AMAP: Hierarchical Multi-label Prediction of Biologically Active and Antimicrobial Peptides

Sadaf Gull, Nauman Shamim and Fayyaz Minhas*

Biomedical Informatics Research Laboratory, Department of Computer and Information

Sciences (DCIS), Pakistan Institute of Engineering and Applied Sciences (PIEAS), Nilore,

Islamabad, Pakistan.

*Corresponding author email: fayyazafsar@gmail.com   or afsar@pieas.edu.pk

ABSTRACT

Due to increase in antibiotic resistance in recent years, development of efficient and accurate techniques for discovery and design of biologically active peptides such as antimicrobial peptides (AMPs) has become essential. The screening of natural and synthetic AMPs in the wet lab is a challenge due to time and cost involved in such experiments. Bioinformatics methods can be used to speed up discovery and design of antimicrobial peptides by limiting the wet-lab search to promising peptide sequences. However, most such tools are typically limited to the prediction of whether a peptide exhibits antimicrobial activity or not and they do not identify the exact type of the biological activities of these peptides. In this work, we have designed a machine learning based model called AMAP for predicting biological activity of peptides with a specialized focus on antimicrobial activity prediction. AMAP used multi-label classification to predict 14 different types of biological functions of a given peptide sequence with improved accuracy in comparison to existing state of the art techniques. We have performed stringent performance analyses of the proposed method. In addition to cross-validation and performance comparison with existing AMP predictors, AMAP has also been benchmarked on recently published experimentally verified peptides that were not a part of our training set. We have also analyzed features used in this work and our analysis shows that the proposed predictor can generalize well in predicting biological activity of novel peptide sequences. A webserver of the proposed method is available at the URL: http://faculty.pieas.edu.pk/fayyaz/software.html#AMAP

Keywords: Biologically active peptides, Antimicrobial peptide prediction, multi-label classification, Antibiotic resistance, Antibiotic peptide prediction.

1. Introduction

Antimicrobial peptides (AMPs) are short length peptide sequences which can perform antimicrobial activity, help in fighting infectious diseases and protect hosts from pathogenic bacteria [1–4]. Due to the emergence of antibiotic resistance, AMPs have become a very active area of research. Identification of naturally occurring AMPs and design of synthetic ones is challenging due to the time and cost involved in the design and execution of biochemical assays for testing or screening candidate peptides [5–7]. As a consequence, development of computational techniques for prediction of antimicrobial and other significant biological activities of peptide sequences is very important. An ideal computational method in this domain should be able to predict possible biological activities (antimicrobial, antibacterial, antiviral, antifungal, anti-cancerous, etc.) of a given peptide sequence and correctly identify the effect of mutations in such peptides.

In the last few years, many databases of anti-microbial peptides have become available such as the Antimicrobial Peptide Database (APD3) [1] which contains experimentally verified natural and synthetic peptides with over 20 different biological activities. Collection of Anti-Microbial Peptides (CAMP) [3] has also been developed which contains a large number of experimentally verified antimicrobial peptides. Database of Antimicrobial Activity and Structure of Peptides (DBAASP) [8] contains detailed information about structure and antimicrobial/cytotoxic activity of different peptides. dbAMP [9] is a database of experimentally verified AMPs with potent biological activity in a variety of different species. The development of these databases has accelerated the pace of development of data-driven predictive models for predicting biological activity of peptides. A number of different predictors of peptide biological activity are available in the literature. However, most existing methods are limited to predicting antimicrobial activity, i.e., they can only predict whether a given peptide sequence is anti-microbial or not. For example,

AMPA [10] takes a protein sequence as input and predicts its antimicrobial activity and the peptide region responsible for such activity. Both CAMPR3 [3] and AmPEP [11] also take a peptide sequence and predict whether it is an antimicrobial peptide or not but, like AMPA, they do not provide information about other biological activities or the type of antimicrobial activity (anti-bacterial, anti-fungal, anti-viral, etc.) a peptide may have. Similarly, AntiMPmod [12] predicts antimicrobial activity of a peptide from its tertiary structure. However, the use of peptide structure instead of sequence limits the practical use of this method as structure information is typically not available for peptides. Vishnepolsky et al. have designed a model which predicts antimicrobial potency for some specific strains of Gram negative bacteria [13]. However, their method is not generalized for other species or targets. One of the most interesting approaches in this domain is Multi-label Anti-Microbial Peptides predictor (MLAMP) [2] as it can predict five different biological activities (antibacterial, antifungal, anticancer, antiviral and anti-HIV) of peptides. However, its accuracy is low on certain classes. Gabere and Noble [4] have recently performed a comparison of existing predictors and found CAMPR3 to have state of the art predictive performance for AMP prediction even though the accuracy of CAMPR3 [3] was reported to be inferior to MLAMP in the original MLAMP [2] paper. However, there is significant room for improvement in the predictive performance of existing methods.

In this paper, we have attempted to overcome the problems associated with existing AMP predictors by developing a hierarchical multi-label predictor called AMAP that can simultaneously predict whether a peptide sequence is an AMP or not, the type of its biological activity of a peptide and the effect of mutations on its biological activity. We have performed a stringent performance evaluation and comparison with existing methods by considering sequence similarity in training and test folds in cross-validation. We have also benchmarked our machine learning model on a

4

number of recently published biologically active peptide sequences that were not a part of our training or cross-validation data set. We have also performed an in-depth analysis of the predictive power of features used for predicting biological activity of peptides. A webserver of our proposed method has also been developed which makes large-scale evaluation of biological activity easier and accessible for biologists working in this domain.

2. Materials and Methods

2.1 Datasets

2.1.1   *Cross-validation dataset*

In line with the recent performance comparison study of different antimicrobial peptide predictors by Gabere and Noble [4], we have also used the peptides in APD3 database. Specifically, we have used a dataset of 2,704 peptides with 14 different types of biological activities collected from APD3 [1] (see Table-1). In the design of our machine learning models these peptides are taken as positive examples. Gabere and Noble extracted 8,563 peptides with no known antimicrobial activity from UniProt. We used these as negative examples to train and evaluate our model. In order to prevent sequence and composition biases from affecting our machine learning model, we removed all peptides with more than 40% sequence identity to each other or to the positive set using CDHIT [14], leaving a total of 5, 156 negative examples. In order to perform an unbiased performance assessment, the sequences are clustered into groups using a 40% sequence similarity threshold with CDHIT for group-wise cross-validation. Our dataset shares significant overlap with the datasets used in other studies as well. Approximately 52% positive examples in our dataset share >40% sequence identity with the dataset used by CAMP-R3(RF) [3]. Similarly, all positive examples used by MLAMP are included in our positive set [2].

