OXFORD UNIVERSITY PRESS | Bioinformatics

Novel 3D Structure Based Model for Activity Prediction and Design of Antimicrobial Peptides

Journal:	*Bioinformatics*
Manuscript ID	BIOINF-2018-0439
Category:	Original Paper
Date Submitted by the Author:	05-Mar-2018
Complete List of Authors:	Liu, Shicai; School of Life Science and Technology, China Pharmaceutical University, Nanjing 210009, China; Bao, Jingxiao; School of Life Science and Technology, China Pharmaceutical University, Nanjing 210009, China Zheng, Heng; School of Life Science and Technology, China Pharmaceutical University, Nanjing 210009, China
Keywords:	antimicrobial peptides, Molecular Dynamics simulations, activity prediction, peptides design

Novel 3D Structure Based Model for Activity Prediction and Design of Antimicrobial Peptides

Shicai Liu, Jingxiao Bao and Heng Zheng*

School of Life Science and Technology, China Pharmaceutical University, Nanjing 210009, China

* To whom correspondence should be addressed.

Associate Professor Heng Zheng

School of Life Science and Technology, China Pharmaceutical University, No. 24 Tongjiaxiang, Nanjing 210009, China; E-mail: zhengh18@hotmail.com

Abstract

Motivation: The emergence and worldwide spread of multi-drug resistant bacteria makes an urgent challenge for the development of novel antibacterial agents. A perspective weapon to fight against severe infections caused by drug-resistant microorganisms is antimicrobial peptides (AMPs). AMPs are a diverse class of naturally occurring molecules that are produced as a first line of defense by all multi-cellular organisms. Limited by the number of experimental determinate 3D structure, most of the prediction or classification methods of AMPs were based on 2D descriptors, including sequence, amino acid composition, peptide net charge, hydrophobicity, amphiphilic, etc. Due to the rapid development of structural simulation methods, predicted models of proteins (or peptides) have been successfully applied in structure based drug design, for example as targets of virtual ligand screening. Here, we establish the activity prediction model based on the predicted 3D structure of AMPs molecule. To our knowledge, it is the first report of prediction method based on 3D descriptors of AMPs. To access the accuracy of the model, novel AMPs were designed by random mutations of parent peptide, and there antibacterial effect was measured by *in vitro* experiments.

Results: In this study, Molecular Dynamics simulations for 84 peptides have been performed. Regression models were built based on peptides' 3D structure and validated rigorously, which were then employed to *in silico* optimization of AMPs. Approximately 1000 peptides were generated by random mutation of parent peptide, and then 30 of potential AMPs were selected by amino acid composition-based method. Then, 3D structures of these AMPs were constructed, and there antimicrobial activity was predicted by 3D structure model. Four predicted active peptides and one predicted inactive peptide were selected for experimental validation. The results of in vitro antibacterial tests confirmed the predict results. The antibacterial activity of the four new AMPs was higher than that of the parent peptide. No antibacterial activity was observed in the range of concentrations tested for the predicted inactive one.

Conclusion: The 3D descriptor Based Models can be used to predict the activity of AMPs, and can be used to design novel AMPs.

Contacts: zhengh18@hotmail.com

1 Introduction

In 2014, the WHO's (World Health Organization) report about global surveillance of antimicrobial resistance reveals that antibiotic resistance is no longer a prediction for the future(Hansen, et al., 2015). It is happening now, across the world. With the emergence of more and more multi-drug resistant bacteria, the development of new antibacterial drugs turns into an urgent challenge(Andersson, et al., 2016). A perspective weapon to fight against severe infections caused by drug-resistant microorganisms is antimicrobial peptides (AMPs) (Chung and Khanum, 2017; Fox, 2013; Kosikowska and Lesner, 2016; Sierra, et al., 2017; Silva, et al., 2016). AMPs are a diverse class of naturally occurring molecules that are produced as a first line of defense by all multi-cellular organisms(Mondal and Jagdale, 2016). These peptides can have broad activity to kill bacteria, fungi, yeasts, viruses and even cancer cells. In addition, AMPs have been found to display immunomodulatory functions such as wound healing, chemotactic, angiogenic(Mangoni and Bhunia, 2016; Sierra, et al., 2017), which make them even more attractive templates for the new-generation antibiotics.

There are more than 2,500 AMPs found in nature(Zhang and Gallo, 2016), such as single-celled organisms, plants, insects, animals. Most of the AMPs information is included in the DRAMP database(Fan, et al., 2016) established by our laboratory. Although AMPs have become as promising candidates to traditional antibiotics for treatment of bacterial diseases, many potential problems should be solved before they can be put in clinic and commerce, including instable and easy to be digested by enzyme in vivo, relatively low activity comparing with antibiotics, toxicity against eukaryotic cells, high production costs(Bradshaw, 2003). There still needs much effort on designing novel AMPs to overcome these limitations. In recent years, machine learning has been applied in AMPs analysis, which may become useful tool to speed up the classification, prediction and design of AMPs(Lee, et al., 2017). By using the database resources, the AMPs information was extracted to establish the activity prediction model. At present, most of the activity prediction models are established based on the primary structure of AMPs(Chen, et al., 2016; Gupta, et al., 2016; Holton, et al., 2013; Lin and Xu, 2016), the amino acid composition, peptide net charge, hydrophobicity, amphiphilic, helix and other structural parameters are all critical for AMPs' activity. Changing any of these parameters can lead to AMPs' activity reduced or even lost. There is a strong correlation between all parameters, and it is not comprehensive enough to predict its antibacterial activity by a given amino acid sequence of AMPs. Feature extraction of AMPs is an important step in data analysis and machine learning. Even the most sophisticated algorithms would perform poorly if inappropriate features are used, while simple methods can potentially perform well when they are fed with the appropriate features. Therefore, in this study, we will establish the activity prediction model based on 3D structure of the AMPs molecule. However, the 3D structure of most AMPs is unknown, only a small part of the 3D structure of the AMPs is determined. Only 5.5% of the AMPs' 3D structure were determined in General dataset of DRAMP database(Fan, et al., 2016). A variety of methods have been developed for the prediction of proteins' 3D structure in the field of computational biology, including homology modeling(Rodriguez, et al., 1998), folding recognition(Lüthy, et al., 1992), and *ab initio* calculations(Chou and Zhang, 1995; Nakashima, et al., 1986; Pillardy, et al., 2001). The first two methods are based on the known protein structure as a template to generate the structure by sequence alignment. With the advancement of molecular dynamics simulation technology, the modeled structure is generally considered to be reasonable and credible after a period of molecular dynamics simulation(Pirtskhalava, et al., 2016). Therefore, we plan to predict the 3D structure of AMPs by homology modeling and molecular dynamics simulation.

