Computational and Structural Biotechnology Journal 20 (2022) 2921–2927

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Identification of plant vacuole proteins by exploiting deep representation learning features

Shihu Jiao [a], Quan Zou [a,b,c,*]

[a] Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, China
[b] State Key Laboratory of Tree Genetics and Breeding, Northeast Forestry University, Harbin, China
[c] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China

ARTICLE INFO

Article history:
Received 18 February 2022
Received in revised form 30 May 2022
Accepted 1 June 2022
Available online 8 June 2022

Keywords:
Vacuole proteins
Machine learning
Deep representation learning
Feature selection
Light gradient boosting machine

ABSTRACT

Plant vacuoles are the most important organelles for plant growth, development, and defense, and they play an important role in many types of stress responses. An important function of vacuole proteins is the transport of various classes of amino acids, ions, sugars, and other molecules. Accurate identification of vacuole proteins is crucial for revealing their biological functions. Several automatic and rapid computational tools have been proposed for the subcellular localization of proteins. Regrettably, they are not specific for the identification of plant vacuole proteins. To the best of our knowledge, there is only one computational software specifically trained for plant vacuolar proteins. Although its accuracy is acceptable, the prediction performance and stability of this method in practical applications can still be improved. Hence, in this study, a new predictor named iPVP-DRLF was developed to identify plant vacuole proteins specifically and effectively. This prediction software is designed using the light gradient boosting machine (LGBM) algorithm and hybrid features composed of classic sequence features and deep representation learning features. iPVP-DRLF achieved fivefold cross-validation and independent test accuracy values of 88.25 % and 87.16 %, respectively, both outperforming previous state-of-the-art predictors. Moreover, the blind dataset test results also showed that the performance of iPVP-DRLF was significantly better than the existing tools. The results of comparative experiments confirmed that deep representation learning features have an advantage over other classic sequence features in the identification of plant vacuole proteins. We believe that iPVP-DRLF would serve as an effective computational technique for plant vacuole protein prediction and facilitate related future research. The online server is freely accessible at https://lab.malab.cn/~acy/iPVP-DRLF. In addition, the source code and datasets are also accessible at https://github.com/jiaoshihu/iPVP-DRLF.

1. Introduction

Vacuoles are the largest membrane-bound organelles (up to 90 % of plant cells) and play essential roles in plant growth and development. Vacuoles have important cellular functions, such as storage of inorganic ions and metabolites, protein degradation, detoxification, and regulation of cytoplasmic ion homeostasis [1–3]. During seed development, a large number of protein storage vacuoles in the tissue are nutrient reservoirs for embryo development and seed germination. There are many peripheral proteins and transmembrane proteins that are related to vacuole activity, such as channel proteins, proton pumps, transport proteins, and various solution carrier proteins [4]. Changes in the abundance and activity of these proteins determine the specific functions of vacuoles. Another important function of vacuoles and lysosomes is the hydrolysis of intracellular proteins and membrane proteins and turnover of organelles (e.g., plastids, mitochondria, peroxides, and partial nuclei). This function helps to remove excess or damaged organelles, which are key factors in cell homeostasis and survival [5,6].

Research on the biochemical properties and physiological functions of plant vacuole proteins (PVPs) is the basis for our understanding of the mechanisms underlying vacuole biogenesis and maintenance [7–9]. Experimental subcellular localization methods are the most reliable means for characterizing of the biological activities of vacuolar proteins; however, they are usually costly and time-consuming. Therefore, it is important to develop compu-

* Corresponding author at: Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, China.
E-mail address: zouquan@nclab.net (Q. Zou).

https://doi.org/10.1016/j.csbj.2022.06.002

tational methods for the identification of PVPs. Recently, computational prediction algorithms for protein subcellular localization have emerged [10–12]. However, most of them have not been specifically developed for PVPs [13], and therefore, they perform poorly in identifying PVPs. To the best of our knowledge, there is only one machine learning tool—a support vector machine (SVM)-based model named VacPred—designed for PVP identification. VacPred has provided two of the best prediction models based on the SVM algorithm using two commonly used classic feature extraction methods: dipeptide composition (DPC) and the position-specific scoring matrix (PSSM)-based feature descriptor K-PSSM. The PSSM-based model showed slightly better performance than the DPC-based model on the blind dataset with an accuracy of approximately 63 %. Therefore, although these predictors have markedly promoted research on PVP prediction, there is still a need to develop high-performance PVP predictive tools.