5

2.1.2   External Validation dataset

We have also used an external validation dataset which contains experimentally verified biologically active peptides collected from different recently published research articles. The similarity of these sequences with our training dataset is calculated using CDHIT [14]. We have found no significant sequence similarity between peptides in the external validation dataset with our training dataset. The maximum percentage identity of a test peptide with the peptides in the dataset used for cross-validation is given in supplementary Table-S1.

2.2 Feature extraction

We use simple sequence-based feature representation to ensure large-scale applicability as characterization of peptide structures is difficult and costly. For feature extraction from the collected examples, two representations are used as explained below.

2.2.1   Amino Acid Composition (AAC)

AAC is the frequency count of 20 amino acids forming a vector of length 20. This representation is useful for capturing information about the frequency of different amino acids in a sequence.

2.2.2   3-mer Composition

Amino acid composition does not model the local sequence properties of a peptide. As a consequence, we have used 3-mer counts as additional features. First all amino acids are divided into 7 groups based on their physiochemical properties. The grouping of amino acids is based on dipole moment, side chain volume and their ability to form di-sulphide bonds [15] as shown in Table-3. In the second step, all possible 3-mer counts of group labels of amino acids in a given peptide sequence is used to create a $7^3 = 343$ dimensional feature vector [16]. This representation

captures information about physiochemical properties of amino acids with 3-mer patterns in the sequences.

2.3 Prediction Models

We have used the following prediction techniques for development of the proposed predictor.

2.3.1 Support Vector Machines (SVM)

SVM is a supervised learning algorithm for binary classification that maximizes the margin or separation between two classes in the training data [17]. Given a set of N training examples $x_i, i = 1 \dots N$ with associated labels $y_i \in \{+1, -1\}$, an SVM finds an optimal linear decision function $f(x) = \langle w, x \rangle$ by maximizing the distance of the linear decision boundary from examples of the positive and negative classes (margin) and minimizing the number of misclassifications or margin violations through the following objective function.

$$min_{w,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i$$

Subject to

$$y_i \langle w, x_i \rangle \geq 1 - \xi_i, \xi_i \geq 0, \forall i = 1 \dots N$$

Here ξ_i is the extent of margin violation with penalty C on the violations. In our model, we used class-specific margin violation penalties to counter the effect of class imbalance. SVMs can use kernel functions to model non-linear classification boundaries as well. In this work, we used both linear and radial basis function kernels for SVMs.

2.3.2 XGBoost

We have also used extreme Gradient Boosting (XGBoost) [18] in our study. It is based on boosted trees for learning by minimizing the objective function given below:

$$L(\Phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

Where,

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda||\boldsymbol{u}||^2.$$

Here $l(\hat{y}_i, y_i)$ is the loss function of predicted output $\hat{y}_i$ of the model and actual output y_i for all examples and $\Omega(f_k)$ is a regularization function that is based on the number of trees T and the norm of the vector of scores $\boldsymbol{u}$ at the k leaves of the trees. The regularization parameters γ and λ control the relative contribution of the two regularization factors in contrast to minimization of the loss function.

2.3.3 Hierarchical Multi-label Prediction

Since a single peptide can be associated simultaneously with a number of different biological activities, we have modeled this prediction problem as multi-label classification. Multi-label predictions can be obtained from a binary classifier by using one-vs-rest classifier fusion [19]. We designed our proposed model in two steps: a peptide is first checked whether it is an AMP or not and then the type of biological activity it may have is predicted using multi-label prediction.

2.3.4 Baseline Evaluation using BLAST and Comparison with other methods

In order to establish a baseline, we used BLAST [20] for prediction of biological activity of peptides using our dataset. In this approach, the minimum e-value score of BLAST alignment of a peptide sequence against the set of known non-redundant peptides with known activities is used as a discriminant function score for predicting biological activity. This approach corresponds to a simple sequence-based homology search for biological activity prediction.

We have also compared the predictive performance of the proposed scheme with state of the art sequence-based predictive methods: CAMPR3-RF [3] and MLAMP [2]. For this purpose, we have used the publicly-available webservers of these methods.

2.4 Cross-validation Strategy

Previously designed models used different techniques for evaluating the performance of their models such as Leave-one-out (LOO) or k-fold cross-validation, etc. The problem the use of such cross-validation schemes for performance assessment is that test examples may have high sequence similarity with the training set which can result in overestimation of prediction accuracy and poor generalization in case of sequences with low sequence similarity to training data [21]. To avoid this issue, we used two different strategies for cross-validation. The first technique is Leave-one-cluster-out (LOCO) [22] cross-validation in which examples are first clustered based on sequence similarity through CD-HIT [14] with a sequence identity threshold of 40% (see Table-2). The examples of one cluster are used for testing while the model is trained on examples in remaining clusters. This process is repeated for all other clusters to obtain performance metric statistics [22]. The second technique is clustered 5 fold cross-validation which is computationally more efficient. In this approach, the set of sequence clusters is divided into 5 folds so that all examples in a single cluster occur in a single fold to limit sequence similarity between training and testing folds. We have also performed cross-validation of our dataset using standard LOO and k-fold and found their scores to be consistently higher than LOCO cross-validation due to the presence of homologous proteins in the training set (results not shown here). As a consequence, we have used the more stringent performance evaluation protocol outlined above. The hyper-parameters of different classification schemes such as the margin violation penalty and kernel parameters were selected through nested validation.

2.5 Performance Metrics

We have used the following performance metrics to evaluate the predictive accuracy of the proposed scheme.

2.5.1 Area under the ROC curve (AUC-ROC)

AUC-ROC captures the area under the Receiver Operating Characteristic (ROC) Curve which plots the sensitivity or true positive rate of a predictor vs. its false positive rate at various decision threshold levels [23].

2.5.2 Area under the Precision-Recall curve (AUC-PR)

Recall is the ratio of the number of correctly predicted positive examples to the total number of positive examples. Precision is the ratio of the number of correctly predicted positive examples to the number of total predicted positive examples [23]. In highly imbalanced datasets, the area under the precision recall curve gives a more informative picture of the predictive performance and is used here.

2.5.3 Mathews Correlation Coefficient (MCC)

MCC is a measure used in machine learning to assess classification performance of imbalanced datasets. The range of coefficient is between +1 and -1, where +1 shows perfect prediction between observed and predicted class labels using following formula [24]. Here, TP, TN, FP and FN are the number of true positives, true negatives, false positives and false negatives generated by a classifier.

