PTPAMP: Prediction Tool for Plant-derived Antimicrobial Peptides

Mohini Jaiswal
National Institute of Plant Genome Research (NIPGR)
Ajeet Singh
National Institute of Plant Genome Research (NIPGR)
Shailesh Kumar (shailesh@nipgr.ac.in)
National Institute of Plant Genome Research (NIPGR)

Research Article

Keywords: Bioactive peptides, antimicrobial peptide, classification, machine learning, prediction tool, plant-derived

Posted Date: May 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1678740/v1

Abstract

The emergence of antimicrobial peptides (AMPs) as a potential alternative to conventional antibiotics has led to the development of efficient computational methods for predicting AMPs. Amongst all organisms, the presence of multiple genes encoding AMPs in plants demands development of a plant-based prediction tool. To this end, we developed models based on multiple peptide features like amino acid composition, dipeptide composition, and physicochemical attributes for predicting plant-derived AMPs. The selected compositional models are integrated in a web server termed PTPAMP. The designed web server is capable to classify a query peptide sequence into four functional activities i.e. antimicrobial (AMP), antibacterial (ABP), antifungal (AFP), and antiviral (AVP). PTPAMP achieved an average area under the curve (AUC) of 0.95, 0.91, 0.85, and 0.88 for AMP, ABP, AFP, and AVP, respectively on benchmark datasets, which were ~ 6.75% higher than the state-of-the-art methods. Moreover, our analysis indicates the abundance of cysteine residue in plant-derived AMPs and distribution of other residues like G, S, K, and R which differs as per the peptide structural family. Finally, the developed web server is made user-friendly and presently available at http://www.nipgr.ac.in/PTPAMP/. We expect the substantial input of this predictor for high throughput identification of plant-derived AMPs followed by the additional insights into their functions.

1. Introduction

Plants produce a range of antimicrobial peptides which is supposed to be their pivotal molecular stratagem to ward off any microbial intrusions. Such peptides have their role defined at the entry level which designates them as an important weapon of host innate immunity. This emphasizes plant-derived peptides being a plausible efficient agent to deal with the intensive increase of microbial resistance towards antibiotics and drugs (Tam et al. 2015; Santos-Silva et al. 2020). Over the last few years, peptides are getting more attention as an alternative to antibiotics due to their small structure, cationic, and amphipathic nature (Kamysz et al. 2003; Chen et al. 2007; Barashkova and Rogozhin 2020). Multiple organisms, extending from prokaryotes to humans are involved in developing antimicrobial peptides and plants are one of them. These molecules are supposed to be evolutionarily conserved throughout the aforesaid range (Zhang and Gallo 2016). For plants, AMPs represent a mean of its defense response along with certain other toxic molecules that acts on pathogens by interacting with their phospholipids followed by membrane permeabilization leading to cell death (Nawrot et al. 2014). In other words, mainly fundamental features were attacked like constraining cell wall formation or protein synthesis as AMPs can bind to DNA, RNA, or protein (Ganz 2003; KA 2005; Hancock and Sahl 2006; JD and RE 2007). It has been reported earlier for AMPs are constitutively expressed in specific susceptible organs or are microbially induced at infection sites rather than circulating (Sels et al. 2008). Cysteine is the predominant amino acid content for the majority of plant-derived AMPs (Hammami et al. 2009), which is accountable for the presence of a myriad of disulfide bonds to provide the peptides a compact structure helpful in enduring the stress conditions (Nawrot et al. 2014). Also, this feature enables peptides to be more resistant to proteolytic and chemical degradation (Tam et al. 2015). Additionally, AMPs are equipped with multiple low-affinity targets for microbes to thwart the advancement in microbial resistance (Maróti et al. 2011).

Based on the beneficial characteristics of AMPs, numerous peptide repositories have been developed, such as APD3 (Wang et al. 2016), PhytAMP (Hammami et al. 2009), and PlantPepDB (Das et al. 2020), the latter is exclusively for plant-derived peptides with diverse activities. Still, several peptides are awaiting to be functionally and structurally characterized. In the post-genomic era, the avalanche of recently found protein and peptide sequences is responsible for the advent of numerous prediction tools based on machine learning (ML), such as AntiBP2 (Lata et al. 2010), AVPpred (Thakur et al. 2012), IAMP-2L (Xiao et al. 2013), amppred (Meher et al. 2017), antifp (Agrawal et al. 2018), Deep-AmPEP30 (Yan et al. 2020), and amPEPpy 1.0 (Lawrence et al. 2020) to configure the properties of peptide sequences. AntiBP2 integrates terminus-wise compositional features to develop models, specifically to predict the antibacterial peptides with an accuracy of 98.95%, but not able to conclude about sequences having a length of less than 15 amino acids. AVPpred and Antifp use the support vector machine (SVM) algorithm to develop models with the respective accuracy of 86% and 84.88% based on the compositional and physicochemical features, respectively for antiviral and antifungal peptides. Besides having high accuracy these models are not able to classify plant-derived peptides appropriately as of not received training for plant datasets. However, IAMP-2L, amppred, Deep-AmPEP30, and amPEP work for antimicrobial peptides, and to the best of our understanding, amPEP is the only predictor based on large and diverse AMP and Non-AMP datasets. Based on selected physicochemical features, amPEP is a specifically designed AMP classification method that uses a model developed by a random forest (RF) algorithm. It achieved an accuracy of 96% highest among all prediction tools developed so far. It classifies a query sequence as AMP or Non-AMP and assigns a score that lies between 0 to1. The prediction result seems to be quite insufficient to characterize an unknown sequence which justifies the concept of developing a web server armed with models build upon plant-derived AMP datasets with modules to elaborate the properties of a query sequence. In comparison to animals, plants are having more genes for antimicrobial peptides with an ability to generate hypervariable sequences that signifies plants being a formidable repertoire of antimicrobial peptides for pathogenic microbes. They are equally capable to evolve with new specifications (Maróti et al. 2011). Plant-derived AMPs are classified into certain families based on their cysteine content (Farrokhi et al. 2008). Following the earlier study, the probability of getting hundreds of distinct AMPs in certain plant species are quite high (Nawrot et al. 2014). Addressing these properties may lead to more precise predictors aimed at plant-derived peptides.

Henceforth, we put forward a machine learning-based predictor, named PTPAMP, for the classification of given peptide sequences into antimicrobial, antibacterial, antifungal, or antiviral activity. To develop a prediction model for each activity, we encoded sequences with amino acid composition (AAC), dipeptide composition (DPC), physicochemical features, and composition-transition-distribution (CTD). Subsequently, models were developed by using SVM for each encoding and the optimal feature set was identified after observing their results on cross-validation and independent test datasets. The comparative performance of PTPAMP with state-of-the-art methods showed that our proposed models outperformed the existing predictors when checking their results on six benchmark datasets. To the best of our knowledge, PTPAMP is the first plant-based approach for prediction of AMPs. The presented web server is freely accessible at the http://www.nipgr.ac.in/PTPAMP/, which provides a platform for peptide categorization among four defined activities along with a module to generate their mutated sequences in the hope of getting better bioactive peptides. Consequently, we anticipate that our designed framework will assist experimentalists in the discovery of novel plant-derived AMPs.

2. Materials And Methods

2.1 Development of PTPAMP

The notion to develop PTPAMP entails four leading stages, consisting of (i) collection of reliable datasets for training and validating the model; (ii) representation of peptide sequences in a manner that can truly reflect their intrinsic properties; (iii) development of a classifier to supervise the prediction; (iv) assessment of the classifier with relevant cross-validation tests; (v) execution of a web server by which user can simply get their intended result without prior understanding of the classifier algorithm used. A detailed overview of the aforementioned steps has been discussed in further sections.