In this study, molecular dynamics simulations for 84 peptides have been performed, and we establish the activity prediction model based on the predicted 3D structure of the AMPs molecule. To our knowledge, it is the first report of prediction method based on 3D descriptors of AMPs. To access the accuracy of the model, novel AMPs were designed by random mutations of parent peptide, and there antibacterial effect was measured by *in vitro* experiments.

2 METHODS

2.1 Molecular Dynamics Simulations

The starting 3D structure model of AMPs was generated based on homology modeling using MOE (Molecular Operating Environment, https://www.chemcomp.com/). The GB/VI was used as the scoring standard of the model. Other parameters were set as the default values.

The starting 3D structure models were then optimized with molecular dynamics (MD) simulations. MD simulations of AMPs were performed in AMBER package(Case DA, et al., 2014) using the FF14SB force field(Maier, et al., 2015). The starting 3D structure model was first solvated with a truncated octahedron box of TIP3P water molecules(Jorgensen, et al., 1983) that extended 10 Å from the atoms and Na^+ and Cl^- neutralizing counterions. Prior to the start of the production simulation, 5000 steps of energy minimization were performed using steepest descent and conjugate gradient method, respectively. Long range electrostatic interactions were addressed by particle mesh Ewald summation, with a real space cutoff of 1.0 nm.

Production runs were conducted at 300 K for 100 ns with data collected every 100 ps. For all simulations, a time step of 2 fs was employed. A Langevin thermostat was used to maintain temperature and a Monte carlo at 1 atm was used to control pressure.

2.2 Datasets

We have extracted 84 experimentally validated anti-listerial peptides from DRAMP databases(Fan, et al., 2016). All these peptides were unique and considered as positive examples. Since there are very few experimentally proved non-antilisterial peptides, we derived 84 random peptides from SwissProt(UniProt, 2015) proteins with the keywords, "not antimicrobial activity", "not antibactreial activity", "not antilisiterial activity", "a length range of 5–70 amino acids" and "have 3D structure". In this study, we assign these random peptides as non-antilisterial peptides (negative examples), though it is possible that some of these random peptides have antimicrobial properties. After obtaining the positive dataset and negative dataset, the training set and the testing set were screened with CD-Hit(Li and Godzik, 2006), with sequence identity cut-off of 85% in order to remove sequence redundancy in the set.

Then, the screened and unscreened data sets were used to establish the prediction model, respectively (**Fig. 1**).

2.3 Feature extraction

Both local and global descriptors were used to characterize peptide structures. The amino acid descriptors amino acid composition (AAC) was employed as local characterization to parameterize peptides. From some literature we know that the AAC is the most important factor for peptide classification and design, so it may be a good choice. For AAC calculation only 20 naturally amino acids are considered, and it has been successfully used for many protein classification problems(Wang, et al., 2016). AAC can be calculated using the formula below:

$$\mathrm{AAC(i)} = \frac{\text{Total number of amino acid(i)}}{\text{Total number of all possible amino acids}}$$

Global structure characterization named MOE-Descriptors was carried out using MOE(https://www.chemcomp.com/) based on 3D structure of AMPs: The peptide structures were converted to three classes of molecular descriptors as 2D Molecular Descriptors, Protein Property Descriptors and 3D Molecular Descriptors by using MOE program. Some of these features may not be relevant to the prediction of AMPs and they could be also redundant with each other. So, we performed two feature selection methods, the mean decrease in accuracy (MDA) and principal component analysis (PCA), to remove the irrelevant and redundant features, which was calculated using the randomForest package and SciViews package in R (http://cran.r-project.org//), respectively. MDA represents the average decrease of classification accuracy on the OOB samples when the values of a particular feature are randomly permuted. Thus the permutation based MDA can be utilized to evaluate the contribution of each feature to the classification. After excluding collinear and irrelevant descriptors, 90 molecular descriptors selected by MDA and 26 principal components derived by PCA were used for further analysis.

2.4 Regression modeling

Two machine learning methods, support vector machine (SVM)(Karasuyama and

Takeuchi, 2010) and random forest (RF)(Lin and Jeon, 2006), were employed to conduct regression modeling of the multivariate correlation between the peptide structural parameters and antibacterial activity. SVM was implemented by using e1071 package in R (http://cran.r-project.org//). SVM is a classification algorithm based on statistical learning theory, which aims at the structural risk minimization rather than the traditional empirical risk minimization and is especially suitable for small-sample, high-dimensional and strong collinear problems. RF was implemented using random Forest package in R. RF uses an ensemble of unpruned decision trees, each grown using a bootstrap sample of the training data, and randomly selected subsets of predictor variables as candidates for splitting tree nodes, which is to maintain the "strength" of the trees while reducing their correlation with each other.

2.5 Evaluating performance

Once the models were ready, their performance was tested in terms of the sensitivity, specificity, accuracy, and Mathew's Correlation Coefficient (MCC). They can be calculated using the formula below:

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

Where TP, FP, TN and FN stand for the number of true positives, false positives, true negatives and false negatives, respectively.

The performance of the models was evaluated by employing a ten-fold cross-validation technique. The whole dataset was divided into ten sets such that in each round, nine sets were used for training and one was set aside for testing. Repeated ten times, this ensured that each set was used once for testing the model that was trained on the remaining nine.

In order to evaluate the performance of our models, we have created an independent dataset of 8 AMPs randomly selected from the final 84 AMPs and 8 non-AMPs randomly selected from the final 84 non-AMPs, which have not been included in the training, feature selection and parameters optimization of the model.