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

2.6 Analysis of Predictive Features

Machine learning models are typically black boxes and linking their predictions to real-world explanations requires additional steps. In bioinformatics, it is important to analyze the role of different features used in a machine learning model and link them to known patterns or rules from a biological or biochemical perspective. Such analysis can help in understanding the inner working of the machine learning model and add significant confidence to its prediction. Based on this motivation, we have used three different methods for analyzing the predictive performance of the proposed model as discussed below:

2.6.1   t-SNE visualization of feature vectors

t-distributed Stochastic Neighbor Embedding (t-SNE) [25] is an unsupervised technique for manifold learning based dimensionality reduction for visualization of high dimensional data. It works by minimizing the Kullback-Leibler divergence between the high and low dimensional probabilistic representations of a data set. We have used t-SNE to visualize the 20-dimensional amino acid composition features by reducing the high-dimensions of our data set to a two-dimensional scatter plot and studying the separability of different classes in our data. It is important to note that t-SNE does not use the labels of examples.

2.6.2   Analysis of weights of linear SVM

The absolute weight vector of a trained linear support vector machine trained over normalized data can be used to analyze the relative feature importance of different features as high positive or negative values of the weight vector have more impact in determining the output decision score for a given example $f(x) = \langle w, x \rangle$. Here, we have used a bar plot of the 20-dimensional weight

vector corresponding to amino acid composition features to analyze the relative importance of different features.

2.6.3 SHAP Analysis

SHAP (SHapley Additive exPlanations) [26] is a recently developed technique that allows us to explain the output of any machine learning model. It produces a plot of the SHAP scores at different values of all features used in a trained machine learning model. High absolute SHAP scores corresponds to more important features. Specifically, a high positive SHAP value at a certain value of a feature indicates that the feature value will have a positive impact on the output of the machine learning model and vice versa. We have used SHAP analysis to analyze the impact of different amino acids in a peptide on its AMP activity.

2.7 Web server for Anti-Microbial Activity Prediction (AMAP)

We have also developed a webserver for the proposed method (URL: http://faculty.pieas.edu.pk/fayyaz/software.html#AMAP.) The user interface of our webserver is shown in Fig 6. It takes input peptide sequences in FASTA format and displays predicted scores for different activities.

2.8 Evaluation on external datasets

As discussed in the datasets section, we have also evaluated our model on an external validation dataset. Peptides in this dataset are not part of our training dataset and cross-validation. We computed prediction scores for the peptides in this dataset using our webserver and compared them to the predictions from MLAMP and their experimentally observed biological activities. The details of the sequences used in this analysis is given in the results section.

3   Results

In this section, we present the results of different experiments for the proposed model. A detailed discussion of these results is presented in the next section.

3.1 Performance comparison for Antimicrobial Peptide Prediction

Fig-1 and Table-4 present the results of AMP peptide prediction with different classification schemes discussed above using clustered 5-fold cross-validation in comparison to existing methods (MLAMP [2] and CAMPR3-RF [3]). Our BLAST-based baseline predictor gives AUC-ROC, AUC-PR and MCC scores equal to 77%, 72% and 0.27, respectively. We have compared the performance of three different machine learning models while developing AMAP (linear-SVM, non-linear SVM with RBF-kernel and XGBoost). Our machine learning based models give the best overall predictive accuracy with AUC-ROC score of 97%. A more detailed description of the results is given in the discussion section.

3.2 Analysis of amino acid composition features for AMP prediction

3.2.1   t-SNE Analysis

We have used t-SNE for generating two-dimensional scatter plots from our 20-dimensional amino acid composition features for all 7,860 examples in our data set (see Fig. 2). It is important to note that t-SNE is an unsupervised technique, i.e., it does not require labels of the examples used in the analysis. The labels of AMPs and non-AMPs for all examples are added in the figure to help us analyze the separability of the two classes after t-SNE transformation as discussed in the next section.

3.2.2   Analysis of linear SVM weight vector

The plot of the weight vector of the trained linear SVM is shown in Fig. 3. As discussed earlier, the relative importance of different amino acids in predicting AMPs can be inferred from this plot.

3.2.3 SHAP Analysis

The SHAP plot for our trained non-linear XGBoost model is shown in Fig 4. As discussed earlier, SHAP values indicate the relative impact of values of different features on the output of a classifier. A more detailed analysis of this plot is presented in the next section.

3.3 Multi-label classification of biological activity

As discussed earlier, our dataset contains peptides that are involved in 14 different biological activities (see Table-1) and a single peptide can have more than one type of activity. After being trained, our proposed multi-label machine learning model generates decision scores corresponding to each biological activity. The AUC-ROC of the leave-one-cluster-out (LOCO) cross-validation of our multi-label SVM model are shown in Table-5 for all 14 different types of biologically active peptides in comparison to the previous state of the art method MLAMP.

3.4 Evaluation on Independent Dataset

To evaluate the performance of our proposed model, we selected some recently discovered Antimicrobial peptides from latest publications that were not included in our original dataset. The maximum sequence similarity of these sequences with peptides in our training dataset is given in table S1 and it is below 50% for almost all sequences. The scores for these peptides are obtained using the AMAP webserver and compared to experimented findings and prediction scores from MLAMP webserver.

3.4.1   Synthetic Antimicrobial and Antibiofilm Peptides (SAAP) derived from LL-37 peptide

Breij et al. screened LL-37-inspired peptides with bactericidal activity [6]. The most effective peptides with lethal concentration (LC99.9) required for killing 99.9 % bacteria are reported in Table-6. Peptide P276 is more effective than others as its LC is the lowest. It can be clearly seen that AMAP is able to correctly identify it as an effective antimicrobial peptide by generating a high score for it in comparison to others. On the other hand, MLAMP predicts P276 as non-AMP although its probability score should be higher than that for peptides P148, P145 and P159. This demonstrates the effectiveness of the proposed scheme.

3.4.2   Synthetic peptides derived from Temporin-Ali peptide

Yoshida et al. have discovered effective AMPs from a natural peptide Temporin-Ali [7]. These peptides are evaluated for their antimicrobial activities *in vitro* by measuring the half maximal inhibitory concentration (IC50) against *E. coli* (MG1655 strain). The identified AMPs with improved antimicrobial activity are given in Table-7. The scores generated by our proposed model for given peptides show a good correspondence with experimentally observed inhibitory concentrations. It is interesting to note that predictive scores generated by both AMAP and MLAMP correlate well with experimental observations except in the case of the antimicrobial peptide 2C for which both methods generate low prediction scores.

3.4.3   Membrane-targeting antibacterial peptides from Viral Capsid proteins

Dias et al. have identified peptide sequences with strong antibacterial properties [27]. They identified two viral protein-derived peptide sequences vCPP 0769 and vCPP 2319 with high antibacterial activity. The scores of our model for these two sequences are also high whereas MLAMP predicts vCPP2319 as non-AMP as shown in Table-8.

We performed another analysis on the Major Capsid Protein of *Fowl adenovirus A serotype 1*, from which sequence vCPP0769 is derived: we used a sliding sequence window of length 20 over the protein to find the most AMP like sequence in the protein through AMAP as shown in Fig 5. It is interesting to note that the highest AMAP score occurs at the window corresponding to the location of the experimentally identified antibacterial sequence (vCPP0769) within the protein. This shows that AMAP can be used to find AMP regions within proteins as well.

