EXCLI Journal 2022;21:11-29 – ISSN 1611-2156
Received: October 12, 2021, accepted: November 29, 2021, published: January 03, 2022

Review article:

LARGE-SCALE COMPARATIVE REVIEW AND ASSESSMENT OF COMPUTATIONAL METHODS FOR PHAGE VIRION PROTEINS IDENTIFICATION

Muhammad Kabir[a], Chanin Nantasenamat[b], Sakawrat Kanthawong[c], Phasit Charoenkwan[d], Watshara Shoombuatong[b,*]

[a] School of Systems and Technology, Department of Computer Science, University of Management and Technology, Lahore, Pakistan, 54770
[b] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok, Thailand, 10700
[c] Department of Microbiology, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand, 40002
[d] Modern Management and Information Technology, College of Arts, Media and Technology, Chiang Mai University, Chiang Mai, Thailand, 50200

* **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/excli2021-4411

ABSTRACT

Phage virion proteins (PVPs) are effective at recognizing and binding to host cell receptors while having no deleterious effects on human or animal cells. Understanding their functional mechanisms is regarded as a critical goal that will aid in rational antibacterial drug discovery and development. Although high-throughput experimental methods for identifying PVPs are considered the gold standard for exploring crucial PVP features, these procedures are frequently time-consuming and labor-intensive. Thusfar, more than ten sequence-based predictors have been established for the *in silico* identification of PVPs in conjunction with traditional experimental approaches. As a result, a revised and more thorough assessment is extremely desirable. With this purpose in mind, we first conduct a thorough survey and evaluation of a vast array of 13 state-of-the-art PVP predictors. Among these PVP predictors, they can be classified into three groups according to the types of machine learning (ML) algorithms employed (i.e. traditional ML-based methods, ensemble-based methods and deep learning-based methods). Subsequently, we explored which factors are important for building more accurate and stable predictors and this included training/independent datasets, feature encoding algorithms, feature selection methods, core algorithms, performance evaluation metrics/strategies and web servers. Finally, we provide insights and future perspectives for the design and development of new and more effective computational approaches for the detection and characterization of PVPs.

Keywords: Phage virion protein, bioinformatics, classification, machine learning, feature representation, feature select

EXCLI Journal 2022;21:11-29 – ISSN 1611-2156
Received: October 12, 2021, accepted: November 29, 2021, published: January 03, 2022

INTRODUCTION

Bacteriophages are viruses that may infect bacteria and replicates within them. They are obligate intracellular parasites that are widely distributed in areas populated by bacterial hosts, such as soil, water, and animal or human intestines, with a viral population estimated to be higher than 10^{31} particles (Clark and March, 2006; Lekunberri et al., 2017; Lyon, 2017). Phage virions are made up of genetic material (DNA or RNA) and a coat of structural proteins (or virion proteins) that can interact with host cell receptors and insert their genome into the cell via one of two basic strategies: the lytic or lysogenic cycle (Roach and Donovan, 2015). Lytic phages harness the host cell's biological machinery to synthesis their DNA and the remaining proteins needed to produce new phage particles. The new genomes are then packed into the head and phage progeny construct. Finally, phages lyse host cells and release additional phage particles (about 100-200 offspring) into the environment (Doss et al., 2017; Roach and Donovan, 2015). Prophages are lysogenic phages (temperate) that integrate their genome into the bacterial chromosome and become part of the host without killing the cell. The prophage genome is reproduced with the host chromosome and passed on to new daughter cells in a dependent manner. Under severe conditions, the viral genome can be extracted from the chromosome of the host bacterium, and lysogenic phages can go through a lytic cycle to produce new particles (Samson et al., 2013).

Protein arrays and mass spectrometry are two prominent examples of well-known experimental approaches used to discover and characterize PVPs (Lavigne et al., 2009; Yuan and Gao, 2016). As these techniques are time-consuming, labor-intensive and costly nature. As a result, the development of computational models capable of swiftly and accurately identifying PVPs is critical. Currently, there are 13 state-of-the-art predictors that are based on a wide range of machine learning (ML) techniques (Arif et al., 2020; Charoenkwan et al., 2020b; Charoenkwan et al., 2020d; Ding et al., 2014; Fang and Zhou, 2021; Feng et al., 2013; Han et al., 2021; Manavalan et al., 2018; Pan et al., 2018; Ru et al., 2019; Seguritan et al., 2012; Tang et al., 2016) for PVPs identification while two review papers (Meng et al., 2020; Nami et al., 2021) have emerged in this aspect. These review articles provided a good summary on the current state-of-the-art of PVPs identification. In spite of their merit, the overall scope of these articles is quite outdated. These review articles provided limited coverage on important aspects that are beneficial for the development of more accurate PVP predictors including training/independent datasets, core algorithms and webserver.

Herein, we propose the following important issues that needs to be addressed. Firstly, a more updated and comprehensive review is highly needed. These review papers did not provide a comprehensive survey on all of the currently available PVP predictors. A comprehensive review of all existing methods will be very useful for experimental scientists in selecting suitable PVP predictors for identifying investigated unknown sequences. Secondly, exploration on the underpinnings contributing to the development of more accurate PVP predictors would also be immensely useful.

Motivated by these aforementioned considerations, we herein conduct the first comprehensive overview and assessment of a large collection of 13 state-of-the-art PVP predictors. Table 1 summarizes these sequence-based PVP predictors along with their employed feature encoding algorithms, feature selection methods, ML algorithms and performance evaluation metrics/strategies. In particular, we reviewed all datasets used in the development of current PVP predictors. Details on these datasets are summarized in Table 2. Furthermore, we had also examined training/independent datasets, feature encoding algorithms, feature selection methods and ML algorithms. From amongst the current predictors, they can be categorized into three groups including conventional machine learning-based methods (i.e. iVIREONS (Seguritan et al., 2012), Feng et al.'s method (Feng et al., 2013), PVPred (Ding et al., 2014), PVP-SVM (Manavalan et al., 2018), PhagePred (Pan et al., 2018), Tan et al.'s method (Tan et al., 2018), Ru et al.'s method (Ru et al., 2019), Pred-BVP-Unb (Arif et al., 2020) and PVPred-SCM (Charoenkwan et al., 2020b)), ensemble-based methods (i.e. Zhang et al.'s method (Zhang et al., 2015), Meta-iPVP (Charoenkwan et al., 2020b) and iPVP-MCV (Han et al., 2021)) and deep learning-based methods (i.e. VirionFinder (Fang and Zhou, 2021)). Subsequently, we performed a comparative result analysis on three well-known benchmark datasets.

Finally, we summarize some key insights and future perspectives for the design and development of new next-generation computational methods for PVPs identification and characterization.

MATERIALS AND METHODS

Benchmark datasets

In the viewpoint of ML, the construction of a high-quality dataset is one of the quintessential step in the development of reliable computational predictors. The following steps were conducted in the establishment of a high-quality PVP dataset. In the first step, all sequences were experimentally verified as PVPs and non-PVPs. In the second step, PVPs and non-PVPs containing non-stan-

Table 1: A comprehensive list of current PVP predictors reviewed in this study

Type	Predictors/Tools	Algorithm [a]	Feature selection [b]	Evaluation strategy [c]
Conventional ML-based method	iVIREONS (Seguritan et al., 2012)	ANN	No	10CV
	Feng et al.'s method (Feng et al., 2013)	NB	CFS	10CV
	PVPred (Ding et al., 2014)	SVM	Two-step	LOOCV, IND
	PVP-SVM (Manavalan et al., 2018)	SVM	Two-step	10CV, IND
	PhagePred (Pan et al., 2018)	NB	Two-step	10CV, LOOCV
	Tan et al.'s method (Pan et al., 2018)	SVM	Two-step	10CV, IND
	Ru et al.'s method (Ru et al., 2019)	RF	MRMD	10CV
	Pred-BVP-Unb (Arif et al., 2020)	SVM	SVM-RFE	LOOCV, IND
	PVPred-SCM (Charoenkwan et al., 2020b)	SCM	No	10CV, IND
Ensemble-based method	Zhang et al.'s method (Zhang et al., 2015)	SVM	Two-step	10CV, IND
	Meta-iPVP (Charoenkwan et al., 2020b)	SVM	GA-SAR	10CV, IND
	iPVP-MCV (Han et al., 2021)	SVM	No	LOOCV, 10CV, IND
Deep learning-based method	VirionFinder (Fang and Zhou, 2021)	CNN	No	10CV, IND

