REVIEW ARTICLE

In Silico Approaches for the Prediction and Analysis of Antiviral Peptides: A Review

Phasit Charoenkwan[1], Nuttapat Anuwongcharoen[2], Chanin Nantasenamat[2], Md. Mehedi Hasan[3] and Watshara Shoombuatong[2,*]

[1]*Modern Management and Information Technology, College of Arts, Media and Technology, Chiang Mai University, Chiang Mai* 50200*, Thailand;* [2]*Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok, 10700, Thailand;* [3]*Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan*

ARTICLE HISTORY

Received: December 4, 2020
Accepted: March 25, 2020

DOI:
10.2174/1381612826666201102105827

Abstract: In light of the growing resistance toward current antiviral drugs, efforts to discover novel and effective antiviral therapeutic agents remain a pressing scientific effort. Antiviral peptides (AVPs) represent promising therapeutic agents due to their extraordinary advantages in terms of potency, efficacy and pharmacokinetic properties. The growing volume of newly discovered peptide sequences in the post-genomic era requires computational approaches for timely and accurate identification of AVPs. Machine learning (ML) methods such as random forest and support vector machine represent robust learning algorithms that are instrumental in successful peptide-based drug discovery. Therefore, this review summarizes the current state-of-the-art application of ML methods for identifying AVPs directly from the sequence information. We compare the efficiency of these methods in terms of the underlying characteristics of the dataset used along with feature encoding methods, ML algorithms, cross-validation methods and prediction performance. Finally, guidelines for the development of robust AVP models are also discussed. It is anticipated that this review will serve as a useful guide for the design and development of robust AVP and related therapeutic peptide predictors in the future.

Keywords: Therapeutic peptides, antiviral peptide, classification, machine learning, feature representation, feature selection.

1. INTRODUCTION

Nowadays, the emergence and re-emergence of viruses are becoming a great concern owing to the fact that therapeutic options are limited by the availability of specific antiviral agents. Although some of these antiviral agents (*e.g*. nucleotide or nucleoside analogues, reverse transcriptase inhibitors and protease inhibitors, *etc*.) can provide broad-spectrum antiviral effects against a vast array of viruses, these drugs also exhibit multiple adverse effects to patients, which may promote severe complications. Moreover, the increasing number of drug-resistant strains of viruses is found at a much faster pace than the pharmaceutical industry's launch of new and effective therapeutic agents for tackling this threat. To overcome these problems, antiviral peptides represent an attractive strategy that can be attributed to several advantages, including biocompatibility, lower adverse effects and better target selectivity. Interestingly, it is observed that the approval rate of therapeutic peptides is 20%, being faster as compared to those of conventional drugs [1-3].

There is an enormous expansion in the field of AVP research, which can be observed from the large dataset deposited in the antimicrobial database (APD3). As of September 23, 2019, a total number of 3129 AMPs are available, in which, 188 are AVPs [4]. In the meanwhile, several AMPs database are available and these include the DAMPD [5] and CAMP$_{R3}$ [6]. Moreover, there are even online databases dedicated to AVPs such as the AVPdb, which is comprised of 2683 experimentally validated AVPs and 624 modified AVPs targeting 60 medically important viruses [7]. Additionally, several other databases focus on the structure and antimicrobial activity of natural and synthetic peptides [8] as well as other therapeutic peptides [9-11]. The rapid expansion in omics research in concomitant with advancements in high-throughput technology has led to the generation of big biological data that consequently give rise to the rapid growth of newly discovered peptide sequences in the post-genomic era.

Bioinformatics and machine learning (ML) are instrumental for efficient analysis of the rapidly growing biological data. For instance, these computational methods make it possible to generate predictive models that can predict the biological activity of unknown peptide sequence as well as discerning the underlying relationship that exists between peptide features and their corresponding activity. Such capabilities are essential for the development of novel therapeutic peptides. Although the need for such tools is increasing, however only a limited number of predictive models for AVPs are in existence, which represent a promising research area to explore for the development of novel antiviral agents.

To the best of our knowledge, this article represents the first comprehensive review of the utilization of ML algorithms for gaining insights into the bioactivity of AVPs. In this review, we compare the underlying architecture of existing studies [12-17] in terms of the dataset used along with the feature encoding methods, ML algorithms, cross-validation methods and prediction performance. Importantly, we provide general guidelines for the development of robust AVP models, which represent suggestions for overcoming some of the inherent weaknesses of current AVP models. It is anticipated that this review would further contribute to the further growth and expansion of this field by providing an overview of the

*Address correspondence to this author at the Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok, 10700, Thailand; E-mail: watshara.sho@mahidol.ac.th

1381-6128/21 $65.00+.00

current state-of-the-art of the field along with expected future trends and outlook.

2. ANTIVIRAL PEPTIDES

Antiviral peptides (AVPs) are members of antimicrobial peptides (AMPs), which are naturally similar in physicochemical properties but have specific inhibitory activity against viruses [18]. AVPs directly inhibit viruses *via* two major mechanisms, including direct- and indirect-inhibition. The direct-inhibition is defined by the peptide's role in direct interaction with their target proteins, which can either disrupt the envelope membranes of viruses or inhibit viral polymerases, which are essential for viral replication. Furthermore, it can bind to the target receptor of viruses on the host membrane and inhibit the process of viral entry and viral-host membrane fusion. On the other hand, for indirect inhibition, AVPs inhibit viruses by recruiting immune cells (*i.e.* that promote the host defense mechanisms) to help eradicate viruses [18]. In some cases, certain peptides may show more than one bioactivity, such as affording antibacterial and antiviral activities, whereby these are known as promiscuous peptides [15].

Currently, some studies have reported that AVPs could inhibit the fusion of viruses to host cells [19-21], while others have shown that AVPs could interfere with viral replication [22-24] and attachment of viruses to host cells [25-27]. For example, P9 (*i.e.* an AVP derived from mouse β-defensin) acts against various flu strains (*e.g.* H1N1, H3N2, H5N1, H7N7, and H7N9) by binding to viral glycoproteins and inhibiting RNA replication by preventing viral fusion in the endosome [28]. Additionally, protegrin-1 (*i.e.* a cyclical cationic peptide derived from swine white blood cells) showed potent antiviral activity against the dengue virus by inhibiting specific viral proteases that are important for dengue virus replication, namely the NS2B-NS3pro [29].

Important characteristics of AVPs were extensively studied in those involved in the inhibition of membrane fusion by strongly interacting with proteins that are required for viral entry. These peptides exhibit inhibitory activity by engaging in electrostatic and hydrophobic interactions with exposed proteins on membranes [18]. In addition, AVPs have been shown to possess cationic and amphipathic characteristics with positive net charges, all of which are essential for these peptides to work as antimicrobials [30]. Moreover, hydrophobicity seems to be a key property for peptides with activity against enveloped viruses [31, 32]. Nevertheless, the intricate balance between these physicochemical properties is crucial for the lipid selectivity of peptides, which may further contribute to target selectivity, antimicrobial activity and the cytotoxicity of AVPs [18].

3. CONCEPTS OF THE DEVELOPMENT OF ANTIVIRAL PEPTIDE PREDICTOR

In light of prior knowledge of peptide sequence analysis, the prediction of AVPs could be categorized into two main tasks: (i) discriminating AVPs from Non-AVPs and (ii) predicting the antiviral activity of such AVPs [12-17]. The prediction of AVPs serves as an alternative approach to complement and alleviate potential problems of high-throughput experimental approaches in identifying novel AVPs (*i.e.* which are time-consuming and costly). Moreover, the avalanche of newly discovered peptide sequences in the post-genomic era constitutes big data that calls for the development of computational tools that can effectively discriminate AVPs from Non-AVPs□.