3.4.4   Synthetic peptides active against *Escherichia coli*

Pini et al. selected a peptide sequence and performed mutations on it to increase its antimicrobial activity and found that peptide M6 has higher antimicrobial activity than other modified sequences (M4, M5 and the wild-type sequence) [28]. Experimentally determined minimum inhibitory concentrations reported for these sequences against various bacteria are given in Table-9. The wild-type sequence did not have any significant antimicrobial activity.

AMAP prediction scores for these sequence are given in Table-10. It is interesting to note that the wild-type sequence has the lowest score and, even though M6 differs from WT by only one amino acid, the AMAP score for M6 is significantly higher and correlates well with experimental findings. MLAMP also produces the lowest score for the wild-type sequence in comparison to the mutated peptides and the wild-type sequence. However, MLAMP score for the best performing peptide (M6) is lower than M5 which is not as potent as M6 in experimental observations.

3.4.5   ACWWP1 peptide

The Antibacterial peptide ACWWP1(GLSRLFTALK) kills bacterial cells via membrane damage. Pu and Tang showed that ACWWP1 can be used in the treatment of food poisoning caused by

certain bacteria [29]. AMAP generates a high prediction score of 1.10 for this peptide. MLAMP also generates a high probability of 0.95 for this peptide to have antimicrobial activity.

3.4.6 Synthetic antifungal peptides against *F.oxysporum*

Badosal et al. screened a set of peptide sequences in vitro against the fungus *F.oxysporum* [30]. Their reported MIC and AMAP prediction scores for antifungal activity are given in Table-11. All these peptides have shown significant antifungal activity in experiments. AMAP is able to correctly predict their antifungal activity with high scores. However, MLAMP predicts all sequences to have antibacterial activity instead of antifungal activity.

3.4.7 Cp1 and Melittin peptides

Hou et al. studied the antimicrobial activity of protein-derived peptide Cp1 synthesized form Bovine α_{S1}-Casein and Melittin which was purified from bee venom [31]. Their antimicrobial activity is summarized in Table-12 and shows that Melittin is more potent in comparison to Cp1. The scores for Cp1 and Mellitin by our model show good correspondence with experimental findings with a high prediction score for Melittin in comparison to Cp1. The prediction results of AMAP and MLAMP are comparable as shown in Table-13.

3.4.8 Peptides derived from the *Acinetobacter baumannii* phage endolysin LysAB2

Peng et al. synthesized four AMPs from the phage endolysin for antimicrobial activity [32]. Peptide P0 has low antimicrobial activity and the AMAP prediction score is also small for that peptide as sh own in Table-14. The MICs of LysAB2 P1 and P3 are in range of 4–8 μM which are lower than the MIC of P1. The AMAP scores of P1 and P3 are higher than P0 which shows higher antimicrobial activity in P1 and P3. There is a good correspondence between AMAP scores and experimentally observed MICs of peptides. On the other hand, MLAMP predicts all these peptides

as non-AMP except P0 which actually has the lowest experimentally observed antibacterial activity.

4   Discussion

4.1 Performance comparison

Table-4 shows that simple sequence alignment to known AMPs using BLAST gives AUC-ROC and AUC-PR of 77% and 72%, respectively whereas the best performing machine learning method proposed in this work (Radial Basis Function SVMs) offers significant improvement in prediction performance (AUC-ROC and AUC-PR of 97% and 96%, respectively). This clearly shows that the proposed machine learning technique is superior to simple homology search. Furthermore, our comparison to previous state of the art techniques (CAMP-R3 (RF) [3] and MLAMP [2]) on our dataset through their respective webservers also verifies this conclusion. However, the performance of these methods is much lower than the proposed machine learning model. As shown in Figure 1, at 2% False Positive Rate, the sensitivity of CAMPR3-RF is 47% in comparison to 80% by the proposed SVM model. It is also important to note that the performance of non-linear SVMs is also matched by XGBoost classifiers and the performance of linear SVMs is also not much lower than the best performing model. However, in the development of the final AMAP predictor, we have used the nonlinear SVM. It is also interesting to note that global amino acid composition is a better predictor of AMPs in comparison to sequence tripeptide composition for all classification schemes. Just like PR and ROC measures, the result of MCC is also higher of proposed model. CAMPR3 is good in performance as compared to MLAMP but still its performance is lower than that of the proposed model (see Table-4).

4.2 Feature Analysis

18

The t-SNE plot (Figure 2) clearly shows that amino acid composition features can distinguish AMPs and non-AMPs into separate clusters with minimal overlap. This lends support to the high accuracy obtained by our machine learning models especially linear support vector machines. It is important to notice a few clusters of AMPs in regions otherwise dominated by non-AMPs in Fig. 2(a). This shows why non-linear or RBF SVM and XGBoost classifiers perform better than linear SVM. It is interesting to note that antibacterial peptides have more overlap with non-AMPs in comparison to other types.

The plot of weight vector of linear SVM (Fig 3) shows that Cysteine (C), Lysine (K), Valine (V), and Phenalalanine (F) are important for AMP prediction whereas the occurrence of D, E, L, Y, P, R and N is predictive of non-AMP activity. These observations are in line with the findings in the literature which indicate that Lysine (K) is the most commonly occurring amino acid in known AMPs [6]. Cysteine (C) is also an important amino acid in natural antimicrobial peptides of vertebrates, invertebrates and plants [33]. This shows that the output of the proposed machine learning model correlates with known facts about AMPs.

The analysis of important features using SHAP (Fig 4) is in line with our findings from the weight vector plot for the SVM as well. This analysis clearly shows that the proposed model is in line with known biological information about AMPs. Our analysis reveals that our model generates negative labels for peptides enriched in amino acids D, E and L whereas the occurrence of C, K, F and Q positively affects the output of the classifier.

4.3 Multi-label Prediction of biology activity

The results for the prediction of biological activities of peptides are given in Table-5. The decision scores from MLAMP webserver for a given peptide sequence are available for only 5 classes. The performance of the proposed scheme is significantly better than MLAMP for all biological activity

categories. This shows the efficacy of the proposed scheme in predicting a different broad types of biologically active peptides.

4.4   External Evaluation on Independent Validation Set

The evaluation of proposed model on experimentally verified biologically active peptides in the independent validation set taken from eight different recent publications shows good prediction results even though these peptides were significantly different in sequence from the ones in the training dataset. The prediction scores generated by the proposed scheme are in line with experimental observations for a wide variety of peptides and types of biological activities. Furthermore, the proposed method shows improved performance in comparison to MLAMP over this dataset as well.