2.2 Data Assembly

We picked up 2383 antimicrobial, 324 antibacterial, 530 antifungal, and 128 antiviral peptides from PlantPepDB as positive datasets, which shows a range of other activities also like anticancer, hemolytic, cytotoxic, etc. owing to their comprehensive range of actions. Few of them have evidence at transcript or protein level, quite a few are inferred from homology, and the rest of them are predicted, in compliance with the PlantPepDB info. After excluding redundant sequences and peptides with non-natural amino acids (B, J, O, U, Z, X), we finally obtained 1815 antimicrobial, 297 antibacterial, 477 antifungal, and 96 antiviral peptides. To build distinct models for comparative purposes, three sets were drawn up based on the discrepancy between negative datasets. While collating positive and negative data for each set, we maintained a specific sequence length range as 5-255 amino acids (aa), and 10-121aa for antimicrobial, antibacterial, antifungal, and antiviral activities, respectively. For the first set, PlantPepDB was noted as the source. While collecting negative data for each of the mentioned activities, certain peptides like antimicrobial, antibacterial, antifungal, antiviral, anticancer, cytotoxic, insecticidal, hemolytic were not included to circumvent the analogy between positive and negative data, as a majority of reported AMPs exhibit more than one activity and can target a wider range of microbes involving fungi, viruses, bacteria. In furtherance of the second set, we used 1815 sequences generated from UniProt (as employed in the ampep.py method) as negative data of antimicrobial activity after covering the same sequence length space as AMP. For antibacterial activity, 285 sequences of length range 5-94aa from non-experimental negative records of AntiBP2 were used as negative data along with 12 sequences (of length range 95-255aa) of MitPred (a source of non-secretory proteins) (Kumar et al. 2006). We extracted 238 and 203 sequences from Antifp_DS1 and Antifp_DS2 (Agrawal et al. 2018), respectively to use them as negative data for antifungal activity. To keep the sequence count and length balanced, we used 36 UniProt sequences (of length range 108-255aa) as well from the ampep.py (Lawrence et al. 2020) method. For antiviral activity, 89 non-experimental negative peptides (as used in the previous antiviral peptide prediction method (Thakur et al. 2012)) having a length range of 10-40aa were used in addition to the rest sequences taken from negative data of AntiBP2 (Lata et al. 2010) and MitPred (Kumar et al. 2006). Negative sequences for the previously discussed two sets were mutually different with no reported recurring sequences. Conducive to the third set, after shuffling negative data of the two sets stated earlier, we randomly picked an equal number of peptides from both sets and merged them by keeping the sequence count balanced as positive data for each functional activity. A statistical representation of datasets and overlapping distribution of positive data are delineated in Supplementary Table S1 and Supplementary Fig. S1, respectively. All collected datasets were divided into 75% and 25% to train and test the models, respectively.

Separately, we constructed six different benchmark datasets to make a functional comparison between our plant-based prediction server (PTPAMP) and formerly developed prediction tools. All the six datasets used in benchmarking are summarized in Table 1. Due to the efficient performance of iAMP2L (Xiao et al. 2013), 599 non-AMP sequences of their benchmark dataset were employed as negative data for our first set.

Table 1
Statistical representation of benchmark datasets. The value inside bracket {} is the number of sequences of that category

Datasets	Positive	Negative
Dataset 1	AMP, ABP, AFP, AVP {599}	iAMP2L {599}
Dataset 2	AMP {453}	Non-AMP {453}
Dataset 3	ABP {74}	Non-ABP {74}
Dataset 4	AFP {119}	Non-AFP {119}
Dataset 5	AVP {24}	Non-AVP {24}
Dataset 6	AMP, ABP, AFP, AVP {599}	Non-AMP, Non-ABP, Non-AFP, Non-AVP {599}

2.3 Peptide features

Numerous features exist, albeit, to define antimicrobial peptides, and a lot of them have been used earlier for pre-existing antimicrobial peptide prediction tools. However, we had chosen a few of them in reference to the elective studies (Lata et al. 2010; Thakur et al. 2012; Agrawal et al. 2018; Lawrence et al. 2020) for characterizing plant-derived AMPs. Those features are amino acid composition, dipeptide composition, physicochemical properties, and CTD descriptors. Additionally, we looked for motifs of antimicrobial, antibacterial, antifungal, and antiviral peptides separately. For identification of those motifs, we used MERCI program (Vens et al. 2011). A complete list of motifs for each activity has been provided in Supplementary sheet 1.

2.3.1 Amino acid and Dipeptide composition

Having the objective to develop models by employing machine learning algorithms, input features should be of fixed length. Here, we have peptides of varying lengths, thus, we computed their amino acid and dipeptide composition profiles with the values confined in a vector of 20 and 400 dimensions, respectively. The amino acid composition depicts the portion of each amino acid in the respective peptide sequences while dipeptide composition tells about the global and local arrangement of amino acid residues in a sequence (Manavalan et al. 2017). We used a Perl script from the GPSR package (https://webs.iiitd.edu.in/raghava/gpsr/) for calculating amino acid and dipeptide composition profiles.

2.3.2 Physicochemical properties

Besides compositional features, the physicochemical attributes are equally important to describe a peptidic feature. We computed aaDescriptors in addition to charge, mass, isoelectric point, hydrophobicity, aliphatic index, PPI (potential protein interaction) index, and hydrophobic moment by using an R package named 'Peptides' (Osorio et al. 2015). The overall calculated physicochemical properties are encapsulated in a vector of 74 dimensions. The complete list of physicochemical attributes used in this study is provided in Supplementary Table S2.

2.3.3 CTD Descriptors

According to Global Protein Sequence Descriptors (Dubchak et al. 1995), the amino acids are categorized into 3 groups as per 7 physicochemical attributes and three kinds of descriptor sets can be calculated for a particular attribute. We computed all three kinds i.e. composition, transition, and distribution descriptor set individually as well as collectively to look into their importance in classifying positive and negative data. The composition descriptor is encoded in 21 feature spaces to describe the global percentage of 3 groups for each physicochemical attribute. Dimension of transition and distribution descriptor sets are 21 and 105, respectively. In this study, to calculate these three descriptors, the ProtR package of R (Xiao et al. 2015) is used.

In addition to the aforesaid feature spaces, we encoded 641 features for each sequence after considering AAC, DPC, physicochemical, and CTD descriptor altogether. This was done to examine whether this hybrid feature is performing better in comparison to the respective feature models.

2.4 Algorithm Implementation

The ability of models to differentiate between positive and negative data not only depends on the feature representation processes but also relies on the employed ML algorithms. We developed models for the categorization of positive and negative peptides for a list of activities like antimicrobial, antibacterial, antifungal, and antiviral by using an ML algorithm named SVM. We used the freely downloadable SVM^{light} Version 6.02 package (Joachims 1998) to develop prediction models. The basic idea behind using SVM is to map our peptide data into higher dimensional space to get them linearly separated. Many studies recorded the impressive use of SVM for small sample sizes due to their excellent learning and finest generalization abilities (Hongjaisee et al. 2019). We used RBF kernel parameters for procuring the best results. Further, to make a comparative study, we employed other algorithms such as Naïve Bayes (NB), K-Nearest Neighbor (KNN), and RF which are commonly used in supervised classification problems. The WEKA package (Hall et al. 2009) is used for the implementation of these three algorithms.

2.5 Performance assessment and Validation

The five-fold cross-validation technique was adopted for evaluating the execution of models. Under this technique, datasets get randomly distributed into five equally sized sets where each set is used once for testing and the rest four sets for training. To appraise the performance of models, both threshold-dependent and threshold-independent parameters are required. Threshold-dependent parameters include sensitivity (Sn), specificity (Sp), accuracy (Ac), and Matthew's correlation coefficient (MCC). ROC (Receiver Operating Characteristic) comes under threshold-independent parameters. We calculated these parameters in all developed models for assessing their performance. In literature, certain equations are often been used to contemplate the prediction quality:

Sensitivity (S_n) = [TP/TP + FN] × 100

Specificity (S_p) = [TN/TN + FP] × 100

Accuracy (A_c) = [TP + TN/ (TP + FP + TN + FN)] × 100

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

Where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively. MCC value ranges between − 1 to + 1, signifying the correlation between true and predicted queries. Larger MCC and ROC values denote better prediction.

2.6 Peptide structural analysis

This study is focused on to classify plant-derived antimicrobial peptides, hence, their better understanding is also simultaneously important. According to the earlier studies (Li et al. 2021), their composition is quite complex and there is a significant role of their structure in the functional classification (Hammami et al. 2009). So, to understand the structural aspects of plant-derived antimicrobial peptides, 10 sequences that were validated at the protein level were handpicked from the complete positive datasets used in this study. For secondary structure prediction, we used PSIPRED available at http://bioinf.cs.ucl.ac.uk/psipred/ which embodies two feed-forward neural networks for analyzing the results acquired from PSI-BLAST (Jones 1999; Buchan et al. 2013). Additionally, homology modeling to get a three-dimensional (3D) structure for the selected sequences was executed with the help of SWISS-MODEL available on the ExPASy website https://swissmodel.expasy.org/ (Waterhouse et al. 2018). In homology modeling, evolutionarily related protein structures were obtained for query sequences after searching them against the SWISS-MODEL template library (SMTL) (Biasini et al. 2014). Two database search methods BLAST (Altschul et al. 1997; Camacho et al. 2009) and HHblits (Remmert et al. 2011) are used by SWISS-MODEL to perform this task. BLAST is useful and accurate against closely related templates while HHblits increases sensitivity in the case of remote homology. The searched templates are graded according to the entire quality of the resulting model assessed by their GMQE (Global Model Quality Estimate) (Biasini et al. 2014). The sequences were selected by keeping certain points in mind which include (i) peptide length should be less than 180aa; (ii) sequence identity to the template structure should be > 95%; (iii) GMQE should be ≥ 0.9 for the structure model.

