OXFORD

Briefings in Bioinformatics, 00(00), 2021, 1–11

doi: 10.1093/bib/bbaa423
Problem Solving Protocol

Identifying anti-coronavirus peptides by incorporating different negative datasets and imbalanced learning strategies

Yuxuan Pang†, Zhuo Wang†, Jhih-Hua Jhong and Tzong-Yi Lee

Corresponding author: Tzong-Yi Lee, Warshel Institute for Computational Biology, The Chinese University of Hong Kong, 2001 Longxiang Road, Shenzhen 518172, P.R. China. Tel: +86-755-23519551; E-mail: leetzongyi@cuhk.edu.cn
†These authors contributed equally to this work.

Abstract

As the current worldwide outbreaks of the SARS-CoV-2, it is urgently needed to develop effective therapeutic agents for inhibiting the pathogens or treating the related diseases. Antimicrobial peptides (AMP) with functional activity against coronavirus could be a considerable solution, yet there is no research for identifying anti-coronavirus (anti-CoV) peptides with the computational approach. In this study, we first investigated the physiochemical and compositional properties of the collected anti-CoV peptides by comparing against three other negative sets: antivirus peptides without anti-CoV function (antivirus), regular AMP without antivirus functions (non-AVP) and peptides without antimicrobial functions (non-AMP). Then, we established classifiers for identifying anti-CoV peptides between different negative sets based on random forest. Imbalanced learning strategies were adopted due to the severe class-imbalance within the datasets. The geometric mean of the sensitivity and specificity (*GMean*) under the identification from antivirus, non-AVP and non-AMP reaches 83.07%, 85.51% and 98.82%, respectively. Then, to pursue identifying anti-CoV peptides from broad-spectrum peptides, we designed a double-stages classifier based on the collected datasets. In the first stage, the classifier characterizes AMPs from regular peptides. It achieves an area under the receiver operating curve (AUCROC) value of 97.31%. The second stage is to identify the anti-CoV peptides between the combined negatives of other AMPs. Here, the *GMean* of evaluation on the independent test set is 79.42%. The proposed approach is considered as an applicable scheme for assisting the development of novel anti-CoV peptides. The datasets and source codes used in this study are available at https://github.com/poncey/PreAntiCoV.

Key words: imbalanced learning; machine learning; Coronavirus; antimicrobial peptides

Introduction

Since the outbreaks of the SARS-CoV-2 in 2019, people have been in a long and laborious struggle with the severe pandemic [52]. Till now, there is scarcely any effective treatment for the novel coronavirus. Antimicrobial peptides (AMPs) are a family of short-amino acid sequences (usually $<$ 100 residues) which have potent effects for inhibiting different kinds of pathogens [33]. The broad-spectrum inhibition of AMPs mainly comes from their mechanism of actions, which interact with and cause the disorders of the microbes' envelopes. It has been reported that AMPs are also considered as pivotal therapeutic agents against coronaviruses. For example, [55] reported the P9, a

Yuxuan Pang is inWarshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen, P.R. China, and also in the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, P.R. China.
Zhuo Wang is in Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen, P.R. China.
Jhih-Hua Jhong is in Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen, P.R. China, and also in the Department of Computer Science and Engineering, Yuan Ze University, Taoyuan, Taiwan.
Tzong-Yi Lee is in Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen, P.R. China and also with the School of Life and Health Sciences, The Chinese University of Hong Kong, Shenzhen, P.R. China.
Submitted: 27 April 2020; **Received (in revised form):** 30 July 2020

subsequence derived from mouse β-defensin, exhibits potent activity to inhibit the SARS-CoV and MERS-CoV. Another antiviral peptide named Mucroporin-M1, which is an analogue peptide of Mucroporin from the venom of scorpion, also manifest inhibition against SARS-CoV [28]. Lu *et al*. [31] discovered that HR2P, a peptide derived from MERS spike protein, can effectively inhibit MERS-CoV on both fusion and replication. Besides, certain AMPs, such as RTD-1, can alleviate the syndromes related to SARS-CoV by regulating the gene expression levels associated with the innate and adaptive immunity toward viral antigens [51].

Numerous researches have been devoted to identifying or designing novel AMPs with different functional activities and intense selectivity. Computational design can accelerate the procedure of developing novel AMPs by integrating databases, computational tools and machine learning [4]. Several databases have been developed to store the records of experimental-validated AMPs. For instance, the well-known APD3 database [48] contains 2747 AMP records and their functional activities, amino sequences, nomenclature, peptide classification section. Several recent AMP repositories, such as DRAMP [23] or dbAMP [22], are devoted to establishing high-throughput storages of AMPs data with investigating functional activities and physiochemical properties of massive AMP collections. Besides, certain databases or datasets like CancerPPD [44], AntiFP [1] focus on collecting AMPs with particular functional activities. Notably, the AVPdb [40] dedicated to collect the experimentally verified antivirus peptides and their functional activity for inhibition of different virus strains, such as influenza, hepatitis C virus and SARS-CoV. Based on the AMP records provided by the databases, various computer-aided automatic tools for identifying different functional AMPs have been developed. For example, the AVPpred [43] made the first attempt to predict the antivirus peptides with amino acid composition (AAC), physicochemical features and support vector machine. The AntiCP [2] is developed for predicting novel peptides with anticancer functions. The iAMPpred [35] utilizes multiple compositional and physiochemical peptide descriptors and support vector machines for predicting the functional activities of AMP, including antibacteria, antivirus and antifungus. The iAMP-2L [53] presented a two-level multi-label classifier for not only classifying the AMPs but also identifying their types of functional activities toward different pathogens including bacteria, fungus, virus, cancer cells and HIV. The Antimicrobial Peptide Scanner [47] is designed for recognizing antimicrobial activities (mainly antibacterial) with deep learning, which improves the performance of prediction by considering the primal sequence information and removing the dependence of domain experts. For the dbAMP database, it integrated a classifier to predict AMPs with the consideration of the related sources [12]. The AMPfun [11] made a thorough analysis and prediction of several functional peptides, especially for those who are antiparasitic or target-mammals. However, little attention has been paid to analyze and identify particular virus strains with a relatively small size of data.

Therefore, in this article, we initiated in a seek to construct a prediction scheme for identifying novel anti-coronavirus (anti-CoV) peptides. It is based on investigations with several individual classifications that tend to distinguish the anti-CoV peptides from different sets of peptides with broader functional categories, respectively. They are the set of antivirus peptides without anti-CoV activity (antivirus), the set of AMPs without antivirus activities (non-AVP), and the set of peptides without antimicrobial activities (non-AMP). We also investigated on identifying anti-CoV peptides from the combined set of different functional peptides above. There were only a few peptide sequences found to be active against coronavirus, which causes imbalance affecting the performance of prediction. Hence, we employed imbalanced learning strategies for inspecting the improvement to the prediction outcome. Based on the above works, a double-stages prediction scheme is proposed for identifying anti-CoV peptides from a broad-spectrum peptide set. The result shows that imbalanced learning can handle the identification of anti-CoV peptide. The proposed approach provides a solution for identifying peptide with dedicated antimicrobial function on an imbalanced dataset with relatively insufficient positive data.