5   Conclusions

We have developed a predictor called AMAP that can be used to identify antimicrobial peptides as well predict their biological activity. The proposed scheme offers significantly better prediction accuracy in comparison to previously published methods. Our extensive performance evaluation reveals that the proposed method can be very useful in predicting antimicrobial peptides, effect of mutations on the biological activity of such peptides, and the determination of active regions within proteins with antimicrobial activity. The use of sequence information alone in our predictive modeling and a publicly available webserver for the proposed method are expected to accelerate the pace of research aimed at countering the threats posed by the rise of antimicrobial resistance.

Conflict of interest

The authors have no conflict of interest.

References

[1] G. Wang, X. Li, Z. Wang, APD3: the antimicrobial peptide database as a tool for research and education, Nucleic Acids Research. 44 (2015) D1087–D1093.

[2] W. Lin, D. Xu, Imbalanced multi-label learning for identifying antimicrobial peptides and their functional types, Bioinformatics. 32 (2016) 3745–3752.

[3] F.H. Waghu, R.S. Barai, P. Gurung, S. Idicula-Thomas, CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides, Nucleic Acids Research. 44 (2015) D1094–D1097.

[4] M.N. Gabere, W.S. Noble, Empirical comparison of web-based antimicrobial peptide prediction tools, Bioinformatics. 33 (2017) 1921–1929.

[5] A.T. Tucker, S.P. Leonard, C.D. DuBois, G.A. Knauf, A.L. Cunningham, C.O. Wilke, M.S. Trent, B.W. Davies, Discovery of Next-Generation Antimicrobials through Bacterial Self-Screening of Surface-Displayed Peptide Libraries, Cell. 172 (2018) 618–628.

[6] A. de Breij, M. Riool, R.A. Cordfunke, N. Malanovic, L. de Boer, R.I. Koning, E. Ravensbergen, M. Franken, T. van der Heijde, B.K. Boekema, others, The antimicrobial peptide SAAP-148 combats drug-resistant bacteria and biofilms, Science Translational Medicine. 10 (2018) eaan4044.

[7] M. Yoshida, T. Hinkley, S. Tsuda, Y.M. Abul-Haija, R.T. McBurney, V. Kulikov, J.S. Mathieson, S.G. Reyes, M.D. Castro, L. Cronin, Using evolutionary algorithms and machine learning to explore sequence space for the discovery of antimicrobial peptides, Chem. 4 (2018) 533–543.

[8] M. Pirtskhalava, A. Gabrielian, P. Cruz, H.L. Griggs, R.B. Squires, D.E. Hurt, M. Grigolava, M. Chubinidze, G. Gogoladze, B. Vishnepolsky, others, DBAASP v. 2: an enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides, Nucleic Acids Research. 44 (2015) D1104–D1112.

[9] K.-Y. Huang, T.-H. Chang, J.-H. Jhong, Y.-H. Chi, W.-C. Li, C.-L. Chan, K.R. Lai, T.-Y. Lee, Identification of natural antimicrobial peptides from bacteria through metagenomic and metatranscriptomic analysis of high-throughput transcriptome data of Taiwanese oolong teas, BMC Systems Biology. 11 (2017) 131.

[10] M. Torrent, V.M. Nogués, E. Boix, A theoretical approach to spot active regions in antimicrobial proteins, BMC Bioinformatics. 10 (2009) 373.

[11] P. Bhadra, J. Yan, J. Li, S. Fong, S.W. Siu, AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest, Scientific Reports. 8 (2018) 1697.

[12] P. Agrawal, G.P. Raghava, Prediction of Antimicrobial Potential of a Chemically Modified Peptide From Its Tertiary Structure, Frontiers in Microbiology. 9 (2018) 2551.

[13] B. Vishnepolsky, A. Gabrielian, A. Rosenthal, D.E. Hurt, M. Tartakovsky, G. Managadze, M. Grigolava, G.I. Makhatadze, M. Pirtskhalava, Predictive Model of Linear Antimicrobial Peptides Active against Gram-Negative Bacteria, Journal of Chemical Information and Modeling. 58 (2018) 1141–1151.

[14] Y. Huang, B. Niu, Y. Gao, L. Fu, W. Li, CD-HIT Suite: a web server for clustering and comparing biological sequences, Bioinformatics. 26 (2010) 680–682.

[15] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, H. Jiang, Predicting protein–protein interactions based only on sequences information, Proceedings of the National Academy of Sciences. 104 (2007) 4337–4341.

[16] C. Leslie, E. Eskin, W.S. Noble, The spectrum kernel: A string kernel for SVM protein classification, in: Biocomputing 2002, World Scientific, 2001: pp. 564–575.

[17] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning. 20 (1995) 273–297.

[18] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, ACM, 2016: pp. 785–794.

[19] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, Machine Learning. 85 (2011) 333.

[20] T. Madden, The BLAST sequence analysis tool, (2013).

[21] Z. John Lu, The elements of statistical learning: data mining, inference, and prediction, Journal of the Royal Statistical Society: Series A (Statistics in Society). 173 (2010) 693–694.

[22] W.A. Abbasi, F.U.A.A. Minhas, Issues in performance evaluation for host–pathogen protein interaction prediction, Journal of Bioinformatics and Computational Biology. 14 (2016) 1650011.

[23] J. Davis, M. Goadrich, The relationship between Precision-Recall and ROC curves, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006: pp. 233–240.

[24] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochimica et Biophysica Acta (BBA)-Protein Structure. 405 (1975) 442–451.

[25] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research. 9 (2008) 2579–2605.

[26] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, 2017: pp. 4765–4774.

[27] S.A. Dias, J.M. Freire, C. Pérez-Peinado, M.M. Domingues, D. Gaspar, N. Vale, P. Gomes, D. Andreu, S.T. Henriques, M.A. Castanho, others, New potent membrane-targeting antibacterial peptides from viral capsid proteins, Frontiers in Microbiology. 8 (2017) 775.

[28] A. Pini, A. Giuliani, C. Falciani, Y. Runci, C. Ricci, B. Lelli, M. Malossi, P. Neri, G.M. Rossolini, L. Bracci, Antimicrobial activity of novel dendrimeric peptides obtained by phage display selection and rational modification, Antimicrobial Agents and Chemotherapy. 49 (2005) 2665–2672.

[29] C. Pu, W. Tang, Affinity and selectivity of anchovy antibacterial peptide for Staphylococcus aureus cell membrane lipid and its application in whole milk, Food Control. 72 (2017) 153–163.

[30] E. Badosa, R. Ferré, J. Francés, E. Bardají, L. Feliu, M. Planas, E. Montesinos, Sporicidal activity of synthetic antifungal undecapeptides and control of Penicillium rot of apples, Applied and Environmental Microbiology. 75 (2009) 5563–5569.