As elaborated in a series of publications [15, 33-60] and summarized in several comprehensive review papers [61, 62], in the development of an efficient and interpretable sequence-based tool for predicting and analyzing peptide functions, six prime steps should be considered that are as follows: (i) establishing a reliable dataset containing experimentally validated sequences for training and validating the model; (ii) extracting peptides sequences using interpretable feature scheme; (iii) using interpretable learning algorithms; (iv) evaluating the model using standard cross-validation tests; (v) analyzing important features derived from the constructed model; and (vi) establishing a user-friendly web-server for obtaining the prediction without the need to understand complex mathematical and statistical details.

3.1. Benchmark Dataset

In order to obtain high-quality benchmark dataset that is crucial for the development of reliable models, the following procedures [41-43, 56, 63, 64] are recommended. Firstly, only peptides with experimentally determined biological activities were considered. Secondly, peptides containing ambiguous residues (*e.g.* X, B, U and Z) were excluded. Thirdly, duplicate peptide sequences were removed. Ideally, redundancy in the dataset should be removed because it affects the performance of the prediction method. However, this process may also lead to the loss of information of AVPs.

The benchmark dataset used in the study of Thakur *et al.*, [14] consisted of two training ($T^{544p+407n}$ and $T^{544p+544n}$) and two independent ($T^{60p+45n}$ and $T^{60p+60n}$) sets. Such data have been compiled from various research articles as well as patents indexed by PubMed and Patent Lens. This led to a set of 1245 peptide sequences with reported antiviral activity against human viruses consisting of HIV, HCV, SARS and Influenza. For the two training sets, $\mathrm{T}^{544\mathrm{p}}$ and $\mathrm{T}^{407\mathrm{n}}$ that correspondingly represent collections of 544 and 407 experimentally validated AVP and Non-AVPs while the $\mathrm{T}^{544\mathrm{n}}$ set represents a collection of 544 non-experimentally validated Non-AVPs. As for the two independent sets, $\mathrm{V}^{60\mathrm{p}}$ and $\mathrm{V}^{45\mathrm{n}}$ represent the collections of 60 and 45 experimentally validated AVP and Non-AVPs, respectively, while the $\mathrm{V}^{60\mathrm{n}}$ set represents a collection of 60 non-experimentally validated Non-AVPs.

3.2. Feature Representation

One of the most challenging problems in computational biology is the development of a sequence-based predictor for rationalizing the pivotal property of biological samples (such as protein, peptide, DNA or RNA). In the development of an effective prediction model, it is necessary to represent biological samples with an effective mathematical expression that can accurately reflect the intrinsic correlation with the desired target [54, 55, 65-75]. For peptide sequences, the most widely used features consist of amino acid composition (AAC), dipeptide composition (DPC) and physicochemical property (PCP) [55, 76-81].

A peptide sequence (**P**) can be represented as:

$$\mathbf{P} = \mathrm{p}_1\mathrm{p}_2\mathrm{p}_3 \cdots \mathrm{p}_\mathrm{N} \tag{1}$$

where p_i and N denote the i^{th} residue in the peptide **P** and the peptide length, respectively.

AAC and DPC are the proportions of each amino acid and dipeptide that are present in a peptide sequence P expressed as fixed lengths of 20 and 400, respectively. Thus, in terms of AAC and DPC features, a peptide **P** can be expressed by vectors of 20D and 400D (dimension) spaces, respectively, as formulated by:

$$\mathbf{P} = [\mathrm{aa}_1, \mathrm{aa}_2, \dots, \mathrm{aa}_{20}]^\mathbf{T} \tag{2}$$

$$\mathbf{P} = [\mathrm{dp}_1, \mathrm{dp}_2, \dots, \mathrm{dp}_{400}]^\mathbf{T} \tag{3}$$

Where T is the transposed operator while aa_1, aa_2, ..., aa_{20} and dp_1, dp_2, ..., dp_{400} are occurrence frequencies of 20 and 400 native amino acids and dipeptides, respectively, in a peptide sequence **P**.

PCP is one of the most intuitive features associated with biophysical and biochemical reactions. In fact, there are over 544 PCPs that can be computed for amino acids extracted from the amino acid index database (AAindex) [82], which is a collection of published literature as well as different biochemical and biophysical properties of amino acids. Each physicochemical property constitutes a set of 20 numerical values for amino acids. After the removal of 13 PCPs that contain not applicable (NA) as their amino acid indices, a total of 531 PCPs can be used for further peptide analysis.

In addition, many research groups have employed other feature encoding properties such as relative frequency of 20 amino acids (Rfre) [13], residues composition of peptides (PEP) [13], aggregation propensity [12], and class feature [15, 16], as summarized in Table **1**.

3.3. Computational Models

Two popular ML methods are widely used for developing AVP predictors, namely random forest (RF) and support vector machine (SVM). Herein, we briefly describe the basic concepts of these two classifiers.

RF models [83] are developed by growing many weak classifications and regression tree (CART) classifiers whereby each classifier is generated using a random vector sampled independently from the input vector so as to enhance the prediction performance of CART [83, 84]. RF has been widely used to model various biological problems [47, 48, 85-89]. In the RF method, the out-of-bag (OOB) approach is utilized for assessing the feature importance as follows: (1) two-thirds of the training data are utilized to construct the predictive classifier while the remaining are used for evaluating the performance of such classifier and (2) the feature importance of each feature can be evaluated by measuring the decrease in the prediction performance.

SVM is a supervised learning model based on the principles of structural risk minimization and kernel method, as proposed by Vapnik [90, 91]. This method has been widely used in computational biology [46, 48-50, 85, 89]. SVM model can deal with the problem of over-fitting arising from the use of small training datasets by mapping the input samples to a higher dimensional space followed by searching for the maximum-margin hyperplane that is used for constructing the classifier. In order to perform linear separation on high-dimensional samples, SVM employs one of the many well-known kernel functions to transform inputs from the sample space having a *p*-dimensional feature vector into the feature space with an *n*-dimensional feature vector where $p < n$. Radial basis function is a popular kernel that is applied to non-linearly transform the feature space, defined as follows:

$$K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right), \gamma > 0 \tag{4}$$

The kernel parameter γ represents how samples are transformed to the feature space while the cost parameter C of SVM adjusts the penalty of the total error.

3.4. Performance Evaluation

From the point of view of binary classification (AVPs and Non-AVPs), there exist three commonly used methods for empirically assessing the predictive model for its robustness in practical applications consisting of a sub-sampling test (2-, 5- or 10-fold cross-validation; 2-, 5-, 10-fold CV), jackknife test and (iii) independent test [40-45].

In order to evaluate the predictive ability of models, four metrics are widely used for binary classification as follows:

$$\begin{cases} Sn = 1 - \frac{N_-^+}{N^+} \\ Sp = 1 - \frac{N_+^-}{N^-} \\ Ac = \Lambda = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \\ MCC = \dfrac{1 - \left(\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}} \end{cases} \tag{5}$$

where N^+ represents the total number of positive samples investigated, while N_-^+ is the number of positive samples incorrectly predicted to be of negative one; N^- the total number of negative samples investigated, while N_+^- the number of the negative samples is incorrectly predicted to be of a positive one. Moreover, a threshold-independent parameter namely the receiver operating characteristic (ROC) curves is suggested to be a robust metric for evaluating the predictive performance. The area under the ROC curve (auAUC) is also a popular metric for assessing the prediction performance where AUC values of 0.5 and 1 are indicative of random and perfect models, respectively.

3.5. Web Server Development

In an effort to maximize the utility of the prediction model by the scientific community, authors are recommended to deploy their best predictive models as a web server. In this aspect, there are only three available web servers were developed for AVP prediction namely the AVPpred, PEPred-Suite and Meta-iAVP that are correspondingly available at the following URL: http://crdd.osdd.net/servers/avppred/, http://server.malab.cn/PEPred-Suite/Server.html and http://codes.bio/meta-iavp/.