2.7 Development of PTPAMP web server

The web server was developed by using Apache HTTP server (version 2.4.6) integrated with PHP (version 7.3.3) on a server machine with Centos 7 Linux as the operating system. CSS and HTML were used to make the template responsive. The best predictive model as described in the result section was deployed in the web server. Firstly, the web server accepts peptide FASTA sequence as an input. Secondly, after submission by invoking the submit button, the query sequences are subjected for the compositional feature calculation and subsequently fed to the predictive models described previously. The prediction for each activity along with the SVM score for query sequences are displayed in the prediction output box. The displayed results are also available as a CSV file upon invoking the download button found directly above the result table.

3. Results And Discussion

In this study, a web-based method is proposed to classify the plant-derived peptide among antimicrobial, antibacterial, antifungal, or antiviral activity. To select an appropriate classifying feature, a comparative performance was conducted between eight types of feature encodings which include AAC (20D), DPC (400D), physicochemical (74D), composition (21D), transition (21D), distribution (105D), CTD (147D), and hybrid (641D). Further, another thing was to determine the best performing ML algorithm from the four commonly used algorithms for supervised classification problems. At last, to help the users in classifying their query sequence with ease, PTPAMP is executed smoothly as a free plant-derived antimicrobial peptide prediction server. The complete workflow of this study starting from the dataset collection to the development of prediction server is summarized in Fig. 1.

3.1 Performance comparison of used feature spaces

To evaluate the efficacy of each feature in accurately predicting AMPs, the five-fold cross-validation, and independent tests were carried out for each activity by using SVM as a classification algorithm. The obtained AUCs for the best performing feature model on cross-validation datasets are 0.98, 0.97, 0.93, and 0.97 for antimicrobial DPC model, antibacterial AAC model, antifungal DPC model, and antiviral hybrid model, respectively. While AUCs of the respective feature models on independent test datasets are recorded as 0.98, 0.99, 0.90, and 0.86 for antimicrobial, antibacterial, antifungal, and antiviral activities, respectively. The relative performance of eight feature encodings across cross-validation and independent test datasets is illustrated by Figs. 2 and 3, respectively. After comparing the results of each feature model, it can be seen that the performance of the DPC model developed after optimizing the parameters on cross-validation dataset is relatable with their performance on the respective independent test datasets for each functional activity. Therefore, to assess the comparative execution of four ML algorithms, DPC is used. To give a detailed picture, results for cross-validation and independent test sets are assembled in Supplementary Tables S3-S6.

Apart from AAC and DPC, the rest six feature encodings also achieved an equitable performance with AUCs ranging from 0.80–0.97, 0.69–0.95, 0.72–0.92, and 0.74–0.97 for antimicrobial, antibacterial, antifungal, and antiviral, respectively. The respective AUC range indicates the utility of the remaining six features in peptide prediction due to their complementary feature depiction from another viewpoint.

3.2 Choosing the finest classification algorithm

In general, there is no professed best algorithm for any classification problems, hence, a rule of thumb is followed in an attempt to find the most optimized algorithm for our data classification. As mentioned before, DPC is the most relevant feature to differentiate between positive and negative data of each activity. Thus, to corroborate the use of DPC as a classifying feature, four different ML models are trained with DPC feature and their execution is examined on five-fold cross-validation and independent test datasets. The detailed performance of all models for each activity is listed in Table 2. As stated in Table 2, the model trained with set 2 datasets of antimicrobial, antibacterial, and antiviral achieved AUC of (0.98, 0.90, 0.97, and 0.87), (0.96, 0.67, 0.95, and 0.92), and (0.95, 0.88, 0.94, and 0.94) for SVM, KNN, RF, and NB algorithms, respectively. While, for antifungal activity, the DPC model developed with set 1 dataset achieved better performance on a five-fold cross-validation dataset with AUC of 0.94, 0.73, 0.92, and 0.82 for SVM, KNN, RF, and NB algorithms, respectively. Therefore, independent test datasets of set 2 and set 1 were selected of antimicrobial, antibacterial, antiviral, and antifungal activity, respectively for benchmarking purposes as briefed in Table 1. To manifest the applicability and effectiveness of the proposed model, we compared its result with the other three aforementioned algorithms. As evident from Table 2, the overall accuracy and MCC of SVM models are higher than those resulting from KNN, RF, and NB on five-fold cross-validation and independent test datasets. It can be affirmed that SVM-derived models are more powerful and remarkably efficient for the classification problems proposed in this study.

Table 2
Performance comparison of SVM derived DPC models with three other ML algorithms (KNN, RF, and NB) derived DPC models across cross-validation (5-fold) and independent test datasets

Activity	Sets	Algorithms	Cross validation dataset					Independent test dataset				
			Sn	Sp	Ac	MCC	AUC	Sn	Sp	Ac	MCC	AUC
Antimicrobial	Set 1	SVM	91.34	89.32	90.38	0.81	0.93	84.77	91.36	87.88	0.76	0.93
		KNN	86.40	76.60	81.80	0.63	0.82	82.60	67.40	75.40	0.50	0.76
		RF	88.80	91.90	90.20	0.80	0.95	85.90	89.40	87.50	0.75	0.93
		NB	85.00	43.00	65.20	0.31	0.65	87.00	49.40	69.20	0.39	0.70
	Set 2	SVM	88.33	96.99	92.66	0.86	0.98	87.42	98.01	92.72	0.86	0.98
		KNN	90.90	89.60	90.30	0.80	0.90	88.10	87.40	87.70	0.75	0.87
		RF	91.20	94.40	92.80	0.85	0.97	90.30	93.60	91.90	0.83	0.97
		NB	88.10	75.30	81.70	0.64	0.87	85.20	83.20	84.20	0.68	0.85
	Set 3	SVM	85.68	93.02	89.35	0.79	0.95	84.11	94.70	89.40	0.79	0.96
		KNN	85.20	82.20	83.70	0.67	0.83	82.80	79.90	81.30	0.62	0.81
		RF	86.70	91.60	89.10	0.78	0.94	85.00	92.70	88.90	0.77	0.94
		NB	84.70	42.60	63.60	0.30	0.68	77.50	66.90	72.20	0.44	0.79
Antibacterial	Set 1	SVM	82.51	89.69	86.10	0.72	0.91	98.65	71.62	85.14	0.73	0.90
		KNN	84.80	74.00	79.40	0.59	0.80	73.00	66.20	69.60	0.39	0.69
		RF	79.40	87.90	83.60	0.67	0.91	85.10	70.30	77.70	0.56	0.82
		NB	70.90	79.40	75.10	0.50	0.80	71.60	82.40	77.00	0.54	0.80
	Set 2	SVM	93.27	92.38	92.83	0.86	0.96	97.30	97.30	97.30	0.95	0.99
		KNN	93.70	39.00	66.40	0.39	0.67	77.00	82.40	79.70	0.59	0.81
		RF	88.80	93.70	91.30	0.82	0.95	87.80	93.20	90.50	0.81	0.95
		NB	89.20	88.80	89.00	0.78	0.92	90.50	86.50	88.50	0.77	0.94
	Set 3	SVM	80.72	89.24	84.98	0.70	0.89	94.59	70.27	82.43	0.67	0.89
		KNN	92.40	48.40	70.40	0.45	0.69	73.00	63.50	68.20	0.36	0.66
		RF	85.20	83.90	84.50	0.69	0.89	81.10	77.00	79.10	0.58	0.83
		NB	73.50	77.10	75.30	0.50	0.80	56.80	82.40	69.60	0.40	0.78
Antifungal	Set 1	SVM	89.94	81.56	85.75	0.72	0.94	78.99	92.44	85.71	0.72	0.90
		KNN	73.70	74.90	74.30	0.48	0.73	75.60	60.50	68.10	0.36	0.69
		RF	84.10	85.50	84.80	0.69	0.92	77.30	80.70	79.00	0.58	0.88
		NB	77.70	78.20	77.90	0.55	0.82	68.10	83.20	75.60	0.51	0.80
	Set 2	SVM	80.73	90.5	85.61	0.72	0.92	68.07	95.8	81.93	0.66	0.89
		KNN	83.80	70.10	77.00	0.54	0.77	81.50	66.40	73.90	0.48	0.75
		RF	81.80	86.00	83.90	0.67	0.91	84.90	83.20	84.00	0.68	0.89
		NB	77.40	74.30	75.80	0.51	0.82	73.90	75.60	74.80	0.49	0.78
	Set 3	SVM	81.56	83.24	82.4	0.65	0.91	79.83	73.11	76.47	0.53	0.84
		KNN	79.30	71.80	75.60	0.51	0.75	82.40	50.40	66.40	0.34	0.66
		RF	79.60	81.80	80.70	0.61	0.89	75.60	70.60	73.10	0.46	0.80
		NB	77.90	71.20	74.60	0.49	0.77	62.20	69.70	66.00	0.32	0.69
Antiviral	Set 1	SVM	86.11	81.94	84.03	0.68	0.80	95.83	70.83	83.33	0.69	0.82
		KNN	86.10	75.00	80.60	0.61	0.79	87.50	79.20	83.30	0.66	0.83
		RF	80.60	81.90	81.30	0.62	0.88	83.30	75.00	79.20	0.58	0.82
		NB	81.90	79.20	80.60	0.61	0.84	75.00	70.80	72.90	0.45	0.78