[31] J. Hou, Z. Liu, S. Cao, H. Wang, C. Jiang, M.A. Hussain, S. Pang, Broad-Spectrum Antimicrobial Activity and Low Cytotoxicity against Human Cells of a Peptide Derived from Bovine αS1-Casein, Molecules. 23 (2018) 1220.

[32] S.-Y. Peng, R.-I. You, M.-J. Lai, N.-T. Lin, L.-K. Chen, K.-C. Chang, Highly potent antimicrobial modified peptides derived from the Acinetobacter baumannii phage endolysin LysAB2, Scientific Reports. 7 (2017) 11477.

[33] J.-L. Dimarcq, P. Bulet, C. Hetru, J. Hoffmann, Cysteine-rich antimicrobial peptides in invertebrates, Peptide Science. 47 (1998) 465–477.

List of tables and Figures

Table 1: Biological activities and number of peptides in each category in our positive dataset

Activity	No of peptides
antibacterial	2,446
antifungal	1,048
anticancer	210
antiviral	180
anti-HIV	109
chemotactic	57
antiparasital	43
antibiofilm	31
insecticidal	28
antimalarial	25
inhibitory	25
antioxidant	22
spermicidal	13
anti-protist	4

Table 2: Statistics of the number of AMPs, non-AMPs and corresponding number of clusters

	AMPs	non-AMPs	Total
Dataset	2,704	5,156	7,860

No of Clusters (>=40% identity)	464	3,264	3,728

Table 3: Division of amino acids in groups based on their physiochemical properties

Group#	1	2	3	4	5	6	7
Amino acids	A,V,G	I,F,L,P	M,S,T,Y	H,N,Q,W	K,R	D,E	C
Dipole moment (Debye)	<1.0	<1.0	(1.0 , 2.0)	(2.0 , 3.0)	>3.0	>3.0 (with opposite orientation)	<1.0
Volume ($Å^3$)	<50	>50	>50	>50	>50	>50	<50
Disulfide Bond Formation	No	No	No	No	No	No	Yes

Table 4: ROC and PR results of LOCO cross validation technique using different machine

learning models and feature representations

Model	Features	AUC-ROC (%)	AUC-PR (%)	MCC
AMAP Linear SVM	1-mer	96	94	0.80
	3-mer	94	91	0.75
AMAP Non-linear SVM	1-mer	**97**	**96**	**0.84**
	3-mer	95	94	0.79
AMAP XGBoost	1-mer	**97**	**96**	**0.84**
	3-mer	96	94	0.79
MLAMP [2]	1-mer	88	81	0.60
CAMP-R3(RF)[3]	1-mer	94	90	0.73
BLAST (Baseline)		77	72	0.27

Table 5: AUC-ROC results for different biological activities in our dataset by AMAP and

MLAMP

Biological activity	No of peptides	AMAP	MLAMP
Antibacterial	2,446	93.4	81.8
Antifungal	1,048	86.6	65.5
Anticancer	210	84.0	56.2
Antiviral	180	81.9	64.3
Anti-HIV	109	81.6	60.0
Chemotactic	57	80.2	-
Antiparasital	43	89.9	-
Antibiofilm	31	85.2	-
Insecticidal	28	87.1	-
Antimalarial	25	75.6	-
Inhibitory	25	80.8	-
Antioxidant	22	81.5	-
Spermicidal	13	73.2	-
Anti-protist	4	85.0	-

Table 6: AMAP and MLAMP scores for predicting biological activities on experimentally

verified peptides derived from LL-37 peptide

Peptide	Sequence	LC99.9 (µM)	AMAP Scores	MLAMP scores
P276 (SAAP-276)	LKRVWKAVFKLLKR YWRQLKKPVR	6.4	1.70	Non-AMP

P148	LKRVWKRVFKLLKR	12.8	1.74	0.88
(SAAP-148)	YWRQLKKPVR			
P145	LKRLYKRLAKLIKRL	12.8	1.38	0.97
(SAAP-145)	YRYLKKPVR			
P159	LKRLYKRVFRLLKR	12.8	1.32	0.88
(SAAP-159)	YYRQLRRPVR			

Table 7: AMAP and MLAMP scores for predicting biological activities on experimentally

verified peptides derived from Temporin-Ali peptide

Peptide	Sequence	IC50 (µM)	AMAP score	MLAMP scores
2	FLPIVKKLLRGLF	0.50	2.43	0.94
1	FFPIVKKLLSGLF	0.75	2.31	0.94
1C	FLPIVKKLLRKLF	1.30	2.55	0.92
2C	FFPIFGKLLRGLF	1.37	2.04	0.87
3C	FFPIVGKLLRKLF	1.39	2.41	0.94
3	VLPIVKKLLKGLF	2.01	2.96	0.95
A	FFPIVGKLLSGLF	21.1	1.83	0.90
WT	FFPIVGKLLSGLL	81.0	2.07	0.90

Table 8: AMAP and MLAMP scores for predicting biological activities on experimentally

verified peptides derived from Viral Capsid proteins

Peptide	Protein-derived Sequence	MIC(µM)				AMAP score	MLAMP score
		S. aureus	MRSA	*E. coli*	*P. aeruginosa*		

vCPP 0275	KKRYKKKYKA YKPYKKKKKF	25-50	50	12.5	100	2.07	Non-AMP
vCPP 0769	RRLTLRQLLGL GSRRRRRSR	3.13	3.13	25	3.13	1.36	0.88
vCPP 2319	WRRRYRRWRR RRRWRRRPRR	1.56	1.56	3.13	3.13	1.20	Non-AMP
vCPP 0417	SPRRRTPSPRR RRSQSPRRR	>100	>100	25	100	0.56	0.81
vCPP 1779	GRRGPRRANQ NGTRRRRRRT	>100	>100	25	25	0.54	0.88
vCPP 0667	RPRRRATTRRR ITTGTRRRR	50	100	12.5	25	0.24	0.9

Table 9: Reported MICs of M4, M5 and M6 against various bacteria

Species and strain	MIC (µg/ml)		
	M4	M5	M6
Escherichia coli ATCC 25922	128	16	8
Escherichia coli W99FI0077	16	128	8
Pseudomonas aeruginosa ATCC 27853	32	16	4
Pseudomonas aeruginosa 885149	64	32	8
Pseudomonas aeruginosa 891	64	16	8
Klebsiella pneumoniae W99FI0057	64	>128	4
Staphylococcus aureus ATCC 25923	64	128	>128

	Staphylococcus aureus MIU-68A	>128	128	128

Table 10: AMAP and MLAMP scores for predicting biological activities on experimentally

verified peptides active against *E.coli*

Peptide	Sequence	AMAP score	MLAMP score
M6	QKKIRVRLSA	0.66	0.94
M5	KIRVRLSA	0.56	0.97
M4	QAKIRVRLSA	-0.20	0.92
WT	QEKIRVRLSA	-0.70	0.91