[a] ANN: artificial neural network; CNN: convolutional neural network, LR: logistic regression, NB: naive bayes, RF: random forest, SCM: scoring card matrix, SVM: support vector machine
[b] CFS: Correlation-based feature selection, MRMD: maximum-relevance-maximum-distance, GA-SAR: Genetic-algorithm based self-assessment-report, SVM-RFE: support vector machine methods based on recursive feature elimination, Two-step: a two-step feature selection algorithm
[c] 10CV: 10-fold cross-validation, IND: independent test, LOO-CV: leave-one-out cross-validation

dard letters (e.g. “B”, “X” or “Z”) were excluded. In the final step, the sequence identity threshold was set to be in the range of 0.3-0.4 in order to avoid sequence redundancy. Details on different datasets used for constructing currently available PVP predictors are summarized in Table 2. From amongst these datasets, there are three popular benchmark datasets consisting of *Feng2013* dataset (Feng et al., 2013), *Manavalan2018* (Manavalan et al., 2018) and *Charoenkwan2020_2.0* (Charoenkwan et al., 2020d), which are frequently used for the development of existing PVP predictors consisting of PVPred (Ding et al., 2014), Zhang et al.’s method (Zhang et al., 2015), PVP-SVM (Manavalan et al., 2018), PhagePred (Pan et al., 2018), Tan et al.’s method (Pan et al., 2018), Pred-BVP-Unb (Arif et al., 2020), PVPred-SCM (Charoenkwan et al., 2020b), Meta-iPVP (Charoenkwan et al., 2020d) and iPVP-MCV (Han et al., 2021). In 2012, the *Seguritan2012* dataset (Seguritan et al., 2012) was released as the first dataset that has been used for the development of a sequence-based predictor for *in silico* PVP identification that consisted of 6303 PVPs and 6303 non-PVPs. Afterwards, the *Feng2013* dataset was introduced by Feng et al. (2013) and it represents the first high-quality dataset to apply a CD-HIT threshold of 0.4 that eventually led to a dataset of 99 PVPs and 208 non-PVPs. Particularly, the Feng2013 dataset can be downloaded from http://lin-group.cn/server/PVPred. In 2018, Manavalan et al. constructed the *Manavalan2018* dataset (Manavalan et al., 2018) by combining the *Feng2013* dataset with a new independent dataset containing 30 PVPs and 64 non-PVPs, which were manually collected from several PVP studies (Feng et al., 2013b; Pan et al.,

Table 2: A summary of training and independent test datasets used in PVP predictors

Name[a]	Training dataset		Independent dataset		CD-HIT threshold	Reference
	PVPs	**non-PVPs**	**PVPs**	**non-PVPs**		
Seguritan2012	5042	5042	1260	1260	0.9	Seguritan et al., 2012
Feng2013	99	208	No	No	0.4	Feng et al., 2013
Ding2014[b]	99	208	11	19	0.4	Ding et al., 2014
Zhang2015	100	100	68	92	0.8	Zhang et al., 2015
Manavalan2018[b]	99	208	30	64	0.4	Manavalan et al., 2018
Pan2018[b]	99	208	No	No	0.4	Pan et al., 2018
Tan2018[b,c]	99	208	30	64	0.4	Tan et al., 2018
Ru2019	6251	6914	No	No	0.8/0.4[c]	Ru et al., 2019
Arif2020[b,c]	99	208	30	64	0.4	Arif et al., 2020
Charoenkwan2020_1.0[b,c]	99	208	30	64	0.4	Charoenkwan et al., 2020b
Charoenkwan2020_2.0	250	250	63	63	0.4	Charoenkwan et al., 2020d
VirionFinder2021	16868	61778	310	766	1.0	Fang and Zhou, 2021
Han2021[b,c]	99/250	208/250	30/63	64/63	0.4/0.4	Han et al., 2021

[a]Datasets’ names are represented using the family name of the first author along with the publication year from the corresponding literature.
[b]Training dataset was directly obtained from the Feng2013 dataset
[c]Independent dataset was directly obtained from the Manavalan2018 dataset
[d]Sequence identity cutoffs were set to 0.8 and 0.4 for PVPs and non-PVPs

2018; Zhang et al., 2015). The *Manavalan2018* dataset can be downloaded from http://www.thegleelab.org/PVP-SVM/PVP-SVM.html. Recently, our group constructed an up-to-date dataset consisting of 313 PVPs and 957 non-PVPs, which were downloaded from the UniProt database (release 2019_11) (Charoenkwan et al., 2020d). To solve the overestimation issue that typically occurs during model optimization, the set of 313 PVPs and 957 non-PVPs were randomly divided into training and independent datasets using the 80/20 split ratio. This led to a training dataset consisting of 250 PVPs and 250 non-PVPs while the independent datasets consisted of 63 PVPs and 63 non-PVPs. These training and independent datasets are referred as the *Charoenkwan2020_2.0* dataset in this review. The Charoenkwan2020*_2.0* dataset can be downloaded from https://github.com/Shoombuatong/Dataset-Code/tree/master/PVP.

Several observations can be made from Table 1. Firstly, the *Feng2013* dataset (Feng et al., 2013) was most frequently used for developing PVP predictors and for assessing their cross-validation performance (i.e. Feng et al.'s method (Feng et al., 2013), PVPred (Ding et al., 2014), PVP-SVM (Manavalan et al., 2018), PhagePred (Pan et al., 2018), Tan et al.'s method (Tan et al., 2018), Pred-BVP-Unb (Arif et al., 2020), PVPred-SCM (Charoenkwan et al., 2020b) and iPVP-MCV (Han et al., 2021)). Secondly, the independent dataset derived from the *Manavalan2018* dataset (Manavalan et al., 2018) was the most frequently used one for assessing the independent test results of variant PVP predictors consisting of PVP-SVM (Manavalan et al., 2018), Tan et al.'s method (Tan et al., 2018), Pred-BVP-Unb (Arif et al., 2020), PVPred-SCM (Charoenkwan et al., 2020b) and iPVP-MCV (Han et al., 2021). Thirdly, the *Charoenkwan2020_2.0* dataset (Charoenkwan et al., 2020d) provided the largest number of PVPs and non-PVPs.

Feature encoding schemes

Machine-learning based PVP predictors require the extraction of feature information from the sequence. PVPs have been encoded into fix-length feature vectors using a variety of features encoding approaches (Arif et al., 2020; Charoenkwan et al., 2020b; Ding et al., 2014; Feng et al., 2013; Han et al., 2021; Manavalan et al., 2018; Pan et al., 2018; Ru et al., 2019; Zhang et al., 2015). In the present PVP predictors, there are five major types of feature descriptors, as shown in Table 3. Composition features (AAC, AKSNG, DPC, GGAP, and SAAC), position features (bi-PSSM, bi-Profile Bayes, DP-PSSM, PSSM, PSSM-AAC, PSSM-Composition, and PSSM Profiles), physicochemical properties (AACPCP, CTD, PAAC, and PCP), meta-based features (i.e. PFs), and structure features (i.e. PFs) were (i.e. Seq-Str). The most widely used descriptors are AAC, CTD, DPC, GGAP, and PSSM, as shown in Table 2 and their definitions are given below.