Materials and Methodology

Data collection

As stated above, the collected data can be divided into four sets according to their functionality: anti-CoV, antivirus, non-AVP and non-AMP. For the anti-CoV peptide set, 137 sequence records were collected. One hundred peptides are experimentally validated data, from which 99 sequence records come from the AVPdb database [40]. The validated dataset includes reported potent coronavirus inhibitors, such as P9, Mucroporin-M1 and HR2P. The rest of the sequence records are the putative functional peptides against coronavirus [38]. Antivirus set (1999 sequence records) and non-AVP set (5217 sequence records) were obtained from several databases or datasets (AVPdb [40], dbAMP [22], DRAMP [23], CancerPPD [44], AntiFP [1]), with excluding the sequences that are redundant with anti-CoV set. The collection of the non-AMP dataset is prepared by following the similar procedures from [12, 53]. We first extracted protein sequences by filtering the annotations of 'membrane', 'toxic,' 'secretory,' 'defensive,' 'antibiotic,' 'anticancer,' 'antiviral' and 'antifungal' properties from Uniprot [14]. To reduce the amount of non-antimicrobial sequences as well as increase their identities, we adopted the CD-HIT [29] to remove the redundancy and sequence homology with 40% threshold. Finally, the non-AMP dataset is consist of 4979 sequence records. In this article, we only considered short peptides with sequence length less or equal to 100.

Sequence encoding and peptide descriptor analysis

To pursue a comprehensive analysis of the discrepancy between the peptide sets with different functional activities, the AAC, dipeptide composition (DiC), the composition of k-spaced amino acid group pairs (CKSAAGP), pseudo amino acid composition (PAAC) and physicochemical features (PHYC) comprised our entire peptide descriptors.

Amino acid composition

AAC is a simple peptide descriptor with 20-dimensions. Each dimension denotes the normalized occurrence of a specific amino acid in the peptide sequence. Take the amino acid sequence 'KTCENLADTFRGPCFATSNC' as an example, the normalized occurrence of alanine (A) is (Number of A)/(Sequence Length) = 0.1. At last, all the normalized occurrences of the amino acid residues were taken as the AAC descriptor.

Dipeptide composition

DiC expands the thoughts of AAC. It calculates the normalized occurrences of paired amino acids. Hence, the dimension of

DiC descriptor is 20 × 20 = 400. The example is to retrieve the occurrence of 'LI' from the sequence 'LFRLIKSLIKRLVSAFK, which is (Number of LI) / (Sequence length − 1) = 0.125. At last, the DiC descriptor takes normalized occurrences of all paired amino acids.

Composition of k-spaced amino acid group pairs

In order to make the feature set abundant, we adopted the CKSAAGP descriptor, which is a further extension from the DiC. The descriptor is a modification of the composition of k-spaced amino acid pairs (CKSAAP), which was adopted in several studies of protein prediction [8, 45] as an effective descriptor to represent the short motifs of the peptide sequence. At first, 20 amino acid residues are categorized into five groups by their physicochemical properties: aliphatic, aromatic, positive-charged, negative-charged and uncharged residues. For each of the $5^2 = 25$ amino acid pairs with grouped annotations, the normalized occurrences of the pairs were separated by k-residues. That is, for example, 'aliphatic.X.X.aromatic,' where the 'X' denotes any residues is the occurrence of 'aliphatic & aromatic' two-spaced amino acid pairs. Then, for a peptide with length L, if the k-spaced residue pair appears n times in the peptide, the occurrence is n / (L — (k + 1)). We chose k = 2 due to the restriction of the shortest peptide length. Finally, the occurrences of 0-spaced, 1-spaced, 2-spaced amino acid group pairs are calculated as the CKSAAGP descriptor with 25 × 3 = 75 dimensions.

Pseudo amino acid composition

PAAC [10] is claimed as an effective peptide descriptor for resolving many proteins/amino acid sequences related problems [15, 35, 42, 50]. The regular AAC or DiC barely consider the sequence-order information. PAAC improves the AAC by introducing a set of discrete factors for handling the sequence order properties. Detailed mathematical description for calculating PAAC features can be found at [24]. There are two key parameters for PAAC: the discrete counted-rank correlation factor λ, and the weight factor ω. The resulted PAAC descriptor has $20 + \lambda$ dimensions. For larger ω, the descriptor is more inclined to the sequence order effect. For the restriction of sequence length and ensure the diversity of different descriptors, we set λ = 4 and ω = 0.4 as the parameters of generating PAAC descriptor.

Physicochemical features

We selected eight physiochemical peptide features that are closely related to the antimicrobial/transmembrane functions, including isoelectric point (IEP), net charge [5, 6], hydrophobicity [20], hydrophobic moment [16], transmembrane propensity [54], Boman index [7], aliphatic index [21] and alpha helical propensity [27].

In this study, feature selection is based on the hypothesis test. There is a severe imbalance between the positive and negative samples within the prediction, which makes the parametric tests, such as t-test, not employable [41]. Hence, we adopted the Wilcoxon rank-sum test [18]. It is a non-parametric test that can handle the imbalance between two groups. For investigating the importance of different peptide descriptors, as well as the discrepancy between anti-CoV peptides to different negative sets, we leave only the significant descriptors with a P-value less than 0.05 given by the rank-sum test.

Model construction with imbalanced learning

Insufficient samples of anti-CoV peptides lead to a dreadful class-imbalance within the dataset, which affects the performance of classifiers. Models without consideration of imbalanced data gravitate toward the majority class and take little notice of the minority. Thus, we introduced imbalanced learning [19] to this study. We compared the performance of two different imbalanced strategies, named NearMiss under-sampling and balanced random forest.

NearMiss under-sampling

A typical scheme for imbalanced learning is called under-sampling, which is to equilibrate the size of the dataset by removing part of samples from the majority class. The difference between the various under-sampling strategies is in how they remove majority samples. NearMiss [34] is a family of under-sampling strategies which choose the samples of majority class based on their distance to the minority class. There are three versions for NearMiss. The *NearMiss-1* selects the samples from the majority class that has the lowest average distance to several nearest-neighbors. In contrast, the *NearMiss-2* selects the majority samples with the smallest average distance to some farthest-neighbors. In *NearMiss-3*, the algorithm first retains the p-nearest-neighbors which belongs to the majority for each sample of the minority class. Then, for those who retained, they are furthered selected for which have the largest average distance to the q-nearest-neighbors belonging to the minority. We illustrated an example of different NearMiss strategies in Supplementary Figure S1. In this study, we adopted *NearMiss-3* approach since the version would probably be less affected by interference within the dataset due to the double-steps selection. Both p and q are set to 3 for the *NearMiss-3* algorithm. After the NearMiss, predictors are constructed by the resampled datasets with incorporating the conventional random forest classifier.