4. MACHINE LEARNING MODELS FOR THE PREDICTION OF ANTIVIRAL PEPTIDES

Several sequence-based computational methods consisting of AVPpred [14], method proposed by Chang *et al*. [12], method proposed by Zare *et al*. [17], AntiVPP 1.0 [13], PEPred-Suite [16] and Meta-iAVP [15] that have been developed will be further discussed in this section. Table **1** summarizes the underlying architecture of these ML models and the corresponding peptide features used in model construction.

4.1. AVPpred

In 2012, Thakur *et al*., first addressed this problem by employing an SVM-based predictor named AVPred [14] and established the benchmark dataset as mentioned in section *3.1. BENCHMARK DATASET*. Various features, including motif, align, AAC and PCP, were used as input for the construction of an SVM model. Particularly, the authors selected RBF kernel in the implementation of AVPred. The best prediction model was obtained by using SVM in conjunction with PCP features. AVPpred afforded 85.0% and 90.0% accuracies over 5-fold CV on $T^{544p+407n}$ and $T^{544p+544n}$ datasets, respectively. The use of the AAC feature provided comparable performance to models using PCP feature with corresponding accuracies of 84.0% and 90.2%. For the independent test, AVPpred yielded 85.7% and 92.5% accuracies on $T^{60p+45n}$ and $T^{60p+60n}$ datasets, respectively.

4.2. Chang *et al*.'s Method

A year later, Chang and Yang [12] utilized three well-known ML methods (*e.g*. linear discriminant analysis (LDA), artificial neural network and RF) together with four peptide features (*e.g*. AAC, PCP and aggregation tendencies of peptides and secondary structure) to construct AVP predictors. To improve the prediction performance, various combinations of these selected four features (AAC+secondary structure, PCP+aggregation, PCP+secondary structure, AAC+ secondary structure+aggregation and PCP+ secondary structure+ aggregation) were considered. In the case of 10-fold CV test, RF model ($T^{544p+407n}$ and $T^{544p+544n}$) with AAC feature (84.1%, 91.1%) was very comparable to the use of the PCP feature (84.2%, 90.0%). Interestingly, using the combination of AAC+ secondary structure and AAC+secondary structure+aggregation gave 85.1% and 91.5% on $T^{544p+407n}$ and $T^{544p+544n}$ datasets, respectively. Meanwhile, test accuracies of 89.5% and 93.3% were obtained from RF models with the combination of AAC+secondary structure and AAC+secondary structure+aggregation, respectively.

4.3. The Method of Zare *et al.*

In 2015, Zare *et al*., [17] collected their own dataset from the Antiviral Peptides prediction Database (http://crdd.osdd.net/server/avppred). This dataset is a non-redundant dataset, and peptides consisting of more than 90% similarity were removed using the CD-HIT webserver [92]. Finally, after pre-processing, the dataset contained 342 AVPs and 312 Non-AVPs. In this study, adaptive boosting (Adaboost) with PseAAC was used in the development of the AVP predictor. The best accuracy of 93.3% was obtained from Adaboost as applied to the J48 algorithm as assessed by a 5-fold CV. However, the authors did not apply the proposed model on an independent dataset. Thus, it is uncertain whether the method would perform well on an independent dataset or not.

4.4. AntiVPP 1.0

In 2019, Lissabet *et al*. ,[13] developed a robust software known as the AntiVPP 1.0. This software is based on two new features (*i.e*. Rfre and PEP features) that were employed as input to the four ML models consisting of SVM, RF, ANN and k-nearest neighbor (kNN). Based on the prediction results from the $T^{60p+60n}$ dataset, RF, SVM, ANN and kNN yielded accuracies of 0.93, 0.79, 0.90 and 0.90%, respectively. As RF was found to outperform other ML models, therefore the AntiVPP 1.0 was developed using RF as the learning algorithm. The AntiVPP 1.0 software is available at https://github.com/bio-coding/AntiVPP.

4.5. PEPred-Suite

Shortly afterward, Wei *et al*., [16] proposed an adaptive feature representation strategy that was applied in predicting different therapeutic properties of peptides. To train and evaluate their proposed predictive models, eight benchmark datasets spanning different bioactivities from previous studies were employed as follows: AVP [14], anti-angiogenic (AAP) [93], antibacterial (ABP) [94], anticancer (ACP) [95], anti-inflammatory Peptides (AIP) [51], cell-penetrating (CPP) [96], quorum sensing peptides (QSP) [97] and polystyrene surface-binding (PSB) peptides [98]. In the PEPred-Suite, input peptides served as input to the feature representation learning scheme for encoding an n-dimensional feature vector. Such feature vectors were further used to feed into the RF models and trained on the $T^{544p+407n}$ dataset. The constructed.

RF models would produce a prediction score for each candidate peptide in which score ranges from 0 to 1. A higher score of the peptide suggests a higher probability that the peptide is likely to be AVPs. Authors considered peptides to be AVPs if their prediction scores were higher than 0.5 and Non-AVPs otherwise. To improve the feature representation ability, the author further optimized the feature representation by means of feature selection techniques based on the minimal Redundancy and Maximal Relevance (mRMR). The optimal feature set as derived from the two-step feature selection strategy was then used to feed into the RF model in order to produce the final model called the PEPred-Suite. Evaluation by 10-fold CV test indicated that the PEPred-Suite (AUC = 0.874) performed better than that of the AntiAngioPred (AUC = 0.820) [93]. In the case of an independent test, the PEPred-Suite (AUC = 0.804) still outperformed the AntiAngioPred (AUC = 0.742).

4.6. Meta-iAVP

Most recently, our group developed the first meta-predictor known as the Meta-iAVP for addressing various aforementioned inefficiencies. In the Meta-iAVP, peptide sequences were encoded by several sets of descriptors including AAC, DPC, pseudo amino acid composition (PseAAC), amphiphilic pseudo amino acid composition (Am-PseAAC), and g-gap dipeptide composition (GDC). Afterward, each feature type was separately fed into six different ML algorithms (*i.e*. RF, SVM, kNN, recursive partitioning and regression trees (rpart), generalized linear model (glm), and extreme gradient boosting (XGBoost)) for generating a new set of feature representation. Subsequently, such an effective feature representation was then used to build a meta-predictor. Prediction results indicated that the average Ac as assessed by five-repeated five-fold CV on $T^{544p+407n}$ and $T^{544p+544n}$ datasets correspondingly gave the following metric values {78.52%, 78.72%, 79.69%, 78.68%, and 77.04%} and {84.91%, 84.88%, 85.28%, 82.19%, and 86.44%} for ACC, PseAAC, Am-PseAAC, DPC and GDC, respectively. Meanwhile, the performance comparisons by means of independent sets $V^{60p+45n}$ and $V^{60p+60n}$ afforded values of {80.29%, 83.17%, 79.01%, 79.49%, and 77.41%} and {86.16%, 86.44%, 85.88%, 86.02%, and 84.59%} for ACC, PseAAC, Am-PseAAC, DPC, and GDC, respectively. Based on these comparative results, it can be deduced that AAC and PseAAC were the most important features for discriminating AVPs from Non-AVPs. The RF model built using the AAC feature performed the best with the highest Ac, Sn, Sp, and MCC of 86.54%, 86.54%, 86.36%, and 0.73, respectively, as evaluated on the independent validation test using the $V^{60p+45n}$ dataset. Meanwhile, the RF model built using the PseAAC feature demonstrated superior discrimination of AVPs from Non-AVPs on the dataset $V^{60p+60n}$ as deduced from the highest Ac, Sn, Sp, and MCC values of 91.53%, 90.00%, 93.10%, and 0.83, respectively.