AAC descriptors represent the occurrence frequency of standard amino acids in a protein sequence (Charoenkwan et al., 2021d, 2020c, 2020d). The percentage composition ($aa(i)$) of the *ith* amino acid is represented by:

$$aa(i) = \frac{AA_i}{L} \qquad (1)$$

where AA_i is the count or occurrence for the *ith* amino acid and L is the length of the protein. DPC descriptors represent the occurrence frequency of all possible dipeptides in a protein sequence (Charoenkwan et al., 2021c, 2020b, 2013, 2020e, 2020f). The percentage composition ($dp(i)$) of the *ith* dipeptide is represented by:

$$dp(i) = \frac{DP_i}{L-1} \qquad (2)$$

Table 3: Different types of features employed for developing the PVP predictors

Feature type	Feature[a]	Dimension	Reference
Composition features	AAC	20	Feng et al., 2013; Manavalan et al., 2018; Seguritan et al., 2012
	AKSNG	400	Ru et al., 2019
	APAAC	23	Charoenkwan et al., 2020b
	DPC	400	Charoenkwan et al., 2020b; Feng et al., 2013; Manavalan et al., 2018
	GGAP	400	Ding et al., 2014; Pan et al., 2018
	GGAPTree	49220	Pan et al., 2018
	SAAC	60	Arif et al., 2020
Position features	bi-PSSM	400	Arif et al., 2020
	bi-Profile Bayes	20	Zhang et al., 2015
	DP-PSSM	200	Han et al., 2021
	PSSM	20	Zhang et al., 2015
	PSSM-AAC	20	Han et al., 2021
	PSSM-Composition	400	Han et al., 2021
	PSSM Profiles	20	Han et al., 2021
Physicochemical properties	AACPCP	180	Ru et al., 2019
	APAAC	23	Charoenkwan et al., 2020b
	CTD	273	Arif et al., 2020; Zhang et al., 2015
	PAAC	25	Zhang et al., 2015
	PCP	11	Manavalan et al., 2018
	PIP	16	Seguritan et al., 2012
Meta-based features	PF	16	Charoenkwan et al., 2020b
Structure features	Seq-Str	27	Ru et al., 2019

[a]AAC: amino acid composition, AACPCP: amino acid composition and physicochemical properties, AKSNG: Adaptive k-skip-n-Gram Algorithm, APAAC: pseudo amino acid composition, ATC: atomic composition, Bi-PSSM: Bigram position-specific scoring matrix, CTD: composition translation and distribution, DPC: dipeptide composition, DP-PSSM: position-specific scoring matric based on dipeptides, GGAP: g-gap dipeptide composition, GGAPTree: g-gap feature tree, PAAC: pseudo amino acid composition, PCP: physicochemical properties, PF: probabilistic features, PIP: protein isoelectric points, PSSM: position-specific scoring matrix, PSSM-AAC: position-specific scoring matrix based on amino acid composition, PSSM-Composition: position-specific scoring matrix based on composition, PSSM Profiles: position-specific scoring matrix based on profiles, SAAC: split amino acid composition, Seq-Str: sequence-structure

where DP_i is the count of occurrences of the i^{th} dipeptide. Final vectors for AAC and DPC descriptors are represented as 20- and 400-dimension (20-D and 400-D, respectively) feature vectors, respectively. The GGAP descriptor is another variation of the DPC descriptor ($g = 0$) by representing the occurrence frequency of any two interval amino acids ($\mathrm{aa_i}, \mathrm{aa_j}; j - i > 1$) in a given protein **P** (Ding et al., 2014; Pan et al., 2018). This descriptor can be formulated as follows:

$$\mathrm{GGAP\ (g)} = \left[f_1^g, f_2^g, \dots f_{400}^g\right] \quad (3)$$

where f_i^g is the percentage composition of the i^{th} ($i = 1,2,\dots,400$) g-gap dipeptide.

$$\boldsymbol{f_i^g} = \frac{\boldsymbol{n_i^g}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{400}} \boldsymbol{n_i^g}} \quad (4)$$

where n_i^g represents the percentage composition of i^{th} g-gap dipeptide in a given protein **P**. The final vector for GGAP is a 400-D feature vector.

The CTD descriptor describes the amino acid characteristics of protein sequences in general (Li et al., 2006). Combination (C), transformation (T), and distribution (D) are three separate feature descriptors provided by

Received: October 12, 2021, accepted: November 29, 2021, published: January 03, 2022

this method (Dubchak et al., 1995). Hydrophobicity, normalized van der Waals volume, polarity, polarization, charge, secondary structure, and solvent accessibility are among the 13 physicochemical properties used to create these three separate feature descriptors (Chen et al., 2018). Particularly, CTDC, CTDD and CTDT represent 39-D, 195-D and 39-D feature vectors, respectively (Arif et al., 2020; Zhang et al., 2015). Further details on CTDC, CTDD and CTDT descriptors are described in the work by Li et al. (2006).

The PSSM descriptor can be extracted by the evolutionary profile feature representation method. This descriptor is a position-based feature encoding that is represented by the characteristics of 20 amino acids at different positions in the protein sequence. Given a protein **P**, the Position-Specific Iterated BLAST (PSI-BLAST) program (Altschul et al., 1997) is often used to extract the PSSM descriptor. The occurrence frequency of amino acid residues at a certain site is generated after running the PSI-BLAST algorithm. Several previous studies have indicated that using the PSSM descriptor improves performance in a variety of biological classification studies (Arif et al., 2020; Charoenkwan et al., 2020d; Zhang et al., 2015).

Machine learning algorithms

As indicated in Table 1, there are three commonly utilized machine learning algorithms (NB, RF, and SVM) in this work. SVM was chosen as the algorithm of choice for creating the current PVP predictors among various ML methods. Meanwhile, in computational biology challenges, SCM-based and ensemble-based approaches are common solutions. The core notions of these five approaches are briefly discussed below.

To determine an unknown sample, the NB method uses the Bayes theorem and a set of conditional independence assumptions (Kumar et al., 2015). NB is known as a probabilistic-based classifier as it computes the predicted class with the maximum probability of investigated features. Given an unknown protein sequence **P**, it is represented with feature vector $F = (f_1, f_2, \dots, f_n)$. Subsequently, its class is predicted by finding out the class C that can maximize the likelihood $P(F|C) = P(f_1, f_2, \dots, f_n)$ where $C = \{0,1\}$ that is 1 and 0 represent PVP and non-PVP classes, respectively (Altschul et al., 1997; Kawashima and Kanehisa, 2000; Truong et al., 2015; Wei et al., 2020). As can be seen in Table 3, NB algorithm was employed in Feng et al.'s method (Feng et al., 2013) and PhagePred (Pan et al., 2018).

SVM is well-known as one of the most effective machine learning algorithms for dealing with binary classification problems, and it has been effectively applied in a variety of domains (Dao et al., 2019; Feng et al., 2019; Lai et al., 2019; Su et al., 2018; Xu et al., 2019; Zhang et al., 2020; Zhu et al., 2019). The Vapnik-Chervonenkis theory of statistical learning was first developed in 1995 (Cortes and Vapnik, 1995; Vapnik, 2013; Vapnik, 1999). It was then expanded to handle the multiclass classification task. Unlike other machine learning algorithms, SVM can reliably generalize the underlying data. In the instance of binary classification, SVM creates a classifier by determining the hyperplane with the greatest distance between two classes (i.e. PVP and non-PVP). In the meanwhile, the kernel function is used to transform the sample space with p-dimensional feature vector into the feature space with n-dimensional feature vector, where $p < n$. As can be seen in Table 1, the SVM algorithm is used to construct several of the existing predictors consisting of PVPred (Ding et al., 2014), PVP-SVM (Manavalan et al., 2018), Tan et al.'s method (Tan et al., 2018), Pred-BVP-Unb (Arif et al., 2020), Zhang et al.'s method (Zhang et al., 2015), Meta-iPVP (Charoenkwan et al., 2020d) and iPVP-MCV (Han et al., 2021).

Conventional RF-based models were often constructed based on the original RF algorithm as introduced by Breiman (2001). These models are constructed by integrating a collection of weak classification and regression tree (CART) classifiers to enhance the predictive performance of CART (Breiman, 2001; Breiman et al., 2017). In RF, the out-of-bag (OOB) approach is used for measuring the invested feature importance. The procedure for the out-of-bag (OOB) approach consists of two main stages as follows: (1) two-thirds of the training sample is employed in the construction of a classifier while the remaining is used for assessing the predictive performance of such classifier and (2) the importance score of each feature is obtained by calculating the decrease in their predictive performance.