Recently, various sequence-based deep representation learning features for proteins have been proposed and have obtained satisfactory results in many protein-sequence analysis applications [14–19]. These methods are also called deep learning embeddings, which are obtained by the multi-dimensional transformation of protein sequences. Deep learning embedding models are always based on unsupervised or semi-supervised learning and trained on large protein sequence databases. These techniques can extract sequence statistics as completely as possible; however, they require considerable time and computing resources to obtain the embedded models. Nonetheless, we can take advantage of these embeddings for PVP prediction by using the idea of transfer learning.

To further establish a new advanced computational PVP predictor with improved accuracy, we used a computational strategy combining deep representation learning and classic sequence features. A two-step feature selection strategy, that is, a light gradient boosting machine (LGBM) combined with sequential forward search (SFS), was subsequently applied to identify the optimal feature subset from each high-dimensional feature. After optimization, we constructed an efficient LGBM-based PVP predictor named iPVP-DRLF. The construction workflow of iPVP-DRLF is shown in Fig. 1. Fivefold cross-validation and independent testing demonstrate that iPVP-DRLF achieves satisfactory overall performance for PVP identification. Furthermore, despite employing fewer features, it outperforms two existing state-of-the-art (SOTA) predictive models on the blind dataset, improving prediction accuracy by approximately 6.6 % and 3.5 %, respectively. Through feature visualization analysis using the uniform manifold approximation and projection (UMAP) algorithm [20], we found that deep representation learning features could represent proteins better than other classic sequence features to distinguish PVPs from non-PVPs. Thus, it helps to improve the performance of iPVP-DRLF and leads to the development of powerful predictive tools.

2. Methods and materials

2.1. Datasets

In this study, the datasets collected by Yadav *et al.* were used for model training and testing [13]. Both PVPs and non-PVPs were derived from the UniProtKB/SwissProt database. The CD-HIT software [21] was then used to remove redundant samples by setting the sequence identity threshold to 60 % [22]. A total of 274 PVPs were obtained as the initial positive samples. Subsequently, 200 PVPs at 40 % identity cut-off were used as positive samples in the training dataset. The remaining 74 sequences were used to form a test dataset to verify the generalization ability of the predictive models. On the other hand, an equal number of negative samples with 40 % identity were collected to construct the balanced training and independent test datasets, respectively. In addition, Yadav *et al.* also created a PVP-blind dataset from cropPAL [23] to further evaluate the performance of developed models. The blind dataset contains 227 vacuole proteins with sequence lengths greater than 50. The above-mentioned datasets are available at https://lab.malab.cn/~acy/iPVP-DRLF.

2.2. Feature extraction

An effective sequence representation approach is crucial for developing satisfactorily performing predictive models. Here, various types of feature descriptors were studied. These sequence representation methods can be roughly divided into two groups:

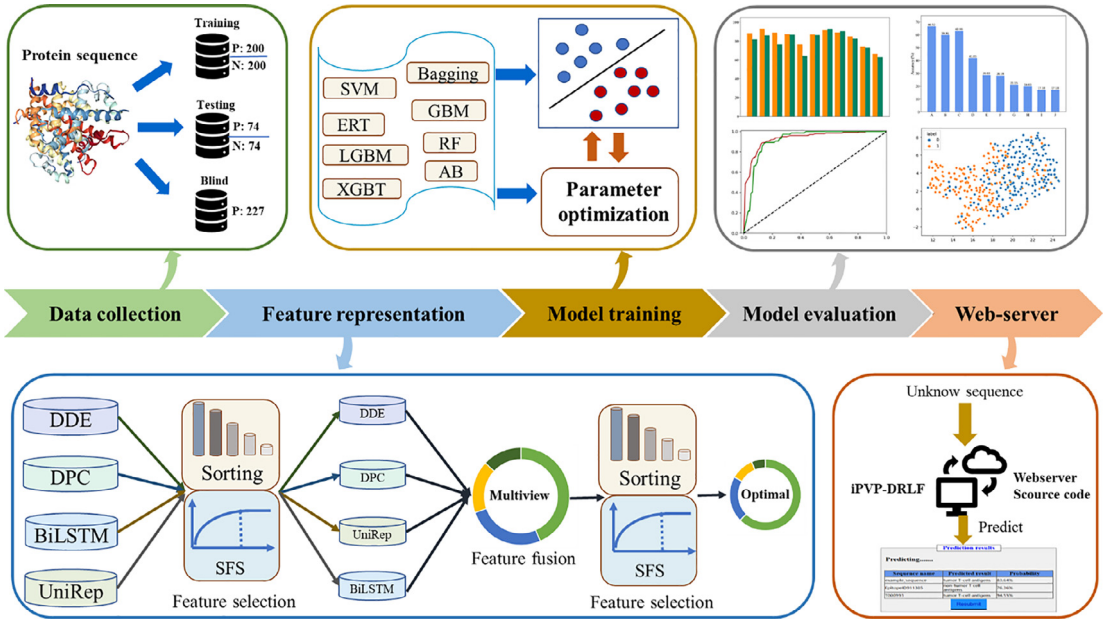

Fig. 1. The workflow of the development and evaluation process for iPVP-DRLF.

classic protein sequence encoding methods and deep representation learning features.

