






Review article:

RECENT DEVELOPMENT OF MACHINE LEARNING-BASED METHODS FOR THE PREDICTION OF DEFENSIN FAMILY AND SUBFAMILY

Phasit Charoenkwan^a, Nalini Schaduagrath^b, S. M. Hasan Mahmud^{b,c},
Orawit Thinnukool^a, Watshara Shoombuatong^{b,*}

^a Modern Management and Information Technology, College of Arts, Media and Technology, Chiang Mai University, Chiang Mai, Thailand, 50200

^b Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand, 10700

^c Department of Computer Science, American International University-Bangladesh (AIUB), Kuratoli, Dhaka 1229, Bangladesh

* **Corresponding author:** Watshara Shoombuatong, Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand, 10700. Phone: +66 2 441 4371; Fax: +66 2 441 4380; E-mail: watshara.sho@mahidol.ac.th

<https://dx.doi.org/10.17179/excli2022-4913>

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>).

ABSTRACT

Nearly all living species comprise of host defense peptides called defensins, that are crucial for innate immunity. These peptides work by activating the immune system which kills the microbes directly or indirectly, thus providing protection to the host. Thus far, numerous preclinical and clinical trials for peptide-based drugs are currently being evaluated. Although, experimental methods can help to precisely identify the defensin peptide family and subfamily, these approaches are often time-consuming and cost-ineffective. On the other hand, machine learning (ML) methods are able to effectively employ protein sequence information without the knowledge of a protein's three-dimensional structure, thus highlighting their predictive ability for the large-scale identification. To date, several ML methods have been developed for the *in silico* identification of the defensin peptide family and subfamily. Therefore, summarizing the advantages and disadvantages of the existing methods is urgently needed in order to provide useful suggestions for the development and improvement of new computational models for the identification of the defensin peptide family and subfamily. With this goal in mind, we first provide a comprehensive survey on a collection of six state-of-the-art computational approaches for predicting the defensin peptide family and subfamily. Herein, we cover different important aspects, including the dataset quality, feature encoding methods, feature selection schemes, ML algorithms, cross-validation methods and web server availability/usability. Moreover, we provide our thoughts on the limitations of existing methods and future perspectives for improving the prediction performance and model interpretability. The insights and suggestions gained from this review are anticipated to serve as a valuable guidance for researchers for the development of more robust and useful predictors.

Keywords: Defensins, sequence analysis, bioinformatics, classification, machine learning, feature selection

INTRODUCTION

Nearly all living species comprise of host defense peptides called defensins, that are crucial for innate immunity. Defensins are

considered as a part of the antimicrobial protein family and are rich in cysteine. Furthermore, defensins offer assistance to cells in combating bacterial (Menendez and Finlay,

2007), viral (Wilson et al., 2013) and fungal infections (Parisi et al., 2019; Sathoff and Samac, 2019) by destroying the structural integrity of bacterial cell membranes (Bun Ng et al., 2013; De Coninck et al., 2013; de Oliveira Dias and Franco, 2015). Precisely, defensins bind to the microbial cell membrane forming a pore-like channel in the membrane which cause ions and nutrients to leak through biphasic permeabilization (Jarczak et al., 2013). Further evidence suggests that a predisposition to diseases (Kim et al., 2015) may be caused by an imbalance or reduction (Albrethsen et al., 2005) of defensins in various organisms.

In addition, defensins are shown to exhibit a wide range of key applications in various industries thus, highlighting the importance of their design to fit specific needs (Whiston et al., 2017). Nevertheless, conventional experimental approaches, such as nuclear magnetic resonance (de Medeiros et al., 2010), are often time-consuming and not cost-effective. On the other hand, there is an increase in the number of new proteins sequenced by next-generation sequencing techniques. As a result, a large number of novel defensin candidates can potentially be found in these proteins. Thus, it is desirable to rapidly and accurately identify defensins from large-scale proteins. Previously, machine-learning (ML) methods were naturally selected to conduct a large-scale identification and prediction of several proteins and peptides (Li et al., 2015; Lin et al., 2010, 2019; Lv et al., 2020; Su et al., 2018; Xu et al., 2019; Zhang et al., 2021; Zulfiqar et al., 2021). These approaches are able to effectively employ protein sequence information without the knowledge of the protein's three-dimensional structure. Furthermore, the general machine learning framework used for the prediction of defensins involves four major steps as summarized in Figure 1, including, the preparation of training and independent test datasets, feature extraction, feature optimization, and model development and evaluation. Currently, there

are six state-of-the-art computational approaches that have been developed for the *in silico* prediction of defensins, including, Karnik's method (Karnik et al., 2009), ID_RAAA (Zuo and Li, 2009), Defensinpred (Ramya Kumari et al., 2012), iDPF-PseRAAAC (Zuo et al., 2015), iDEF-PseRAAC (Zuo et al., 2019) and DEF-PRED (Kaur et al., 2021) as summarized in Table 1.

We categorize these computational methods in Table 1 into two groups according to their predictive applications. The first group is comprised of computational methods developed for the *in silico* prediction of defensins, which make up two out of six existing methods (i.e., Karnik's method (Karnik et al., 2009) and DEF-PRED (Kaur et al., 2021)). The second group is focused on those computational methods which have been developed for the *in silico* prediction of the defensin peptide family and subfamily, and comprise of four out of the six existing methods (i.e., ID_RAAA (Zuo and Li, 2009), Defensinpred (Ramya Kumari et al., 2012), iDPF-PseRAAAC (Zuo et al., 2015) and iDEF-PseRAAC (Zuo et al., 2019)).

Motivated by the above-mentioned considerations, we provide a comprehensive comparison and analysis of the current state-of-the-art computational methods. Major contributions of this review article could be summarized as follows: (i) to the best of our knowledge, this article provides the first comprehensive review on the development of computational approaches for the *in silico* identification of the defensin peptide family and subfamily; (ii) we have provided several important aspects that play a crucial role for the development of reliable and stable prediction models, covering, their dataset quality, feature encoding methods, feature selection schemes, ML algorithms, cross-validation methods and web server availability/usability and (iii) we have discussed the limitations as well as the advantages and disadvantages of existing methods and provided future perspectives for improving the prediction performance and model interpretability.