5. ANALYSIS OF PERFORMANCE COMPARISON OF EXISTING PREDICTORS

Based on the six prime steps, as mentioned previously, the effectiveness of a proposed method can be determined by subjecting it to comparison with existing models. Several researchers in the field have made efforts to develop computational methods for discriminating AVPs from Non-AVPs and this included the AVPpred [14], the method **by** Chang *et al*.[12] and Meta-iAVP [15] as summarized in Table **1**. Amongst the existing methods, there were three computational methods (*i.e*., AVPpred [14], the method by Chang *et al*. [12], and Meta-iAVP [15]) performed on the two training ($T^{544p+407n}$ and $T^{544p+544n}$) and independent ($T^{60p+45n}$ and $T^{60p+60n}$) sets as described in *3.1. BENCHMARK DATASET* section. Therefore, in this section, we performed a comparative analysis of these three methods. Particularly, the details of the comparative analysis of these three methods, as assessed from the cross-validation and independent test, are provided in Table **2**.

In the case of cross-validation test results, Meta-iAVP was found to afford the highest accuracy of 88.2% and 93.2% for $T^{544p+407n}$ and $T^{544p+544n}$ datasets, respectively. Meanwhile, AVPpred was shown to be very comparable to the work of the method by Chang *et al*. On the basis of independent test results, Meta-iAVP still outperformed both AVPpred and the method of Chang *et al*.. Furthermore, Meta-iAVP yielded test accuracies of 95.2% and 94.9% for $T^{60p+45n}$ and $T^{60p+60n}$ datasets, respectively. As for the aforementioned performance comparison, consistent performance comparison over cross-validation and independent tests indicated that Meta-iAVP could accurately discriminate AVPs from non-AVPs for unknown peptides. The robust performance of Meta-iAVP over both AVPpred and the method of Chang *et al*. can be attributed to the following aspects: (i) Amongst the various types of features employed for developing AVP predictors, PseAAC and Am-PseAAC features were employed for the first time in AVP prediction. Particularly, several studies reported that these two features have been successfully implemented to predict many peptides and proteins [17, 64, 99-101]. (ii) Optimal performance parameters of Meta-iAVP were obtained from 5-repeated 5-fold CV (*i.e*. indicating that the estimated parameters were more stable and accurate) [80]; (iii) Meta-iAVP was developed using only six-dimensional (6D) feature vectors that provided not only sufficient but also comprehensive information for AVP prediction.

Table 1. Summary of existing methods for predicting antiviral peptides.

Method (Year)	Classifier[a]	Size of Training/Independent Set[b]	Sequence Feature[c]	Cross-Validation (CV) Method
AVPpred (2012) [14]	SVM	$(544^P+407^N, 544^P+544^N)/(60^P+45^N, 60^P+60^N)$	PCP	5-fold CV/ independent test
Chang *et al.*'s method (2013) [12]	RF	$(544^P+407^N, 544^P+544^N)/(60^P+45^N, 60^P+60^N)$	AAC, aggregation	10-fold CV/ independent test
Zare *et al.*'s method (2015) [17]	Adaboost	(342^P+312^N)/NA	PseAAC	5-fold CV/NA
AntiVPP 1.0 (2019) [13]	RF	$(544^P,544^N)/(60^P,60^N)$	Rfre, PEPP	5-fold CV/ independent test
PEPred-Suite (2019) [16]	RF	$(544^P+407^N)/(60^P+45^N)$	Class feature	5-fold CV/ independent test
Meta-iAVP (2019) [15]	RF	$(544^P+407^N, 544^P+544^N)/(60^P+45^N, 60^P+60^N)$	Probabilistic feature	5-fold CV/ independent test

[a]RF: Random forest and SVM: Support vector machine. [b] P: A number of AVPs, N: Aa number of Non-AVPs, NA: Data is not provided. [c]AAC: Amino acid composition, aggregation: Aggregation propensity, Am-PseAAC: Amphiphilic pseudo amino acid composition, PCP: Physicochemical properties, Rfre: Relative frequency of 20 amino acids, PEP: Residues composition of peptides.

Table 2. Performance comparison of existing methods and their web servers.

Method (Year)	Accuracy (CV Test) $(544^P+407^N, 544^P+544^N)$	Accuracy (Independent Test) $(60^P+45^N, 60^P+60^N)$	Web Server Availability
AVPpred (2012) [14]	(85.0%, 90.2%) [a]	(85.7%, 92.5%)	http://crdd.osdd.net/servers/avppred/
Chang *et al.*'s method (2013) [12]	(85.1%, 91.5%) [b]	(89.5%, 93.3%)	-
AntiVPP 1.0 (2019) [13]	(NA, 93.0%) [a]	(NA, 93.0%)	-
PEPred-Suite (2019) [16]	(86.2%, NA) [a]	(86.7%, NA)	http://server.malab.cn/ PEPred-Suite/Server.html
Meta-iAVP (2019) [15]	(88.2%, 93.2%) [c]	(95.2%, 94.9%)	http://codes.bio/meta-iavp/

CV: the cross-validation method and NA: Data is not provided.

6. BIOLOGICAL INSIGHTS FROM MACHINE LEARNING MODELS

The analysis of feature importance can provide a better understanding of the mechanistic details governing the antiviral activity of peptides. To the best of our knowledge, only two studies had made efforts in performing feature importance analysis [12, 15]. Chang and Yang [12] reported that Leu and Lys residues were found to be predominant in AVPs. Additionally, Thr, Pro and Val residues were also found to be prevalent in AVPs [12]. Schaduangrat *et al.*, [15] employed the value of the mean decrease of Gini index (MDGI) to rank and estimate the importance of each AAC and DPC feature. This study reported that the ten informative amino acids with the highest MDGI values consisted of Lys, Thr, Leu, Ile, Ser, Trp, Asn, Arg, Cys, and Glu (49.27, 46.27, 35.06, 34.52, 30.95, 30.93, 30.19, 28.52, 26.33, and 24.87, respectively) and Lys, Pro, Cys, Thr, Ser, Trp, Val, Ala, Gly, and Leu (77.11, 68.87, 57.68, 46.84, 39.57, 36.83, 25.69, 24.40, 24.25, and 23.80, respectively) for 544^P+407^N, 544^P+544^N datasets, respectively. Moreover, the five top-ranked dipeptides according to their MDGI value consisted of LL, RK, LV, WI, and EI for the $T^{544p+407n}$dataset and KR, KK, GP, AS, and SA for the $T^{544p+544n}$ dataset.

7. GUIDELINES FOR DEVELOPMENT OF ROBUST AVP MODELS

A survey of existing work pertaining to the discrimination of AVPs from Non-AVPs suggested that these models provided reasonably high prediction accuracies. However, there is still ample room for further improvement pertaining to model performance and interpretability. Hereafter, five recommendations are provided.

Firstly, almost all of the existing models were trained and tested *via* the use of a benchmark dataset containing high homologous sequences that consequently leads to potential sequence homology bias. It should be noted that lower thresholds of the sequence identity (less than 0.5) might reduce the sequence homology bias and could therefore improve the model reliability [48]. However, using higher threshold is necessary owing to the inherently small dataset size. In the case of predicting peptide functions, several studies have suggested using cutoff thresholds of 0.8-0.9 as acceptable criteria for reducing the sequence homology bias [16, 33, 47, 48, 51, 54, 85, 87, 102].

Secondly, the advantages of using computational models to predict the bioactivity of unknown data is inherently dependent on the number of samples in the dataset. Thus, it could be stated that if the proposed model is developed from a small number of samples in the dataset, the proposed model would likely possess a narrow applicability domain and consequently lead to low generalization capability. To resolve such issue, it is required to increase the size of the peptide dataset by combining all data sources together so as to capture as much as possible the pattern of peptide data for alleviating uncertainties stemming from the prediction system.

Thirdly, as can be seen from existing studies [12-17], authors are mainly focused on increasing both the complexity of the prediction model as well as the number of feature types for enhancing their prediction results. However, the mechanism of existing methods [12-17] has from low interpretability and is, therefore of little use for biologists. In 2012, Huang *et al.*, proposed a scoring card method (SCM) for alleviating such problems [103, 104]. The motivation for the development of the SCM method arises mainly from the following reasons: (i) the features of amino acid and dipeptide composition are important for predicting and analysing protein functions; (ii) it is desirable to deduce the relationship that exists between the protein function with biochemical and biophysical properties of amino acid residues; (ii) the widely used SVM-based classifiers can provide prediction accuracy, but they suffer from low interpretability; and (iv) a simple and easily interpretable classifier with an acceptable level of accuracy is desirable. Previously, the SCM-based classifier has been widely used to address several biological problems [78, 103-110].