The original SCM and ensemble-based SCM methods was firstly introduced by Huang et al. (2012) and Charoenkwan et al. (2013) for predicting and analyzing the protein solubility and protein crystallization, respectively. Recently, Charoenkwan et al. developed an improved version of SCM method known as the flexible scoring card method (FSCM) (Charoenkwan et al., 2021c) that provides improved prediction and characterization of anticancer peptides. The procedure for the SCM-based predictor development consists of five main stages (Charoenkwan et al., 2021a, 2020a, 2020b, 2013, 2020f):(i) preparing training and independent datasets, (ii) calculating initial propensity scores of 20 amino acids and 400 dipeptides, (iii) using the genetic algorithm for obtaining optimal propensity scores of 20 amino acids and 400 dipeptides, (iv) constructing a scoring function based on the optimal propensity scores of 400 dipeptides and (v) predicting the biological functions of unknown protein sequences.

Feature selection algorithms

Feature selection is an important step in building an effective and robust machine learning model. As indicated in Table 1, the most popular methodology for selecting the ideal feature sets to create 5 out of 13 existing PVP predictors is a two-step feature selection strategy, which includes PVPred (Ding et al., 2014), PVP-SVM (Manavalan et al., 2018), PhagePred (Pan et al., 2018), Tan et al.'s method (Tan et al., 2018) and Zhang et al.'s method (Zhang et al., 2015). This strategy's procedure is outlined below. The first step is to rate all attributes in order of importance. Analysis of variance (ANOVA) (used in PVPred (Ding et al., 2014), PhagePred (Pan et al., 2018) and Tan et al.'s method (Pan et al., 2018)), relief algorithm (used in Zhang et al.'s method (Zhang et al., 2015)) and RF algorithm (used in PVP-SVM (Manavalan et al., 2018)). The most significant characteristic is the one with the greatest importance scores. The next step is to choose the best feature set. It is also worth noting that all five PVP predictors make use of the incremental feature selection (IFS) for determining the best feature set. Particularly, the IFS strategy's procedure for determining the optimal number of features has two stages: (i) the first feature subset is built using the feature with the highest importance scores and (ii) the second feature subset is built by integrating the first feature subset with features with the second highest importance scores. This method was repeated until all of the researched traits were incorporated, starting with the higher score and ending with the lower score. The feature set with the highest performance is deemed to be the best.

Performance evaluation and evaluation strategy

To date, the performance of the 13 state-of-the-art PVP predictors has been assessed using three well-known performance evaluation strategies: *K*-fold cross-validation, jackknife validation test/LOOCV, and independent test (Arif et al., 2020; Charoenkwan et al., 2020b, 2020d; Ding et al., 2014; Fang and Zhou, 2021; Feng et al., 2013; Han et al., 2021; Manavalan et al., 2018; Pan et al., 2018; Ru et al., 2019; Seguritan et al., 2012; Tang et al., 2016). The performance of these PVP predictors is assessed using the accuracy (ACC), sensitivity (Sn), specificity (Sp), Matthew's correlation coefficient (MCC), and area under receiver operating characteristic (AUC) curves (Arif et al., 2020; Charoenkwan et al., 2020b, 2020d; Ding et al., 2014; Fang and

Zhou, 2021; Feng et al., 2013; Han et al., 2021; Manavalan et al., 2018; Pan et al., 2018; Ru et al., 2019; Seguritan et al., 2012; Tang et al., 2016). These performance metrics are defined as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})} \quad (5)$$

$$\mathrm{Sn} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \quad (6)$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{(\mathrm{TN} + \mathrm{FP})} \quad (7)$$

where TP and TN are true positive and true negative, which represent the number of correctly predicted PVPs and non-PVPs. FP is the false positive, which represents the number of non-PVPs predicted as PVPs. FN is the false negative, which represents the number of PVPs predicted as non-PVPs. As for the Sn and Sp metrics, they are used to measure the model's predictive ability in PVPs and non-PVPs, respectively. Moreover, ACC and MCC are used to measure the model's predictive ability for two class problems.

MACHINE LEARNING-BASED PVP PREDICTORS

Table 1 summarizes 13 state-of-the-art ML-based PVP predictors in terms of ML algorithms, feature selection techniques and performance evaluation strategies. These PVP predictors can be divided into three groups based on the types of machine learning algorithms used. The first group is made up of methods that are based on traditional machine learning algorithms consisting of ANN, NB, SCM, SVM, and RF. An approach based on ensemble learning constitutes the second group. Particularly, the ensemble approach uses two strategies: majority voting and meta-predictor approaches. A deep learning (DL)-based approach makes up the third group.

Conventional machine learning-based method

Seguritan et al. developed the first PVP predictor based on ANN algorithm (called iVIREONS (Seguritan et al., 2012)) for determining viral structure proteins using the primary sequence information namely making use of AAC and PIP descriptors. A year later, Feng et al. developed an NB-based PVP predictor (referred herein as the Feng et al.'s method (Feng et al., 2013)) that makes use of AAC and DPC feature descriptors as applied to the *Feng2013* dataset that contained 99 PVPs and 208 non-PVPs. To improve the precision of PVP identification, Feng et al. used the CFS algorithm for determining m informative features from 420 features. The optimal feature set having m informative features consists of V, T, A, H, K, E, R, S, LE, VT, VG, MK, TA, TS, AT, HI, KL, KI, KH, KN, KK, KD, KE, KW, KR, DK, EF, EL, EV, EK, EE, EW, CE, WK, RE, SG, GV and GG. The LOOCV performance (ACC, AUC) was (0.756, 0.758) and (0.792, 0.855), respectively, which made use of a total of 420 features as well as optimal features.

In 2014, Ding et al. introduced an SVM-based PVP predictor named PVPred (Ding et al., 2014). Particularly, PVPred makes use of the GGAP descriptor for distinguishing PVPs from non-PVPs on the *Ding2017* dataset. Specifically, the GGAP's parameter (*g*) was set to be in the range of 0 to 9. Finally, the protein sequence P is represented by a 400-D feature vector for each g. Subsequently, Ding et al. used the ANOVA approach together with the IFS process for determining important features that leads to improvement in the prediction ability of the model. Finally, the optimal feature set was inputted in to SVM algorithm to construct the final model. PVPred achieved a maximum ACC of 0.850 by using the 160 top-ranked GGAP (*g=1*) features. Ding et al. established the first independent dataset containing 11 PVPs and 19 non-PVPs. Particularly, PVPred correctly identified the 9 PVPs and 17 non-PVPs.

In 2018, Manavalan et al. (2018) proposed a novel predictor called PVP-SVM for accurately recognizing PVPs in 2018. PVP-SVM was a PVP predictor based on SVM that worked with AAC, DPC, CTD, ATC, and PCP. To find the best feature set, PVP-SVM used the SVMQA method, which was a systematic feature selection strategy. There were 136 informative features in the best feature set. They were obtained from 8 AAC features, 1 ATC feature, 25 CTD features, 98 DPC features, and 4 PCP features among the 136 relevant features. Cross-validation and independent test (ACC, MCC) results from PVP-SVM were (0.870, 0.695) and (0.798, 0.531), respectively.

PhagePred (Pan et al., 2018) and Tan et al.'s method (Tan et al., 2018) were developed using the GGAP descriptor. Unlike that of PVPred (Ding et al., 2014), PhagePred (Pan et al., 2018) is an NB-based PVP predictor built with the GGAPTree descriptor. The final vector for GGAPTree is represented as a 49220-D feature vector. Particularly, the ANOVA approach together with the IFS process was employed for determining important features as well as for improving the prediction ability of the model. As for Tan et al.'s method, it is an SVM-based PVP predictor that combines the use of ten best feature subsets, which were obtained from the ANOVA and mRMR feature selection methods. The cross-validation and independent test ACC of PhagePred (Pan et al., 2018) and Tan et al.'s method (Tan et al., 2018) provided corresponding values of (0.981, N/A) and (0.880, 0.755), respectively.

Ru et al.'s method (Ru et al., 2019) is an RF-based PVP predictor that is built with AACPCP, AKSNG and Seq-Str. Particularly, this method employs the MRMD approach for determining m informative features from a set of 661 features. This led to identification of the best m number that was found to be 256. The method was found to achieve values of 0.879, 0.963, 0.935 and 0.853 for Sn, Sp, Ac and MCC, respectively, using the 10-fold cross-validation test. In addition, this study also reported that the charge property was the most important physicochemical property for PVP identification.