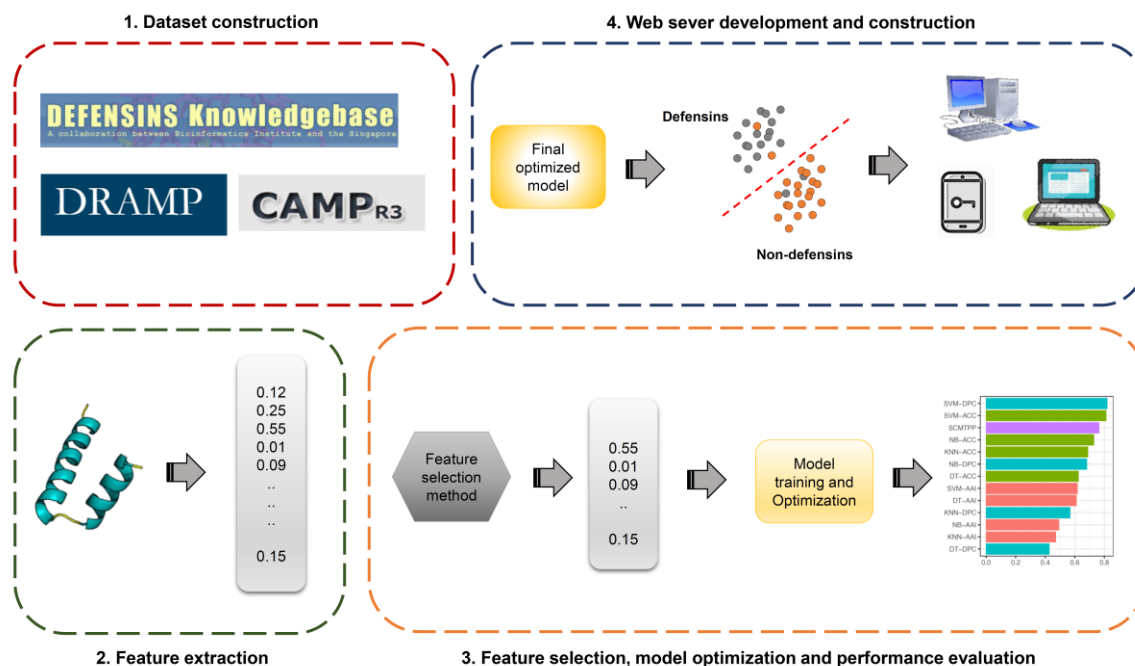


Figure 1: The general machine learning framework of the prediction of defensins and their family/sub-family

Table 1: A list of currently available machine learning-based methods for the predictions of defensins and their family/subfamily

Method	Classifier ^a	Feature ^b	Feature selection	Evaluation strategy ^c	Web sever availability, status
Karnik's method (Karnik et al., 2009)	RF	RQA	N/A	10CV/IND	No
ID_RAAA (Zuo and Li, 2009)	ID	RAAA	N/A	Jackknife test	No
Defensinpred (Ramya Kumari et al., 2012)	SVM	PAAC	N/A	Jackknife test	Yes, inactive
iDPF-PseRAAAC (Zuo et al., 2015)	SVM	RAAA	N/A	Jackknife test	Yes, inactive
iDEF-PseRAAC (Zuo et al., 2019)	SVM	RAAA	F-score	Jackknife test	Yes, active
DEFRED (Kaur et al., 2021)	SVM	More than 20 descriptors, including AAC, APAAC, DPC, CTD, PSSM, et al.	SVC-L1	5CV/IND	Yes, active

^a ID: increment of diversity, RF: random forest, SVM: support vector machine

^b AAC: amino acid composition, APAAC: pseudo amino acid composition, CTD: composition-transition-distribution, DPC: dipeptide composition, PAAC: pseudo amino acid composition, PSSM: position-specific scoring matrix, RAAA: reduced amino acid alphabet, RQA: recurrence quantification analysis

^c 5CV: 5-fold cross-validation, 10CV: 10-fold cross-validation, IND: independent test

MATERIALS AND METHODS

General machine learning framework of the predictions of defensins and their family/subfamily

Until now, a number of computational approaches for *in silico* prediction of defensins and their family/subfamily have been developed (Karnik et al., 2009; Kaur et al., 2021; Ramya Kumari et al., 2012; Seebah et al., 2007; Zuo and Li, 2009; Zuo et al., 2015). The general machine learning framework used for the prediction of defensins involves four major steps as summarized in Figure 1. The first step is the preparation of training and independent test datasets. The training and independent test datasets are used for cross-validation and model validation purposes, respectively. The second step is the feature extraction. There are many feature encoding schemes that are used to encode variable-length proteins and peptides into fixed-length feature vectors. However, using the original feature dimensions might include irrelevant/redundant information as well as require additional computational resources during model optimization. On the other hand, the performance is not robust in many cases. Therefore, the third step is to select a set of important features. The fourth step is to train

and evaluate a prediction model. The effectiveness and robustness of the prediction models are assessed on the independent test dataset. Finally, the optimal prediction model is employed to establish a web server.

Datasets

We reviewed all the datasets used for developing the existing methods (Karnik et al., 2009; Kaur et al., 2021; Zuo and Li, 2009; Zuo et al., 2015, 2019). The detailed information of these datasets are provided in Table 2. As seen in Table 2, the datasets of Zou2015 (Zuo et al., 2015) and Zou2019 (Zuo et al., 2019) derived from the defensins knowledge-base (Seebah et al., 2007) applied a lower CD-HIT threshold of 0.8 in order to exclude all homologous sequences. For the Zou2015 dataset (Zuo et al., 2015), it contained 333 defensin proteins, which were classified into 60 insect defensins, 34 invertebrate defensins, 42 plant defensins, 157 vertebrate defensins and 40 unclassified defensins. Among the 157 vertebrate defensins, they were also classified as alpha-, beta- and theta-defensins. In the case of the Zou2019 dataset (Zuo et al., 2019), it contained 333 defensin proteins, which were classified as 60 insect defensins, 31 invertebrate defensins, 42 plant defensins, 157 vertebrate defensins and 38 unclassified defensins.