Balanced random forest

Random forest [25] is an ensemble learning algorithm derived from bootstrap aggregating. It can handle the classification tasks. The bootstrap aggregating, utilizing the thoughts of resampling, select a certain number of samples from a given dataset with replacement. For each time of performing bootstrap aggregating, the selected samples are used for establishing a base decision-tree classifier. Finally, the classifier with the best performance is chosen by majority vote. Based on the bootstrap aggregating, the random forest also randomly selects part attributes of the nodes in the decision tree as subsets, and then select the best attributes from the subsets for classification. Since the random forest classifier is established based on combining several base classifiers, it has considerable robustness to the interference. As increasing the number of base estimators, classifier tends to converge to a low prediction error. The balanced random forest makes modification by performing random under-sampling at bootstrap aggregating. The conventional under-sampling only considers a particular strategy, for which the different distribution of the positive/negative samples could affect the results. For the balanced random forest, it could combine multiple under-sampling results. Hence, compared to the conventional under-sampling, it may improve the performance of prediction by ensemble learning.

With the collected data, for each peptide set with a dedicated function, 70% is divided for training the predictors while

Table 1. Summary of data within the investigated classifications.

Positive class	Negative class	Training set		Test set	
		(+)	(−)	(+)	(−)
Anti-CoV	Antivirus	95	1399	42	600
	non-AVP		3746		1566
	non-AMP		3485		1494
	Antivirus, non-AVP, non-AMP		8535		3660
	Antivirus, non-AVP		5050		2166
Anti-CoV, antivirus, non-AVP	non-AMP	5145	3485	2208	1494

remaining are utilized for evaluating the prediction outcome. A summary of the positive and negative dataset for all classification investigated in this article is given in Table 1. We adopted model selection with 5-fold cross-validation on the training set to obtain the best model with the highest sensitivity.

Evaluation

Under the circumstance of imbalanced learning, the regular error rate is not capable of evaluating the performance. We use Sensitivity (*SEN*), Specificity (*SPEC*), F2-measure (F_2), Geometric mean (*GMean*) and Matthew's correlation coefficient (MCC) to evaluate the prediction results. Denote TP as the number of true positives, TN as the number of true negatives, FP as the false positives, FN as the false negatives, the metrics mentioned above are defined as follows:

$$SEN = \frac{TP}{TP + FN} \tag{1}$$

$$SPEC = \frac{TN}{TN + FP} \tag{2}$$

$$F_2 = 5 \times \frac{PREC \times SEN}{4 \times PREC + SEN} \tag{3}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

$$GMean = \sqrt{SEN \times SPEC} \tag{5}$$

where the precision score (PREC) is defined as:

$$PREC = \frac{TP}{TP + FP} \tag{6}$$

In this study, the MCC and F2-measure are adopted for comparison of different imbalanced learning strategies within the same investigated task. However, those measurements consider the negative and positive samples in the same or near weights, which result in a curse of bias evaluation in the severe label-imbalance circumstances. For measuring the pragmatic performance of identifying anti-CoV with different negative sets, sensitivity, specificity and geometric mean are more precise in consideration of the unbiased evaluation.

There is no apparent class-imbalanced issue within the last classification in Table 1. Hence, we also give the regular accuracy (ACC), the precision, and the MCC for evaluation. We also plot the receiver operating curves (ROC) [17, 49] and calculate the area under the curves (AUCROC) for assessing the performance of the balanced prediction and making comparison of the state-of-art models.

Implementation

The analysis and model construction were implemented at a server with CentOS Linux 7.6 and python 3.6.10. Some of the methods were completed based on integrating computational biology or machine learning packages, which are mentioned as follows. For peptide sequence encoding, the calculation of PAAC, CKSAAGP descriptors were executed by the codes within the iFeature [9]; in the eight physiochemical features, IEP and net charge were calculated by the Biopython 1.75 [13] and the remaining were implemented by the modlAMP 4.1.2 [37]. For the classification, we applied the random forest classifier by the scikit-learn 0.22.1 [39]; the two imbalanced learning strategies were executed with imbalanced learn 0.6.2 [26].

Results

Investigations on sequence-based descriptors of different functional AMPs

The distribution of amino acid sequence length among four different sets of peptides is shown in Figure 1. AMPs tend to have shorter amino acid sequences than non-AMPs. There is no significant difference in the length distribution among AMPs, yet the antivirus peptide sequences (not include anti-CoV) are concentrated in < 20 amino acids length. Besides, the sources and targets of the collected anti-CoV sequence records are shown in Supplementary Figure S2. Dimensional reduction with t-SNE [46] is performed to inspect the distribution under different source or target domains.

The bar chart of mean AACs among different sets of peptides is given in Figure 2. Amino acids are categorized into five groups by their chemical properties [30]. It is illustrated that the acidic amino acids of non-AMPs, including aspartic acid (D) and glutamic acid (E), are more than those of AMPs. For alkaline amino acids, including arginine (R), histidine (H), and lysine (K), those of the non-AMPs are less than those of the antiviruses and non-AVPs, but more than the anti-CoVs. The asparagine (N) and arginine (R) of anti-CoVs differ slightly from those of the others.

Table 2. Summary (mean ± SD) of eight physiochemical features, significant features apart from anti-CoV are marked as '*'

PHYC features	Anti-CoV	Antivirus	Non-AVP	Non-AMP
IEP	7.03 ± 2.421	7.75 ± 2.703*	9.34 ± 1.882*	7.60 ± 2.374*
Net charge	−0.04 ± 3.312	0.98 ± 3.074*	4.14 ± 4.419*	1.01 ± 6.771
Hydrophobic moment	0.25 ± 0.137	0.25 ± 0.169	0.26 ± 0.185	0.12 ± 0.065*
Hydrophobicity	−0.06 ± 0.404	−0.04 ± 0.537	−0.01 ± 0.458	0.11 ± 0.332*
Transmembrane propensity	−0.43 ± 0.374	−0.44 ± 0.494	−0.44 ± 0.476	−0.55 ± 0.305*
Alpha helical propensity	1.02 ± 0.068	1.03 ± 0.083	1.02 ± 0.079	1.04 ± 0.049*
Aliphatic index	90.47 ± 34.571	87.66 ± 45.079	86.72 ± 46.392*	82.09 ± 23.219*
Boman index	1.43 ± 1.120	1.47 ± 1.714	1.29 ± 1.653	1.82 ± 0.874*