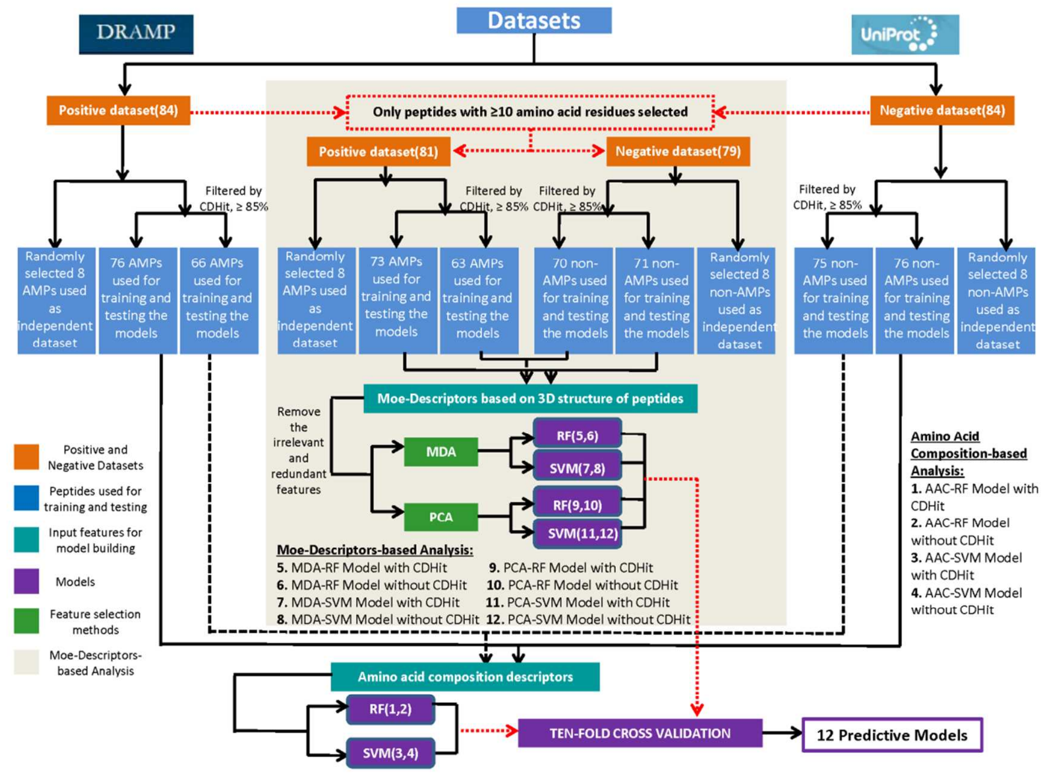

Fig.1. Flowchart depicting the overall approach implemented as the establishment of activity prediction model. The flowchart gives an overview of the steps followed in building the predictive models.

2.6 *In silico* optimization of AMPs

By using database resources, more than 2500 natural AMPs from DRAMP were used for sequence alignment, and a 7-amino-acid consensus sequence (**short peptides 1**, FLRRIRV-NH_2) was apparent in some peptides (**Fig. 2**), and was selected as seed peptide. The second position of most AMPs is tryptophan(Piotto, et al., 2012), contributing to the anchoring of AMPs on the cell membrane. The leucine at position 2 of the consensus sequence was transformed into tryptophan, resulting in **short peptides 2**(FWRRIRV-NH_2). We argued that smaller peptides would be less expensive to produce and that a reduction in the number of amino acids would allow a more comprehensive understanding of the amino acid sequence responsible for antimicrobial activity. Therefore, we screened the sequence from DRAMP, with sequence length less than fifteen, complete sequence information and anti-listeria activity. Finally, we get the parental peptide DRAMP00228 (Fan, et al., 2016; Ghalfi, et al., 2010) (TPVVNPPFLQQT-NH_2). We link short peptides and parental peptide, and random single-point was imposed on the hybrid peptides, resulting in random mutant. The mutation introduces only natural amino acids to the peptide. After mutation the antimicrobial activity of the mutant was predicted by using the predictive models, and the activity test was carried out (**Fig. 3**).

DRAMP02975 VRRFPWWWPFLRR
DRAMP03575 FKRIVQRIKDFLR
DRAMP02980 GLICESCRKIIQKLEDMVGPQPNEDTVTQAASRVCDKMKILRGVCKKIMRTFLRRISKDILTGKKPQAICVDIKICKE
DRAMP02530 MVVIVYVIVAYDVNVERVNRVKKFLRRYLNWVQNSLFEGELSSADLEEVKMGLREIINEDEDMVVIYRFRSADAFKREVMGIEKGLGGEEVI
DRAMP03430 FLRRSVSGFQECHSKGGYCYRYYCPRPHRRLGSCYPYAANCCRRRR
DRAMP04578 KRFGRLAKSFLRMRILLPRRKILLAS
DRAMP02851 RLCRIVVIRVCR
DRAMP03462 ATELEKALSNVIEVYHNYSGIKGNHHALYRDDFRKMVTTECPQFVQNKNTESLFKELDVNSDNAINFEEFLVLVIRVGVAAHKDSHKE
DRAMP02040 FLPAVIRVAANVLPTAFCAISKKC
DRAMP02073 FLPAVIRVAANVLPTVFCAISKKC
DRAMP02748 GVVPHDFRI
DRAMP02910 LRRIIRKIIHIIKK
DRAMP02348 WLRRIGKGVKIIGGAALDHL
DRAMP02349 GRRKRKWLRRIGKGVKIIGGAALDHL
DRAMP02664 ACYCRIPACLAGERRYGTCFYLRRVWAFCC
DRAMP03545 RVIEVVQGACRAIRHIPRRIRQGLERIL
DRAMP03462 ATELEKALSNVIEVYHNYSGIKGNHHALYRDDFRKMVTTECPQFVQNKNTESLFKELDVNSDNAINFEEFLVLVIRVGVAAHKDSHKE
... ...

Fig.2. Sequence alignment of the natural AMPs from DRAMP. Underlining represents consensus sequence amino acids.

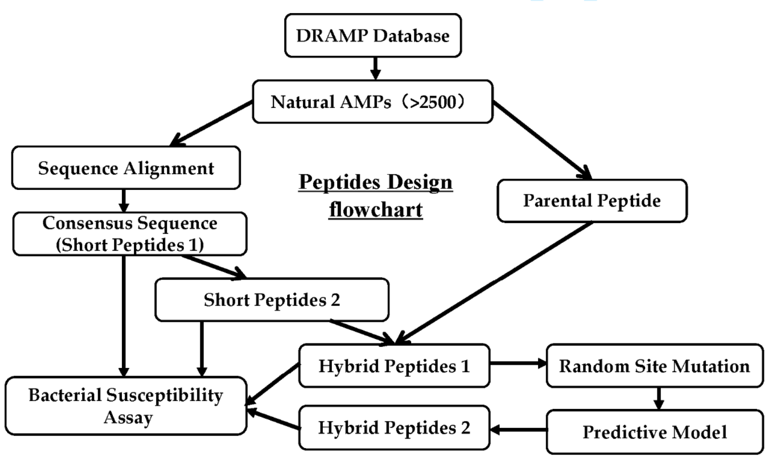

Fig.3. The flowchart gives an overview of the steps followed in designing novel bioactive peptides.

2.7 Bacterial susceptibility assay

All peptides used in this study were synthesized by ChinaPeptides (ChinaPeptides Co., Ltd) using 9-fluorenylmethoxy carbonyl (Fmoc) chemistry and purified to a purity of >95% using high-performance liquid chromatography (HPLC). Peptide mass was confirmed by mass spectrometry.

The experimentally determined strains are as follows: *Listeria monocytogenes* (ATCC 19115), *Staphylococcus aureus* (CMCC(B)26003), *Bacillus subtilis* (CMCC(B)63501), *Escherichia coli* (CMCC(B) 44102), *Pseudomonas aeruginosa* (CMCC(B)10104), *Enterococcus faecalis*（clinical strains）from China Pharmaceutical University Microbiology Laboratory.