Set 2	SVM	90.28	98.61	94.44	0.89	0.95	95.83	100.00	97.92	0.96	0.99
	KNN	94.40	79.20	86.80	0.74	0.88	95.80	83.30	89.60	0.79	0.87
	RF	90.30	97.20	93.80	0.87	0.94	87.50	100.00	93.80	0.88	0.94
	NB	80.60	97.20	88.90	0.78	0.94	95.80	91.70	93.80	0.87	0.95
Set 3	SVM	90.28	77.78	84.03	0.69	0.87	95.83	75.00	85.42	0.72	0.92
	KNN	86.10	41.70	63.90	0.31	0.62	75.00	62.50	68.80	0.37	0.72
	RF	83.30	80.60	81.90	0.63	0.85	87.50	75.00	81.30	0.63	0.78
	NB	70.80	79.20	75.00	0.50	0.83	70.80	75.00	72.90	0.45	0.81

3.3 Comparison with existing antimicrobial peptide predictors

Although, there is no such prediction method available so far, fully accountable to plant-derived antimicrobial peptides. Still, in terms of predicting plant-derived AMPs, we made a comparison of PTPAMP with existing approaches such as amPEPpy1.0 (Lawrence et al. 2020), Antifp (Agrawal et al. 2018), Meta-iAVP (N et al. 2019), and iAMPpred (Meher et al. 2017). The AUC obtained with datasets outlined in Table 1 was examined for all eight feature models across each activity. The antimicrobial model based on DPC is comparatively efficient in distinguishing between positive and negative data of benchmark datasets as reflected by Fig. 4. On the contrary, AAC based model achieved higher AUCs for another three activities named antibacterial, antifungal, and antiviral. Thereby, the best-performed model of all activities for six benchmark datasets was selected for their implementation into the web server. The selected model was also compared with the state-of-art predictors and the results are displayed in Fig. 5. After observing the comparative performance results listed in Supplementary Table S7, the relatively effective execution of our proposed model for each activity in classifying plant-derived antimicrobial peptides has been proven.

Even though negative data of this study was taken from the previously developed tools, the performance of respective tools is not satisfactory when plant-derived peptides were given as queries. The improved performance of our proposed models can be justified by the following aspects: (i) the existing predictors were developed on datasets of multiple organisms causing models to learn generalized features. In contrast, we developed models especially intended for plant-derived datasets; (ii) the parameters used for our proposed model were optimized on five-fold cross-validation datasets indicating these parameters to be more accurate and stable; (iii) amongst distinct features utilized for constructing models in this study, we found significant differences among positive and negative data during compositional analysis i.e. AAC and DPC. Many studies reported the successful implementation of these two features in predicting peptides and proteins (Bhasin and Raghava 2004; Raghava and Han 2005; Garg and Raghava 2008; Kumar et al. 2008, 2015; Lata et al. 2010; Thakur et al. 2012; Gautam et al. 2013; Gupta et al. 2016). Analysis of the literature revealed the successful applicability of binary features in classifying antifungal peptides (Agrawal et al. 2018), however, a flaw of this feature is the requirement of equal length peptides to be used for the development of prediction models. As described in Section 3.2, our dataset sequence count is not much higher so the binary feature was not included in this study to prevent the reduction of our dataset size. Additionally, the prediction models were developed by keeping sequence count and length balanced thus beneficial in classifying peptides of varied lengths precisely.

3.4 Structure analysis of the selected sequences

The secondary structure predicted by PSIPRED is shown in Fig. 6. With the increase of alpha-helical structures in peptides, its hydrophobicity also gets increased in aqueous environments (Chen et al. 2007). Peptide hydrophobicity is believed to be an important factor in their antimicrobial activity. In lieu of higher or lower hydrophobicity of a peptide, an optimum hydrophobicity window is accountable for its potent antimicrobial function. There will be a decreased antimicrobial activity with the increase in hydrophobicity. The clarification for this trend is the self-association of peptides which acts as a hindrance for peptide entry into the bacterial cell (Chen et al. 2007). The 3D structure generated by SWISS-MODEL for the selected sequences is represented in Fig. 7. To assess the quality of structure, GMQE, QMEANDisCo global scores (Studer et al. 2020), and the percentage of residues favored by Ramachandran were calculated for each sequence as listed in Table 3. The total percentage of amino acid residues of each sequence found in Ramachandran favored and allowed region is > 90%, making the constructed structures reasonable and convincing. The ideal case would be if > 98% of amino acid residues are present in the favored region (Chen et al. 2010). In this study, the respective ideology is fulfilled by the structures of viscotoxin A3, HsAFP1, beta hordothionin, and ginkbilobin-2 with 100%, 98.08%, 100%, and 99.06% residues, respectively. Moreover, the structures presented in Fig. 7 are determined by X-ray diffraction with high resolution which is a preferable method in homology modeling.

Table 3
Description of structure quality for selected sequences

PPepDB ID	Peptide name	Sequence
PPepDB_1622	Hellethionin-D	KSCCRNTLARNCYNACRFTGGSQPTCGILCDCIHVTTTTCPSSHPS
PPepDB_1543	Snakin-1	GSNFCDSKCKLRCSKAGLADRCLKYCGICCEECKCVPSGTYGNKHECPCYRDKKNSKGKSKCP
PPepDB_1491	Griffithsin	SLTHRKFGGSGGSPFSGLSSIAVRSGSYLDAIIIDGVHHGGSGGNLSPTFTFGSGEYISNMTIRSGDYIDNISFETNMGRRFGPYGGSG
PPepDB_2071	Ep-AMP1	CVLIGQRCDNDRGPRCCSGQGNCVPLPFLGGVCAV
PPepDB_3948	Viscotoxin-A3	KSCCPNTTGRNIYNACRLTGAPRPTCAKLSGCKIISGSTCPSDYPK
PPepDB_2152	PsDef1	RMCKTPSGKFKGYCVNNTNCKNVCRTEGFPTGSCDFHVAGRKCYCYKPCP
PPepDB_621	Ginkbilobin-2	ANTAFVSSACNTQKIPSGSPFNRNLRAMLADLRQNTAFSGYDYKTSRAGSGGAPTAYGRATCKQSISQSDCTACLSNLVNRIFSICNN
PPepDB_2077	HsAFP1	DGVKLCDVPSGTWSGHCGSSSKCSQQCKDREHFAYGGACHYQFPSVKCFCKRQC
PPepDB_3974	Beta-hordothionin	KSCCRSTLGRNCYNLCRVRGAQKLCANACRCKLTSGLKCPSSFPK
PPepDB_2206	NaD1	RECKTESNTFPGICITKPPCRKACISEKFTDGHCSKILRRCLCTKPC

The structures recorded by SWISS-MODEL for beta-hordothionin and NaD1 have 4 and 2 salt-bridges in them, respectively. The formation of salt bridges in a peptide confers proteolytic stability to them along with assisting in peptide folding but is not vital for microbicidal action (Andersson et al. 2012).