Table 2: The detailed information of the existing datasets used for analyzing in this review

Dataset	CD-HIT threshold	No of samples	Dataset availability
Karnik2009 (Karnik et al., 2009)	1.0	238 defensins and 238 non-defensins	No
Zou2009 (Zuo and Li, 2009)	1.0	286 defensins (37 insect defensins, 48 plant defensins, 190 vertebrate defensins, 11 unclassified defensins)	No
Zou2015 (Zuo et al., 2015)	0.8	333 defensins (60 insect defensins, 34 invertebrate defensins, 42 plant defensins, 157 vertebrate defensins, 40 unclassified defensins)	No
Zou2019 (Zuo et al., 2019)	0.8	328 defensins (60 insect defensins, 31 invertebrate defensins, 42 plant defensins, 157 vertebrate defensins, 38 unclassified defensins)	Yes
Kaur2021 (Kaur et al., 2021)	1.0	1035 defensins and 1035 non-defensins (main dataset), 1035 defensins and 1054 non-defensins (alternative dataset)	Yes

Recently, Kaur et al. established two up-to-date datasets (Kaur2021 (Kaur et al., 2021)) containing a main and alternative datasets. In Kaur2021, the defensin samples were collected from various sources, including literature (Zuo and Li, 2009; Zuo et al., 2015, 2019), DRAMP2.0 (Kang et al., 2019) and CAMPR3 (Waghu et al., 2016). The samples in the work of (Kaur et al., 2021) were experimentally validated defensins exhibiting antimicrobial activity and the number of residues were in the range of 10-60. However, sequences containing non-natural or non-standard amino acids (B, J, O, U, X, and Z) were excluded. As a result, a total of 1,036 unique defensins were obtained and used to create the main and alternative datasets. For the main dataset, it contained 1,036 positives and 1,036 negatives, where positives and negatives are experimentally validated defensins and antimicrobial peptides (AMPs), respectively. In case of the alternative dataset, it contained 1,036 positives and 1,054 negatives, where positives and negatives are experimentally validated defensins and selected peptides from Swiss-Prot (UniProt Consortium, 2017), respectively.

Machine learning algorithms used for the prediction of defensins and their family/subfamily

As can be seen from Table 1, SVM is the most popular ML algorithm for building computational models in the prediction of defensins and their family/subfamily, used in Defensinpred (Ramya Kumari et al., 2012), iDPF-PseRAAAC (Zuo et al., 2015), iDEF-PseRAAC (Zuo et al., 2019) and DEFRED (Kaur et al., 2021). In the meanwhile, the RF and ID methods were used to develop Karnik's method (Karnik et al., 2009) and ID_RAAA (Zuo and Li, 2009), respectively. Hereafter, we provide the basic concepts of SVM and RF algorithms.

SVM is a well-known and powerful ML algorithm that is commonly employed to deal with binary classification problems (Vapnik, 1999, 2000). In particular, SVM maps the given input features into a higher dimensional

space using kernel functions and finds optimal hyperplanes that can separate positive samples from negative samples. To date, there are several kernel functions used for developing SVM classifiers, such as linear function, polynomial function, sigmoid function and gaussian radial basis function (RBF). Amongst the several kernel functions, the RBF kernel is the most commonly used one. In order to enhance the performance of SVM classifiers, a grid search strategy was utilized to optimize the two important aspects of the RBF kernel, including C (controls the trade-off between the misclassification rate and margin) and γ (the kernel width parameter). Although SVM often yields satisfactory prediction performances, this method is known as a black-box computation method (Ahmad et al., 2022; Charoenkwan et al., 2021d; Li et al., 2021a; Wei et al., 2021).

RF is another powerful and widely employed ML algorithm for dealing with binary classification problems (Charoenkwan et al., 2020d; Hasan et al., 2020, 2021; Manavalan et al., 2019a, b; Su et al., 2020). RF is an ensemble-based method originally introduced by Leo Breiman (2001) that is created by integrating a number of decision trees. Each decision tree consists of a single root node, leaf nodes and a number of intermediate nodes (Breiman, 2001). An if-then rule is derived from the path connecting the root node to the leaf node. As a result, RF is able to provide a collection of if-then rules. Therefore, this method is known as a white-box computation method. In order to enhance the performance of RF classifiers, a grid search strategy was employed to optimize two key parameters: *mtree* (the number of decision trees) and *mtry* (the number of selected features).

Performance evaluation and evaluation strategy

Here, we employed five commonly used performance measures to comprehensively evaluate and analyze the performance of the six state-of-the-art predictors (Karnik et al., 2009; Kaur et al., 2021; Ramya Kumari et al., 2012; Seebah et al., 2007; Zuo and Li, 2009;

Zuo et al., 2015), including ACC, Sn, Sp, MCC and OA. The definitions of these performance measures are defined as follows:

$$\text{ACC} = \frac{\text{TP}(i) + \text{TN}(i)}{(\text{TP}(i) + \text{TN}(i) + \text{FP}(i) + \text{FN}(i))} \quad (1)$$

$$\text{Sn} = \frac{\text{TP}(i)}{(\text{TP}(i) + \text{FN}(i))} \quad (2)$$

$$\text{Sp} = \frac{\text{TN}(i)}{(\text{TN}(i) + \text{FP}(i))} \quad (3)$$

$$\text{OA} = \frac{1}{N} \sum_{i=1}^M \text{TP}(i) \quad (4)$$

where TP(i), TN(i), FP(i) and FN(i) denote true positives, true negatives, false positives and false negatives of the i^{th} class or the i^{th} family. And, M and N are the number of subsets and the number of samples, respectively.

RESULTS AND DISCUSSION

State-of-the-art computational approaches for the prediction of defensins

In this section, we conducted a performance comparison for the two existing methods for the prediction of defensins (Karnik's method (Karnik et al., 2009) and DEF-PRED (Kaur et al., 2021)). The performance comparison results are provided in Table 3. In 2009, Karnik et al. developed the first ML-based predictor (called Karnik's method (Karnik et al., 2009)) to discriminate defensins from non-defensins. The dataset consisting of 238 defensins and 238 non-defensins, was randomly partitioned into train (80 %) and independent test (20 %) datasets. Specifically, Karnik's method was created using the RQA descriptor coupled with RF algorithm by tuning the *mtry* parameter based on the 10-fold cross-validation scheme and the training dataset.

Recently, Kaur et al. created DEF-PRED (Kaur et al., 2021) for specifically discriminating defensins from AMPs (antimicrobial peptides, or non-defensins). DEF-PRED was developed using various types of features, including AAC, APAAC, DPC, CTD, PSSM, etc. In their study, the support vector classifier (SVC) coupled with linear kernel penalized with L1 regularization (SVC-L1) method was employed to determine the optimal feature subset. In addition, several ML algorithms were used to develop the models, including RF, SVM, extra tree (ET), logistic regression (LR), k-nearest neighbors (KNN) and multilayer perceptron (MLP). The model that achieved the highest predictive performance was considered as the optimal one (DEF-PRED). A user-friendly web server is publicly available at <https://webs.iitd.edu.in/raghava/defpred/>. As seen in Tables 1 and 3, DEF-PRED outperforms Karnik's method in terms of generalization ability, robustness and utility.