Minimal inhibitory concentration(MIC) of peptides were determined using broth microdilution method. Two-fold serial dilutions of eight peptides were prepared from 1024.0 to 1.0μg/ml in 96-well microtiter plate (100.0μl of each well). Then peptide dilutions were mixed with LB broth and bacterial culture (100.0μl) containing 2.0 $\times 10^5$CFU/ml. Final peptide concentrations ranged from 0.5 to 512.0μg/ml. The final bacterial concentration was approximately 1.0 $\times 10^5$ CFU/ml. Positive controls were incubated with Cefuroxime instead of peptide, at concentrations from 0.5 to 512.0μg/ml. Negative and blank controls were incubated, respectively, with sterile deionized water or only LB broth. Microtiter plates were incubated at 37℃ for 24h under normal atmospheric conditions. OD_{600} was measured using a microplate spectrophotometer (Multiskan GO, Thermo Scientific, USA). MIC was recorded as the endpoint where no difference of OD_{600} could be detected with respect to the blank LB broth(Song, et al., 2012). MIC assays were performed three times for all strains.

3 RESULTS AND DISCUSSION

3.1 Molecular Dynamics Simulations

MD simulations for 84 peptides (the positive dataset, **Fig. 1**) have been performed. In these 84 peptides there were only five peptides have experimental determined structures. However, we still carried out homology modeling and MD simulation for the five peptides, to valid the structure prediction method by comparison of the predict model with the known crystal structure. For example, the crystal structure (PDB ID: 2m60) of DRAMP18261 is compared with the representative structure (**Fig. 4**) obtained after the structure simulation to obtain the RMSD value of 1.968 Å, indicating that the simulation result is feasible.

A simplified similarity measure, C^{α} torsion angle, was used to analyze and present the MD simulation results. The C^{α} torsion angle is defined as the non-bonded torsion angle arising from four consecutive C^{α} atoms along the chain of the peptide. For each frame of the MD simulation, an array of C^{α} torsion angles for each of the amino acids in the peptide was created (Devadoss and Paul Raj, 2014; Flocco and Mowbray, 1995; Victor Paul Raj and Exner, 2014). The "representative structure" was identified as the simulation frame whose array has the smallest mean Root Mean Square Deviation to all the other frames in the MD simulation, and a PDB file for this frame was generated (**Fig. 4. A**). In addition, the C^{α} torsion angle arrays were used to create a heat map plot, showing the simulation frames groupings with similar structures and suggesting the number of different structures' types arising during the MD simulation. The heat map was produced by first re-ordering all of the simulation frames according to increasing distance of their corresponding C^{α} torsion angle array. The C^{α} torsion angle distance is calculated between all pairs of frames in the MD simulation trajectory. The heat map was then constructed with each axis corresponding to all the simulation frames ordered as described above. Each element of the heat map represents the color-coded difference between the arrays for the two corresponding frames (**Fig. 4. B**), Note that the heat map is symmetrical above and below the diagonal, the latter corresponding to the comparison between each simulation frame and itself. The propensity of each amino acid position along the peptide to assume a secondary structure type (helix, sheet, bend or coil) over the course of the MD simulation was determined using the program "AmberTools", and the secondary structure type for each amino acid versus simulation frame have been plotted (**Fig. 4. C**). The resulting "representative structures" were used to establish the activity prediction models as per the procedure illustrated in the methods section. The peptides having crystal structure are using crystal structure to extracting feature.

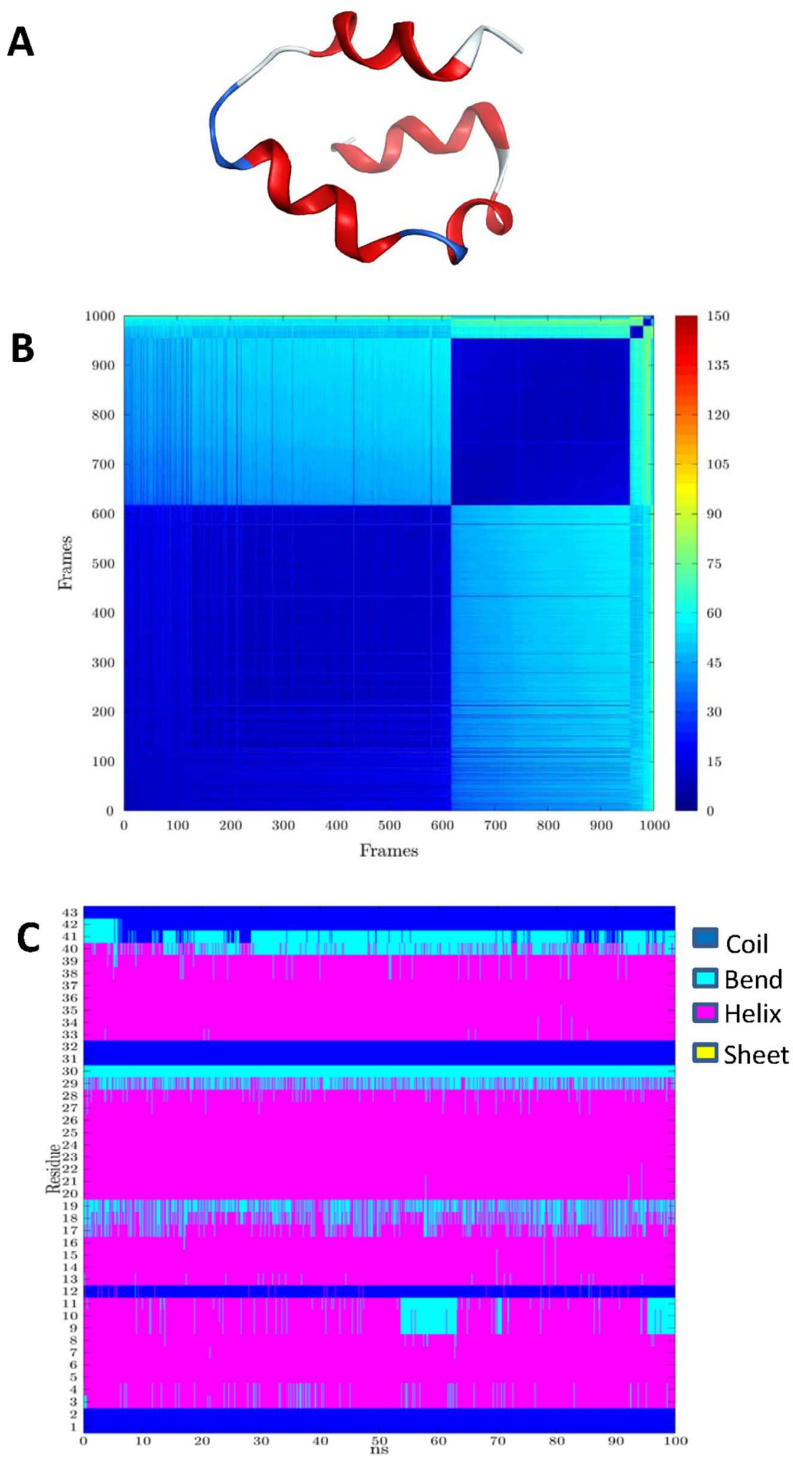

Fig.4. The results of MD simulation. (A) Representative Structure produced by MD simulation for DRAMP18261. (B) Heat map plot produced by MD simulation for DRAMP18261. (C) Secondary Structure assumed by each amino acid throughout the MD simulation for DRAMP18261