3.5 Structural insights based on amino acid residue preferences

Most of the plant-derived AMPs belong to β or αβ family such as cyclotides and defensins. In response to said structural family, the dominant amino acid in each group differs. For instance, β and αβ family is dominated by cysteine (C) as the hydrophobic amino acid prerequisite in peptide folding. The abundance of cysteine residue in plant peptides indicates the frequently occurring disulfide bonds which confer metabolic stability to them (Cole and Cole 2008). The presence of disulfide bonds also directs the majority of plant-derived peptides to be disulfide-bonded defensin-like molecules (Li et al. 2021). Other than C, arginine (R) and lysine (K) is equally distributed among αβ family. While, glycine (G) and serine (S) are the two favored amino acids in entire families (Mishra and Wang 2012). Moreover, the presence of short-chain residues like glycine and serine in antimicrobial peptides is accountable for their disordered region. These regions provide structural flexibility to such peptides (Tavares et al. 2012). Plant cyclotides are mainly rich in C, G, T/S, and K which regulate a widespread β-sheet containing scaffold in them (Wang 2010). Concisely, plant-derived peptides are mostly rich in cysteine and glycine. The respective presence of C, R, K, G, and S in our peptide datasets can be visualized by Fig. 8. As per Wang et al. (Wang et al. 2009) amino acid composition of natural AMPs is directly related to their 3D structures.

3.6 PTPAMP web server

In an effort to maximize the utility of plant-derived AMPs, we established a web server, PTPAMP, which is targeted at reaching a wide plant research community and enables them to predict plant-derived AMPs. To validate our proposed work, the datasets used in this study can be downloaded from the server. Hereafter, we are providing summarized steps for using this server. Firstly, the peptide screening module, in which user can enter their query peptide sequences (in FASTA format) in the provided text-box or can upload a peptide FASTA file from their system. This module will help the users to categorize their queries in four functional activities (AMP, ABP, AFP, or AVP) based on the score computed by integrated feature models. Results are shown in tabular form. Along with prediction scores, there are four columns representing charge, molecular weight, isoelectric point, and hydrophobicity values for each sequence. Additionally, there is a link to see peptide properties where some relevant graphs were plotted like AAC (regarding physicochemical properties), hydrophobicity, charge, and hydrophobic moment plot. In next column BLAST (SF et al. 1990) option is given to search the similar sequences from PlantPepDB. Secondly, the peptide designing module in which user can enter their query peptide sequence and obtain the probable peptide mutants with single amino acid substitution. The said module will lead to generate peptide mutants. Based on the predicted scores and peptide property gained after using pepcalc (https://pepcalc.com/) placed on our server, users can ultimately select the best peptide mutant. Thirdly, there is a protein scan module that helps to generate peptides of desired length and overlapping residues provided, and each peptide will be displayed with its prediction scores and motif search results. Users can paste the query protein sequence in the given text box or may upload a FASTA file. This module is helpful in getting possible regions of protein sequence to be classified as AMPs. In all the three modules there is an option of selecting SVM threshold and users are suggested to keep its value high if they want results of high specificity or may keep it low if needed results with high sensitivity.

4. Conclusions

To delve into the properties of plant-derived peptides, we developed a predictor named PTPAMP to trace a peptide activity regarding antimicrobial, antibacterial, antifungal, and antiviral ones. In this predictor, an optimal feature set was selected after individually observing the performance of eight feature encodings. The compositional feature set i.e. AAC and DPC are supposed to integrate multiple aspects of sequence information thereby showed consistent performance across cross-validation and independent test datasets. To increase the robustness of these models we trained them with multiple ML algorithms and after comparing found SVM as the best classifier. Furthermore, to make sure that our proposed models are sufficiently effective in classifying plant-

derived AMPs, we carried out a comparative analysis with existing AMP predictors. Most of the existing AMP predictors were limited to classify a peptide into a single activity at a time, while by using PTPAMP users can classify their query sequences into four different activities on a single platform. As there is a shred of evidence regarding the broad-spectrum activity of antimicrobial peptides (Ageitos et al. 2017; Campos et al. 2018), PTPAMP can support the research going on towards understanding and sub-classifying the functional activity of plant-derived peptides. To maximize the convenience of users, the models were deployed as a web server which is made freely available and is user-friendly. Besides functional activity prediction, our proposed framework could be used to generate peptide variants either after producing mutant analogs of the given peptide sequence or after scanning given protein sequences and creating overlapping peptides of a specific window size provided by the user. Additionally, with the increase of experimentally verified plant peptides and novel features, the prediction of plant-derived AMPs will also get simultaneously enhanced. The goal of PTPAMP is to be proved as a valuable tool for the identification of plant-derived AMPs in a high-throughput and cost-effective manner trailed by characterization as per their physicochemical properties.

Declarations

Acknowledgments

The authors acknowledge the Council of Scientific and Industrial Research (CSIR), India, for the Senior Research Fellowship. The authors are thankful to DBT (Department of Biotechnology)-eLibrary Consortium (DeLCON), India for providing access to e-resources. The authors are also thankful to the Distributed Information Sub-Centre (Sub-DIC) of the Department of Biotechnology (DBT) at NIPGR. SK acknowledge the BT/PR40146/BTIS/137/4/2020 project grant from the Department of Biotechnology (DBT), Government of India for this study.

Author Contributions

M.J. performed the data analysis and developed the webpage. A.S. helped in the development of the web-server. M.J. and S.K. wrote the manuscript. S.K. conceived the idea and coordinated the project. S.K. agrees to serve as the author responsible for contact and ensures communication.

Declaration of Competing Interest

The authors declare that they have no competing interests.

References

1. Ageitos JM, Sánchez-Pérez A, Calo-Mata P, Villa TG (2017) Antimicrobial peptides (AMPs): Ancient compounds that represent novel weapons in the fight against bacteria. Biochem. Pharmacol. 133:117–138
2. Agrawal P, Bhalla S, Chaudhary K, et al (2018) In Silico Approach for Prediction of Antifungal Peptides. Front Microbiol 9:323. https://doi.org/10.3389/fmicb.2018.00323
3. Altschul SF, Madden TL, Schäffer AA, et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/NAR/25.17.3389
4. Andersson HS, Figueredo SM, Haugaard-Kedström LM, et al (2012) The α-defensin salt-bridge induces backbone stability to facilitate folding and confer proteolytic resistance. Amino Acids 43:1471–1483. https://doi.org/10.1007/S00726-012-1220-3/FIGURES/7
5. Barashkova AS, Rogozhin EA (2020) Isolation of antimicrobial peptides from different plant sources: Does a general extraction method exist? Plant Methods 16:143
6. Bhasin M, Raghava GPS (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res 32:. https://doi.org/10.1093/nar/gkh350
7. Biasini M, Bienert S, Waterhouse A, et al (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 42:W252–W258. https://doi.org/10.1093/NAR/GKU340
8. Buchan DWA, Minneci F, Nugent TCO, et al (2013) Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Res 41:W349–W357. https://doi.org/10.1093/NAR/GKT381
9. Camacho C, Coulouris G, Avagyan V, et al (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421
10. Campos ML, De Souza CM, De Oliveira KBS, et al (2018) The role of antimicrobial peptides in plant immunity. J. Exp. Bot. 69:4997–5011
11. Chen VB, Arendall WB, Headd JJ, et al (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr Sect D Biol Crystallogr 66:12. https://doi.org/10.1107/S0907444909042073
12. Chen Y, Guarnieri MT, Vasil AI, et al (2007) Role of Peptide Hydrophobicity in the Mechanism of Action of-Helical Antimicrobial Peptides Downloaded from. Antimicrob Agents Chemother 51:1398–1406. https://doi.org/10.1128/AAC.00925-06
13. Cole AM, Cole AL (2008) REVIEW ARTICLE: Antimicrobial Polypeptides are Key Anti-HIV-1 Effector Molecules of Cervicovaginal Host Defense. Am J Reprod Immunol 59:27–34. https://doi.org/10.1111/J.1600-0897.2007.00561.X
14. Das D, Jaiswal M, Khan FN, et al (2020) PlantPepDB: A manually curated plant peptide database. Sci Rep 10:1–8. https://doi.org/10.1038/s41598-020-59165-2
15. Dubchak I, Muchnikt I, Holbrook SR, Kim S-H (1995) Prediction of protein folding class using global description of amino acid sequence. Biophysics (Oxf) 92:8700–8704