State-of-the-art computational approaches for the prediction of defensin family and subfamily

As mentioned above, there are four existing methods developed for the predictions of defensins family and vertebrate defensins subfamily, including ID_RAAA (Zuo and Li, 2009), Defensinpred (Ramya Kumari, et al., 2012), iDPF-PseRAAC (Zuo et al., 2015) and iDEF-PseRAAC (Zuo et al., 2019). The performance comparison results are provided in Tables 4-5. In 2009, Zuo and Li first proposed ID_RAAA (Zuo and Li, 2009), an ID-based approach in conjunction with RAAA descriptor. The ID is a similarity-based approach used for measuring the similarity score of two diversity samples. ID_RAAA was created for the prediction of the defensin family (vertebrate defensins, plant defensins, insect defensins and others) and subfamily (alpha-type, beta-type and theta-type). In the RAAA descriptor, there are two main parameters:

Table 3: Performance comparison of Karnik’s method and DEFPRED for the prediction of defensins

Cross-validation	Method	No of samples ^d	No of features	ACC (%)	Sn (%)	Sp (%)	MCC
10-fold CV	Karnik’s method ^a	190P,190N	7	78.20	-	-	-
	DEFPRED ^b	828P,828N	60	93.00	89.26	96.74	0.86
	DEFPRED ^c	828P,843N	50	93.96	90.82	97.10	0.88
Independent test	Karnik’s method ^a	190P,190N	7	79.16	73.60	81.20	0.56
	DEFPRED ^b	208P,207N	60	96.59	95.17	97.98	0.93
	DEFPRED ^c	208P,211N	50	98.09	97.10	99.05	0.96

^a Performance results were obtained from Karnik et al. (2009).

^b Results based on the main dataset (Kaur et al., 2021).

^c Results based on the alternative dataset (Kaur et al., 2021).

^d P: positive samples, N: negative samples

Table 4: Performance comparison of ID_RAAA, iDPF-PseRAAAC and iDEF-PseRAAC for the prediction of defensins family

Family	Method ^a	Sn (%)	Sp (%)	MCC
Insect	ID_RAAA	79.36	-	-
	iDPF-PseRAAAC	90.00	97.07	0.86
	iDEF-PseRAAC	96.67	98.13	0.93
Invertebrate	ID_RAAA	-	-	-
	iDPF-PseRAAAC	61.76	97.32	0.64
	iDEF-PseRAAC	74.19	97.64	0.73
Plant	ID_RAAA	85.33	-	-
	iDPF-PseRAAAC	90.48	98.97	0.90
	iDEF-PseRAAC	92.86	98.60	0.91
Vertebrate	ID_RAAA	95.74	-	-
	iDPF-PseRAAAC	99.36	88.64	0.88
	iDEF-PseRAAC	97.45	97.08	0.95
Unclassified	ID_RAAA	74.73	-	-
	iDPF-PseRAAAC	40.00	96.63	0.46
	iDEF-PseRAAC	68.42	97.23	0.69

^a Performance of ID_RAAA was obtained from Karnik et al. (2009), while Performance of iDPF-PseRAAAC and iDEF-PseRAAC was obtained from Zuo et al. (2019).

different sizes (S = 5, 8, 9, 11, 13, 20) and N-peptide compositions (N = 1, 2, 3). For the prediction of the defensin family, an OA of 91.36 % was obtained by using a combination of N = 2 and S = 13 as indicated by the jack-knife test. The combination of N = 2 and S =

13 still achieved an OA of 94.21 % for the prediction of the defensin subfamily.

Zuo et al. (2015) introduced iDPF-PseRAAAC, a multi-class SVM predictor coupled with the RAAA descriptor. Zuo’s study was proposed to address a small number of samples (Zuo and Li, 2009). As a result, Zuo et al.

collected more than 500 defensin proteins from the defensins knowledgebase (Seebah et al., 2007). Then, the CD-HIT threshold of 0.8 was used to exclude sequence redundancy. Finally, Zuo et al. obtained a dataset containing 333 defensin proteins. These defensin proteins could be classified into five families, including insect defensins, invertebrate defensins, plant defensins, vertebrate defensins and unclassified defensins. The SVM classifier coupled with a combination of $N = 2$ and $S = 13$ (called iDPF-PseRAAAC) yielded an OA of 99.36 %. In addition, the 10-fold cross-validation results were also performed to assess the predictive ability of iDPF-PseRAAAC. The 10-fold cross-validation and jackknife test results were 83.78 % and 85.59 %, respectively. For the vertebrate defensin subfamily prediction, iDPF-PseRAAAC provided an OA of 98.39 %, while the MCC for the prediction of alpha-type, beta-type and theta-type was 0.97, 0.96 and 0.89, respectively (Table 5).

In 2019, Zuo et al. presented iDEF-PseRAAC (Zuo et al., 2019) by applying SVM algorithm and the F-score method. In iDEF-PseRAAC, it was developed based on a brand-new descriptor (i.e., reduced amino acid resource) containing more than 600 types of features. Their comparative results showed that the DPC of type 5 and cluster 19 ($T = 5$, $C = 19$) provided an OA of 91.16 %. To improve the predictive performance, the F-score method was employed to select informative features. Then, the 329 selected informative features were obtained and they achieved an

OA of 92.38 %. The SVM classifier in conjunction with the 329 selected informative features was considered as iDEF-PseRAAC in the work of Zuo et al. (2019). In the case of the defensin family prediction, iDEF-PseRAAC gave an OA of 92.38 %. Meanwhile for the vertebrate defensin subfamily prediction, iDEF-PseRAAC gave an OA of 98.79 %, an Sn of 0.99, and an Sp of 0.99.

From Table 2, it can be observed that, iDPF-PseRAAAC and iDEF-PseRAAC were developed for predicting the five defensin protein families (i.e., insect defensins, invertebrate defensins, plant defensins, vertebrate defensins and unclassified defensins), while ID_RAAA were developed for predicting the four defensin protein families (i.e., insect defensins, plant defensins, vertebrate defensins and unclassified defensins). Therefore, we conducted a performance comparison between iDPF-PseRAAAC and iDEF-PseRAAC in order to make a fair conclusion. As can be seen from Table 4, iDEF-PseRAAC achieves the best overall performance as compared with iDPF-PseRAAAC for all the five families of defensins in terms of Sn and MCC. To be specific, the MCC of iDEF-PseRAAC was 0.86, 0.64, 0.90, 0.88 and 0.46 respectively, which were 7 %, 9 %, 1 %, 7 % and 23 % higher than that of iDPF-PseRAAAC for insect defensins, invertebrate defensins, plant defensins, vertebrate defensins and unclassified defensins, respectively (Table 4). Taken together, iDEF-PseRAAC outperforms ID_RAAA and iDPF-PseRAAAC in terms of predictive performance and robustness.