3.2 Machine learning regression modeling

The statistics of twelve models were summarized in **Table 1** and **Table 2**. The accuracies of the **AAC-RF with CDHit (1)** and **AAC-SVM with CDHit (3)** based models were 89.26% and 85.71%, with MCC values of 0.79 and 0.72 respectively, while the **AAC-RF without CDHit (2)** and **AAC-SVM without CDHit (4)** based models performed with accuracies of 80.00% and 86.67%, the corresponding MCC values being 0.60 and 0.74 respectively. To get best prediction results, only the **AAC-RF with CDHit (1)** based models with accuracy 89.26% and MCC 0.79 were selected.

The MOE-Descriptors of the peptides' 3D structure to be used as input features were selected for building the RF and SVM-based models (**Fig. 1**). Performances of MOE Descriptors-based models were summarized in **Table 2**. The models were evaluated using a ten-fold cross validation technique as per the procedure illustrated in the Methods section. As might be expected, overall, MOE-Descriptors of the peptides' 3D structure performed much well as compared to amino acid composition descriptor in sensitivity, specificity, accuracy, and MCC (**Table 1, Table 2**). In addition, the models of dataset screened with CD-Hit performed much well as compared to the models based on unscreened dataset. In the models based on MOE-Descriptors, although scoring function results of the **MDA-RF with CDHit (5)** and **MDA-RF without CDHit (6)** based model were all 1.00 (**Table 2**), their independent dataset results did not perform well (**Table 3**), which is occuring overfitting phenomenon. The **MDA-SVM with CDHit (7)** (accuracy of 92.59% with sensitivity, specificity and MCC of 90.00, 94.12 and 0.84, respectively) and **PCA-SVM with CDHit (11)** (accuracy of 92.59% with sensitivity, specificity and MCC of 90.00, 94.12 and 0.84, respectively) based model exhibit the similar profile in sensitivity, specificity, accuracy and MCC, although some difference between them on independent dataset results can be observed. Compare to the **MDA-SVM with CDHit (7)**, the **PCA-SVM with CDHit (11)** based model has the comparable fitting ability on training set but worse predictive power on independent dataset. In all the models **MDA-SVM with CDHit (7)** seems to have the best performance in internal stability and external predictability with accuracy 92.59%, MCC 0.84(on training set), and accuracy 100.00%, MCC 1.00(on independent dataset), suggesting that the combination of SVM and MOE descriptors processed with MDA on CDHit-screened datasets is a good choice that exhibits high internal stability and strong external predictive power.

In order to validate our *in silico* methods, performances of our models were evaluated on independent dataset. Positive and negative independent datasets of 8 peptides each were used to judge the predictive capacity of the twelve models（**Fig. 1**）. All these models performed reasonably good as shown in **Table 3,** demonstrating that these models are useful or effective in real life. The **MDA-SVM with CDHit(7)** performed with the highest accuracy (with accuracy, sensitivity, specificity and MCC of 100.00%, 100.00%, 100.00% and 1.00, respectively) among all these models. Performances on both the training and independent datasets were considered to select the best models for the design of novel AMPs.

Table 1. Performance of the models based on amino acid composition of the peptides on training datasets

Classifier	Sensitivity	Specificity	Accuracy	MCC
AAC-RF with CDHit**(1)**	72.73	100.00	89.26	0.79
AAC-RF without CDHit**(2)**	71.43	87.50	80.00	0.60
AAC-SVM with CDHit**(3)**	63.64	100.00	85.71	0.72
AAC-SVM without CDHit**(4)**	78.57	93.75	86.67	0.74

AAC: amino acid composition; **RF**: Random Forest algorithm; **SVM**: support vector machine algorithm; **with CDHit**: CDHit-screened datasets; **without CDHit**: CDHit-unscreened datasets.

Table 2. Performance of the models based on MOE-Descriptors of the peptides' 3D structure on training datasets

Classifier	Sensitivity	Specificity	Accuracy	MCC
MDA-RF with CDHit**(5)**	100.00	100.00	100.00	1.00

MDA-RF without CDHit(**6**)	100.00	100.00	100.00	1.00
MDA-SVM with CDHit(**7**)	90.00	94.12	92.59	0.84
MDA-SVM without CDHit(**8**)	92.86	86.67	89.66	0.80
PCA-RF with CDHit(**9**)	80.00	88.24	85.19	0.68
PCA-RF without CDHit(**10**)	78.57	86.67	82.76	0.66
PCA-SVM with CDHit(**11**)	90.00	94.12	92.59	0.84
PCA-SVM without CDHit(**12**)	100.00	66.67	82.76	0.70

MDA: mean decrease in accuracy; **PCA**: principal component analysis; **RF**: Random Forest algorithm; **SVM**: support vector machine algorithm; **with CDHit**: CDHit-screened datasets; **without CDHit**: CDHit-unscreened datasets.

Table 3. Performance of the models on independent datasets

Classifier	Sensitivity	Specificity	Accuracy	MCC
AAC-RF with CDHit(**1**)	71.43	87.50	80.00	0.60
AAC-RF without CDHit(**2**)	87.50	75.00	81.25	0.63
AAC-SVM with CDHit(**3**)	87.51	75.00	80.00	0.61
AAC-SVM without CDHit(**4**)	87.50	75.00	81.25	0.63
MDA-RF with CDHit(**5**)	100.00	87.50	93.33	0.88
MDA-RF without CDHit(**6**)	100.00	87.50	93.75	0.88
MDA-SVM with CDHit(**7**)	100.00	100.00	100.00	1.00
MDA-SVM without CDHit(**8**)	87.50	100.00	93.75	0.88
PCA-RF with CDHit(**9**)	100.00	87.50	93.33	0.88
PCA-RF without CDHit(**10**)	87.50	87.50	87.50	0.75
PCA-SVM with CDHit(**11**)	87.51	100.00	93.33	0.87
PCA-SVM without CDHit(**12**)	87.50	100.00	93.75	0.88