16. Farrokhi N, Whitelegge JP, Brusslan JA (2008) Plant peptides and peptidomics. Plant Biotechnol J 6:105–134. https://doi.org/10.1111/j.1467-7652.2007.00315.x
17. Ganz T (2003) Defensins: antimicrobial peptides of innate immunity. Nat Rev Immunol 2003 39 3:710–720. https://doi.org/10.1038/nri1180
18. Garg A, Raghava GPS (2008) A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. In Silico Biol 8:129–140
19. Gautam A, Chaudhary K, Kumar R, et al (2013) In silico approaches for designing highly effective cell penetrating peptides. J Transl Med 11:1–12. https://doi.org/10.1186/1479-5876-11-74/FIGURES/6
20. Gupta S, Sharma AK, Jaiswal SK, Sharma VK (2016) Prediction of biofilm inhibiting peptides: An In silico Approach. Front Microbiol 7:949. https://doi.org/10.3389/FMICB.2016.00949/BIBTEX
21. Hall M, Frank E, Holmes G, et al (2009) The WEKA data mining software. ACM SIGKDD Explor Newsl 11:10–18. https://doi.org/10.1145/1656274.1656278
22. Hammami R, Ben Hamida J, Vergoten G, Fliss I (2009) PhytAMP: a database dedicated to antimicrobial plant peptides. Nucleic Acids Res 37:D963-8. https://doi.org/10.1093/nar/gkn655
23. Hancock REW, Sahl H-G (2006) Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. Nat Biotechnol 2006 2412 24:1551–1557. https://doi.org/10.1038/nbt1267
24. Hongjaisee S, Nantasenamat C, Carraway TS, Shoombuatong W (2019) HIVCoR: A sequence-based tool for predicting HIV-1 CRF01_AE coreceptor usage. Comput Biol Chem 80:419–432. https://doi.org/10.1016/J.COMPBIOLCHEM.2019.05.006
25. JD H, RE H (2007) Alternative mechanisms of action of cationic antimicrobial peptides on bacteria. Expert Rev Anti Infect Ther 5:951–959. https://doi.org/10.1586/14787210.5.6.951
26. Joachims T (1998) Advances in Kernel Methods-Support Vector Learning
27. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202. https://doi.org/10.1006/JMBI.1999.3091
28. KA B (2005) Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria? Nat Rev Microbiol 3:238–250. https://doi.org/10.1038/NRMICRO1098
29. Kamysz W, Okrój M, Łukasiak J (2003) Novel properties of antimicrobial peptides. Acta Biochim. Pol. 50:461–469
30. Kumar M, Thakur V, Raghava GPS (2008) COPid: Composition based protein identification. In Silico Biol 8:121–128
31. Kumar M, Verma R, Raghava GPS (2006) Prediction of mitochondrial proteins using support vector machine and hidden Markov model. J Biol Chem 281:5357–5363. https://doi.org/10.1074/jbc.M511061200
32. Kumar R, Chaudhary K, Singh Chauhan J, et al (2015) An in silico platform for predicting, screening and designing of antihypertensive peptides. Sci Rep 5:12512. https://doi.org/10.1038/srep12512
33. Lata S, Mishra NK, Raghava GPS (2010) AntiBP2: Improved version of antibacterial peptide prediction. BMC Bioinformatics 11:S19. https://doi.org/10.1186/1471-2105-11-S1-S19
34. Lawrence TJ, Carper DL, Spangler MK, et al (2020) amPEPpy 1.0: a portable and accurate antimicrobial peptide prediction tool. Bioinformatics. https://doi.org/10.1093/bioinformatics/btaa917
35. Li J, Hu S, Jian W, et al (2021) Plant antimicrobial peptides: structures, functions, and applications. Bot Stud 2021 621 62:1–15. https://doi.org/10.1186/S40529-021-00312-X
36. Manavalan B, Basith S, Shin TH, et al (2017) MLACP: machine-learning-based prediction of anticancer peptides. Oncotarget 8:77121. https://doi.org/10.18632/ONCOTARGET.20365
37. Maróti G, Kereszt A, Va Kondorosi ´, Mergaert P (2011) Natural roles of antimicrobial peptides in microbes, plants and animals. https://doi.org/10.1016/j.resmic.2011.02.005
38. Meher PK, Kumar Sahu T, Saini V, et al (2017) Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC OPEN. Nat Publ Gr. https://doi.org/10.1038/srep42362
39. Mishra B, Wang G (2012) The importance of amino acid composition in natural amps: An evolutional, structural, and functional perspective. Front Immunol 3:221. https://doi.org/10.3389/FIMMU.2012.00221/BIBTEX
40. N S, C N, V P, W S (2019) Meta-iAVP: A Sequence-Based Meta-Predictor for Improving the Prediction of Antiviral Peptides Using Effective Feature Representation. Int J Mol Sci 20:. https://doi.org/10.3390/IJMS20225743
41. Nawrot R, Barylski J, Nowicki G, et al (2014) Plant antimicrobial peptides. Folia Microbiol (Praha) 59:181. https://doi.org/10.1007/S12223-013-0280-4
42. Osorio D, Rondón-Villarreal P, Torres R (2015) Peptides: A package for data mining of antimicrobial peptides. R J 7:4–14. https://doi.org/10.32614/RJ-2015-001
43. Raghava GPS, Han JH (2005) Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein. BMC Bioinformatics 6:. https://doi.org/10.1186/1471-2105-6-59
44. Remmert M, Biegert A, Hauser A, Söding J (2011) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 2011 92 9:173–175. https://doi.org/10.1038/nmeth.1818
45. Santos-Silva CA dos, Zupin L, Oliveira-Lima M, et al (2020) Plant Antimicrobial Peptides: State of the Art, In Silico Prediction and Perspectives in the Omics Era. Bioinform. Biol. Insights 14

46. Sels J, Mathys J, De Coninck BMA, et al (2008) Plant pathogenesis-related (PR) proteins: A focus on PR peptides. Plant Physiol. Biochem. 46:941–950
47. SF A, W G, W M, et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
48. Studer G, Rempfer C, Waterhouse AM, et al (2020) QMEANDisCo—distance constraints applied on model quality estimation. Bioinformatics 36:1765–1771. https://doi.org/10.1093/BIOINFORMATICS/BTZ828
49. Tam JP, Wang S, Wong KH, Tan WL (2015) Antimicrobial peptides from plants. Pharmaceuticals 8:711–757
50. Tavares LS, Rettore JV, Freitas RM, et al (2012) Antimicrobial activity of recombinant Pg-AMP1, a glycine-rich peptide from guava seeds. Peptides 37:294–300. https://doi.org/10.1016/J.PEPTIDES.2012.07.017
51. Thakur N, Qureshi A, Kumar M (2012) AVPpred: Collection and prediction of highly effective antiviral peptides. Nucleic Acids Res 40:W199–W204. https://doi.org/10.1093/nar/gks450
52. Vens C, Rosso MN, Danchin EGJ (2011) Identifying discriminative classification-based motifs in biological sequences. Bioinformatics 27:1231–1238. https://doi.org/10.1093/BIOINFORMATICS/BTR110
53. Wang G (2010) Antimicrobial peptides: Discovery, design, and novel therapeutic strategies. Antimicrob Pept Discov Des Nov Ther Strateg 1–230. https://doi.org/10.1079/9781845936570.0000
54. Wang G, Li X, Wang Z (2016) APD3: the antimicrobial peptide database as a tool for research and education. Nucleic Acids Res 44:D1087-93. https://doi.org/10.1093/nar/gkv1278
55. Wang G, Li X, Wang Z (2009) APD2: the updated antimicrobial peptide database and its application in peptide design. Nucleic Acids Res 37:D933–D937. https://doi.org/10.1093/NAR/GKN823
56. Waterhouse A, Bertoni M, Bienert S, et al (2018) SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 46:W296–W303. https://doi.org/10.1093/NAR/GKY427
57. Xiao N, Cao DS, Zhu MF, Xu QS (2015) protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics 31:1857–1859. https://doi.org/10.1093/BIOINFORMATICS/BTV042
58. Xiao X, Wang P, Lin WZ, et al (2013) IAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal Biochem 436:168–177. https://doi.org/10.1016/j.ab.2013.01.019
59. Yan J, Bhadra P, Li A, et al (2020) Deep-AmPEP30: Improve Short Antimicrobial Peptides Prediction with Deep Learning. Mol Ther - Nucleic Acids 20:882–894. https://doi.org/10.1016/j.omtn.2020.05.006
60. Zhang LJ, Gallo RL (2016) Antimicrobial peptides. Curr. Biol. 26:R14–R19