2.2.1. Classic sequence encoding methods

In this study, seven types of sequence-based classic feature encoding schemes were investigated to convert protein sequences into feature vectors, including pseudo-amino acid composition (PAAC); amino acid composition (AAC); dipeptide composition (DPC); adaptive skip dipeptide composition (ASDC); quasi-sequence order (QSO); composition, transition, and distribution model of physicochemical properties (CTD) and dipeptide deviation from expected mean (DDE). These feature encoding algorithms can be implemented conveniently using various published state-of-the-art platforms, such as BioSeq-Analysis [24], iFeature [25], and iLearn [26].

2.2.2. Deep representation learning features

Deep learning has achieved marked success in biological sequence processing due to its powerful sequence representation ability and automatic feature extraction capabilities. It is a special type of machine learning method that can capture parameters in neural networks to automatically learn feature representation [27]. Based on the principle of transfer learning, we can utilize pre-trained deep learning models for feature extraction of new data or migrate applications to other similar tasks. Typical examples of deep learning protein sequence embedding methods include UniRep [28], bidirectional long short-term memory embedding model (BiLSTM) [29], TAPE [30], SSA [14] and language embedding model (LM) [31]. These techniques have been demonstrated to be powerful tools for many protein engineering task applications. In particular, unified representation (UniRep), proposed by Alley *et al.*, was based on a multiplicative long short-term memory architecture. It was trained on UniRef50 (a dataset with $\sim$ 24 million protein sequences) and has the ability to extract the biological, chemical, and evolutionary information within the protein sequences. Protein sequences can be represented as 1900-dimensional feature vectors (average final hidden state output) using UniRep. More detailed information can be found at https://github.com/churchlab/UniRep. The BiLSTM embedding model was trained on the full set of protein domain sequences in the Pfam database, approximately 22 million protein sequences. BiLSTM embedding feature can be effectively combined with the global structural similarity between proteins and pairwise residue contact maps for individual proteins, allowing the vector matrix mapped from the protein sequences to be fully characterized [29].

2.3. Classifier

LightGBM is a distributed and efficient gradient-boosting framework developed by Microsoft Research [24]. This algorithm is based on decision tree and can be used in various machine learning tasks, including classification, sorting, and regression. The traditional gradient boosting decision tree needs to scan all data samples when estimating the information gain of all possible split points. This process is very time-consuming. LightGBM uses two engineering optimization novelties to overcome this problem. The first is gradient-based one-side sampling (GOSS). The GOSS algorithm excludes most of the data samples with small gradients and only uses the remaining ones to estimate the information gain. This leads to a more accurate gain estimation and significantly reduces the number of data instances without losing much training accuracy. The second is the exclusive feature bundling algorithm, which reduces the feature number by bundling mutually exclusive features. Therefore, it provides satisfactory efficiency and scalability when large datasets or high-dimensional features are used. More details can be found in Ref. [32]. We also compared seven other commonly used classifiers to find the best machine learning algorithm for PVP identification, including adaboost (AB), bagging, extremely randomized trees (ERT), gradient boosting machine (GBM), support vector machine (SVM), random forest (RF) and extreme gradient boosting (XGBT). We utilized scikit-learn to implement these efficient algorithms and tuned the hyperparameters via grid search (https://scikit-learn.org/). The search range for each classifier is presented in Supplementary Table S1.

2.4. Feature selection

To overcome overfitting and acquire the most significant feature space for modeling improvement, feature selection is commonly utilized. In recent years, scientists have proposed many methods to evaluate feature importance, such as analysis of variance (ANOVA), Chi2, and maximum-relevance-maximum-distance (MRMD) [33–36]. In this study, LGBM was used to rank features, and SFS was subsequently applied to search for the best feature subset [37]. A brief introduction is provided here. The training data with true labels was first fed into and fitted to the LGBM model. Next, we could obtain the importance value of each feature according to the built-in function of the LGBM model. A feature ranking list was then generated based on the feature importance values. A higher-ranked feature in the list indicates that it is more informative. The second step was using SFS to search for the optimal feature set from the sorted feature list. Features were added one-by-one from a low index to a high index to form feature subsets with different dimensions. The feature subsets were then fed into the classifier to construct predictive models and evaluated by fivefold cross validation. Finally, the subset with which the prediction model achieves the best performance was considered optimal.

