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.*, **34**, W32–W37.
30. Dubchak,I., Muchink,I., Holbrook,S.R. and Kim,S.H. (1995) Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl Acad. Sci. USA*, **92**, 8700–8704.
31. Dubchak,I., Muchink,I., Mayor,C., Dralyuk,I. and Kim,S.H. (1999) Recognition of a protein fold in the context of the SCOP classification. *Proteins*, **35**, 401–407.
32. Breiman,L. (2001) Random forests. *Machine Learning*, **45**, 5–32.
33. R Development Core Team. (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
34. Liaw,A. and Wiener,M. (2002) Classification and regression by random forest. *R News*, **2**, 18–22.

35. Vapnik,V. (1995) *The Nature of Statistical Learning Theory*, 1st edn. Springer, NY.
36. Muller,K.R., Mika,S., Ratsch,G., Tsuda,K. and Scholkopf,B. (2001) An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Netw.*, **12**, 181–201.
37. Gunn,S. (1998) Support vector machines for classification and regression. *ISIS Technical Report ISIS-1-98*. Image Speech & Intelligent Systems Research Group, University of Southapton.
38. Kulkarni,A., Kulkarni,B.D. and Jayaraman,V.K. (2004) Support vector classification with parameter tuning assisted by agent-based technique. *Comput. Chem. Eng.*, **28**, 311–318.
39. Karatzoglou,A., Smola,A., Hornik,K. and Zeileis,A. (2004) Kernlab - An S4 package for Kernel methods in R. *J. Stat. Softw.*, **11**, 1–20.
40. Norusis,M.J. (1988) *SPSS/PC+Advanced StatisticsTM V2.0*. ISBN 0-918469-57-0, SPSS Inc., USA.
41. Baldi,P., Brunak,S., Chauvin,Y., Andersen,C.A.F. and Nielsen,H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.
42. Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
43. Altschul,S.F., Madden,T.L., Schaffer,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.
44. Ikai,A.J. (1980) Thermostability and aliphatic index of globular proteins. *J. Biochem.*, **88**, 1895–1898.
45. Guruprasad,K., Reddy,B.V.B. and Pandit,M.W. (1990) Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng.*, **4**, 155–161.
46. Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
47. Brahmachary,M., Krishnan,S.P., Koh,J.L., Khan,A.M., Seah,S.H., Tan,T.W., Brusic,V. and Bajic,V.B. (2004) ANTIMIC: a database of antimicrobial sequences. *Nucleic Acids Res.*, **32**, D586–D589.
48. 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.
49. Lata,S., Sharma,B.K. and Raghava,G.P.S. (2007) Analysis and prediction of antibacterial peptides. *BMC Bioinformatics*, **8**, 263.