Figures

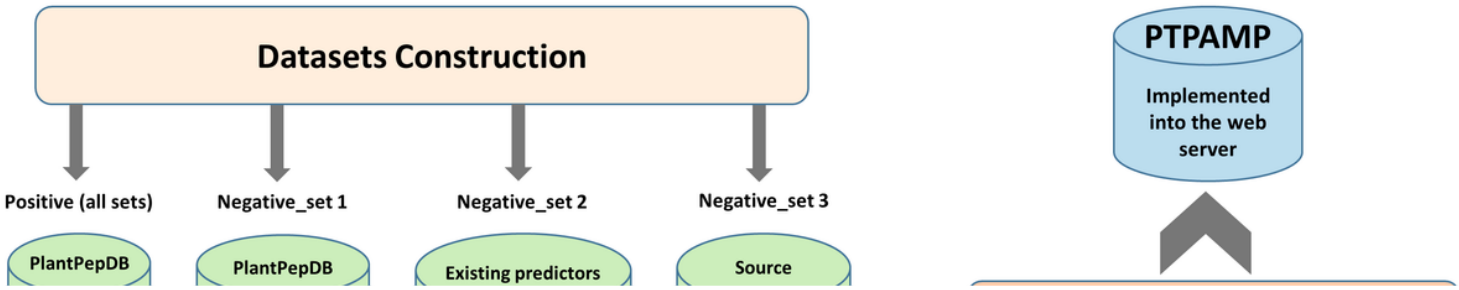

Figure 1

Schematic framework of PTPAMP. It includes four steps: (i) dataset construction, (ii) represent peptide sequences by eight feature encodings, (iii) construction of prediction models, and (iv) performance assessment and implementation of best model in PTPAMP

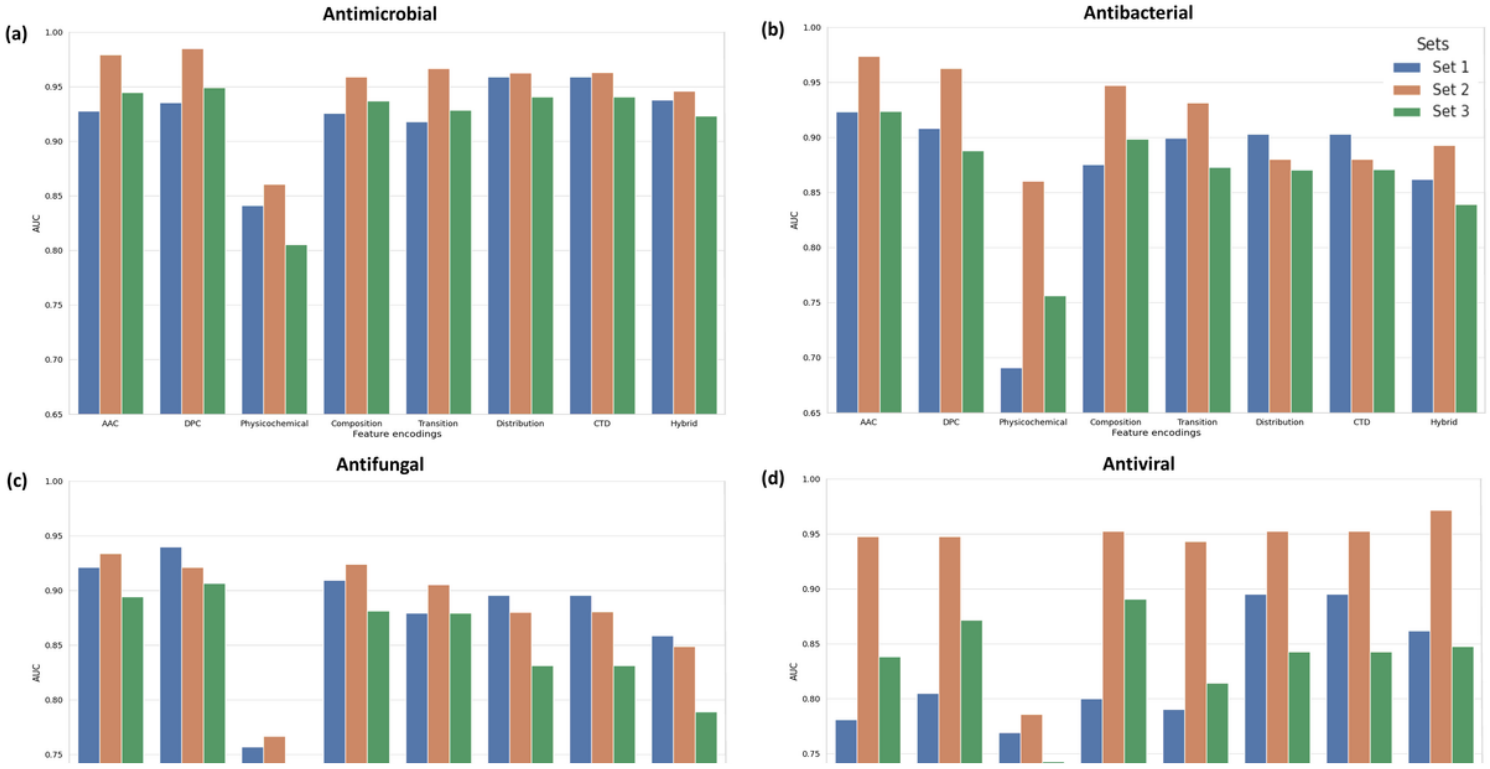

Figure 2

Five-fold cross-validation performance of SVM models developed with three types of datasets with respect to eight feature encodings for **(a)** antimicrobial, **(b)** antibacterial, **(c)** antifungal, and **(d)** antiviral functional activity

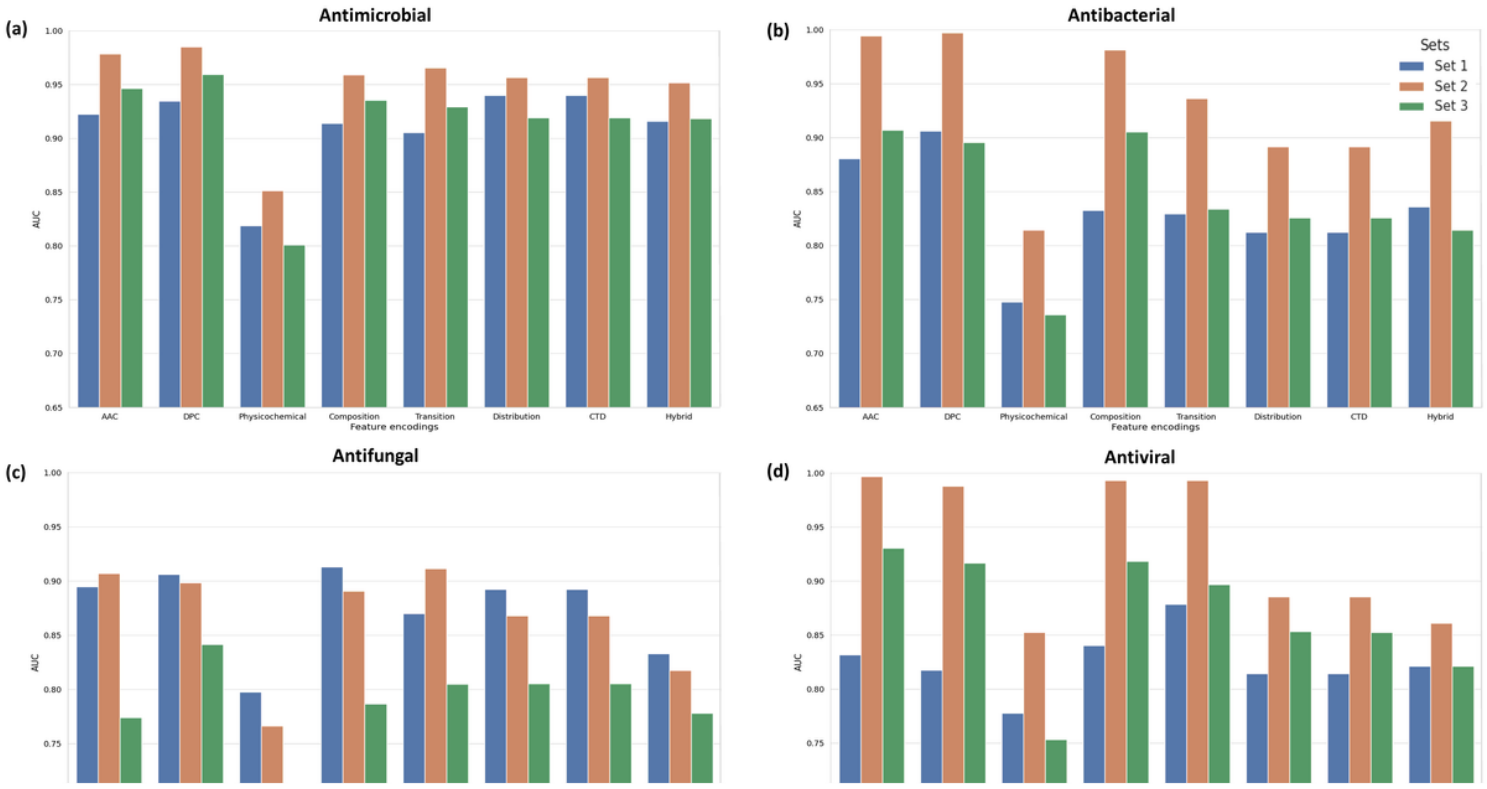

Figure 3

Performance comparison of SVM models on respective independent test data of three sets and eight feature encodings for **(a)** antimicrobial, **(b)** antibacterial, **(c)** antifungal, and **(d)** antiviral functional activity