3.3 *In silico* optimization of AMPs

As per the procedure illustrated in the methods section (**Fig. 3**), after obtaining the short peptide 1(FLRRIRV-NH_2) and short peptide 2(FWRRIRV-NH_2), the short peptides are bound to the parental peptide DRAMP00228 (TPVVNPPFLQQT-NH_2), respectively, resulting the hybrid peptide 1 (FLRRIRV-TPVVNPPFLQQT-NH_2 and FWRRIRV-TPVVNPPFLQQT-NH_2). Random mutational point of the hybrid peptide 1 is performed, resulting in nearly 1000 new peptides. Due to these new peptides do not have 3D structure, a preliminary prediction based on amino acid composition (**AAC-RF with CDHit(1)**) was used, and approximately 350 of these new peptides are predicted to be active. Then, we randomly chose 30 peptides from the preliminary selection result for 3D structure simulation. Then **MDA-SVM with CDHit(7)** based model was used to predict the peptides' activity after getting 3D structure. Finally, we selected five peptides (including the hybrid peptide 1) from the prediction results for experimental validation. Although the predictive models are established based on anti-listeria activity of AMPs, several of the strains were tested in the case of experimental validation.

The results were summarized in **Table 4**. Consequently, the short peptides FLRRIRV-NH_2 showed an ability to inhibit *Listeria monocytogenes* (ATCC 19115) and *Staphylococcus aureus* (CMCC(B)26003) with MIC 128μg/ml. Moreover, FWRRIRV-NH_2 displayed higher antibacterial activity across *Listeria monocytogenes* (ATCC 19115), *Staphylococcus aureus* (CMCC(B)26003) and *Bacillus subtilis* (CMCC(B)63501) with MIC 32μg/ml, 64μg/ml, 128μg/ml, respectively. The leucine at position 2 of FLRRIRV-NH_2 transformed into tryptophan makes it significantly more active. The short peptides adopt an amphipathic conformation on Helical wheel projection diagrams (**Fig. 5**). We argue that amino acid change at position 2 of FLRRIRV-NH_2 increased amphiphilicity. Tryptophan is a hydrophobic amino acid containing a benzene ring, which can effectively promote the anchoring of AMPs on the cell membrane, resulting in activity of the peptide increased. The parent peptide TPVVNPPFLQQT-NH_2 has only antilisteria activity with MIC 512μg/ml.

In the five designed peptides, except for FLRRIRVTPVVNPPFLQQT-NH_2 with the predicted result, no anti-listeria activity, and the other four predicted results are all active, which is suggesting that the predicted results are consistent with the experimental verification results. FLRRIRVTPWVNPPFLQQT-NH_2 showed a ability to inhibit *Listeria monocytogenes* (ATCC 19115) and *Staphylococcus aureus*

(CMCC(B)26003) with MIC 128μg/ml, 256μg/ml, respectively. FWRRIRVTPVVNPPFLQQT-NH_2 showed a ability to inhibit *Listeria monocytogenes* (ATCC 19115) and *Staphylococcus aureus* (CMCC(B)26003) with MIC 256μg/ml. FWRRIRVTPWVNPPFLQQT-NH_2 showed a ability to inhibit *Listeria monocytogenes* (ATCC 19115), *Staphylococcus aureus* (CMCC(B)26003) and Bacillus subtilis (CMCC(B)63501) with MIC 64μg/ml, 64μg/ml, 256μg/ml, respectively. The designed peptide FWRRIRVTPVVNPWFLQQT-NH_2 showed a marked ability to inhibit *Listeria monocytogenes* (ATCC 19115), *Staphylococcus aureus* (CMCC(B)26003), *Bacillus subtilis* (CMCC(B)63501), *Escherichia coli* (CMCC(B) 44102) and *Pseudomonas aeruginosa* (CMCC(B)10104) compared with the parental peptides, with MIC 32μg/ml, 32μg/ml, 64μg/ml, 256μg/ml, 256μg/ml, respectively. These assays confirmed that these designed peptides displayed approximately 2–16-fold higher antibacterial activity across *Listeria monocytogenes* (ATCC 19115) in comparison to their parent peptide TPVVNPPFLQQT-NH_2 and showed an extended antibacterial spectrum.

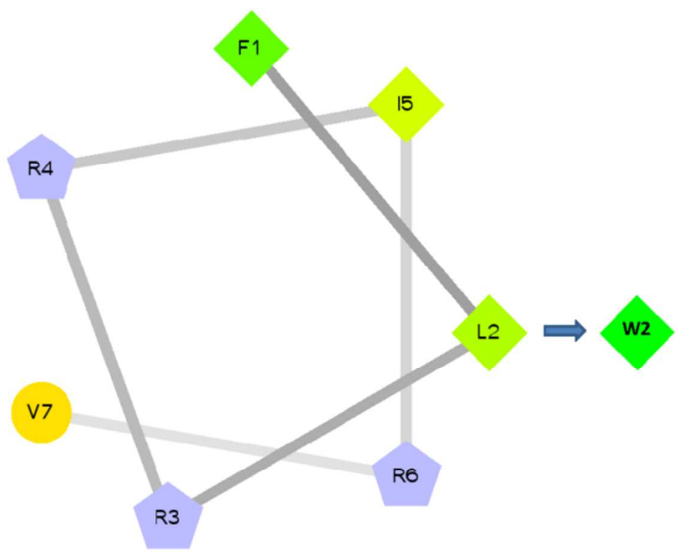

Fig.5. Helical wheel projection diagrams of the short peptides. Arrows indicate the substituted-amino acids. Hydrophilic residues are represented in circles, hydrophobic residues in diamonds, positive charged residues in pentagons. The most hydrophobic residue is green, and the amount of green is decreasing proportionally to the hydrophobicity.

Table 4. Antibacterial activity of short peptides, parental peptide, and designed peptides (MIC µg/ml)

	Listeria monocytogenes	*Staphylococcus aureus*	*Bacillus subtilis*	*Escherichia coli*	*Pseudomonas aeruginosa*	*Enterococcus faecalis*
Short Peptides						
FLRRIRV-NH2	128	128	—	—	—	—
FWRRIRV-NH2	32	64	128	—	—	—
Parental Peptide						
TPVVNPPFLQQT-NH2	512	—	—	—	—	—
Designed Peptides						
FLRRIRVTPVVNPPFLQQT-NH2	—	—	—	—	—	—
FLRRIRVTPWVNPPFLQQT-NH2	128	256	—	—	—	—
FWRRIRVTPVVNPPFLQQT-NH2	256	256	—	—	—	—
FWRRIRVTPVVNPWFLQQT-NH2	32	32	64	256	256	—
FWRRIRVTPWVNPPFLQQT-NH2	64	64	256	—	—	—
Positive Control						
Cefuroxime	4	4	≤0.5	8	—	16

"— " indicates that the peptide is inactive at 512µg/ml.