In 2019, Arif et al. developed an unbiased predictor called the Pred-BVP-Unb (Arif et al., 2020). Particularly, the model was built using the SVM algorithm with the synthetic minority oversampling technique (SMOTE) for solving the imbalance problem on the training dataset (i.e. 99 PVPs and 208 non-PVPs). Moreover, multi-view features containing AAC, SAAC and bi-PSSM were employed to capture the wide array of information of PVPs. An optimal feature set of 86 top-ranked features was selected from amongst an initial set of 502 features for the construction of the final model. Pred-BVP-Unb yielded cross-validation and independent test ACC of 0.925 and 0.831, respectively.

Unlike previous existing PVP predictors, PVPred-SCM (Charoenkwan et al., 2020b) is a simple and highly interpretable PVP predictor. Particularly, PVPred-SCM was developed using the SCM method together with DPC descriptors. Furthermore, propensity scores of 400 dipeptides for PVPs were generated and optimized for predicting and characterizing PVPs. Experimental results demonstrated that the performance of PVPred-SCM as evaluated by cross-validation and independent test was found to achieve an ACC of 0.925 and 0.777, respectively, when compared to those of existing SVM-based method (i.e. PVP-SVM (Manavalan et al., 2018)) and could outperform a few other PVP predicators such as PVPred (Ding et al., 2014) and Tan et al.'s method (Tan et al., 2018).

Ensemble-based PVP method

In 2015, Zhang et al. proposed the first stacking-based PVP predictor (called Zhang et al.'s method (Zhang et al., 2015)). In their stacking-based PVP predictor, they employed hybrid features consisting of CTD, bi-profile Bayes, PAAC and PSSM. From amongst these four feature descriptors, the bi-profile Bayes descriptor could achieve the best cross-validation performance with ACC of 0.795, MCC of 0.595 and AUC of 0.835. Particularly, bi-profile Bayes afforded the best performance from amongst the four feature spaces with an accuracy of 0.795, an MCC of 0.595 and an AUC of 0.835. In addition, the Relief method was used to construct and rank the 4 feature types. As a result, the optimal feature subsets for CTD, bi-profile Bayes, PseAAC

Received: October 12, 2021, accepted: November 29, 2021, published: January 03, 2022

and PSSM consisted of top 79, 55, 32 and 50 features, respectively. The four RF models trained on four optimal feature subsets were integrated and used in the development of the final model using the LR algorithm. On the independent dataset, Zhang et al.'s method provided Sn of 0.853, Sp of 0.815, ACC of 0.831 and MCC of 0.662.

In 2020, our group had developed a novel meta-predictor called the Meta-iPVP (Charoenkwan et al., 2020b). Unlike that of Zhang et al.'s method (Zhang et al., 2015), Meta-iPVP combined four different ML algorithms (ANN, NB, RF and SVM) and seven different feature descriptors (AAC, APAAC, DPC, CTDC, CTDD, CTDT and PAAC) for generating 28 baseline models. These baseline models were used to generate 28 PFs. To improve the representation ability of PFs, the GA-SAR algorithm was used to determine the best m number out of 28 PFs. Finally, the 16 selected PFs were used as inputs for training the final meta-predictor using the SVM algorithm. Cross-validation and independent test results (ACC, MCC) of Meta-iPVP were (0.846, 0.698) and (0.817, 0.642), respectively.

Most recently, another ensemble-based PVP predictor (named iPVP-MCV) was proposed by Han et al. (2021). Particularly, three PSSM-based descriptors were found to perform well as compared to that of the Seq-AAC descriptor in terms of four out of five metrics (i.e. ACC, SN, SP, and MCC). In the base layer, three SVM-based models were generated using three different feature encodings (i.e. PSSM-AAC, DP-PSSM and PSSM-composition). In the meta layer, iPVP-MCV integrated predicted classes as derived from these baseline models via the use of the majority voting strategy. The iPVP-MCV approach was applied on the *Manavalan2018* dataset and gave the following results ACC, Sn, Sp and MCC of 0.840, 0.667, 0.922 and 0.621, respectively, as evaluated by independent test. In the meanwhile, when applied on the *Charoenkwan2020_2.0* dataset, iPVP-MCV gave rise to ACC, Sn, Sp and MCC of 0.833, 0.889, 0.778 and 0.671, respectively, as evaluated by independent test.

Deep learning-based PVP method

To the best of our knowledge, there is one PVP predictor in existence that was developed using the DL algorithm and this is VirionFinder (Fang and Zhou, 2021). Particularly, Fang and Zhou employed the CNN algorithm for model building using one-hot representation and 20 PCPs. Interestingly, this method was effective in identifying both complete and partial PVP from the virome data. Their comparative results with related PVP predictors (i.e. PVPred (Ding et al., 2014), PVP-SVM (Manavalan et al., 2018), PVPred-SCM (Charoenkwan et al., 2020b) and Meta-iPVP (Charoenkwan et al., 2020d)) using their datasets showed that the Sn of VirionFinder was higher than those of the compared PVP predictors on both the complete and partial datasets.

PERFORMANCE COMPARISON AND ANALYSIS

As can be seen from Table 3, almost all of the existing PVP predictors were developed and optimized using the three well-known benchmark datasets (i.e. *Feng2013* (Feng et al., 2013), *Manavalan2018* (Manavalan et al., 2018) and *Charoenkwan2020_2.0* (Charoenkwan et al., 2020b)), with only few exceptions (i.e. iVIREONS (Seguritan et al., 2012), Ru et al.'s method (Ru et al., 2019) and VirionFinder (Fang and Zhou, 2021)). Herein, we conducted a comparative analysis of these existing PVP predictors. Cross-validation and independent test results are summarized in Tables 4 and 5, respectively. Several of these existing PVP predictors including Feng et al.'s method (Feng et al., 2013), PVPred (Ding et al., 2014), PVP-SVM (Manavalan et al., 2018), PhagePred (Pan et al., 2018), Tan et al.'s method (Tan et al., 2018), Pred-BVP-Unb (Arif et al., 2020), PVPred-SCM (Charoenkwan et al., 2020b) and iPVP-MCV (Han et al., 2021), were developed and evaluated using the benchmark dataset through the

Received: October 12, 2021, accepted: November 29, 2021, published: January 03, 2022

10-fold cross-validation test. As can be seen from Table 4 and Figure 1, PhagePred achieved the highest ACC and MCC of 0.970 and 0.963 while PVPred-SCM (0.938, 0.866) and Pred-BVP-Unb (0.925, 0.850) performed well with the second and third highest ACC and MCC, respectively. For the *Manavalan2018* dataset, six out of eleven existing PVP predictors were built using this dataset (i.e. PVPred (Ding et al., 2014), PVP-SVM (Manavalan et al., 2018), Tan et al.'s method (Tan et al., 2018), Pred-BVP-Un (Arif et al., 2020), PVPred-SCM (Charoenkwan et al., 2020b), iPVP-MCV (Han et al., 2021)). Table 5 and Figure 2A show that iPVP-MCV and Pred-BVP-Unb could achieve the best independent test results as evaluated by ACC (0.836-0.840) and MCC (0.621-0.660). PVP-SVM could perform well with the second highest ACC and MCC of 0.798 and 0.531, respectively. In the meanwhile, five out of eleven existing PVP predictors were evaluated by the *Charoenkwan2020_2.0* dataset and this included PVPred (Ding et al., 2014), PVP-SVM (Manavalan et al., 2018), PVPred-SCM (Charoenkwan et al., 2020b), Meta-iPVP (Charoenkwan et al., 2020b) and iPVP-MCV (Han et al., 2021). Meta-iPVP and iPVP-MCV could achieve the best independent test results in terms of ACC (0.817-0.833) and MCC (0.642-0.671) (Figure 2B).

Table 4: Cross-validation results for different PVP predictors evaluated on the *Feng2013* dataset

Method	ACC	Sn	Sp	MCC	AUC
Feng et al.'s method[a]	0.758	0.808	0.792	-	0.855
PVPred[a]	0.758	0.894	0.850	-	0.899
PVP-SVM[a]	0.737	0.933	0.870	0.695	0.900
PhagePred[a]	0.970	0.986	0.981	0.963	0.990
Tan et al.'s method[a]	0.838	0.899	0.880	0.761	0.915
Pred-BVP-Unb[b]	0.925	0.938	0.913	0.850	-
PVPred-SCM[a]	0.938	0.948	0.933	0.866	0.960
iPVP-MCV[c]	0.879	0.778	0.928	0.720	

[a] Results were reported from the work of PVPred-SCM (Charoenkwan et al., 2020b).
[b] Results were reported from the work of Pred-BVP-Unb (Arif et al., 2020).
[c] Results were reported from the work of iPVP-MCV (Han et al., 2021).