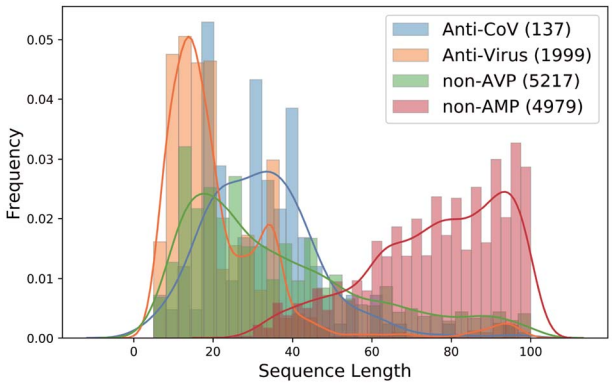

Figure 1. The distribution of amino acid sequence length among four different sets of peptides.

We made three investigations that distinguish anti-CoV peptides from a peptide set with dedicated function (anti-CoV versus antivirus, anti-CoV versus non-AVP, anti-CoV versus non-AMP). For each investigation, we selected the significant features between anti-CoVs and each of the negative sets by 0.05 *P*-value threshold of the Wilcoxon rank-sum test. Feature selections are conducted on the training sets only. The summary of physiochemical features is presented in Table 2. Supplementary Table S1 tabulated the *P*-values of physiochemical features under different investigated classifications. Isoelectric point and net charge of anti-CoVs tend to differ from those of the antiviruses and non-AVPs. The aliphatic index also has a difference between the anti-CoVs and non-AVPs. All physiochemical features except for net charge have a significant difference between anti-CoVs and non-AMPs. The number of distinct physiochemical features conforms to the intuition of their discrepancies: peptides with anti-CoV function shared similar physiochemical properties with antiviruses/non-AVP peptides. The anti-CoV peptides tend to have a similar mode of actions to the antivirus peptides or regular AMPs.

The summary of the feature importance of different peptide descriptors under the rank-sum results of three investigations is illustrated in Figure 3. The number of selected features within different descriptor categories are tabulated in Supplementary Table S2. The negative-logarithmic *P*-value can represent the importance of each feature. It is observed that the negative-log *P*-value under anti-CoV versus non-AMP can reach up to $>$ 40. The same indexes under anti-CoV versus non-AVP and anti-CoV versus antivirus are $<$ 20 and $<$ 15. The overall negative-logarithmic *P*-value also confirms their different discrepancies to the anti-CoV.

The peptide descriptors present different importance and influence on each of the different investigations. It is observed that the relative proportion of selected PAAC after the rank-sum ranks high under each of the three comparisons, while many selected PAAC features have high negative-log p-value under the rank-sum of antivirus and non-AMP. For physiochemical features, relative proportion after the rank-sum is low for the antivirus and non-AVP, but high in the case of non-AMP. Nonetheless, two of the physiochemical features, isoelectric point and net charge rank the highest negative-log *P*-value in the non-AVP case. The DiC has the lowest relative ratio after the feature selection, yet the remaining features still compose the most in the descriptor set. It is observed that the AAC might be an informative descriptor for distinguishing the anti-CoV peptide from AMPs without antivirus function since the relative proportion of the descriptor ranks the highest. Besides, the number of selected AAC features ranks under the non-AVP is more than that of the antivirus and non-AMP. For CKSAAGP, relative ratio after rank-sum is not much as the AAC and PAAC under each investigation, yet part of the features still gives a high negative-log *P*-values, especially within the non-AVP case.

Performance analysis of identifying anti-CoV peptides from different peptide sets

One of the strategies for identifying functional peptides is to directly distinguish it from a set of peptides with other dedicated characteristics. Hence, the classifier should be built on a positive dataset comprised of the anti-CoV peptides and a negative dataset. The critical issue is to decide an appropriate negative dataset, which should be representative to agree with a pragmatic identification. In this section, we established several classifiers with different negative datasets for identifying anti-CoV peptides. We considered four negative datasets here: antivirus peptides without anti-CoV function, peptides with various antimicrobial functions except for antivirus (non-AVP), regular peptides without any microbial functions (non-AMP) and the combination of three together (All-Neg). Due to the class-imbalance, all the classifiers are built based on imbalanced learning algorithms. Here, we made a comparative analysis of two different imbalanced learning strategies: NearMiss undersampling and balanced random forest. Tables 3–5 present the performances of the predictions constructed by anti-CoV and three other individual sets: antivirus, non-AVP and non-AMP. We also assessed the predictions built with features selected by the rank-sum test.

Anti-CoV versus antivirus

For the identification of anti-CoV peptides from normal antivirus, the performance achieved 78.57% sensitivity at the