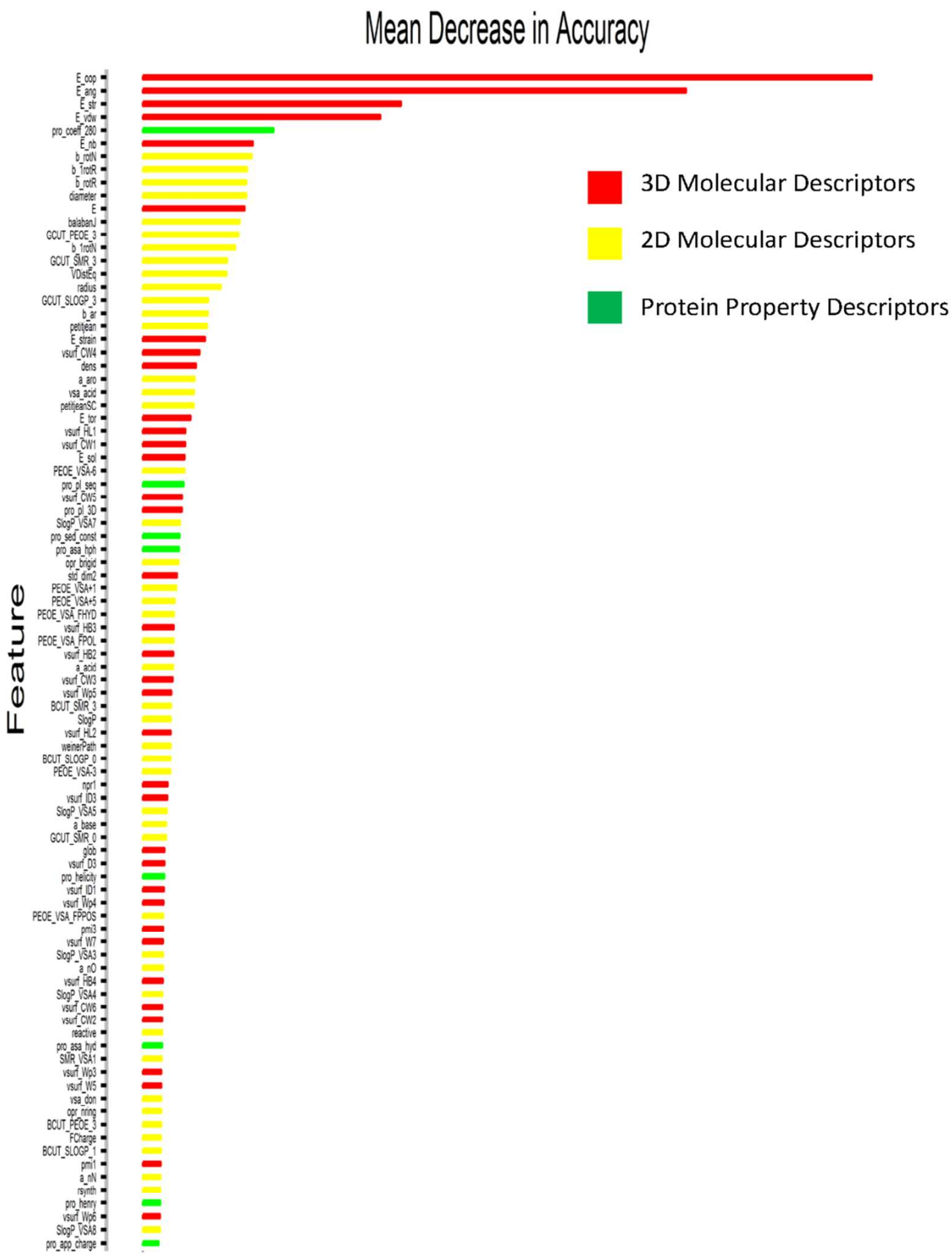

Fig.6. The top 90 features after feature selection based MDA on MOE-Descriptors.

4 CONCLUSION

An integrated in silico–in vitro discovery of bioactive peptides was described to perform computer-aided rational design of antimicrobial peptides. In the procedure, regression models were built based on peptides' 3D structure and validated rigorously. The performance of **MDA-SVM with CDHit (7)** based model was measured with an accuracy of 92.59% and a MCC of 0.84 on the training and testing dataset. Additionally, **MDA-SVM with CDHit (7)** was evaluated using an independent dataset resulting in an accuracy of 100.00%, which were then employed to direct *in silico* optimization of AMPs, attempting to obtain a new AMPs population with improved antimicrobial potency. During the process of feature selection based MDA, we selected the top 90 features in MOE-Descriptors to constitute model which achieved the best result. The top 90 features were shown in **Fig. 6**. From **Fig. 6**, we drew a conclusion that 3D Molecular features were more important for the modeling, mainly including Potential Energy Descriptors, Surface Area, Volume and Shape Descriptors. In the top 90 features, the first 4 all belong to Potential Energy Descriptors, and the abscissa value of the four features is much larger than the others, indicating that their importance is much higher than other features. 3D structure descriptors are closer to reality, and better reflect the essence of the peptides drug.

Five AMPs were successfully designed and synthesized, and their antibacterial activity was tested against six bacteria. Consequently, the results predicted by regression model are consistent with the experimental verification results. The designed peptide FWRRIRVTPVVNPWFLQQT-NH_2 exhibited the highest activity in all the tested candidates, showing a marked ability and an extended antibacterial spectrum to inhibit *Listeria monocytogenes* (ATCC 19115), *Staphylococcus aureus* (CMCC(B)26003), *Bacillus subtilis* (CMCC(B)63501), *Escherichia coli* (CMCC(B) 44102) and *Pseudomonas aeruginosa* (CMCC(B)10104) compared with the parental peptides, with MIC 32μg/ml, 32μg/ml, 64μg/ml, 256μg/ml, 256μg/ml, respectively. In structure prediction, we successfully obtained the 3D structure of 84 peptides by MD simulations, and later we will integrate these results (including the trajectory data of MD simulation, PDB file of representative structure, heat map, secondary Structure assumed by each amino acid) into our DRAMP database(Fan, et al., 2016). By analyzing the structure of these peptides, 73 out of the 84 peptides contain a stable helical structure. In the five designed peptides, FLRRIRVTPVVNPPFLQQT-NH_2 shows a coil structure, while the other four contain stable helical structure, implying that the structure of helix is crucial for the peptides' activity. In the process of sequence alignment, we identified a consensus sequence (FLRRIRV-NH_2) present in several antimicrobial peptides. Meanwhile, the leucine at position 2 of the consensus sequence was transformed into tryptophan to effectively promote the anchoring of AMPs on the cell membrane, resulting in the peptide FWRRIRV-NH_2. FWRRIRV-NH_2 displayed higher antibacterial activity across *Listeria monocytogenes* (ATCC 19115) and *Staphylococcus aureus* (CMCC(B)26003) with MIC 32μg/ml, 64μg/ml, respectively, which will serve as a basis for iterative design of improved peptides. Based on the strengths of these designed peptides, this type of rational design will be useful for future assessments to develop and apply these peptides as novel antibiotics.