Table 5: Independent test results from different PVP predictors as evaluated on *Manavalan2018* and *Charoenkwan2020_2.0 datasets*

Dataset[a]	Method	ACC	Sn	Sp	MCC	AUC
Manavalan2018	PVPred[b]	0.713	0.600	0.765	0.357	0.742
	PVP-SVM[b]	0.798	0.667	0.859	0.531	0.844
	Tan et al's method[b]	0.755	0.700	0.781	0.464	0.651
	Pred-BVP-Unb[c]	0.836	0.867	0.797	0.660	-
	PVPred-SCM[b]	0.777	0.767	0.781	0.523	0.781
	iPVP-MCV[d]	0.840	0.667	0.922	0.621	-
Charoenkwan2020_2.0	PVPred[e]	0.730	0.892	0.663	0.505	0.857
	PVP-SVM[e]	0.746	0.816	0.701	0.505	0.844
	PVPred-SCM[e]	0.714	0.745	0.690	0.432	0.781
	Meta-iPVP[f]	0.817	0.889	0.746	0.642	0.870
	iPVP-MCV[d]	0.833	0.889	0.778	0.671	-

[a]The independent test datasets (PVPs, non-PVPs) of Manavalan2018 and Charoenkwan2020_2.0 consisted of (30, 64) and (63, 63), respectively
[b] Results were reported from the work of PVPred-SCM (Charoenkwan et al., 2020b).
[c] Results were reported from the work of Pred-BVP-Unb (Arif et al., 2020).
[d] Results were reported from the work of iPVP-MCV (Han et al., 2021).
[e] Results were reported from the work of Meta-iPVP (Charoenkwan et al., 2020b).
[f] Results were reported from the work of Meta-iPVP (Charoenkwan et al., 2020d).

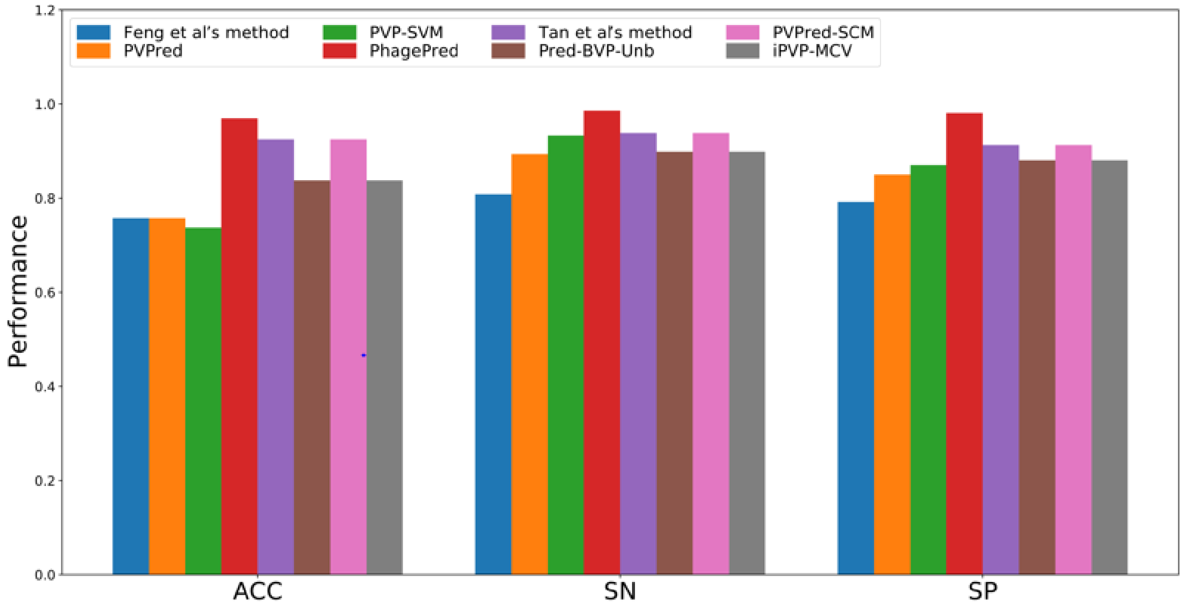

Figure 1: Performance evaluation on the *Feng2013* dataset as deduced from 10-fold cross validation test

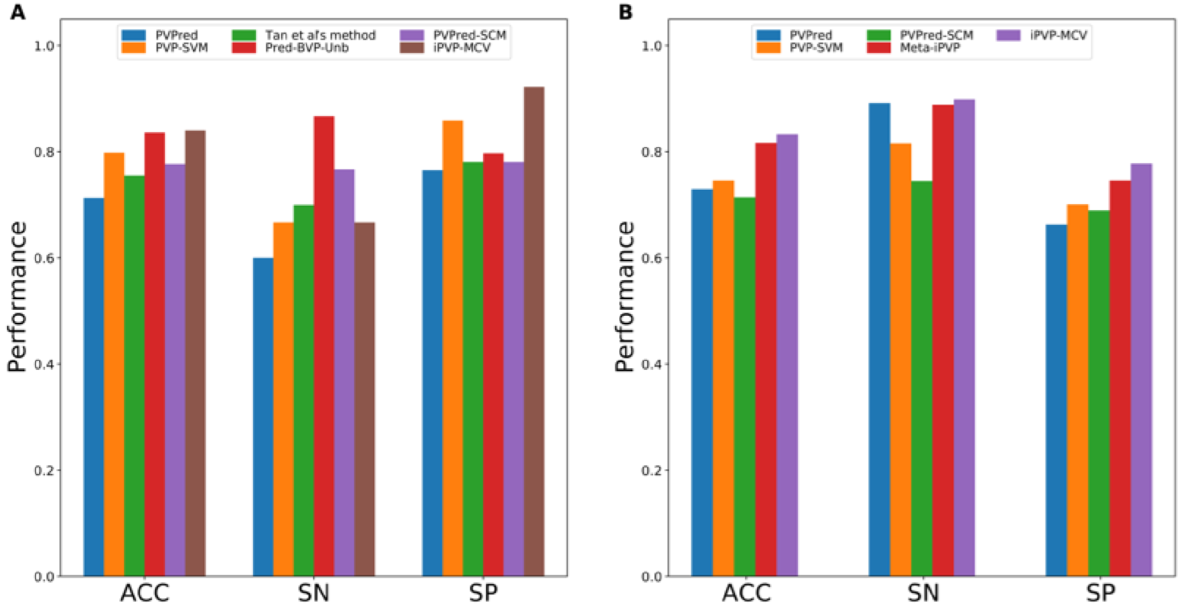

Figure 2: Performance evaluation on Manavalan2018 (**A**) and Charoenkwan2020_2.0 (**B**) datasets as deduce from independent test

From Tables 4 and 5, several observations can be made. Firstly, iPVP-MCV and Pred-BVP-Unb were found to provide the best independent test results as evaluated on both Manavalan2018 and *Charoenkwan2020_2.0* datasets. However, no web server was provided from these two PVP predictors. Hence, their utility and usage is quite limited. Second, although, PhagePred achieved the best cross-validation results on the *Feng2013* dataset, this predictor did not provide the independent test. It could be stated that PhagePred might not be a suitable tool for identifying candidate PVPs from large-scale proteins.

Received: October 12, 2021, accepted: November 29, 2021, published: January 03, 2022

Third, PVP-SVM and Meta-iPVP yielded relatively predictive performance to iPVP-MCV and Pred-BVP-Unb on both Manavalan2018 and Charoenkwan2020_2.0, respectively. In the meanwhile, these two PVP predictors were deployed as a user-friendly web server (http://www.thegleelab.org/PVP-SVM/PVP-SVM.html and http://camt.pythonanywhere.com/Meta-iPVP). Altogether, these comparative results indicated that PVP-SVM and Meta-iPVP could outperform iPVP-MCV and Pred-BVP-Unb as well as other existing PVP predictors in terms of their prediction results and community utility.