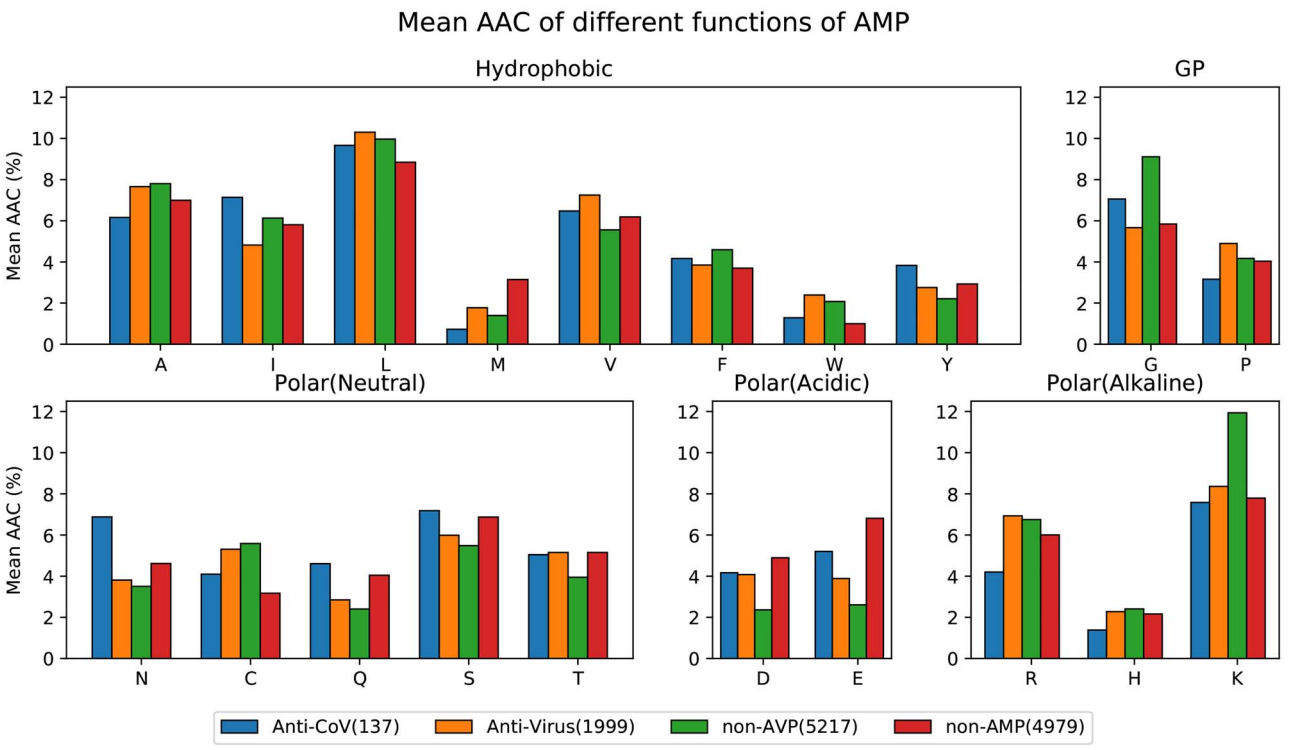

Figure 2. Mean AAC of different sets of peptides. The amino acids are categorized by their physiochemical features.

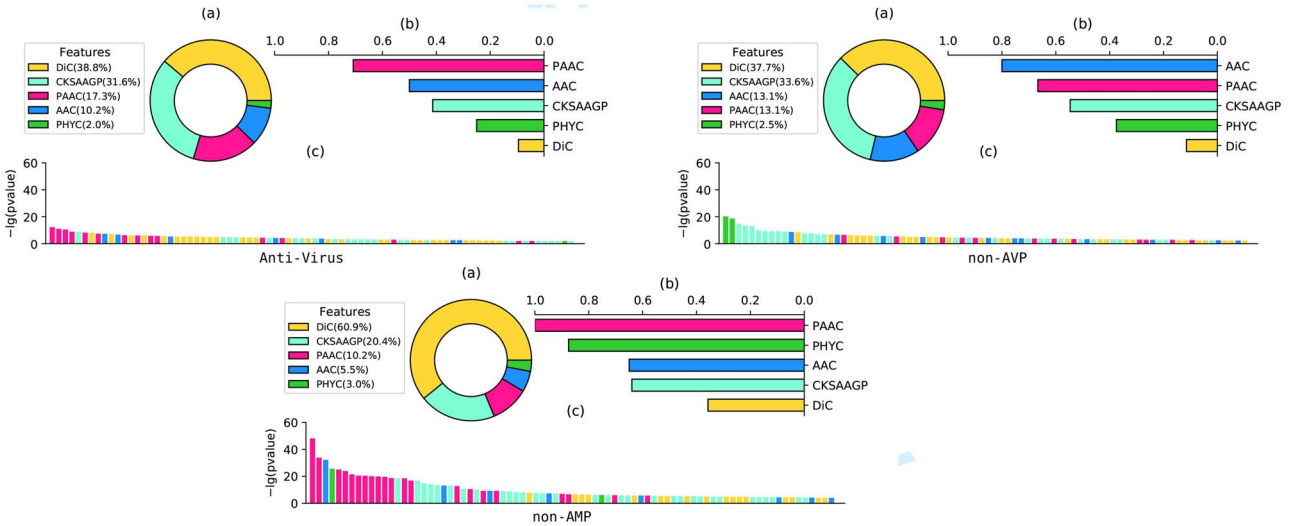

Figure 3. Summarized feature importance of three investigations. For each investigation, (**A**) is the absolute proportions of different descriptors after rank-sum; (**B**) ranks the ratio of selected features relative to their dimension of the descriptors before rank-sum (**C**) depicts the top-100 features ranked by their negative logarithm of the *P*-value.

Table 3. Performance of the predictions under anti-CoV versus antivirus

ML approach	Rank-sum	SEN(%)	SPEC(%)	GMean(%)	F_2(%)	MCC(%)
NearMiss+RF	No	80.95	57.83	68.42	37.36	19.29
	Yes	**78.57**	**87.83**	**83.07**	**60.22**	**44.22**
Balanced RF	No	78.57	84.33	81.40	55.93	39.05
	Yes	76.19	84.17	80.08	54.24	37.46

random forest classifier with NearMiss under-sampling and feature selection. For the same machine learning approach without feature selection, however, the low specificity suggests that it failed to predict the antivirus samples. The performances under whether adopting feature selection tend to be closer to each other under the balanced random forest, though their

Table 4. Performance of the predictions under anti-CoV versus non-AVP

ML approach	Rank-sum	SEN(%)	SPEC(%)	GMean(%)	F_2(%)	MCC(%)
NearMiss + RF	No	85.71	79.18	82.38	33.96	24.78
	Yes	80.95	80.97	80.96	34.00	24.40
Balanced RF	**No**	**85.71**	**85.31**	**85.51**	**41.47**	**30.49**
	Yes	76.19	84.61	80.29	36.28	25.83

Table 5. Performance of the predictions under anti-CoV versus non-AMP

ML approach	Rank-sum	SEN(%)	SPEC(%)	GMean(%)	F_2(%)	MCC(%)
NearMiss + RF	No	100	96.99	98.48	82.35	68.43
	Yes	**100**	**97.66**	**98.82**	**85.71**	**72.98**
Balanced RF	No	100	95.78	97.87	76.92	61.90
	Yes	100	94.78	97.35	72.92	57.60

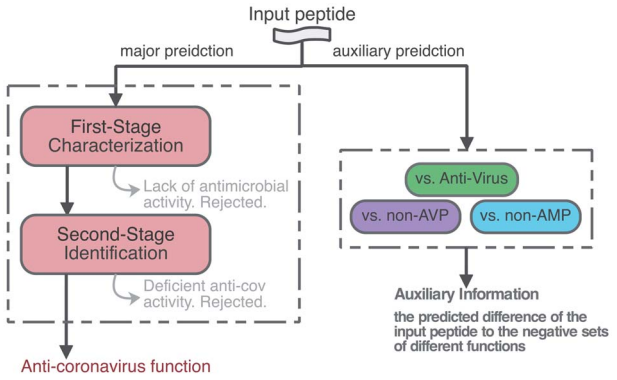

Figure 4. Framework of the proposed scheme for computational-assisted anti-CoV peptide design.

accuracies are slightly lower than the NearMiss approach. We also attempted to search for the common subsequences of the collected anti-CoV peptides by MEME [3] and compared to the common subsequences of the regular antivirus peptides. The results are shown in Supplementary Figure S3. Several functional motifs identified in Supplementary Figure S3(a) may be related to some specific anti-CoV mode of actions. Moreover, there is no similarity of obtained sequence motifs between the anti-CoV and antivirus, which suggests the feasibility of distinguishing between each other.