Table 11: AMAP and MLAMP scores for predicting biological activities on experimentally

verified antifungal peptides

Peptide	Sequence	MIC (μM)	AMAP scores
BP33	LKLFKKILKVL	0.3-0.6	1.26
BP16	KKLFKKILKKL	0.6-1.2	1.72
BP76	KKLFKKILKFL	0.6-1.2	1.53
BP15	KKLFKKILKVL	0.6-1.2	1.49
BP20	WKLFKKILKYL	0.6-1.2	1.20
BP17	WKLFKKILKKL	1.2-2.5	1.47
BP13	FKLFKKILKVL	1.2-2.5	1.26
BP19	WKLFKKILKFL	1.2-2.5	1.24
BP14	YKLFKKILKVL	1.2-2.5	1.22
BP18	WKLFKKILKWL	1.2-2.5	1.16
Pep3	WKLFKKILKVL	2.5-5.0	1.18

Table 12: Reported MICs of Cp1 and Melittin peptides against various bacteria

	MIC (μM)	
Peptides	**Cp1**	**Melittin**
E. coli ATCC 25922	64	1
E. coli UB1005	128	2
Salmonella pullorum C7913	256	8
Salmonella enterica subsp enterica CMCC 50071	256	2
Staphylococcus aureus ATCC 29213	640	2
L. monocytogenes CMCC 54004	64	1

Table 13: AMAP and MLAMP scores for predicting biological activities on experimentally

verified Cp1 and Melittin peptides

Peptide	Sequence	AMAP scores	MLAMP scores
Cp1	LRLKKYKVPQL	0.96	Non-AMP
Melittin	GIGAVLKVLTTGLPALISWIKRKRQQ	1.36	0.98

Table 14: Reported MICs of peptides and scores generated by AMAP and MLAMP for

predicting biological activities on experimentally verified peptides derived from the

Acinetobacter baumannii phage endolysin LysAB2

Peptide	**Sequence**	MIC (μM)			**AMAP score**	**MLAMP score**
		A. baumannii ATCC17978	*A. baumannii* ATCC19606	colistin-susceptible MDRAB (M3237)		

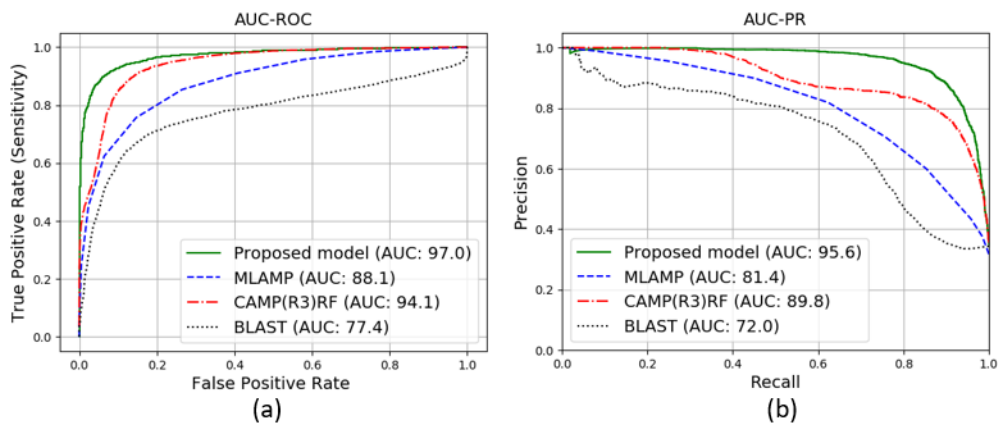| LysAB2 P0	NPEKALEPLIAI QIAIKGMLNGW FTGVGFRRKR	64	64	64	0.09	0.97
LysAB2 P1	EKALEKLIAIQK AIKGMLNGWFT GVGFRRKR	8	8	8	1.57	Non-AMP
LysAB2 P2	EKALEKLIAIQK AIKGMLAGWFT GVGARRKR	16	16	16	1.58	Non-AMP
LysAB2 P3	NPEKALEKLIAI QKAIKGMLNG WFTGVGFRRKR	4	8	16	1.22	Non-AMP

Fig 1: (a) ROC curves on our dataset by our proposed model and other models in comparison, (b) PR curves on our dataset by our proposed model and other models in comparison
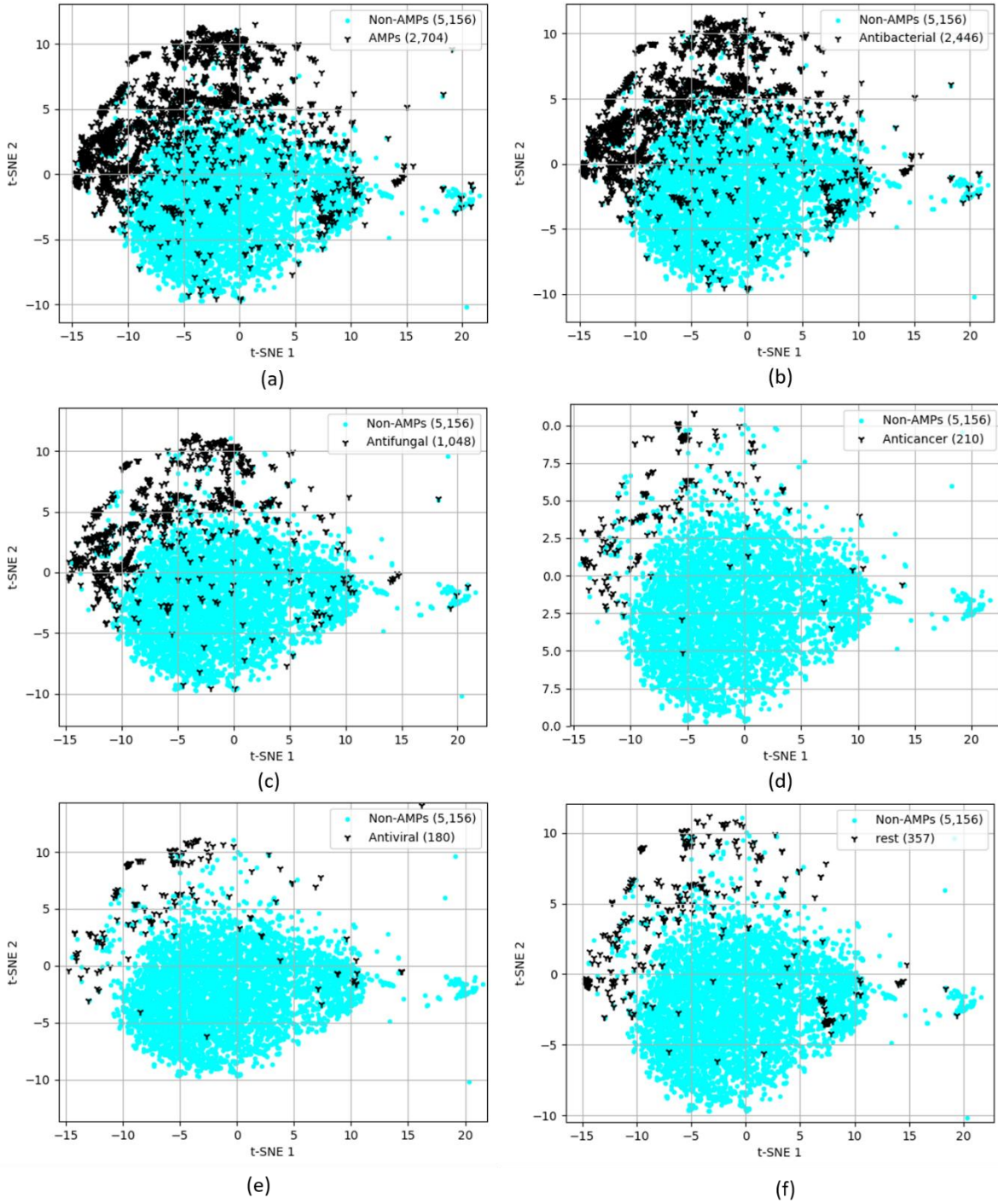
Fig 2: Scatter plots of t-SNE 2-dimensional data of AMPs and non-AMPs. The numbers in parenthesis in the legend are the number of examples of each class. (a) plot of Non-AMPs and