Table 5: Performance comparison of ID_RAAA and iDPF-PseRAAAC for the prediction of vertebrate defensins subfamily

Family	Method ^a	Sn (%)	Sp (%)	MCC
Alpha-type	ID_RAAA	91.67	-	0.91
	iDPF-PseRAAAC	95.83	100.00	0.97
Beta-type	ID_RAAA	96.03	-	0.92
	iDPF-PseRAAAC	100.00	94.81	0.96
Theta-type	ID_RAAA	75.00	-	0.46
	iDPF-PseRAAAC	80.00	100.00	0.89

^a Performance of ID_RAAA and iDPF-PseRAAAC was obtained from Karnik et al. (2009) and Zuo et al. (2015), respectively.

Characterization of defensins based on sequence information

Kaur et al. (2021) provided compositional analysis and preferential position analysis based on the main and alternative datasets. As mentioned above, the main dataset contains 1,036 defensins and 1,036 AMPs, while the alternative dataset contains 1,036 defensins and 1,054 non-defensins. As shown in Figure 2A, it can be observed that Cys, Asp, Glu, Asn, Arg, Thr and Tyr were found to be abundant in defensins as compared to AMPs, while Phe, Ile, Ala, Lys, Leu were found to be abundant in AMPs as compared to defensins. Interestingly, most of the amino acids were significantly different between the classes at the level of $P < 0.05$, with the exception of His ($P = 0.154$), Pro ($P = 0.369$) and Trp ($P = 0.289$). In addition, Figure 2B reveals that Cys, Gly, Arg, and Thr were abundant in defensins as compared to non-defensins, while Asp, Val, Glu, Leu and Ala were abundant in non-defensins as compared to defensins. Furthermore, most of the amino acids were significantly different between the classes at the level of $P < 0.05$, with the exception of His ($P = 0.575$), Pro ($P = 0.341$), Ser ($P = 0.674$) and Thr ($P = 0.755$). Taken together, Cys, Tyr, Arg and Asn might be important amino acids for defensins. In addition, the prevalence of these four amino acids (i.e., Cys, Try, Arg and Asn) are significantly different between defensins and AMPs/non-defensins at the level of $P < 0.05$.

Web server availability and usability

As can be seen from Table 1, among the five state-of-the-art computational approaches, four of them were implemented as web servers for the prediction of defensins (DEF-PRED (Kaur et al., 2021)) and their family/subfamily (Defensinpred (Ramya

Kumari et al., 2012), iDPF-PseRAAAC (Zuo et al., 2015) and iDEF-PseRAAC (Zuo et al., 2019)). Unfortunately, only two web servers (iDEF-PseRAAC (Zuo et al., 2019) and DEF-PRED (Kaur et al., 2021)) were functional during our manuscript preparation (accessed on 13 March 2022).

iDEF-PseRAAC is a multi-class SVM predictor coupled with RAAA descriptor. iDEF-PseRAAC is freely available at <http://bioinfor.imu.edu.cn/idpf/public/>. In the case of the iDEF-PseRAAC web server, the query sequence pertains to five defensin families, including insect defensins, invertebrate defensins, plant defensins, vertebrate defensins and unclassified defensins.

DEF-PRED, on the other hand, is an *in silico* method developed for the prediction and design of defensins. The web server provides users with two options for the prediction of the query sequence: (i) discriminating defensins from AMPs (obtained from Model 1) and (ii) discriminating defensins from any random protein sequences (obtained from Model 2). In particular, Model 1 and Model 2 were trained using SVM-based models coupled with the 50 and 60 selected important features, respectively (Kaur et al., 2021). Moreover, DEF-PRED is able to predict potential defensins from primary protein sequences, but the length of the query sequence should be in the range of 10 to 60 residues. Furthermore, the web server provides users with a protein scan module that plays an important role in identifying specific regions in a protein. These regions can contribute in the designing of defensins. The DEF-PRED web server is freely available at <https://webs.iiitd.edu.in/raghava/defpred/>.