2.5. Performance measurement

Fivefold cross-validation and independent testing were employed to evaluate the performance of the proposed machine learning models comprehensively and quantitatively [38–42]. Fivefold cross-validation randomly splits the original training data into five subsets. Each time, four of them were used for training, and the remaining one was used as the validation dataset. The performance metrics on the five subsets were averaged to obtain the overall fivefold cross-validation results. In addition, an independent test was used to demonstrate the generalization ability of the proposed models. Four standard confusion matrix-based metrics in binary classification tasks were used to measure the recognition ability. These included accuracy (Acc), specificity (SP), sensitivity (SE), and Matthew's correlation coefficient (MCC). They were calculated using Equations (1) – (4). Moreover, receiver operating characteristic (ROC) curves and the area under the ROC curve (AUC) were also used to make an intuitive performance comparison of constructed models.

$$\begin{cases} SE = \frac{TP}{TP+FN} * 100\% (1) \\ SP = \frac{TN}{TN+FP} * 100\% (2) \\ \text{Acc} = \frac{TP+TN}{TP+FP+TN+FN} * 100\% (3) \\ \text{MCC} = \frac{(TP \times TN)-(FP \times FN)}{\sqrt{(TP+FP)\times(TN+FN)\times(TP+FN)\times(TN+FP)}} (4) \end{cases}$$

TP: true positive; FP: false positive; TN: true negative; FN: false negative.

3. Experimental results

3.1. Performance of classic sequence encoding methods

To determine the best PVP sequence feature representation type, we first developed predictive models using classic protein sequence descriptors based on the LGBM classifier. Notably, several of them are high-dimensional features, which are always redundant, noisy, and computationally expensive. Thus, a feature selection procedure based on the LGBM classifier was performed to remove redundant and irrelevant features, where only the most discriminative features were retained to construct predictive models with upgraded efficiency. The feature selection SFS curves of each descriptor are shown in Figure S1. The fivefold cross-validation performance results of models based on optimal feature subsets for different descriptors are summarized in Table 1.

As can be observed from Table 1, four types of sequence descriptors, namely DPC, CTD, ASDC and DDE, achieve very similar performance in terms of five performance measures. And DDE seems to be the most powerful classic feature encoding method for PVP prediction. The associated model achieves the highest Acc, AUC, SP, and MCC values of 79.25 %, 0.867, 79.50 %, and 0.585, respectively. Following the DDE, the DPC feature outperforms the other remaining features in terms of Acc and MCC, and most indicators are only slightly lower than the DDE. Furthermore, DPC has an SE score of 81.00 %, the highest among all features. AAC performs the worst among all the descriptors, with Acc and MCC values of 68.25 % and 0.365, respectively. The QSO and PAAC encodings also achieve a comparatively lower performance with an accuracy value of about 70 %, possibly because they can only capture limited informative patterns from PVP sequences. These observations are consistent with previous studies suggesting that the DPC features are critical and essential for prediction of PVPs. Based on these facts, we then integrated the two top-performing features, DDE and DPC, into our hybrid feature to explore more comprehensive and discriminative feature encoding strategies for PVP prediction.

3.2. Performance of deep representation learning features

Based on the pre-trained deep learning models, we investigated five types of effective deep representation learning features. Notably, most deep representation learning features have higher dimensionality than classical feature descriptors. Therefore, we applied the same two-step feature selection method to determine the optimal feature space for each descriptor. The feature selection results are shown in Figure S2. Table 2 summarizes the performance of the optimal feature subsets for each descriptor on the LGBM classifier.

Deep representation learning features achieve a better and more stable performance compared with the classic sequence features. Except for SSA, the other deep representation learning features show significantly improved performance. Among all the deep representation learning features, BiLSTM achieves the overall best performance in terms of the three metrics: Acc (85.25 %), SE (84.00 %) and MCC (0.705). UniRep outperforms all other descriptors on the remaining two metrics, AUC and SP, with values of 0.915 and 87.50 %, respectively. Furthermore, the Acc and MCC scores of UniRep are only slightly worse than BiLSTM. The accuracy of BiLSTM and UniRep achieves a marked improvement of approximately 6 % compared with the best performing classic sequence features. Even TAPE and LM achieve better performance than DDE on five metrics. These results indicate that deep representation learning embedding methods can capture more informative patterns for identifying vacuole proteins from sequences. Considering the overall performance on the training sets, UniRep and BiLSTM were selected for the next feature analysis experiment.