Fourthly, existing methods [12-17] could only discriminate AVPs from Non-AVPs. However, none of these can predict both (1) AVPs from Non-AVPs as well as (2) the degree of AVP activity (high or low) from the given peptide sequences. Hence, their practical usage is quite limited. The idea of model development for predicting the class of peptide and its efficacy activity has been previously used for the investigation of several protein and peptide functions. For example, in 2018, Manavalan *et al*., developed a two-layer prediction framework named MLCPPs for predicting cell-penetrating peptides and their uptake efficiency [53]. Such a two-layer framework entails the use of the first layer to predict whether the given peptide can or cannot elicit investigated property while the second layer makes a more refined prediction for those that can elicit the investigated property by making a second prediction pertaining to the relative degree of the investigated property whether it can afford a high or low property of interest.

CONCLUSION

In the present review, we comprehensively surveyed the existing literature on the use of computational methods for identifying AVPs. The collective literature on the prediction and characterization of AVPs *via* the use of ML approaches serves as a useful, high throughput and cost-effective tool for large-scale analysis of AVPs that would further help contribute to a series of interesting follow-up research studies on antiviral peptides as well as other related therapeutic peptides. It is anticipated that this review would help contribute to further growth and expansion of the field by providing readers with the current state-of-the-art of the field as well as expected future trends and outlook.

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

This work was fully supported by the TRF Research Grant for New Scholar (No. MRG6180226) and by the College of Arts, Media and Technology, Chiang Mai University. This work was partially supported by Chiang Mai University, the TRF Research Career Development Grant (No. RSA6280075) from the Thailand Research Fund.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Mäde V, Els-Heindl S, Beck-Sickinger AG. Automated solid-phase peptide synthesis to obtain therapeutic peptides. Beilstein J Org Chem 2014; 10: 1197-212.
http://dx.doi.org/10.3762/bjoc.10.118 PMID: 24991269

[2] Fotouhi N. Peptide therapeutics Peptide chemistry and drug design. 1st ed. New York: WH Freeman & Co. 2015; pp. 1-8.

[3] Fox JL. Rare-disease drugs boosted by new prescription drug user fee act.ed^eds. Nature Publishing Group 2012.
http://dx.doi.org/10.1038/nbt0812-733

[4] Wang G, Li X, Wang Z. APD3: the antimicrobial peptide database as a tool for research and education. Nucleic Acids Res 2016; 44(D1): D1087-93.
http://dx.doi.org/10.1093/nar/gkv1278 PMID: 26602694

[5] Seshadri Sundararajan V, Gabere MN, Pretorius A, *et al.* DAMPD: a manually curated antimicrobial peptide database. Nucleic Acids Res 2012; 40(Database issue): D1108-12.
http://dx.doi.org/10.1093/nar/gkr1063 PMID: 22110032

[6] Waghu FH, Barai RS, Gurung P, Idicula-Thomas S. CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. Nucleic Acids Res 2016; 44(D1): D1094-7.
http://dx.doi.org/10.1093/nar/gkv1051 PMID: 26467475

[7] Qureshi A, Thakur N, Tandon H, Kumar M. AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses. Nucleic Acids Res 2014; 42(Database issue): D1147-53.
http://dx.doi.org/10.1093/nar/gkt1191 PMID: 24285301

[8] Pirtskhalava M, Gabrielian A, Cruz P, *et al.* DBAASP v.2: an enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides. Nucleic Acids Res 2016; 44(D1): D1104-12.
http://dx.doi.org/10.1093/nar/gkv1174 PMID: 26578581

[9] Singh S, Chaudhary K, Dhanda SK, *et al.* SATPdb: a database of structurally annotated therapeutic peptides. Nucleic Acids Res 2016; 44(D1): D1119-26.
http://dx.doi.org/10.1093/nar/gkv1114 PMID: 26527728

[10] Rajput A, Thakur A, Sharma S, Kumar M. aBiofilm: a resource of anti-biofilm agents and their potential implications in targeting antibiotic drug resistance. Nucleic Acids Res 2018; 46(D1): D894-900.
http://dx.doi.org/10.1093/nar/gkx1157 PMID: 29156005

[11] Sharma D, Priyadarshini P, Vrati S. Unraveling the web of viroinformatics: computational tools and databases in virus research. J Virol 2015; 89(3): 1489-501.
http://dx.doi.org/10.1128/JVI.02027-14 PMID: 25428870

[12] Chang KY, Yang J-R. Analysis and prediction of highly effective antiviral peptides based on random forests. PLoS One 2013; 8(8)e70166
http://dx.doi.org/10.1371/journal.pone.0070166 PMID: 23940542

[13] Beltrán Lissabet JF, Belén LH, Farias JG. AntiVPP 1.0: A portable tool for prediction of antiviral peptides. Comput Biol Med 2019; 107: 127-30.
http://dx.doi.org/10.1016/j.compbiomed.2019.02.011
PMID: 30802694

[14] Thakur N, Qureshi A, Kumar M. AVPpred: collection and prediction of highly effective antiviral peptides. Nucleic Acids Res 2012; 40(Web Server issue)W199-204
http://dx.doi.org/10.1093/nar/gks450 PMID: 22638580

[15] Schaduangrat N, Nantasenamat C, Prachayasittikul V, Shoombuatong W. 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.
http://dx.doi.org/10.3390/ijms20225743 PMID: 31731751

[16] Wei L, Zhou C, Su R, Zou Q. PEPred-Suite: improved and robust prediction of therapeutic peptides using adaptive feature representation learning. Bioinformatics 2019; 35(21): 4272-80.
http://dx.doi.org/10.1093/bioinformatics/btz246 PMID: 30994882

[17] Zare M, Mohabatkar H, Faramarzi FK, Beigi MM, Behbahani M. Using Chou's pseudo amino acid composition and machine learning method to predict the antiviral peptides. Open Bioinform J 2015; •••: 9.
http://dx.doi.org/10.2174/1875036201509010013

[18] Gomes B, Augusto MT, Felício MR, *et al.* Designing improved active peptides for therapeutic approaches against infectious diseases. Biotechnol Adv 2018; 36(2): 415-29.
http://dx.doi.org/10.1016/j.biotechadv.2018.01.004
PMID: 29330093

[19] Henriques ST, Craik DJ. Cyclotides as templates in drug design. Drug Discov Today 2010; 15(1-2): 57-64.
http://dx.doi.org/10.1016/j.drudis.2009.10.007 PMID: 19878736

[20] Nawae W, Hannongbua S, Ruengjitchatchawalya M. Molecular dynamics exploration of poration and leaking caused by Kalata B1 in HIV-infected cell membrane compared to host and HIV membranes. Sci Rep 2017; 7(1): 3638.

http://dx.doi.org/10.1038/s41598-017-03745-2 PMID: 28620219

[21] Vigant F, Santos NC, Lee B. Broad-spectrum antivirals against viral fusion. Nat Rev Microbiol 2015; 13(7): 426-37.
http://dx.doi.org/10.1038/nrmicro3475 PMID: 26075364

[22] Ngai PH, Ng TB. Phaseococcin, an antifungal protein with antiproliferative and anti-HIV-1 reverse transcriptase activities from small scarlet runner beans. Biochem Cell Biol 2005; 83(2): 212-20.
http://dx.doi.org/10.1139/o05-037 PMID: 15864329

[23] Huang Y, Zhang J, Zhao Y-Y, *et al.* SPARC expression and prognostic value in non-small cell lung cancer. Chin J Cancer 2012; 31(11): 541-8.
http://dx.doi.org/10.5732/cjc.012.10212 PMID: 23114088