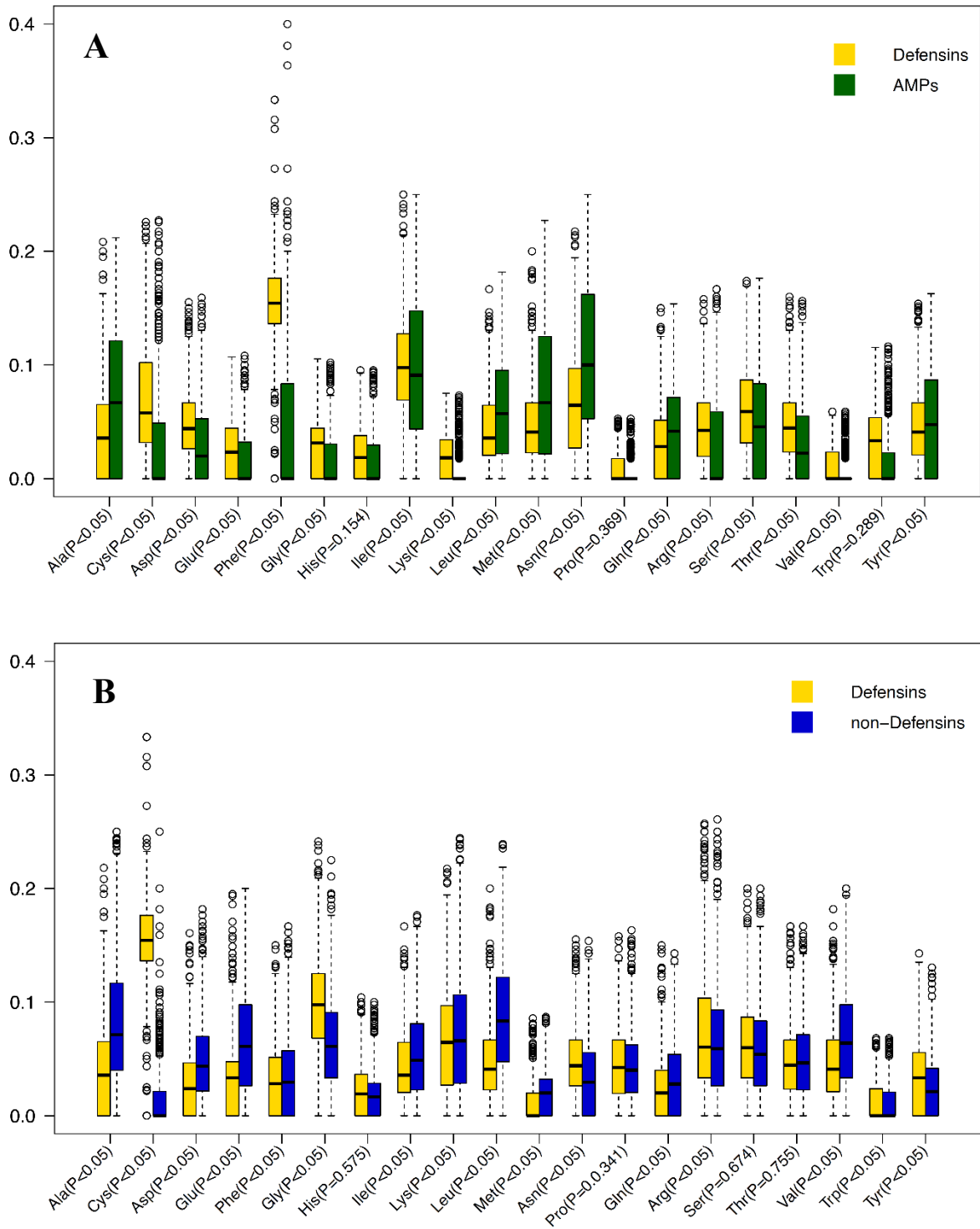


Figure 2: Boxplots of average amino acid compositions of 20 amino acids of Defensins vs AMPs (A) and Defensins vs non-Defensins (B). X- and Y-axes represent 20 amino acids along with their p-value and average amino acid composition.

SHORTCOMINGS OF EXISTING PREDICTORS AND FUTURE PERSPECTIVES FOR IMPROVING THE PREDICTION PERFORMANCE

To date, there are six ML-based predictors in existence that have been developed to predict defensins and their family/subfamily as summarized in Table 1. These ML-based predictors could effectively facilitate the prediction of the defensin family/subfamily and exhibit promising predictive performance. However, several shortcomings remain in these approaches that need to be addressed to develop more accurate and useful models as summarized hereafter.

First, the existing training datasets were relatively limited, especially for the defensin family and subfamily. Thus, the prediction performance of the defensin family and subfamily was not satisfying for real-life applications. In the future, when more defensin family and subfamily become available, more samples should be gathered and then employed to train a more comprehensive model (Charoenkwan et al., 2021a, 2022a; Kabir et al., 2022).

Second, all the existing ML-based predictors were trained on the training datasets with high homologous sequences based on the CD-HIT threshold of 0.8-1.0 (Table 2). It could be stated that iDPF-PseRAAC (Zuo et al., 2015) and iDEF-PseRAAC (Zuo et al., 2019) could achieve promising prediction performances when these two models were evaluated based on the dataset having high sequence identity. On the other hand, the prediction performance for those models was unsatisfactory based on the dataset having low sequence identity. Therefore, in order to construct a high-quality dataset, the CD-HIT threshold should be set as 0.3-0.4 to avoid over-estimation of the model's performance (Dao et al., 2019; Feng et al., 2019; Lai et al., 2019; Lv et al., 2019; Su et al., 2018; Xu et al., 2019).

Third, almost all of the existing ML-based predictors, including PseRAAC (Zuo et al., 2015), iDEF-PseRAAC (Zuo et al., 2019) and DEF-PRED (Kaur et al., 2021), were develo-

ped based on a black-box computational method (SVM) (Table 1). Motivated by this limitation, our group has proposed the scoring card method (SCM) which is a simple and interpretable model (Charoenkwan et al., 2013; Huang et al., 2012). The SCM method has been effectively applied to characterize and predict a variety of biological activities of proteins and peptides (Charoenkwan et al., 2021b, c, 2020b, c, d, e, f, 2013; Huang et al., 2012). The main contribution of the SCM method can be described into two major aspects: (i) the SCM method outperforms the well-known SVM and RF methods in terms of simplicity and interpretability. Specifically, this method identifies desired proteins using only the weighted sum between the composition and propensity scores and (ii) the SCM method provides propensity scores for 20 amino acids and 400 dipeptides that could help to provide insight into the characteristics of the proteins and peptides. Another solution for overcoming the limitations of the black-box method is to make use of the Shapley Additive explanation (SHAP) algorithm (Lundberg and Lee, 2017). The SHAP approach can provide both, the feature importance scores and the directionality of features.

Fourth, there is a lack of comprehensive assessment of the state-of-the-art feature encoding methods and ML algorithms in the prediction of defensins and their family/subfamily. Nevertheless, this comprehensive assessment provided could serve and facilitate users to select appropriate feature encodings/ML algorithms and provide useful guidelines for the development of more accurate and robust models in the future (Li et al., 2021b; Liang et al., 2021).

Finally, as described in many articles (Basith et al., 2020; Charoenkwan et al., 2021a, 2022a, b; Kabir et al., 2022), user-friendly web servers are considered as useful tools that are able to identify defensins and their family/subfamily without the use of experimental evidence. Although there are six computational approaches in existence, only two of them (i.e., iDEF-PseRAAC (Zuo et al.,

2019) and DEFPRED (Kaur et al., 2021)) were deployed as freely available web servers for the prediction of defensins and their family/subfamily.

CONCLUSIONS

In this study, we have conducted a comprehensive review and assessment of six current state-of-the-art computational approaches for predicting defensins and their family/subfamily in terms of different important aspects, covering the dataset quality, feature encoding methods, feature selection schemes, ML algorithms, cross-validation methods and web server availability/usability. In addition, we performed a comparative analysis of the existing computational approaches for predicting defensins and their family/subfamily. For the prediction of defensins, DEFPRED outperforms Karnik's method in terms of generalization ability, robustness and utility. In case of the defensin family and subfamily prediction, iDEF-PseR-AAC outperforms ID_RAAA and iDPF-PseRAAAC in terms of their predictive performance and robustness. These computational approaches are able to facilitate the identification of defensins and their family/subfamily. However, several shortcomings remain in these approaches that need to be addressed. Herein, five crucial aspects to develop more accurate and useful models have been listed as follows: (i) compiling an up-to-date dataset, (ii) excluding highly homologous sequences, (iii) using an interpretable method, (iv) performing a comprehensive assessment of the state-of-the-art feature encoding methods and (v) constructing a web server. We anticipate that this comprehensive review will provide useful guidance to researchers interested in developing cutting-edge computational approaches for the prediction of defensins and their family/subfamily.