CHARACTERIZATION OF PHAGE VIRION PROTEINS

From amongst the eleven current PVP predictors, PVPred-SCM as introduced by Charoenkwan et al. (2020b) represents a simple and easily interpretable approach (Charoenkwan et al., 2021a, 2020a, 2020b, 2013, 2020f). Particularly, the PVPred-SCM model was built using 99 PVPs and 208 non-PVPs as derived from *Feng2013* dataset and their PVP scores were calculated using the scoring function S(P). Results from Charoenkwan et al. (2020b) indicated that four of ten proteins having the highest PVP scores were capsid protein (capsid protein G8P, capsid protein G8P, G VIII capsid protein precursor, and major coat protein). In addition, the SCM-derived propensity score of 20 amino acids and 400 dipeptides for PVPs were determined for analyzing the biochemical and biophysical properties of PVPs (Charoenkwan et al., 2020b). Charoenkwan et al. (2020b) reported that Ala, Thr, Val, Gly and Ser were the five top-ranked amino acids with the highest propensity scores of 529.50, 511.43, 506.88, 506.68 and 504.63, respectively, while Leu, Arg, His, Glu, and Lys were found to be amongst the five top-ranked amino acids with the lowest propensity scores. This finding was consistent with results reported by Ding et al. (2014). Particularly, in their study it was found that from amongst the GGAP (*g=1*) features, Ala, Gly, Pro, Ser, Thr were beneficial for PVPs while Glu, Lys, Leu and Arg were beneficial for non-PVPs. Moreover, informative PCPs from the AA index were also determined for analyzing important characteristics of PVPs. Results showed that alpha-helix propensity (KOEP990101) and hydrophobicity index (WOLR790101) were crucial properties of PVPs. Particularly, KOEP990101 and WOLR790101 properties exhibited high positive correlations of 0.502 and 0.484, respectively, while the side-chain of amino acids exhibited high negative correlation of -0.516. Two of five top-ranked amino acids having the highest propensity scores (i.e. Gly and Thr) were found to have high alpha-helix propensity with ranks of propensity scorers (PS, alpha-helix) of (4, 1) and (2, 3), respectively. Moreover, the helix propensity of amino acids was mentioned to be amongst the important contributors of protein stability as rationalized by strong H-bond and Van der Waals interactions (Pace and Scholtz, 1998). Moreover, important amino acids (i.e. Ala, Val and Gly) were found from amongst the ten top-ranked highest propensity scorers and hydrophobicity index as (1, 5), (3, 4) and (4, 1), respectively. Several studies have highlighted the importance of hydrophobic side chain amino acids in the stability of procapsids and phage virions. Gly and Ala were discovered in 50-residues, which were necessary for the inclusion of the M13 filamentous bacteriophage coat (Roth et al., 2002). Furthermore, Ala substitution at Glu52, Glu59, and Glu72 in the coat protein E-loop enhanced the stability of procapsids and virions of bacteriophages P22 (Asija and Teschke, 2019). According to these results, PVPs favored amino acids with high alpha-helix propensity and hydrophobic side index. It was also found that the side-chain property of amino acids were negatively correlated with PVPs. Charoenkwan et al. (2020b) reported that Ala, Thr, Val, Gly and Ser had propensity scorer (PS, side-chain) ranks as follows: (1,19), (2,15), (3,16), (4,20) and (5,18), respectively. The role of each amino acid in this protein was studied by performing random mutations at the N-terminal region of fusion phage protein containing the β-galactosidase-binding peptide. The findings revealed that short amino acids play a crucial role in providing a high binding affinity for the principal coat protein's domain C (Kuzmicheva et al., 2009). PVPs favored short amino acids because they had a low radius of octapeptide composing domain C, which can build alpha helix areas and have low steric hindrances, resulting in a low conformation number.

PROSPECTIVE STRATEGIES FOR IMPROVING THE PREDICTION PERFORMANCE OF PVPS

To date, there are 13 ML-based PVP predictors which have been proposed and developed for predicting and analyzing PVPs using primary sequence information only. From amongst these predictors, there were only four PVP predictors that were deployed as a web server (i.e. PVPred (Ding et al., 2014), PVP-SVM (Manavalan et al., 2018), PVPred-SCM (Charoenkwan et al., 2020b) and Meta-iPVP (Charoenkwan et al., 2020b)). In the meanwhile, only one PVP predictor (i.e. PVPred-SCM (Charoenkwan et al., 2020b)) could provide mechanistic understanding on the underlying properties governing PVPs and this was made possible via the use of SCM-derived propensity scores of 20 amino acids and 400 dipeptides. Although current PVP predictors could achieve an accurate and stable performance, there are still under explored aspects that can help to improve the identification of PVPs. Firstly, sufficient size of training datasets are often needed to enhance the predictive performance of the model. Although several research groups have made efforts in constructing up-to-date PVP datasets (Charoenkwan et al., 2020b; Ding et al., 2014; Manavalan et al., 2018), the relative size of PVPs is not of satisfactory level. Secondly, a number of sequence-based feature descriptors were used for the development of current PVP predictors. However, these feature descriptors had certain shortcomings (Charoenkwan et al., 2021d). Thus, there is the need to employ a built-in feature extractor for encoding PVPs. Several previous studies have demonstrated a natural language processing (NLP)-based technique (such as TF-IDF, Pep2Vec and FastText) that is known to be an effective built-in feature extractor technique, which is able to achieve an outstanding level of performance when compared to well-known sequence-based feature encodings (Charoenkwan et al., 2021d; Le et al., 2019; Li et al., 2020; Nguyen et al., 2020a,). Thirdly, it is highly desirable to utilize a feature representation learning (FRL) algorithm that can combine variant sequence-based feature descriptors together with ML algorithms in providing class information or probabilistic information. The original FRL algorithm was proposed by Wei et al. (2018), which was developed using several single feature-based SVM-based models. Recently, Charoenkwan et al. (2021b), Basith et al. (2021) and Hasan et al. (2021a,) extended the FRL algorithm of Wei et al. (2018) by integrating various ML algorithms such as ANN, KNN, NB and SVM. Fourthly, DL techniques have been demonstrated to be powerful ML techniques that could achieve good level of performance for various biological and chemical classification problems. Although VirionFinder (Fang and Zhou, 2021) is a DL-based PVP predictor that was developed using the CNN algorithm, however its DL structure is quite simple. Thus, there is the need to develop a more comprehensive DL structure.

CONCLUSIONS

This study surveyed and evaluated all currently available ML-based predictors for PVP prediction and characterization. We examined, assessed, and ranked all known PVP predictors in terms of their training/independent datasets, feature encoding algorithms, feature selection methods, core algorithms, performance evaluation metrics/strategies, and website. We used three benchmark datasets in a comparison analysis to find the best PVP predictor. PVP-SVM and Meta-iPVP were found to exceed other existing PVP predictors in terms of effectiveness, according to a comparison investigation. PVP-SVM and Meta-iPVP were found to exceed other existing PVP predictors in terms of effectiveness, acceptability and community utility. This research also provides useful information and future directions for the design and development of new and more sophisticated computational approaches for identifying and characterization of PVPs. We hope that this review will be useful to researchers in identifying the best PVP predictor for their needs as well as assisting in the rapid identification of phage virion proteins.

Ethical statement

This review paper does not include animal or human experiments.

Conflict of interest

The authors declare no conflict of interest.

Author contributions‘ statement

W.S. - conceptualization, project administration, supervision, investigation, manuscript preparation and revision

M.K. - data analysis; data interpretation, manuscript preparation

S.K. and P.C. - manuscript preparation

C.N. - manuscript revision.

All authors reviewed and approved the manuscript.

Acknowledgments

This work was fully supported by College of Arts, Media and Technology, Chiang Mai University and partially supported by Chiang Mai University and Mahidol University.

REFERENCES

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389-402.

Arif M, Ali F, Ahmad S, Kabir M, Ali Z, Hayat M. Pred-BVP-Unb: Fast prediction of bacteriophage Virion proteins using un-biased multi-perspective properties with recursive feature elimination. Genomics. 2020;112:1565-74.