[24] Rothan HA, Bahrani H, Rahman NA, Yusof R. Identification of natural antimicrobial agents to treat dengue infection: In vitro analysis of latarcin peptide activity against dengue virus. BMC Microbiol 2014; 14: 140.
http://dx.doi.org/10.1186/1471-2180-14-140 PMID: 24885331

[25] Quintero-Gil C, Parra-Suescún J, Lopez-Herrera A, Orduz S. In-silico design and molecular docking evaluation of peptides derivatives from bacteriocins and porcine beta defensin-2 as inhibitors of Hepatitis E virus capsid protein. Virusdisease 2017; 28(3): 281-8.
http://dx.doi.org/10.1007/s13337-017-0383-7 PMID: 29291214

[26] Chiang AW, Wu WY, Wang T, Hwang MJ. Identification of Entry Factors Involved in Hepatitis C Virus Infection Based on Host-Mimicking Short Linear Motifs. PLOS Comput Biol 2017; 13(1)e1005368
http://dx.doi.org/10.1371/journal.pcbi.1005368 PMID: 28129350

[27] Yin P, Zhang L, Ye F, *et al.* A screen for inhibitory peptides of hepatitis C virus identifies a novel entry inhibitor targeting E1 and E2. Sci Rep 2017; 7(1): 3976.
http://dx.doi.org/10.1038/s41598-017-04274-8 PMID: 28638089

[28] Nyanguile O. Peptide antiviral strategies as an alternative to treat lower respiratory viral infections. Front Immunol 2019; 10: 1366.
http://dx.doi.org/10.3389/fimmu.2019.01366 PMID: 31293570

[29] Rothan HA, Abdulrahman AY, Sasikumer PG, Othman S, Abd Rahman N, Yusof R. Protegrin-1 inhibits dengue NS2B-NS3 serine protease and viral replication in MK2 cells. BioMed Research International 2012.

[30] Bulet P, Stöcklin R, Menin L. Anti-microbial peptides: from invertebrates to vertebrates. Immunol Rev 2004; 198: 169-84.
http://dx.doi.org/10.1111/j.0105-2896.2004.0124.x
PMID: 15199962

[31] Badani H, Garry RF, Wimley WC. Peptide entry inhibitors of enveloped viruses: the importance of interfacial hydrophobicity. Biochim Biophys Acta 2014; 1838(9): 2180-97.
http://dx.doi.org/10.1016/j.bbamem.2014.04.015 PMID: 24780375

[32] Wang CK, Shih LY, Chang KY. Large-Scale Analysis of Antimicrobial Activities in Relation to Amphipathicity and Charge Reveals Novel Characterization of Antimicrobial Peptides. Molecules 2017; 22(11): 22.
http://dx.doi.org/10.3390/molecules22112037 PMID: 29165350

[33] Schaduangrat N, Nantasenamat C, Prachayasittikul V, Shoombuatong W. ACPred: A Computational Tool for the Prediction and Analysis of Anticancer Peptides. Molecules 2019; 24(10): 1973.
http://dx.doi.org/10.3390/molecules24101973 PMID: 31121946

[34] Pratiwi R, Malik AA, Schaduangrat N, *et al.* Protegrin-1 inhibits dengue NS2B-NS3 serine protease and viral replication in MK2 cells. BioMed Research International 2017.
http://dx.doi.org/10.1155/2017/9861752

[35] Win TS, Malik AA, Prachayasittikul V, S Wikberg JE, Nantasenamat C, Shoombuatong W. HemoPred: a web server for predicting the hemolytic activity of peptides. Future Med Chem 2017; 9(3): 275-91.
http://dx.doi.org/10.4155/fmc-2016-0188 PMID: 28211294

[36] Hongjaisee S, Nantasenamat C, Carraway TS, Shoombuatong W. HIVCoR: A sequence-based tool for predicting HIV-1 CRF01_AE coreceptor usage. Comput Biol Chem 2019; 80: 419-32.
http://dx.doi.org/10.1016/j.compbiolchem.2019.05.006 PMID: 31146118

[37] Charoenkwan P, Schaduangrat N, Nantasenamat C, Piacham T, Shoombuatong W. Correction: Shoombuatong, W., et al. iQSP: A Sequence-Based Tool for the Prediction and Analysis of Quorum Sensing Peptides via Chou's 5-Steps Rule and Informative Physicochemical Properties. *Int. J. Mol. Sci.* 2020, *21*, 75. Int J Mol Sci 2020; 21(7): 75.
http://dx.doi.org/10.3390/ijms21072629 PMID: 32290041

[38] Win TS, Schaduangrat N, Prachayasittikul V, Nantasenamat C, Shoombuatong W. PAAP: a web server for predicting antihypertensive activity of peptides. Future Med Chem 2018; 10(15): 1749-67.
http://dx.doi.org/10.4155/fmc-2017-0300 PMID: 30039980

[39] Laengsri V, Nantasenamat C, Schaduangrat N, Nuchnoi P, Prachayasittikul V, Shoombuatong W. TargetAntiAngio: A Sequence-Based Tool for the Prediction and Analysis of Anti-Angiogenic Peptides. Int J Mol Sci 2019; 20(12): 2950.
http://dx.doi.org/10.3390/ijms20122950 PMID: 31212918

[40] Su R, Hu J, Zou Q, Manavalan B, Wei L. Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. Brief Bioinform 2019.
PMID: 30649170

[41] Su Z-D, Huang Y, Zhang Z-Y, *et al.* iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. Bioinformatics 2018; 34(24): 4196-204.
http://dx.doi.org/10.1093/bioinformatics/bty508 PMID: 29931187

[42] Wei L, Su R, Luan S, *et al.* Iterative feature representations improve N4-methylcytosine site prediction. Bioinformatics 2019; 35(23): 4930-7.
http://dx.doi.org/10.1093/bioinformatics/btz408 PMID: 31099381

[43] Xu Z-C, Feng P-M, Yang H, Qiu W-R, Chen W, Lin H. iRNAD: a computational tool for identifying D modification sites in RNA sequence. Bioinformatics 2019; 35(23): 4922-9.
http://dx.doi.org/10.1093/bioinformatics/btz358 PMID: 31077296

[44] Zhang Z-Y, Yang Y-H, Ding H, Wang D, Chen W, Lin H. Design powerful predictor for mRNA subcellular location prediction in Homo sapiens. Brief Bioinform 2020.bbz177
http://dx.doi.org/10.1093/bib/bbz177 PMID: 31994694

[45] Zhu X-J, Feng C-Q, Lai H-Y, Chen W, Hao L. Predicting protein structural classes for low-similarity sequences by evaluating different features. Knowl Base Syst 2019; 163: 787-93.
http://dx.doi.org/10.1016/j.knosys.2018.10.007

[46] Manavalan B, Basith S, Shin TH, Choi S, Kim MO, Lee G. MLACP: machine-learning-based prediction of anticancer peptides. Oncotarget 2017; 8(44): 77121-36.
http://dx.doi.org/10.18632/oncotarget.20365 PMID: 29100375

[47] Manavalan B, Basith S, Shin TH, Lee DY, Wei L, Lee G. 4mCpred-EL: An Ensemble Learning Framework for Identification of DNA N^4-methylcytosine Sites in the Mouse Genome. Cells 2019; 8(11): 1332.
http://dx.doi.org/10.3390/cells8111332 PMID: 31661923

[48] Manavalan B, Basith S, Shin TH, Wei L, Lee G. mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. Bioinformatics 2019; 35(16): 2757-65.
http://dx.doi.org/10.1093/bioinformatics/bty1047
PMID: 30590410

[49] Manavalan B, Basith S, Shin TH, Wei L, Lee G. Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. Mol Ther Nucleic Acids 2019; 16: 733-44.
http://dx.doi.org/10.1016/j.omtn.2019.04.019 PMID: 31146255