Ethical statement

This review paper does not include animal or human experiments.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions statement

WS: Conceptualization, project administration, supervision, investigation, manuscript preparation and revision. PC: Data analysis; data interpretation, investigation and manuscript preparation. NS: Manuscript revision. SMHM, OT and NS: Manuscript preparation. All authors reviewed and approved the manuscript.

Acknowledgments

This work was fully supported by the College of Arts, Media and Technology, Chiang Mai University, and partially supported by the Chiang Mai University and Mahidol University. In addition, computational resources were supported by the Information Technology Service Center (ITSC) of Chiang Mai University.

REFERENCES

- Ahmad S, Charoenkwan P, Quinn JM, Moni MA, Hassan MM, Lio P, et al. SCORPION is a stacking-based ensemble learning framework for accurate prediction of phage virion proteins. *Sci Rep.* 2022;12:4106.
- Albrethsen J, Bøgebo R, Gammeltoft S, Olsen J, Winther B, Raskov H. Upregulated expression of human neutrophil peptides 1, 2 and 3 (HNP 1-3) in colon cancer serum and tumours: a biomarker study. *BMC Cancer.* 2005;5:8.
- 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; 40:1276-314.
- Breiman L. Random forests. *Machine learning.* 2001; 45:5-32.
- Bun Ng T, Chi Fai Cheung R, Ho Wong J, Juan Ye X. Antimicrobial activity of defensins and defensin-like peptides with special emphasis on those from fungi and invertebrate animals. *Curr Prot Pept Sci.* 2013;14:515-31.
- Charoenkwan P, Shoombuatong W, Lee H-C, Chaijaruwanich J, Huang H-L, Ho S-Y. SCMCRYST: 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.

- Charoenkwan P, Kanthawong S, Nantasenamat C, Hasan MM, Shoombuatong W. iDPPIV-SCM: A sequence-based predictor for identifying and analyzing dipeptidyl peptidase IV (DPP-IV) inhibitory peptides using a scoring card method. *J Prot Res.* 2020b;19:4125-36.
- Charoenkwan P, Kanthawong S, Schaduagrath N, Yana J, Shoombuatong W. PVPred-SCM: Improved prediction and analysis of phage virion proteins using a scoring card method. *Cells.* 2020c;9(2):353.
- 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.* 2020d;34:1105-16.
- Charoenkwan P, Yana J, Nantasenamat C, Hasan MM, Shoombuatong W. iUmami-SCM: A novel sequence-based predictor for prediction and analysis of umami peptides using a scoring card method with propensity scores of dipeptides. *J Chem Inf Model.* 2020e;60:6666-78.
- Charoenkwan P, Yana J, Schaduagrath 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.* 2020f;112:2813-22.
- Charoenkwan P, Anuwongcharoen N, Nantasenamat C, Hasan M, Shoombuatong W. In silico approaches for the prediction and analysis of antiviral peptides: A review. *Curr Pharm Des.* 2021a;27:2180-8.
- Charoenkwan P, Chiangjong W, Lee VS, Nantasenamat C, Hasan MM, Shoombuatong W. Improved prediction and characterization of anticancer activities of peptides using a novel flexible scoring card method. *Sci Rep.* 2021b;11:3017.
- Charoenkwan P, Kanthawong S, Nantasenamat C, Hasan MM, Shoombuatong W. iAMY-SCM: Improved prediction and analysis of amyloid proteins using a scoring card method with propensity scores of dipeptides. *Genomics.* 2021c;113:689-98.
- Charoenkwan P, Nantasenamat C, Hasan MM, Moni MA, Manavalan B, Shoombuatong W. StackDPPIV: a novel computational approach for accurate prediction of dipeptidyl peptidase IV (DPP-IV) inhibitory peptides. *Methods.* 2021d; epub ahead of print.
- Charoenkwan P, Chiangjong W, Hasan MM, Nantasenamat C, Shoombuatong W. Review and comparative analysis of machine learning-based predictors for predicting and analyzing anti-angiogenic peptides. *Curr Med Chem.* 2022a;29:849-64.
- Charoenkwan P, Schaduagrath N, Hasan MM, Moni MA, Lió P, Shoombuatong W. Empirical comparison and analysis of machine learning-based predictors for predicting and analyzing of thermophilic proteins. *EXCLI J.* 2022b;21:554-70.
- Dao F-Y, Lv H, Wang F, Feng C-Q, Ding H, Chen W, et al. Identify origin of replication in *Saccharomyces cerevisiae* using two-step feature selection technique. *Bioinformatics.* 2019;35:2075-83.
- De Coninck B, Cammue BP, Thevissen K. Modes of antifungal action and in planta functions of plant defensins and defensin-like peptides. *Fungal Biol Rev.* 2013;26:109-20.
- de Medeiros LN, Angeli R, Sarzedas CG, Barreto-Bergter E, Valente AP, Kurtenbach E, et al. Backbone dynamics of the antifungal Psd1 pea defensin and its correlation with membrane interaction by NMR spectroscopy. *Biochim Biophys Acta.* 2010;1798:105-13.
- de Oliveira Dias R, Franco OL. Cysteine-stabilized $\alpha\beta$ defensins: from a common fold to antibacterial activity. *Peptides.* 2015;72:64-72.
- Feng C-Q, Zhang Z-Y, Zhu X-J, Lin Y, Chen W, Tang H, et al. iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. *Bioinformatics.* 2019;35:1469-77.
- Hasan M, Schaduagrath 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;36:3350-6.
- Hasan MM, Basith S, Khatun MS, Lee G, Manavalan B, Kurata H. Meta-i6mA: an interspecies predictor for identifying DNA N 6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. *Brief Bioinform.* 2021;22(3):bbaa202.
- Huang H-L, Charoenkwan P, Kao T-F, Lee H-C, Chang F-L, Huang W-L, et al. , editors. Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition. *BMC Bioinformatics.* 2012;13(Suppl 17):S3.
- Jarczak J, Kościuczuk EM, Lisowski P, Strzałkowska N, Józwick A, Horbańczuk J, et al. Defensins: natural component of human innate immunity. *Hum Immunol.* 2013;74:1069-79.