ACKNOWLEDGEMENTS

This research was supported by the Top-Notch Academic Programs Project of Jiangsu Higher Education Institutions (No. PPZY2015A057), a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), and the Fundamental Research Funds for the Central Universities (No.2632018ZD04).

Conflict of Interest: none declared.

References

Case DA, Darden TA, Cheatham TE III et al (2014) Amber 14.0.University of California, San Francisco.

Molecular Operating Environment (MOE) 2016 Chemical Computing Group, Montreal, Canada, https://www.chemcomp.com/.

Andersson, D.I., Hughes, D. and Kubicek-Sutherland, J.Z. Mechanisms and consequences of bacterial resistance to antimicrobial peptides. *Drug Resist Updat* 2016;26:43-57.

Bradshaw, J.P. Cationic antimicrobial peptides - Issues for potential clinical use. *Biodrugs* 2003;17(4):233-240.

Chen, W., *et al.* iACP: a sequence-based tool for identifying anticancer peptides. *Oncotarget* 2016;7(13):16895-16909.

Chou, K. and Zhang, C. Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology* 1995;30(4):275-349.

Chung, P.Y. and Khanum, R. Antimicrobial peptides as potential anti-biofilm agents against multidrug-resistant bacteria. *J Microbiol Immunol Infect* 2017;50(4):405-410.

Devadoss, F.R. and Paul Raj, V. Analysis and visual summarization of molecular dynamics simulation. *J Cheminform* 2014;6(Suppl 1):O16.

Fan, L., *et al.* DRAMP: a comprehensive data repository of antimicrobial peptides. *Sci Rep* 2016;6:24482.

Flocco, M.M. and Mowbray, S.L. C alpha-based torsion angles: a simple tool to analyze protein conformational changes. *Protein Sci* 1995;4(10):2118-2122.

Fox, J.L. Antimicrobial peptides stage a comeback. *Nat Biotechnol* 2013;31(5):379-382.

Ghalfi, H., *et al.* Production of three anti-listerial peptides by Lactobacillus curvatus in MRS broth. *Food Research International* 2010;43(1):33-39.

Gupta, S., *et al.* Prediction of Biofilm Inhibiting Peptides: An In silico Approach. *Front Microbiol* 2016;7:949.

Hansen, M.P., *et al.* Antibiotic Resistance: What are the Opportunities for Primary Care in Alleviating the Crisis? *Front Public Health* 2015;3:35.

Holton, T.A., *et al.* CPPpred: prediction of cell penetrating peptides. *Bioinformatics* 2013;29(23):3094-3096.

Jorgensen, W.L., *et al.* Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 1983;79(2):926.

Karasuyama, M. and Takeuchi, I. Multiple incremental decremental learning of support vector machines. *IEEE Trans Neural Netw* 2010;21(7):1048-1059.

Kosikowska, P. and Lesner, A. Antimicrobial peptides (AMPs) as drug candidates: a patent review (2003-2015). *Expert Opin Ther Pat* 2016;26(6):689-702.

Lüthy, R., Bowie, J. and Eisenberg, D. Assessment of protein models with three-dimensional profiles. *Nature* 1992;356(6364):83-85.

Lee, E.Y., *et al.* What can machine learning do for antimicrobial peptides, and what can antimicrobial peptides do for machine learning? *Interface Focus* 2017;7(6):20160153.

Li, W. and Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;22(13):1658-1659.

Lin, W. and Xu, D. Imbalanced multi-label learning for identifying antimicrobial peptides and their functional types. *Bioinformatics* 2016;32(24):3745-3752.

Lin, Y. and Jeon, Y. Random Forests and Adaptive Nearest Neighbors. *Journal of the American Statistical Association* 2006;101(474):578-590.

Maier, J.A., *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* 2015;11(8):3696-3713.

Mangoni, M.L. and Bhunia, A. Editorial: Antimicrobial Peptides in Medicinal Chemistry: Advances and Applications. *Curr Top Med Chem* 2016;16(1):2-3.

Mondal, N.R. and Jagdale, D.M. Antimicrobial Peptides A Review on its Types, Mechanism of Action, Synthesis and Therapeutic Applications. *International Journal For Pharmaceutical Research Scholars* 2016;5(2):72-81.

Nakashima, H., Nishikawa, K. and Ooi, T. The folding type of a protein is relevant to the amino acid composition. *J Biochem* 1986;99(1):153-162.

Pillardy, J., *et al.* Recent improvements in prediction of protein structure by global optimization of a potential energy function. *Proc Natl Acad Sci U S A* 2001;98(5):2329-2333.

Piotto, S.P., *et al.* YADAMP: yet another database of antimicrobial peptides. *Int J Antimicrob Agents* 2012;39(4):346-351.

Pirtskhalava, M., *et al.* DBAASP v.2: an enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides. *Nucleic Acids Res* 2016;44(D1):D1104-1112.

Rodriguez, R., *et al.* Homology modeling, model and software evaluation: three related resources. *Bioinformatics* 1998;14(6):523-528.

Sierra, J.M., *et al.* An overview of antimicrobial peptides and the latest advances in their development. *Expert Opin Biol Ther* 2017;17(6):663-676.

Silva, J.P., Appelberg, R. and Gama, F.M. Antimicrobial peptides as novel anti-tuberculosis therapeutics. *Biotechnol Adv* 2016;34(5):924-940.

Song, J., *et al.* Curcumin suppresses Streptococcus mutans adherence to human tooth surfaces and extracellular matrix proteins. *Eur J Clin Microbiol Infect Dis* 2012;31(7):1347-1352.

UniProt, C. UniProt: a hub for protein information. *Nucleic Acids Res* 2015;43(Database issue):D204-212.

Victor Paul Raj, F.R. and Exner, T.E. C(alpha) torsion angles as a flexible criterion to extract secrets from a molecular dynamics simulation. *J Mol Model* 2014;20(4):2196.

Wang, G., Li, X. and Wang, Z. APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res* 2016;44(D1):D1087-1093.

Zhang, L.J. and Gallo, R.L. Antimicrobial peptides. *Curr Biol* 2016;26(1):R14-19.