Anti-CoV versus non-AVP

All performance metrics achieve the best level under the balanced random forest with all the peptide descriptors. Performances of feature selection are closed to those of all-descriptors prediction with observed about 80% sensitivity, specificity, and geometric mean, which indicates that the selected features are capable of capturing the discrepancies between anti-CoV peptides and the non-antivirus AMP. Note that the F_2 and *MCC* are low here. It is because of the drastic imbalance within the dataset (42 positives, 1566 negatives). The number of negative samples increase the false negatives and affects the precision.

Anti-CoV versus non-AMP

It is observed that the specificity and geometric mean is >94% for each of the four approaches. All methods achieve 100% sensitivities, indicated that all anti-CoV peptides are successfully identified from non-AMP. The performance reached the highest level under the NearMiss with random forest and feature selection.

The result of the predictions above also concurs with the different discrepancies between the anti-CoV peptides and three other different types of peptides. Classifiers reached superior performance under the circumstance of anti-CoV and non-AMP due to the significant disparity between two classes. Classifiers could not identify the anti-CoV peptides as well from antivirus and non-AVP since the peptides share similar physiochemical properties among different sets. Besides, the overall sensitivities under anti-CoV versus antivirus are lower than those versus non-AVP, which indicates that classifiers are more capable of predicting the anti-CoV peptides from non-AVP, rather than the antivirus. It is in accordance with that anti-CoV peptides share the most physicochemical properties with antivirus peptides.

Anti-CoV versus all-Neg

The above predictors investigated are focus on identifying anti-CoV peptides from another peptide set with dedicated functions. They may be employable for some specific circumstances (e.g. identifying the capability of eliminating coronavirus for a peptide with known antivirus function). Nonetheless, they may fail to correctly predict an anti-CoV peptide from a set composed of different functional peptides. To develop a comprehensive predictor that recognizes the anti-CoV peptides from broad-spectrum amino acid sequences, we first attempt to construct the negative dataset that combined antiviral, non-AVP and non-AMP for prediction. The results of the predictions are shown in Table 6. The balanced random forest with all peptide descriptors achieves the best geometric mean, while NearMiss with random forest and feature extraction achieves the best specificity, F_2 and *MCC*. Then, we also investigated the classification errors under different subnegative classes. Subclass specificity, true negatives and false positives of NearMiss with random forest and balanced random forest are presented in Supplementary Table S3. Although the total specificity can reach >80%, all the predictors failed to classify antivirus peptides. Most of the subclass specificity of antivirus dropped to <60%, especially for the NearMiss with random forest under entire feature set, which has the highest sensitivity for predicting anti-CoV but only 32% specificity for antivirus. The failure of predicting antivirus peptides might cause by the relatively small data size of antivirus

Table 6. Performance of predictions under anti-CoV versus all-negative combined set

ML approach	Rank-sum	SEN(%)	SPEC(%)	GMean(%)	F_2(%)	MCC(%)
NearMiss + RF	No	88.10	75.98	81.82	17.07	15.73
	Yes	76.19	88.01	81.89	25.04	20.40
Balanced RF	**No**	**85.71**	**81.64**	**83.65**	**20.55**	**18.14**
	Yes	80.95	79.75	80.35	18.03	15.80

Table 7. Performance of the first-stage characterization

Rank-sum	ACC(%)	SEN(%)	SPEC(%)	PREC(%)	MCC(%)	AUCROC(%)
No	91.19	91.26	91.1	93.81	81.87	97.31
Yes	91.33	91.17	91.57	94.11	82.18	97.21

Table 8. Performance of the second-stage identification

ML approach	Rank-sum	SEN(%)	SPEC(%)	GMean(%)	F_2(%)	MCC(%)
NearMiss + RF	No	83.33	71.14	77.00	17.07	16.26
	Yes	73.81	74.28	74.05	25.04	14.86
Balanced RF	**No**	**73.81**	**85.46**	**79.42**	**20.55**	**22.27**
	Yes	71.43	80.70	75.92	18.03	17.71

compared to other negative sets, and their similarity to the anti-CoV peptides.

Moreover, we also perform all the classifications above under random forest without imbalanced learning approach (default strategy). The resulting sensitivities are shown in Supplementary Table S6. It is observed that the classifiers without imbalanced learning failed to predict the positive samples. It is the imbalanced learning strategies that relieve the error of predicting anti-CoV peptides.

Double-stages classifier for identifying anti-CoV peptides from broad-spectrum peptides

To relieve the biased results of broad-spectrum prediction, we adopted a double-stages classification scheme. The first stage is to characterize the AMPs, which use the combined set of anti-CoV, antivirus and non-AVP set as the positive dataset, non-AMP as the negative dataset. The second stage classifier identifies the anti-CoV peptides from AMPs, which use the anti-CoV as the positive set. The negative dataset in the second-stage classifier is comprised of antivirus and non-AVP.

First-stage characterization

Since the prediction task here does not suffer from class-imbalance, we adopted the random forest classier without imbalanced learning. We also evaluate the performance of the prediction under the feature selection. The results of the classifier on the test dataset are shown in Table 7. It is observed that the regular accuracy, sensitivity and specificity are >91%, precision is >93%, MCC is >81% for the random forest classifier in this task. Subclasses true positives, false negatives and sensitivities for the characterization are tabulated in Supplementary Table S4. Sensitivities of anti-CoV and antivirus reached >96%, while that of non-AVP is about 89%. The predictor has an excellent capability of characterizing the AMPs with and without feature selection. For the first-stage classification, the classifier is designed for distinguishing between regular peptides and peptides with antimicrobial functions, including anti-CoV, regular antivirus and other AMPs. Hence, we made a comparison with state-of-art AMP predictors. The proposed classifiers achieve the best performance with the highest AUCROC scores, with 97.31% under the normal (entire feature set) and 97.21% under the rank-sum. The receiver operating curves for different classifiers are given in Supplementary Figure S4(a). Supplementary Figure S4(b-e) shows the sensitivities of the anti-CoV, antivirus and non-AVP test sets and the specificities of the non-AMP test set. It is observed that the performances of dbAMP-integrated predictor [12] and AMPfun [11] bias toward positive samples and have low specificities in predicting regular peptides. The iAMP-2L [53] has high specificity but low sensitivities for positive subclasses. The APScanner [47] performs well at only non-AVP test set, considering that it mainly focuses on the prediction of antibacterial peptides. Our proposed classifiers have more balanced results toward both AMPs and regular peptides, and they are the most suitable preclassification models for this study. Besides, the paired scatter plot of the eight physiochemical features for comparing the positive/negative sets in the first-stage characterization is shown in Supplementary Figure S5. It is observed that the physiochemical properties of the AMPs are quite different from the regular peptides.