3.3. Performance of hybrid features

To further explore the best model for PVP identification, we investigated the performance of hybrid features that integrated sequence information from multiple aspects. Therefore, we directly combined the aforementioned four best individual descriptors, namely DDE, DPC, UniRep and BiLSTM, to obtain a 176-dimensional hybrid feature vector (named F176). Preliminary experiments showed that this hybrid feature still contains redundant information. Hence, we performed the same feature selection as described in section 2.4. Here, we systematically evaluated the SFS results based on eight widely used machine learning classifiers. The SFS curves of each classifier are shown in Figure S3. The fivefold cross-validation performance results for different classifiers are summarized in Table 3. In this step, the testing dataset was employed to further estimate the comprehensive performance of these models, and the prediction results are also presented in Table 3.

We found that the performance of different classifiers varied greatly. Moreover, it can be observed from Table 3 that there is no distinct regularity between the validation dataset and the training dataset. The AB classifier achieves the overall best performance in terms of five metrics, and the accuracy score is up to 89.50 %. But its performance on the test dataset is not satisfactory, with an accuracy of 82.43 %. We observed that the LGBM classifier significantly outperforms the other classifiers on three metrics (Acc, SP, and MCC) on the independent test set. Furthermore, the Acc, AUC, SP and MCC of the LGBM classifier on the training dataset are 88.25 %, 0.933, 87.50 % and 0.765, respectively, which is only slightly worse than the AB classifier. Comprehensively considering the comparison results among the different classifiers, the consistently competitive performance on both datasets demonstrates that the LGBM is the most suitable algorithm for developing predictive models for vacuole protein identification. The new 63-dimensional fused feature vector (named F63) selected using LGBM and the SFS method is considered as the optimal subset. Therefore, the LGBM model trained on the F63 was determined to be the final model for use in the iPVP-DRLF implementation.

Table 1
5-fold cross-validation results of different classic sequence descriptors.

Feature	Acc (%)	AUC	SE (%)	SP (%)	MCC
DPC (34D)	79.00	0.840	**81.00**	77.00	0.580
CTD (51D)	78.50	0.845	78.00	79.00	0.570
ASDC (57D)	78.25	0.829	77.50	79.00	0.565
DDE (38D)	**79.25**	**0.867**	79.00	**79.50**	**0.585**
PAAC (22D)	72.00	0.775	67.50	76.50	0.442
AAC (20D)	68.25	0.742	67.00	69.50	0.365
QSO (44D)	71.00	0.774	70.00	72.00	0.420

Note: The best performance value of each column is highlighted in bold for clarification. Numbers in parentheses represent feature dimensions after feature selection.

Table 2
Fivefold cross-validation results of different deep representation learning features.

Features	Acc (%)	AUC	SE (%)	SP (%)	MCC
BiLSTM (44D)	**85.25**	0.908	**84.00**	86.50	**0.705**
LM (40D)	82.00	0.888	81.00	83.00	0.640
SSA (28D)	75.25	0.815	74.50	76.00	0.505
TAPE (55D)	83.00	0.893	**84.00**	82.00	0.660
UniRep (60D)	85.00	**0.915**	82.50	**87.50**	0.701

Note: The best performance value of each column is highlighted in bold for clarification. Numbers in parentheses represent feature dimensions after feature selection.

Table 3
Performance comparison of eight machine learning models based on the corresponding optimal feature subset of F176.

Classifier	Fivefold cross-validation					Independent testing				
	Acc (%)	AUC	SE (%)	SP (%)	MCC	Acc (%)	AUC	SE (%)	SP (%)	MCC
AB	**89.50**	**0.943**	**89.00**	**90.00**	**0.790**	82.43	0.885	85.14	79.73	0.650
Bagging	84.25	0.895	82.50	86.00	0.685	83.11	0.898	85.14	81.08	0.663
ERT	83.50	0.888	80.00	87.00	0.672	85.81	**0.936**	**90.54**	81.08	0.719
GBM	86.25	0.925	87.50	85.00	0.725	83.11	0.899	85.14	81.08	0.663
LGBM	88.25	0.933	**89.00**	87.50	0.765	**87.16**	0.916	89.19	**85.14**	**0.744**
RF	84.25	0.899	83.00	85.50	0.685	84.46	0.926	87.84	81.08	0.691
SVM	86.75	0.922	85.50	88.00	0.735	80.41	0.871	**90.54**	70.27	0.621
XGBT	88.25	0.926	88.00	88.50	0.765	83.78	0.900	85.14	82.43	0.676

Note: The best performance value of each column is highlighted in bold for clarification.