[50] Manavalan B, Lee J. SVMQA: support-vector-machine-based protein single-model quality assessment. Bioinformatics 2017; 33(16): 2496-503.
http://dx.doi.org/10.1093/bioinformatics/btx222 PMID: 28419290

[51] Manavalan B, Shin TH, Kim MO, Lee G. AIPpred: sequence-based prediction of anti-inflammatory peptides using random forest. Front Pharmacol 2018; 9: 276.
http://dx.doi.org/10.3389/fphar.2018.00276 PMID: 29636690

[52] Manavalan B, Shin TH, Lee G. PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine. Front Microbiol 2018; 9: 476.
http://dx.doi.org/10.3389/fmicb.2018.00476 PMID: 29616000

[53] Manavalan B, Subramaniyam S, Shin TH, Kim MO, Lee G. Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. J Proteome Res 2018; 17(8): 2715-26.

http://dx.doi.org/10.1021/acs.jproteome.8b00148
PMID: 29893128

[54] Khatun MS, Hasan MM, Kurata H. PreAIP: computational prediction of anti-inflammatory peptides by integrating multiple complementary features. Front Genet 2019; 10: 129.
http://dx.doi.org/10.3389/fgene.2019.00129 PMID: 30891059

[55] Khatun S, Hasan M, Kurata H. Efficient computational model for identification of antitubercular peptides by integrating amino acid patterns and properties. FEBS Lett 2019; 593(21): 3029-39.
http://dx.doi.org/10.1002/1873-3468.13536 PMID: 31297788

[56] Lai H-Y, Zhang Z-Y, Su Z-D, *et al.* iProEP: a computational predictor for predicting promoter. Mol Ther Nucleic Acids 2019; 17: 337-46.
http://dx.doi.org/10.1016/j.omtn.2019.05.028 PMID: 31299595

[57] Li W-C, Deng E-Z, Ding H, Chen W, Lin H. iORI-PseKNC: a predictor for identifying origin of replication with pseudo k-tuple nucleotide composition. Chemom Intell Lab Syst 2015; 141: 100-6.
http://dx.doi.org/10.1016/j.chemolab.2014.12.011

[58] Lin H, Ding H, Guo F-B, Huang J. Prediction of subcellular location of mycobacterial protein using feature selection techniques. Mol Divers 2010; 14(4): 667-71.
http://dx.doi.org/10.1007/s11030-009-9205-1 PMID: 19908156

[59] Lin H, Liang Z-Y, Tang H, Chen W. Identifying sigma70 promoters with novel pseudo nucleotide composition. IEEE/ACM Trans Comput Biol Bioinformatics 2017.
PMID: 28186907

[60] Lv H, Zhang Z-M, Li S-H, Tan J-X, Chen W, Lin H. Evaluation of different computational methods on 5-methylcytosine sites identification. Brief Bioinform 2019.
PMID: 31157855

[61] Shoombuatong W, Prathipati P, Prachayasittikul V, *et al.* ES Wikberg J, Paul Gleeson M, Spjuth O. Towards predicting the cytochrome P450 modulation: from QSAR to proteochemometric modeling. Curr Drug Metab 2017; 18(6): 540-55.
http://dx.doi.org/10.2174/1389200218666170320121932 PMID: 28322159

[62] Shoombuatong W, Schaduangrat N, Nantasenamat C. Towards understanding aromatase inhibitory activity via QSAR modeling. EXCLI J 2018; 17: 688-708.
PMID: 30190660

[63] Dao F-Y, Lv H, Wang F, *et al.* Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. Bioinformatics 2019; 35(12): 2075-83.
http://dx.doi.org/10.1093/bioinformatics/bty943
PMID: 30428009

[64] Feng C-Q, Zhang Z-Y, Zhu X-J, *et al.* iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. Bioinformatics 2019; 35(9): 1469-77.
http://dx.doi.org/10.1093/bioinformatics/bty827 PMID: 30247625

[65] Hasan MM, Khatun MS, Kurata H. Large-scale assessment of bioinformatics tools for lysine succinylation sites. Cells 2019; 8(2): 95.
http://dx.doi.org/10.3390/cells8020095 PMID: 30696115

[66] Hasan MM, Khatun MS, Mollah MNH, Yong C, Dianjing G, Dianjing G. NTyroSite: Computational identification of protein nitrotyrosine sites using sequence evolutionary features. Molecules 2018; 23(7): 1667.
http://dx.doi.org/10.3390/molecules23071667 PMID: 29987232

[67] Hasan MM, Guo D, Kurata H. Computational identification of protein S-sulfenylation sites by incorporating the multiple sequence features information. Mol Biosyst 2017; 13(12): 2545-50.
http://dx.doi.org/10.1039/C7MB00491E PMID: 28990628

[68] Hasan MM, Khatun MS, Kurata H. A comprehensive review of in silico analysis for protein S-sulfenylation sites. Protein Pept Lett 2018; 25(9): 815-21.
http://dx.doi.org/10.2174/0929866525666180905110619
PMID: 30182830

[69] Hasan MM, Khatun MS, Mollah MNH, Yong C, Guo D. A systematic identification of species-specific protein succinylation sites using joint element features information. Int J Nanomedicine 2017; 12: 6303-15.
http://dx.doi.org/10.2147/IJN.S140875 PMID: 28894368

[70] Hasan MM, Kurata H. GPSuc: Global Prediction of Generic and Species-specific Succinylation Sites by aggregating multiple sequence features. PLoS One 2018; 13(10)e0200283
http://dx.doi.org/10.1371/journal.pone.0200283 PMID: 30312302

[71] Hasan MM, Manavalan B, Khatun MS, Kurata H. i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome. Int J Biol Macromol 2019.
http://dx.doi.org/10.1016/j.ijbiomac.2019.12.009
PMID: 31805335

[72] Hasan MM, Manavalan B, Khatun MS, Kurata H. Prediction of S-nitrosylation sites by integrating support vector machines and random forest. Molecular omics 2019; 15: 451-8.

[73] Hasan MM, Rashid MM, Khatun MS, Kurata H. Computational identification of microbial phosphorylation sites by the enhanced characteristics of sequence information. Sci Rep 2019; 9(1): 8258.
http://dx.doi.org/10.1038/s41598-019-44548-x PMID: 31164681

[74] Hasan MM, Yang S, Zhou Y, Mollah MNH. SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties. Mol Biosyst 2016; 12(3): 786-95.
http://dx.doi.org/10.1039/C5MB00853K PMID: 26739209

[75] Hasan MM, Zhou Y, Lu X, Li J, Song J, Zhang Z. Computational identification of protein pupylation sites by using profile-based composition of k-spaced amino acid pairs. PLoS One 2015; 10(6)e0129635
http://dx.doi.org/10.1371/journal.pone.0129635 PMID: 26080082

[76] Charoenkwan P, Nantasenamat C, Hasan MM, Shoombuatong W. Meta-iPVP: a sequence-based meta-predictor for improving the prediction of phage virion proteins using effective feature representation. J Comput Aided Mol Des 2020; 34(10): 1105-16.
http://dx.doi.org/10.1007/s10822-020-00323-z PMID: 32557165

[77] Charoenkwan P, Nantasenamat C, Hasan MM, Shoombuatong W. iTTCA-Hybrid: Improved and robust identification of tumor T cell antigens by utilizing hybrid feature representation. Anal Biochem 2020; 599113747
http://dx.doi.org/10.1016/j.ab.2020.113747 PMID: 32333902

[78] Charoenkwan P, Yana J, Schaduangrat N, Nantasenamat C, Hasan MM, Shoombuatong W. iBitter-SCM: Identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides. Genomics 2020; 112(4): 2813-22.
http://dx.doi.org/10.1016/j.ygeno.2020.03.019 PMID: 32234434