- Kabir M, Nantasenammat C, Kanthawong S, Charoenkwan P, Shoombuatong W. Large-scale comparative review and assessment of computational methods for phage virion proteins identification. *EXCLI J.* 2022; 21:11-29.
- Kang X, Dong F, Shi C, Liu S, Sun J, Chen J, et al. DRAMP 2.0, an updated data repository of antimicrobial peptides. *Sci Data.* 2019;6:148.
- Karnik S, Prasad A, Diwevedi A, Sundararajan V, Jayaraman VK. Identification of Defensins employing recurrence quantification analysis and random forest classifiers. In: Chaudhury S, Mitra S, Murthy CA, Sastry PS, Pal SK (eds): *Pattern recognition and machine intelligence. PReMI 2009* (pp 152-7). Berlin: Springer, 2009. (Lecture Notes in Computer Science, Vol. 5909).
- Kaur D, Patiyal S, Arora C, Singh R, Lodhi G, Raghava GP. In-silico tool for predicting, scanning, and designing Defensins. *Front Immunol.* 2021;12:780610.
- Kim YS, Lee HJ, Yeo JE, Kim YI, Choi YJ, Koh YG. Isolation and characterization of human mesenchymal stem cells derived from synovial fluid in patients with osteochondral lesion of the talus. *Am J Sports Med.* 2015;43:399-406.
- Ramya Kumari S, Badwaik R, Sundararajan V, Jayaraman VK. Defensinpred: defensin and defensin types prediction server. *Protein Pept Lett.* 2012;19:1318-23.
- Lai H-Y, Zhang Z-Y, Su Z-D, Su W, Ding H, Chen W, et al. iProEP: a computational predictor for predicting promoter. *Mol Ther Nucleic Acids.* 2019;17:337-346.
- Li F, Chen J, Ge Z, Wen Y, Yue Y, Hayashida M, et al. Computational prediction and interpretation of both general and specific types of promoters in *Escherichia coli* by exploiting a stacked ensemble-learning framework. *Brief Bioinform.* 2021a;22:2126-40.
- Li F, Guo X, Jin P, Chen J, Xiang D, Song J, et al. Porpoise: a new approach for accurate prediction of RNA pseudouridine sites. *Brief Bioinform.* 2021b;22(6):bbab245.
- 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. *Chemometrics Intell Lab Syst.* 2015;141:100-6.
- Liang X, Li F, Chen J, Li J, Wu H, Li S, et al. Large-scale comparative review and assessment of computational methods for anti-cancer peptide identification. *Brief Bioinform.* 2021;22(4):bbaa312.
- Lin H, Ding H, Guo F-B, Huang J. Prediction of subcellular location of mycobacterial protein using feature selection techniques. *Mol Diversity.* 2010;14:667-71.
- Lin H, Liang Z-Y, Tang H, Chen W. Identifying sigma70 promoters with novel pseudo nucleotide composition. *IEEE/ACM Trans Comput Biol Bioinform.* 2019;16:1316-21.
- Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: von Luxburg U et al. (eds.): *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp 4768-77). Red Hook, NY: Curran Associates Inc., 2017.
- Lv H, Zhang ZM, Li SH, Tan JX, Chen W, Lin H. Evaluation of different computational methods on 5-methylcytosine sites identification. *Brief Bioinform.* 2020;21:982-95.
- 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.* 2019a;35:2757-65.
- 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 Therap Nucleic Acids.* 2019b;16:733-44.
- Menendez A, Finlay BB. Defensins in the immunology of bacterial infections. *Curr Opin Immunol.* 2007;19:385-91.
- Parisi K, Shafee TMA, Quimbar P, van der Weerden NL, Bleackley MR, Anderson MA. The evolution, function and mechanisms of action for plant defensins. *Semin Cell Dev Biol.* 2019;88:107-18.
- Sathoff AE, Samac DA. Antibacterial activity of plant defensins. *Mol Plant Microbe Interact.* 2019;32:507-14.
- Seebah S, Suresh A, Zhuo S, Choong YH, Chua H, Chuon D, et al. Defensins knowledgebase: a manually curated database and information source focused on the defensins family of antimicrobial peptides. *Nucleic Acids Res.* 2007;35(Suppl_1):D265-8.
- Su R, Liu X, Xiao G, Wei L. Meta-GDBP: a high-level stacked regression model to improve anticancer drug response prediction. *Brief Bioinform.* 2020;21:996-1005.
- Su Z-D, Huang Y, Zhang Z-Y, Zhao Y-W, Wang D, Chen W, et al. iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics.* 2018;34:4196-204.

- UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucl Acids Res.* 2017;45(D1):D158-69.
- Vapnik VN. An overview of statistical learning theory. *IEEE Transact Neural Networks.* 1999;10:988-99.
- Vapnik VN. *The nature of statistical learning theory*: New York: Springer Science & Business Media, 2000.
- 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:D1094-7.
- Wei L, He W, Malik A, Su R, Cui L, Manavalan B. Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework. *Brief Bioinform.* 2021;22(4):bbaa275.
- Whiston R, Finlay EK, McCabe MS, Cormican P, Flynn P, Cromie A, et al. A dual targeted β -defensin and exome sequencing approach to identify, validate and functionally characterise genes associated with bull fertility. *Sci Rep.* 2017;7:12287.
- Wilson SS, Wiens ME, Smith JG. Antiviral mechanisms of human defensins. *J Mol Biol.* 2013;425:4965-80.
- 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;35:4922-9.
- 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.* 2021;22:526-35.
- Zulfiqar H, Sun Z-J, Huang Q-L, Yuan S-S, Lv H, Dao F-Y, et al. Deep-4mCW2V: a sequence-based predictor to identify N4-methylcytosine sites in Escherichia coli. *Methods.* 2021;epub ahead of print.
- Zuo Y-C, Li Q-Z. Using reduced amino acid composition to predict defensin family and subfamily: Integrating similarity measure and structural alphabet. *Peptides.* 2009;30:1788-93.
- Zuo Y, Lv Y, Wei Z, Yang L, Li G, Fan G. iDPF-PseR-AAAC: a web-server for identifying the defensin peptide family and subfamily using pseudo reduced amino acid alphabet composition. *PLoS One.* 2015;10(12):e0145541.
- Zuo Y, Chang Y, Huang S, Zheng L, Yang L, Cao G. iDEF-PseRAAC: identifying the defensin peptide by using reduced amino acid composition descriptor. *Evolut Bioinform.* 2019;15:1-9.