To understand the effectiveness of the deep representation learning features, we used the most popular feature analysis strategy, UMAP, for dimension reduction to analyze the distribution characteristics of the training samples. The distribution is shown in Fig. 2, and the positive and negative samples are distributed totally differentially in the five compared feature spaces (A-E). Notably, many PVPs and non-PVPs overlap in the feature space of the DPC and DDE. In contrast, although there are still some samples with overlapping distributions in the feature space of UniRep, BiLSTM, and the F63, marked boundaries appear to exist to separate the most positive and negative samples. Especially for the optimal feature subset F63, the gap between positive and negative clusters is more obvious, and only a very small number of positive and negative samples overlap. This suggests that the information extracted by the deep representation learning embedding methods is more effective in capturing the difference between PVPs and non-PVPs. Thus, the performance of the iPVP-DRLF was enhanced.

3.4. Comparison with existing predictors

In this section, we compared the prediction performance of iPVP-DRLF with two SOTA predictive models: VacPred-DPC and VacPred-PSSM. All of them were trained and validated using the same training and testing datasets. As shown in Table 4, VacPred-PSSM outperforms VacPred-DPC in almost all result metrics, so we mainly compared iPVP-DRLF with VacPred-PSSM. Fig. 3A also visually shows the evaluation metrics comparison. It is clear that iPVP-DRLF achieved better performance in terms of most metrics compared to VacPred-PSSM on both datasets. Especially on the training dataset, the improvements made by our pre-

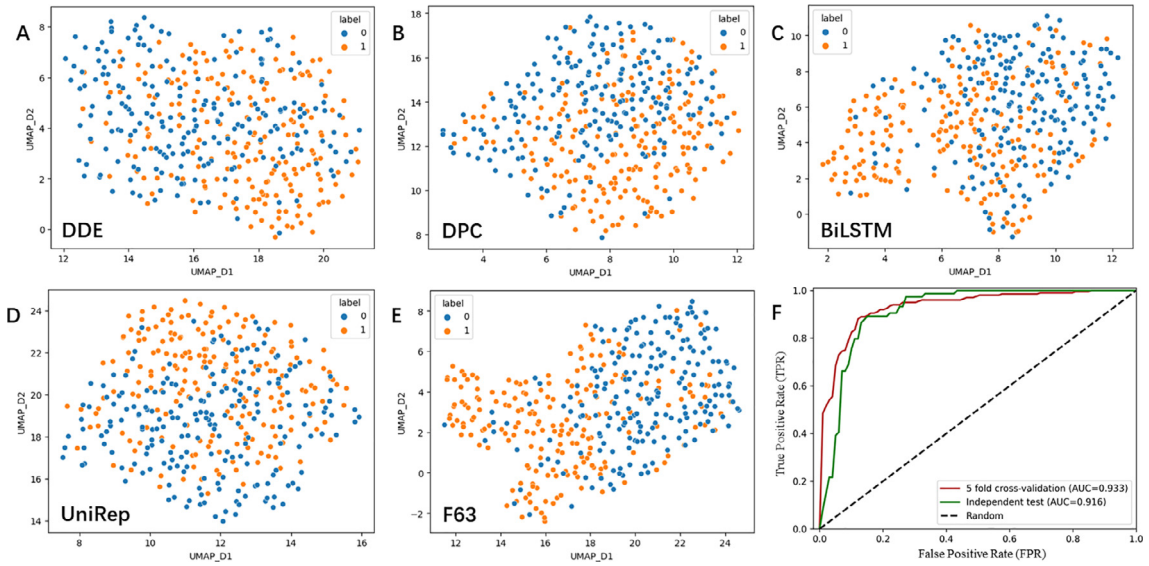

Fig. 2. UMAP distribution of PVPs and non-PVPs using the 63-dimensional vector F63 and four compared individual descriptors. The orange dots represent PVPs and the blue dots represent non-PVPs. (A-E) are the distributions of DDE, DPC, BiLSTM, UniRep and F63, respectively. F presents the ROC curves for iPVP-DRLF on the training and independent test datasets. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Performance comparison of proposed iPVP-DRLF and the SOTA predictors.