AMPs (b) plot of Non-AMPs and Antibacterial peptides (c) plot of non-AMPs and Antifungal peptides (d) plot of non-AMPs and Anti-cancerous peptides (e) plot of non-AMPs and Antiviral peptides (f) plot of non-AMPs and biologically active peptides from rest of the classes.
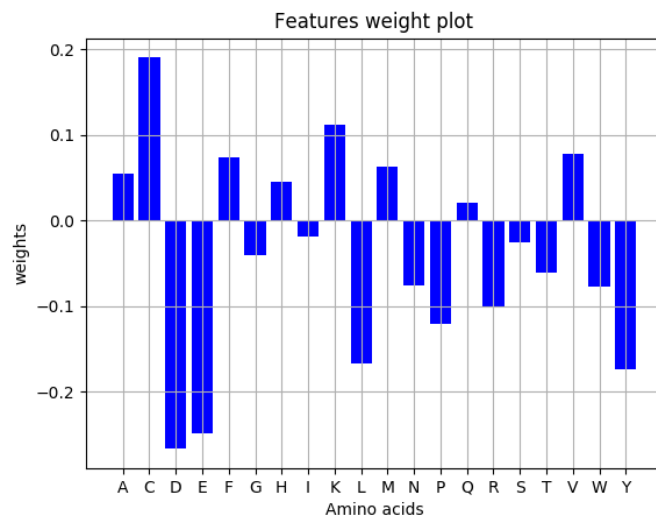
Fig 3: Plot of weight vector of Linear SVM corresponding to different amino acids in the composition feature space
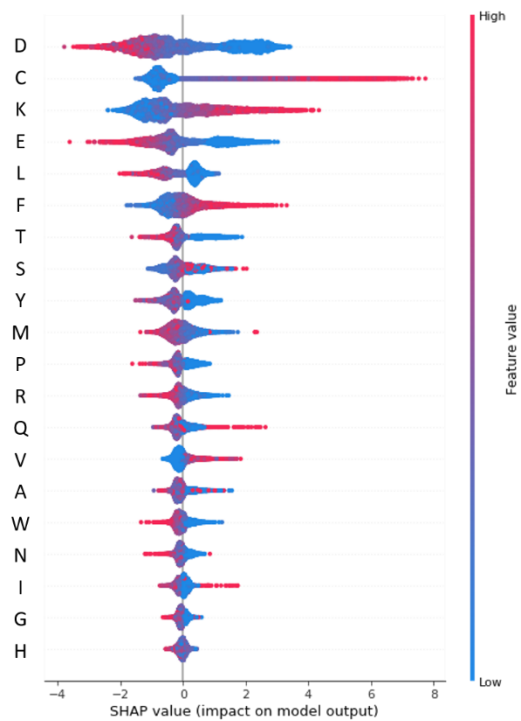
Fig 4: SHAP analysis of amino acid composition for prediction of antimicrobial activity of

peptides
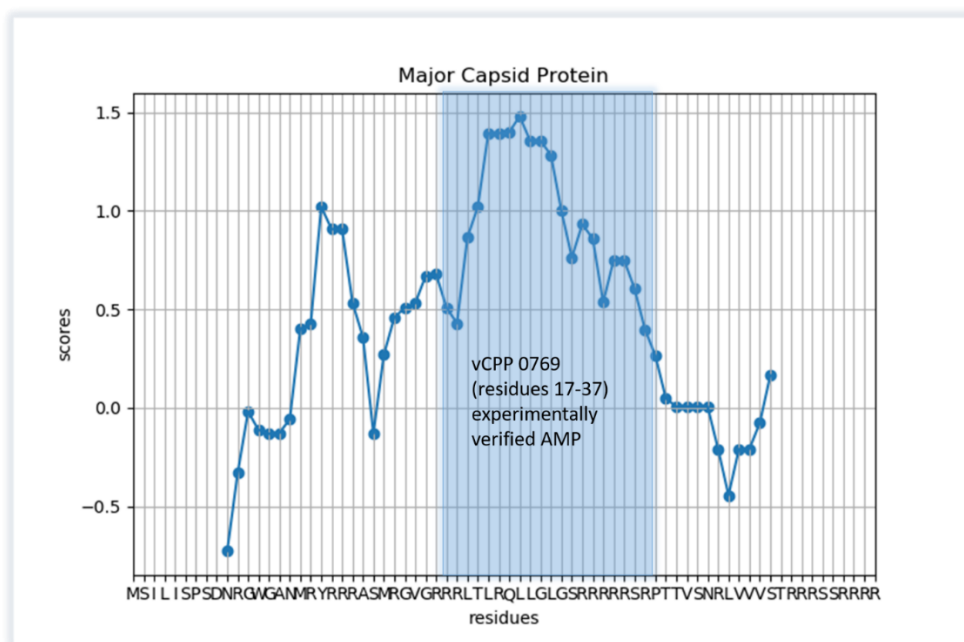
Fig 5: AMP scores of sliding sequence window of length 20 on Capsid protein

Fig 6: User interface of the web server for predicting <mark>biologically active</mark> peptides.