Second-stage identification

Classifiers in the second-stage identification are applied the same imbalanced learning strategies as the mentioned individual classifications. Performances under the feature selection are also evaluated. The evaluation results of classifiers on the test dataset is presented in Table 8. The sensitivity of predicting anti-CoV is about 71–84%, which are lower than those of the prediction with combined negatives. This is because of the difficulty of recognizing the anti-CoV peptides from a peptide set with similar physiochemical properties. We summarized subclass performances of the near-miss with random forest and the balanced

random forest in Supplementary Table S5. Although the near-miss with random forest classifier without feature selection has the highest 83.33% sensitivity for predicting anti-CoV, it failed to classify the antivirus. The highest overall/subclass specificity is achieved by the balanced random forest without feature selection. It tends to be more unbiased for the prediction of each class than the combined-negative classification.

Besides, we also give the resulting sensitivity without imbalanced learning approach in Supplementary Table S6. The default classifier also failed to identify the anti-CoV peptides, which implicates the necessity of employing imbalanced learning.

Conclusion

On the basis of the above works, we provide an integrated approach for pragmatically assisting the design of anti-CoV peptides, which is illustrated in Figure 4. For a given peptide, the major prediction uses the double-stages classification scheme to decide the anti-CoV function. Simultaneously, the auxiliary prediction use three investigated classifiers to measure how it is distinguished from different negative sets if considering the given peptide as an anti-CoV peptide. Moreover, an example of applying the approach for predicting anti-CoV peptides with a sequence record from the anti-CoV test set is given in Supplementary Figure S6. We also use the SHAP explainer [32] to make an attempt for interpreting the prediction result of the anti-CoV peptide characterization.

AMPs are potential therapeutic agents for treating the infections of coronavirus. Computational models for identifying anti-CoV peptides can accelerate the development of these novel drug candidates [36]. In this study, we first investigated several classifications for identifying the anti-CoV peptides from three peptide sets with different functional activities and their combination. Among them, some of the compositional and physiochemical features are highlighted as having significant difference for which might be a considerable descriptor for distinguishing the anti-CoV peptides. Besides, the descriptors can reveal the discrepancies from anti-CoV peptides to other functional amino acid sequences. Due to the drastic class-imbalance of prediction tasks, classifications are based on imbalanced learning strategies. The developed classifiers have shown their capabilities for identifying the anti-CoV peptides from certain broad categories of functional peptides, which may be applicable for some circumstances of identifying novel anti-CoV peptides. The accuracies of different classifiers are affected by the discrepancies between the different positive/negative sets. For distinguishing anti-CoV peptides from the combined datasets, classifier failed to correctly predict the regular antivirus peptides with low specificity. Then, for the purpose of making the classification capable of identifying anti-CoV peptides from broad-spectrum amino acid sequences with a balanced accuracy, we developed a double-stages classification scheme. The proposed scheme relives the prediction error of regular antivirus peptides within the negative sets, although there is a tradeoff of sacrificing certain performance at the positive prediction. In brief, each of the classification proposed in this study has its inclination of prediction. In pragmatic tasks for identifying novel anti-CoV peptides, users can adopt a specific classification scheme by their definite intention. The proposed study succeeds in an attempt to make identification for dedicated functional peptides with small size of data. Nevertheless, with developing more amounts of peptides with validated anti-CoV function, there are more prospects of machine-learning-aided identification of anti-CoV peptides or effective therapeutic agents for other particular virus strains.

Key Points

- The proposed approach investigated the physiochemical and compositional properties of anti-coronavirus peptides (anti-CoV) by comparing with other peptide sets of different functional activities: antivirus peptides without anti-coronavirus functions (antivirus); antimicrobial peptides without antivirus functions (non-AVP); and regular non-antimicrobial peptides (non-AMP).
- Based on the investigations, we conducted ML-based classification to identify the anti-coronavirus peptides from other peptides sets. Samples of anti-coronavirus peptides are relatively deficient. Hence, we adopted the imbalanced learning strategies to relieve the curse of class-imbalance.
- By integrating above works, we established a double-stages classification scheme for predicting anti-coronavirus peptides from broad-spectrum peptide sets. The results show that computational methods are capable for identifying anti-coronavirus peptides.

Funding

This work was supported by the National Natural Science Foundation of China [32070659] and the Warshel Institute of Computational Biology, The Chinese University of Hong Kong, Shenzhen, China.

Author Contributions

Y.X.P and T.Y.L carried out the concept and design of this study. J.H.J and Y.X.P were responsible for the data acquisition. Y.X.P and Z.W participated in the data analysis and drafted the manuscript. All authors contributed to manuscript revision.

Conflict of interest

The authors declare no conflict of interests.

Availability of data and materials

The datasets used and analyzed during the current study are available at https://github.com/poncey/PreAntiCoV.

References

1. Agrawal P, *et al*. In silico approach for prediction of antifungal peptides. *Front Microbiol* 2019; **9**:323–3.
2. Agrawal P, Bhagat D, Mahalwal M, *et al*. AntiCP 2.0: an updated model for predicting anticancer peptides. *Brief Bioinform* 2020;bbaa153.
3. Bailey TL, Boden M, Buske FA, *et al*. MEME suite: tools for motif discovery and searching. *Nucleic Acids Res* 2009; **37**(suppl_2): W202–8.
4. Barreto-Santamaría A, Patarroyo ME, Curtidor H. Designing and optimizing new antimicrobial peptides: all targets are not the same. *Crit Rev Clin Lab Sci* 2019; **56**(6): 351–73.
5. Bjellqvist B, Hughes GJ, Pasquali C, *et al*. The focusing positions of polypeptides in immobilized pH gradients can be