Classifier	Training					Testing					Blind
	Acc (%)	AUC	SE (%)	SP (%)	MCC	Acc (%)	AUC	SE (%)	SP (%)	MCC	Acc (%)
iPVP-DRLF	**88.25**	**0.933**	**89.00**	**87.50**	**0.765**	**87.16**	0.916	89.19	**85.14**	**0.744**	**66.52**
VacPred-DPC	75.50	0.800	70.00	81.00	0.510	80.41	0.840	82.43	78.38	0.610	59.91
VacPred-PSSM	81.75	0.860	76.50	87.00	0.640	86.49	**0.930**	**90.54**	82.43	0.730	62.99

Note: The best performance value of each column is highlighted in bold for clarification.

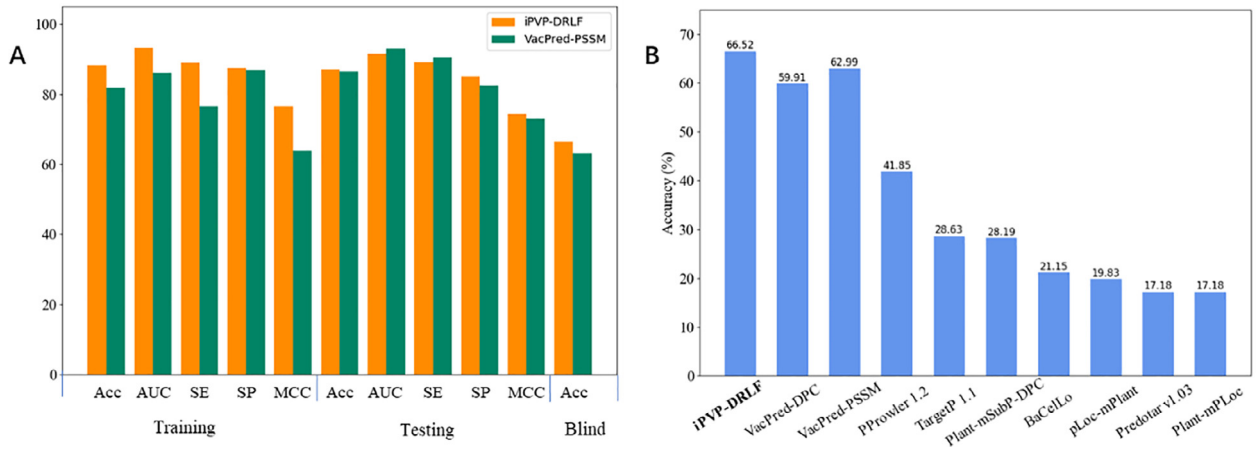

Fig. 3. Performance comparison of different PVPs prediction software. A presents the comparison of iPVP-DRLF with the SOTA predictor VacPred-PSSM on training, test and blind datasets. B shows the benchmark results of iPVP-DRLF and different published software using blind dataset.

dictive model are statistically significant with an accuracy boost of about 6.5 %. Furthermore, the gap between SE and SP of our model is smaller, which indicates that iPVP-DRLF has a more balanced ability to identify positive and negative samples. It is also worth noting that both VacPred-DPC and VacPred-PSSM use 400-dimensional feature vectors, while ours is 63-dimensional, which is approximately-one-sixth of the former. This significantly reduces the computational cost.

To verify the robustness of iPVP-DRLF, we further tested it using a blind dataset. In this section, we compared its performance not only with VacPred, but also with seven SOTA protein subcellular localization prediction tools, including PProwler 1.2, TargetP 1.1, Plant-mSubP-DPC, BaCelLo, pLoc-mPlant, Predotar v1.03 and Plant-mPLoc. For a fair comparison, the blind dataset collected by Yadav et al. was used to test all the compared predictors. The accuracy scores for the seven subcellular localization prediction tools on the blind datasets came from reference [13]. As shown in Fig. 3B, iPVP-DRLF outperformed all the compared models with an accuracy of 66.52 %, which is an approximately 6.6 % and 3.5 % improvement compared with VacPred-DPC and VacPred-PSSM, respectively. Among the protein subcellular localization predictors, PProwler 1.2 achieves the best performance with an Acc score of 41.85 %. Overall, these predictors performed poorly, probably because they were not designed specifically for PVPs. The above analysis clearly suggests the practical applicability of iPVP-DRLF over other methods. ROC curves were drawn (Fig. 2F) to visually depict the predictive efficiency of iPVP-DRLF. The AUC values obtained by fivefold cross validation and independent testing are 0.933 and 0.916, respectively. Although the developed predictor shows favorable performance, we do observe that our model also does not perform as well on the blind data as it does on the training and independent test data. Thus, there is still some room to improve the generalization ability.