[79] Hasan MM, Manavalan B, Shoombuatong W, Khatun MS, Kurata H. i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation. Plant Mol Biol 2020; 103(1-2): 225-34.
http://dx.doi.org/10.1007/s11103-020-00988-y PMID: 32140819

[80] Hasan MM, Schaduangrat N, Basith S, Lee G, Shoombuatong W, Manavalan B. HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. Bioinformatics 2020; 36(11): 3350-6.
http://dx.doi.org/10.1093/bioinformatics/btaa160
PMID: 32145017

[81] Hasan MM, Manavalan B, Shoombuatong W, Khatun MS, Kurata H. i4mC-Mouse: Improved identification of DNA N4-methylcytosine sites in the mouse genome using multiple encoding schemes. Comput Struct Biotechnol J 2020; 18: 906-12.
http://dx.doi.org/10.1016/j.csbj.2020.04.001 PMID: 32322372

[82] Kawashima S, Kanehisa M. AAindex: amino acid index database. Nucleic Acids Res 2000; 28(1): 374-4.
http://dx.doi.org/10.1093/nar/28.1.374 PMID: 10592278

[83] Breiman L. Random forests. Mach Learn 2001; 45: 5-32.
http://dx.doi.org/10.1023/A:1010933404324

[84] Breiman L. Classification and regression trees. Routledge 2017.
http://dx.doi.org/10.1201/9781315139470

[85] Boopathi V, Subramaniyam S, Malik A, Lee G, Manavalan B, Yang D-C. mACPpred: a support vector machine-based metapredictor for identification of anticancer peptides. Int J Mol Sci 2019; 20(8): 1964.
http://dx.doi.org/10.3390/ijms20081964 PMID: 31013619

[86] Basith S, Manavalan B, Shin TH, Lee G. iGHBP: Computational identification of growth hormone binding proteins from sequences using extremely randomised tree. Comput Struct Biotechnol J 2018; 16: 412-20.

http://dx.doi.org/10.1016/j.csbj.2018.10.007 PMID: 30425802

[87] Manavalan B, Shin TH, Kim MO, Lee G. PIP-EL: A new ensemble learning method for improved proinflammatory peptide predictions. Front Immunol 2018; 9: 1783.
http://dx.doi.org/10.3389/fimmu.2018.01783 PMID: 30108593

[88] Basith S, Manavalan B, Hwan Shin T, Lee G. Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening. Med Res Rev 2020; 40(4): 1276-314.
http://dx.doi.org/10.1002/med.21658 PMID: 31922268

[89] Basith S, Manavalan B, Shin TH, Lee G. SDM6A: A web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. Mol Ther Nucleic Acids 2019; 18: 131-41.
http://dx.doi.org/10.1016/j.omtn.2019.08.011 PMID: 31542696

[90] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995; 20: 273-97.
http://dx.doi.org/10.1007/BF00994018

[91] Drucker H, Burges CJ, Kaufman L, Smola AJ, Vapnik V. Support vector regression machines.ed^eds, Advances in neural information processing systems. 1997; pp. 155-61.

[92] Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 2010; 26(5): 680-2.
http://dx.doi.org/10.1093/bioinformatics/btq003 PMID: 20053844

[93] Ettayapuram Ramaprasad AS, Singh S, Gajendra P S R, Venkatesan S. AntiAngioPred: a server for prediction of anti-angiogenic peptides. PLoS One 2015; 10(9)e0136990
http://dx.doi.org/10.1371/journal.pone.0136990 PMID: 26335203

[94] Lata S, Sharma BK, Raghava GP. Analysis and prediction of antibacterial peptides. BMC Bioinformatics 2007; 8: 263.
http://dx.doi.org/10.1186/1471-2105-8-263 PMID: 17645800

[95] Wei L, Zhou C, Chen H, Song J, Su R. ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. Bioinformatics 2018; 34(23): 4007-16.
http://dx.doi.org/10.1093/bioinformatics/bty451 PMID: 29868903

[96] Wei L, Xing P, Su R, Shi G, Ma ZS, Zou Q. CPPred-RF: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency. J Proteome Res 2017; 16(5): 2044-53.
http://dx.doi.org/10.1021/acs.jproteome.7b00019 PMID: 28436664

[97] Rajput A, Gupta AK, Kumar M. Prediction and analysis of quorum sensing peptides based on sequence features. PLoS One 2015; 10(3), e0120066.
http://dx.doi.org/10.1371/journal.pone.0120066 PMID: 25781990

[98] Li N, Kang J, Jiang L, He B, Lin H, Huang J. PSBinder: a web service for predicting polystyrene surface-binding peptides. BioMed research international 2017.
http://dx.doi.org/10.1155/2017/5761517

[99] Shoombuatong W, Schaduangrat N, Nantasenamat C. Unraveling the bioactivity of anticancer peptides as deduced from machine learning. EXCLI J 2018; 17: 734-52.
PMID: 30190664

[100] Chou K-C. Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 2011; 273(1): 236-47.
http://dx.doi.org/10.1016/j.jtbi.2010.12.024 PMID: 21168420

[101] Xiao N, Cao D-S, Zhu M-F, Xu Q-S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics 2015; 31(11): 1857-9.
http://dx.doi.org/10.1093/bioinformatics/btv042 PMID: 25619996

[102] Chen W, Ding H, Feng P, Lin H, Chou K-C. iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget 2016; 7(13): 16895-909.
http://dx.doi.org/10.18632/oncotarget.7815 PMID: 26942877

[103] Huang H-L, Charoenkwan P, Kao T-F, *et al.* Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition.ed^eds, Bmc Bioinformatics BioMed Central. 2012.
http://dx.doi.org/10.1186/1471-2105-13-S17-S3

[104] Charoenkwan P, Shoombuatong W, Lee H-C, Chaijaruwanich J, Huang H-L, Ho S-Y. SCMCRYS: predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of P-collocated amino acid pairs. PLoS One 2013; 8(9)e72368
http://dx.doi.org/10.1371/journal.pone.0072368 PMID: 24019868

[105] Huang H-L. Propensity scores for prediction and characterization of bioluminescent proteins from sequences. PLoS One 2014; 9(5)e97158
http://dx.doi.org/10.1371/journal.pone.0097158 PMID: 24828431

[106] Charoenkwan P, Kanthawong S, Schaduangrat N, Yana J, Shoombuatong W. PVPred-SCM: Improved Prediction and Analysis of Phage Virion Proteins Using a Scoring Card Method. Cells 2020; 9(2): 353.
http://dx.doi.org/10.3390/cells9020353 PMID: 32028709

[107] Vasylenko T, Liou Y-F, Chiou P-C, *et al.* SCMBYK: prediction and characterization of bacterial tyrosine-kinases based on propensity scores of dipeptides. BMC Bioinformatics 2016; 17(Suppl. 19): 514.
http://dx.doi.org/10.1186/s12859-016-1371-4 PMID: 28155663

[108] Liou Y-F, Charoenkwan P, Srinivasulu Y, *et al.* SCMHBP: prediction and analysis of heme binding proteins using propensity scores of dipeptides. BMC Bioinformatics 2014; 15(Suppl. 16): S4.
http://dx.doi.org/10.1186/1471-2105-15-S16-S4 PMID: 25522279

[109] Liou Y-F, Vasylenko T, Yeh C-L, *et al.* SCMMTP: identifying and characterizing membrane transport proteins using propensity scores of dipeptides. BMC Genomics 2015; 16(Suppl. 12): S6.
http://dx.doi.org/10.1186/1471-2164-16-S12-S6 PMID: 26677931

[110] Vasylenko T, Liou Y-F, Chen H-A, Charoenkwan P, Huang H-L, Ho S-Y. SCMPSP: Prediction and characterization of photosynthetic proteins based on a scoring card method.ed^eds, BMC bioinformatics BioMed Central. 2015.

DISCLAIMER: The above article has been published in Epub (ahead of print) on the basis of the materials provided by the author. The Editorial Department reserves the right to make minor modifications for further improvement of the manuscript.