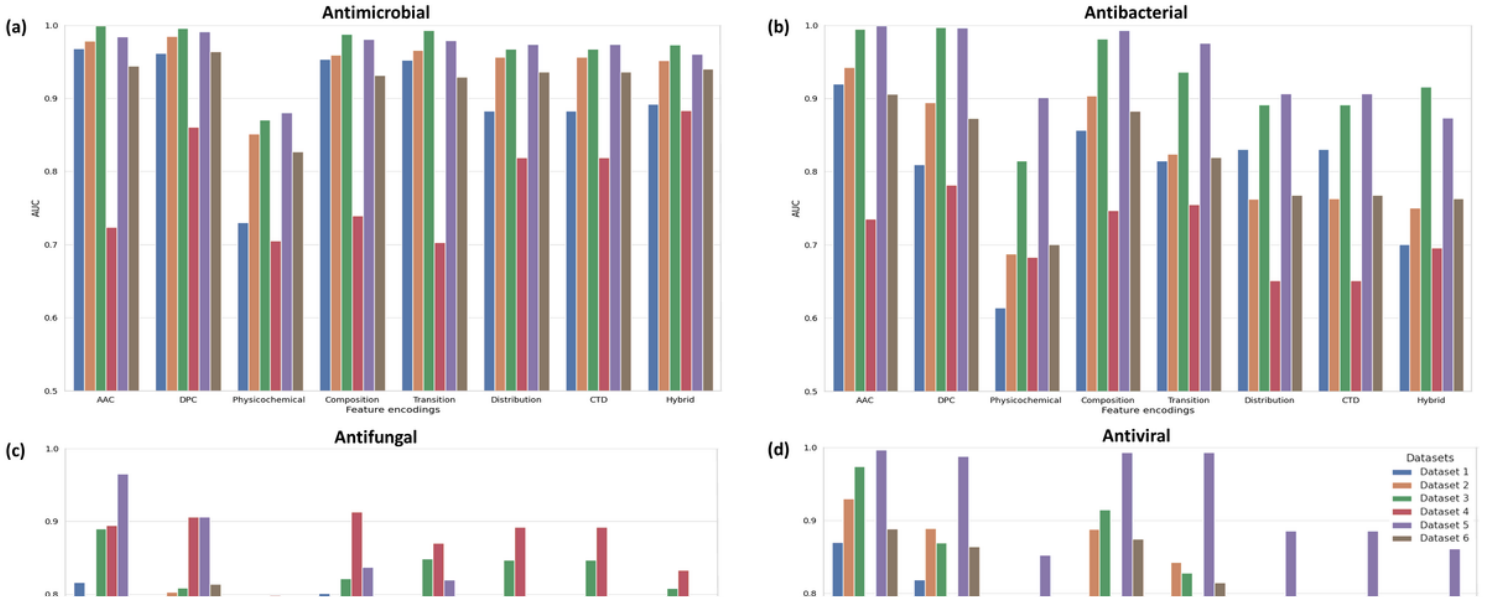

Figure 4

Comparison between the performance of SVM models prepared with respect to eight feature encodings when checked their results on six benchmark datasets. **(a)**, **(b)**, **(c)**, and **(d)** respectively describe the results obtained for antimicrobial, antibacterial, antifungal, and antiviral models

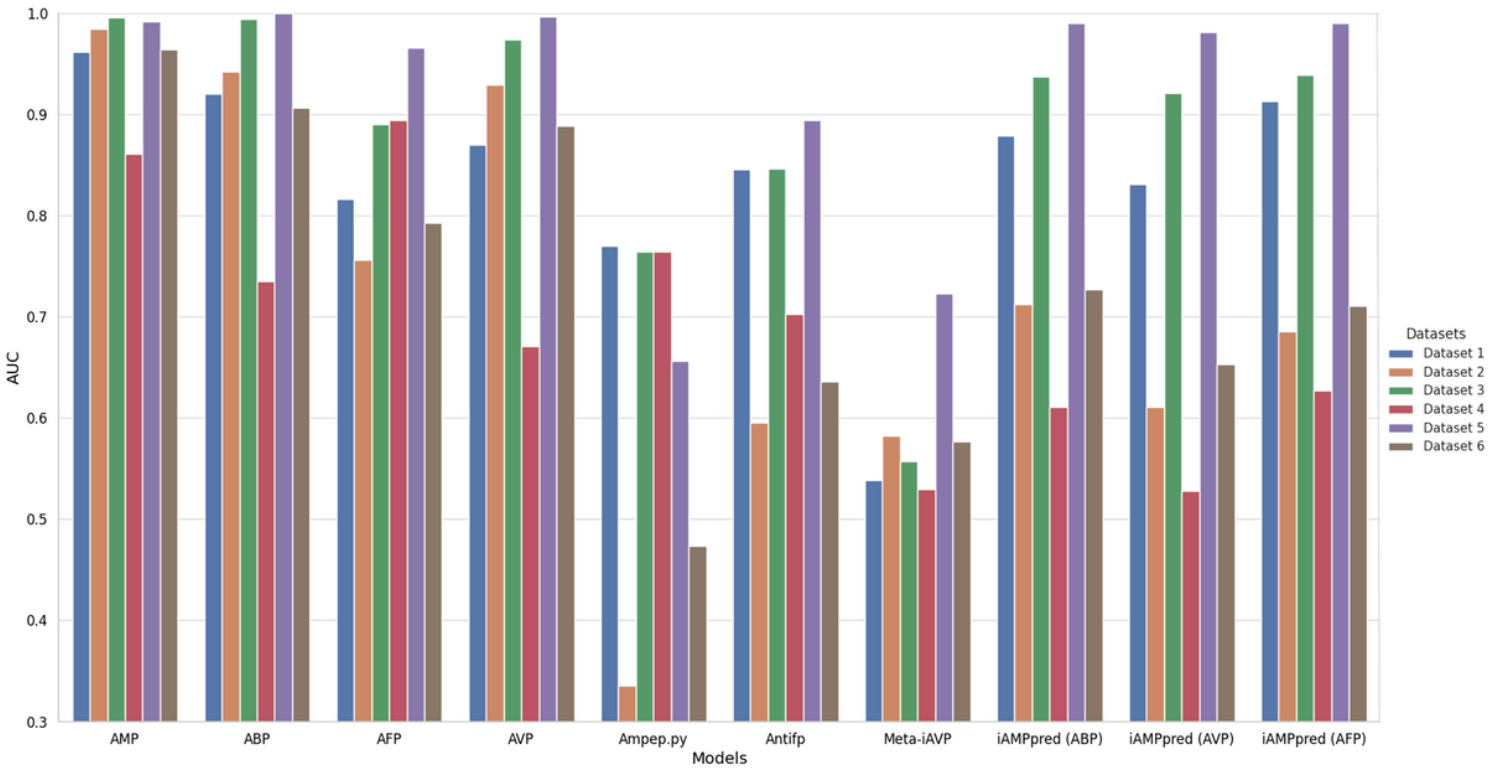

Figure 5

Comparative performance of PTPAMP models and state-of-the-art predictors based on six benchmark datasets

Figure 6

Predicted secondary structures of selected sequences that have been validated at protein level. Secondary structures were predicted using PSIPRED with moderate confidence of prediction. The structure depicts the majority of plant-AMPs belongs to β or αβ family

Figure 7

3D structures of selected antimicrobial peptides that have been validated at protein level. Structures are predicted by SWISS-MODEL. Each structures are labelled with the peptide name followed by the SMTL ID of the recorded template. Positions of N and C-terminal of peptides are also shown. Blue represents the least hydrophobic amino acid residues and red represents the most hydrophobic amino acid residues

Figure 8

Average amino acid compositions for positive and negative data of set 2 for **(a)** antimicrobial, **(b)** antibacterial, **(c)** antifungal, and **(d)** antiviral activity. Amino acids are categorized according to their physicochemical properties

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfile.pdf
- Supplementarysheet1.xls