Asija K, Teschke CM. Of capsid structure and stability: The partnership between charged residues of E-loop and P-domain of the bacteriophage P22 coat protein. Virology. 2019;534:45-53.

Basith S, Hasan MM, Lee G, Wei L, Manavalan B. Integrative machine learning framework for the identification of cell-specific enhancers from the human genome. Brief Bioinform. 2021;22(6): bbab252.

Breiman L. Random forests. Machine Learn. 2001;45 (1):5-32.

Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Boca Raton, FL: Routledge, 2017.

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.

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 Proteome Res. 2020a; 19: 4125-36.

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. 2020b;9(2):353.

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. 2020c;599: 113747.

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 Computer-Aided Mol Design. 2020d;34:1105-16.

Received: October 12, 2021, accepted: November 29, 2021, published: January 03, 2022

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, 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. 2020f;112: 2813-22.

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. 2021a;11(1):1-13.

Charoenkwan P, Chiangjong W, Nantasenamat C, Hasan MM, Manavalan B, Shoombuatong W. StackIL6: a stacking ensemble model for improving the prediction of IL-6 inducing peptides. Brief Bioinform. 2021b;22(6):bbab172.

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, Manavalan B, Shoombuatong W. BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides. Bioinformatics. 2021d; 37:2556–62.

Chen Z, Zhao P, Li F, Leier A, Marquez-Lago TT, Wang Y et al. iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics. 2018; 34:2499-502.

Clark JR, March JB. Bacteriophages and biotechnology: vaccines, gene therapy and antibacterials. Trends Biotechnol. 2006;24:212-8.

Cortes C, Vapnik V. Support-vector networks. Machine Learn. 1995;20:273-97.

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.

Ding H, Feng P-M, Chen W, Lin H. Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. Molecular BioSystems. 2014; 10:2229-35.

Doss J, Culbertson K, Hahn D, Camacho J, Barekzi N. A review of phage therapy against bacterial pathogens of aquatic and terrestrial organisms. Viruses. 2017; 9(3):50.

Dubchak I, Muchnik I, Holbrook SR, Kim S-H. Prediction of protein folding class using global description of amino acid sequence. Proc Natl Acad Sci. 1995;92: 8700-4.

Fang Z, Zhou H. VirionFinder: identification of complete and partial prokaryote virus virion protein from virome data using the sequence and biochemical properties of amino acids. Front Microbiol. 2021;12:9.

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.

Feng P-M, Ding H, Chen W, Lin H. Naïve Bayes classifier with feature selection to identify phage virion proteins. Comput Math Methods Med. 2013;2013: 530696.

Han H, Zhu W, Ding C, Liu TJS. iPVP-MCV: A multi-classifier voting model for the accurate identification of phage virion proteins. Symmetry. 2021;13(8):1506.

Hasan MM, Alam MA, Shoombuatong W, Deng H-W, Manavalan B, Kurata H. NeuroPred-FRL: an interpretable prediction model for identifying neuropeptide using feature representation learning. Brief Bioinform. 2021a;22(6):bbab167.

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. 2021b;22(3):bbaa202.

Huang H-L, Charoenkwan P, Kao T-F, Lee H-C, Chang F-L, Huang W-L, et al. Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition. BMC Bioinformatics. 2012;13(Suppl 17):S3.

Kawashima S, Kanehisa M. AAindex: amino acid index database. Nucleic Acids Res. 2000;28(1):374.

Kumar R, Chaudhary K, Chauhan JS, Nagpal G, Kumar R, Sharma M, et al. An in silico platform for predicting, screening and designing of antihypertensive peptides. Sci Rep. 2015;5(1):1-10.

Kuzmicheva G, Jayanna P, Eroshkin A, Grishina M, Pereyaslavskaya E, Potemkin V, et al. Mutations in fd phage major coat protein modulate affinity of the displayed peptide. Protein Eng Des Sel. 2009;22:631-9.

Received: October 12, 2021, accepted: November 29, 2021, published: January 03, 2022

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;17:337-46.

Lavigne R, Ceyssens PJ, Robben J. Phage proteomics: applications of mass spectrometry. Methods Mol Biol. 2009;502:239-51.

Le NQK, Yapp EKY, Nagasundaram N, Yeh H-Y. Classifying promoters by interpreting the hidden information of DNA sequences via deep learning and combination of continuous fasttext N-grams. Front Bioeng Biotechnol. 2019;7:305.

Lekunberri I, Subirats J, Borrego CM, Balcazar JL. Exploring the contribution of bacteriophages to antibiotic resistance. Environ Pollut. 2017;220(Pt B):981-4.

Li F, Chen J, Leier A, Marquez-Lago T, Liu Q, Wang Y, et al. DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites. Bioinformatics. 2020;36:1057-65.

Li Z-R, Lin HH, Han L, Jiang L, Chen X, Chen YZ. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res. 2006;34 (Suppl_2):W32-7.

Lyon J. Phage therapy's role in combating antibiotic-resistant pathogens. JAMA. 2017;318:1746-8.

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.

Meng C, Zhang J, Ye X, Guo F, Zou Q. Review and comparative analysis of machine learning-based phage virion protein identification methods. Biochim Biophys Acta Proteins Proteom. 2020;1868(6): 140406.

Nami Y, Imeni N, Panahi B. Application of machine learning in bacteriophage research. BMC Microbiol. 2021;21(1):1-8.

Nguyen D, Ho-Quang T, Dinh-Phan V, Ou Y-Y. Use Chou's 5-steps rule with different word embedding types to boost performance of electron transport protein prediction model. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2020a; preprints. doi: 10.1109/TCBB.2020.3010975.

Nguyen T-T-D, Le N-Q-K, Ho Q-T, Phan D-V, Ou Y-Y. TNFPred: identifying tumor necrosis factors using hybrid features based on word embeddings. BMC Med Genomics. 2020b;13(10):1-11.

Pace CN, Scholtz JM. A helix propensity scale based on experimental studies of peptides and proteins. Biophys J. 1998;75:422-7.

Pan Y, Gao H, Lin H, Liu Z, Tang L, Li S. Identification of bacteriophage virion proteins using multinomial naive bayes with g-gap feature tree. Int J Mol Sci. 2018;19(6):1779.

Roach DR, Donovan DM. Antimicrobial bacteriophage-derived proteins and therapeutic applications. Bacteriophage. 2015;5(3):e1062590.

Roth TA, Weiss GA, Eigenbrot C, Sidhu SS. A minimized M13 coat protein defines the requirements for assembly into the bacteriophage particle. J Mol Biol. 2002;322:357-67.

Ru X, Li L, Wang C. Identification of phage viral proteins with hybrid sequence features. Front Microbiol. 2019;10:507.

Samson JE, Magadán AH, Sabri M, Moineau S. Revenge of the phages: defeating bacterial defences. Nat Rev Microbiol. 2013;11:675-87.

Seguritan V, Alves N Jr, Arnoult M, Raymond A, Lorimer D, Burgin AB Jr, et al. Artificial neural networks trained to detect viral and phage structural proteins. PLoS Comput Biol. 2012;8(8):e1002657.

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.

Tan J-X, Dao F-Y, Lv H, Feng P-M, Ding H. Identifying phage virion proteins by using two-step feature selection methods. Molecules. 2018;23(8): 2000.

Tang H, Zou P, Zhang C, Chen R, Chen W, Lin H. Identification of apolipoprotein using feature selection technique. Sci Rep. 2016;6(1):30441.

Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Meth. 2015;12: 902-3.

Vapnik VN. An overview of statistical learning theory. IEEE Trans Neural Netw. 1999;10:988-99.

Vapnik V. The nature of statistical learning theory. Berlin: Springer, 2013.

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;34:4007-16.

Wei L, Hu J, Li F, Song J, Su R, Zou Q. Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. Brief Bioinform. 2020;21(1): 106-19.

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.

Yuan Y, Gao M. Proteomic analysis of a novel bacillus jumbo phage revealing glycoside hydrolase as structural component. Front Microbiol. 2016;7:745.

Zhang L, Zhang C, Gao R, Yang R. An ensemble method to distinguish bacteriophage virion from nonvirion proteins based on protein sequence characteristics. Int J Mol Sci. 2015;16:21734-58.

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;22:526-35.

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. Knowledge-Based Systems. 2019;163:787-93.