predicted from their amino acid sequences. *Electrophoresis* 1993; **14**(1): 1023–31.
6. Bjellqvist B, Basse B, Olsen E, *et al*. Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. *Electrophoresis* 1994; **15**(1): 529–39.
7. Boman H, Wade D, Boman IA, *et al*. Antibacterial and antimalarial properties of peptides that are cecropin-melittin hybrids. *FEBS Lett* 1989; **259**(1): 103–6.
8. Chen Z, Zhou Y, Zhang Z, *et al*. Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features. *Brief Bioinform* 2015; **16**(4): 640–57.
9. Chen Z, Zhao P, Li F, *et al*. iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 2018; **34**(14): 2499–502.
10. Chou K-C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Struct, Funct, Bioinformat* 2001; **43**(3): 246–55.
11. Chung C-R, *et al*. Characterization and identification of antimicrobial peptides with different functional activities. *Brief Bioinform* 2019; **21**(3): 1098–114.
12. Chung C-R, Jhong JH, Wang Z, *et al*. Characterization and identification of natural antimicrobial peptides on different organisms. *Int J Mol Sci* 2020; **21**(3): 986.
13. Cock PJA, Antao T, Chang JT, *et al*. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009; **25**(11): 1422–3.
14. Consortium TU. Uniprot: the universal protein knowledgebase. *Nucleic Acids Res* 2017; **45**:D1158–69.
15. Ding Y-S, Zhang T. Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. *Pattern Recognit Lett* 2008; **29**:1887–92.
16. Eisenberg D, Weiss RM, Terwilliger TC. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature* 1982; **299**(5881): 371–4.
17. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 1982; **143**(1): 29–36.
18. Haynes W, *et al*. Wilcoxon rank sum test. *Ency Syst Biol* 2013;2354–5.
19. He H, Garcia E. Learning from imbalanced data. *IEEE Trans Knowl Data Eng* 2009; **21**(9): 1263–84.
20. Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. *Proc Natl Acad Sci* 1981; **78**(6): 3824–8.
21. Ikai A. Thermostability and aliphatic index of globular proteins. *J Biochem* 1980; **88**(6): 1895–8.
22. Jhong J-H, *et al*. dbAMP: an integrated resource for exploring antimicrobial peptides with functional activities and physicochemical properties on transcriptome and proteome data. *Nucleic Acids Res* 2018; **47**:D1285–97.
23. Kang X, Dong F, Shi C, *et al*. Dramp 2.0, an updated data repository of antimicrobial peptides. *Scientific Data* 2019; **6**(1): 148–8.
24. Kuo-Chen C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr Proteomics* 2009; **6**(4): 262–74.
25. Leo B. Random forests. *Mach Learn* 2001; **45**(1): 5–32.
26. Lema TG, *et al*. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res* 2017; **18**(1): 559–63.
27. Levitt M. Conformational preferences of amino acids in globular proteins. *Biochemistry* 1978; **17**(20): 4277–85.
28. Li Q, Zhao Z, Zhou D, *et al*. Virucidal activity of a scorpion venom peptide variant mucroporin-m1 against measles, SARS-COV and influenza H5N1 viruses. *Peptides* 2011; **32**(7): 1518–25.
29. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006; **22**(13): 1658–9.
30. Lide DR. *CRC Handbook of Chemistry and Physics*, 1991.
31. Lu L, Liu Q, Zhu Y, *et al*. Structure-based discovery of Middle East respiratory syndrome coronavirus fusion inhibitor. *Nat Commun* 2014; **5**(1): 3067.
32. Lundberg SM, *et al*. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2020; **2**(1): 2522–5839.
33. Mahlapuu M, *et al*. Antimicrobial peptides: an emerging category of therapeutic agents. *Front Cell Infect Microbiol* 2016; **6**:194–4.
34. Mani, I. and Zhang, I. (2003). KNN approach to unbalanced data distributions: a case study involving information extraction. *Proceedings of International Conference on Machine Learning (ICML' 2003) workshop on learning from imbalanced datasets*.
35. Meher PK, Sahu TK, Saini V, *et al*. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general pseAAC. *Sci Rep* 2017; **7**(1): 42362–2.
36. Memariani H, Memariani M. Therapeutic and prophylactic potential of antimicrobial peptides against coronaviruses. *Ir J Med Sci* 1971;2020–0.
37. Müller AT, Gabernet G, Hiss JA, *et al*. modlAMP: python for antimicrobial peptides. *Bioinformatics* 2017; **33**(17): 2753–5.
38. Mustafa S, *et al*. Peptide-protein interaction studies of antimicrobial peptides targeting middle east respiratory syndrome coronavirus spike protein: an in silico approach. *Adv Bioinformat* 2019; **2019**:1–16.
39. Pedregosa F, *et al*. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011; **12**:2825–30.
40. Qureshi A, Thakur N, Tandon H, *et al*. AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic Acids Res* 2014; **42**(D1): D1147–53.
41. Rusticus SA, Lovato CY. Impact of sample size and variability on the power and type I error rates of equivalence tests: a simulation study. *Pract Assess Res Eval* 2014; **19**(1): 11–1.
42. Schaduangrat N, Nantasenamat C, Prachayasittikul V, *et al*. Meta-iAVP: a sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation. *Int J Mol Sci* 2019; **20**(22): 5743–3.
43. Thakur N, Qureshi A, Kumar M. AVPpred: collection and prediction of highly effective antiviral peptides. *Nucleic Acids Res* 2012; **40**(W1): W199–204.
44. Tyagi A, Tuknait A, Anand P, *et al*. CancerPPD: a database of anticancer peptides and proteins. *Nucleic Acids Res* 2015; **43**(D1): D837–43.
45. Usman M, Lee JA. AFP-CKSAAP: prediction of antifreeze proteins using composition of k-spaced amino acid pairs with deep neural network. In: *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering*, 2019, 38–43.
46. van derMaaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008; **9**(86): 2579–605.

47. Veltri D, Kamath U, Shehu A. Deep learning improves antimicrobial peptide recognition. *Bioinformatics* 2018; **34**(16): 2740–7.
48. Wang G, *et al*. Apd3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res* 2016; **44**(D1): 1087–93.
49. Wang H-Y, Chung CR, Wang Z, *et al*. A large-scale investigation and identification of methicillin-resistant Staphylococcus aureus based on peaks binning of matrix-assisted laser desorption ionization-time of flight MS spectra. *Brief Bioinform* 2020;bbaa138.
50. Wang P, *et al*. Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. *PLoS One* 2011; **6**(4): 18476–6.
51. Wohlford-Lenane CL, Meyerholz DK, Perlman S, *et al*. Rhesus theta-defensin prevents death in a mouse model of severe acute respiratory syndrome coronavirus pulmonary disease. *J Virol* 2009; **83**(21): 11385–90.
52. Wu D, Wu T, Liu Q, *et al*. The SARS-CoV-2 outbreak: what we know. *Int J Infect Dis* 2020; **94**:44–8.
53. Xiao X, Wang P, Lin WZ, *et al*. iamp-2l: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal Biochem* 2013; **436**(2): 168–77.
54. Zhao G, London E. An amino acid "transmembrane tendency" scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: relationship to biological hydrophobicity. *Protein Sci* 2006; **15**(8): 1987–2001.
55. Zhao H, Zhou J, Zhang K, *et al*. A novel peptide with potent and broad-spectrum anti-viral activities against multiple respiratory viruses. *Sci Rep* 2016; **6**(1): 22008–8.