3.5. Webserver implementation

The iPVP-DRLF is freely available at https://lab.malab.cn/~acy/iPVP-DRLF. The online server helps researchers determine whether their query protein sequences are plant vacuolar proteins. Here, we provide a simple introduction to server usage. The user first needs to paste or type protein sequences into the text box on the left and then click the "Submit" button for prediction. Notably, only FASTA-formatted sequences are supported as inputs for prediction. We have also provided an example of FASTA-formatted sequences in the input box. After the calculation is complete, the prediction results will be shown in a tabular format on the right. To reset the model and start new tasks, the "Resubmit" button can be clicked and the above-mentioned steps can be repeated to obtain new prediction results. Detailed step-by-step instructions on how to use the iPVP-DRLF server are available on the interface of the webserver. Furthermore, the datasets employed in this study can be downloaded from the web server to validate our findings or perform other research.

4. Conclusion

This research covered a rarely explored area of protein sequence analysis in bioinformatics, that is, the computational identification of PVPs. Based on existing methods, we tackled this problem by combining classic sequence features and deep representation features to encode plant vacuole protein sequences. We found that deep representation learning features are more informative and help plant vacuole protein identification than the commonly used classic sequence features. Moreover, we have proposed a more informative feature representation scheme by integrating and learning from both deep learning embedding features (BiLSTM and UniRep) and classic sequence descriptors (DDE and DPC). Sub-

sequently, these multi-view features were fused and fed into the LGBM classifier to construct the final predictive model. Application of the model shows favorable cross-validation, independent test and blind test accuracies of 88.25 %, 87.16 % and 66.52 %, respectively, all outperforming existing PVP predictors. Furthermore, the UMAP feature visualization demonstrates that the deep representation learning feature plays a more important role than the classic features in the model predictions. To facilitate the relevant research community, a user-friendly online webserver was implemented for iPVP-DRLF and made available for public use. Because of the current lack of highly accurate models specifically dedicated to PVP prediction, our study provides a complete methodology and benchmark and lays a foundation for further research in the future. We anticipate that iPVP-DRLF could serve as a powerful technique that could be used as a supplement to hands-on wet experiments for PVP identification and that its use could facilitate the elucidation of associated biological function mechanisms.

5. Author's contribution

Quan Zou conceived and designed the experiment. Shihu Jiao performed the experiment and analyzed the results. Quan Zou and Shihu Jiao wrote and revised the manuscript.

Funding

The work was supported by the National Natural Science Foundation of China (No. 61922020 and No. 62131004), the Sichuan Provincial Science Fund for Distinguished Young Scholars (2021JDJQ0025), and the Special Science Foundation of Quzhou (2021D004).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.06.002.

References

[1] Zhang C, Hicks GR, Raikhel NV. Molecular composition of plant vacuoles: important but less understood regulations and roles of tonoplast lipids. Plants (Basel, Switzerland) 2015;4(2):320–33.
[2] Kolb C et al. FYVE1 is essential for vacuole biogenesis and intracellular trafficking in arabidopsis. Plant Physiol 2015;167(4):1361–U414.
[3] Cui Y et al. Vacuole biogenesis in plants: how many vacuoles, how many models? Trends Plant Sci 2020;25(6):538–48.
[4] Zhang C, Hicks GR, Raikhel NV. Plant vacuole morphology and vacuolar trafficking. Frontiers. Plant Sci 2014;5.
[5] Neuhaus HE, Trentmann O. Regulation of transport processes across the tonoplast. Frontiers. Plant Sci 2014;5.
[6] Wiederhold E et al. The yeast vacuolar membrane proteome. Mol Cell Proteomics 2009;8(2):380–92.
[7] Kataoka T et al. Vacuolar sulfate transporters are essential determinants controlling internal distribution of sulfate in Arabidopsis. Plant Cell 2004;16 (10):2693–704.
[8] Martinoia, E., et al., Vacuolar Transporters in Their Physiological Context, in Annual Review of Plant Biology, Vol 63, S.S. Merchant, Editor. 2012. p. 183-213.
[9] Martinoia E, Maeshima M, Neuhaus HE. Vacuolar transporters and their essential role in plant metabolism. J Exp Bot 2007;58(1):83–102.
[10] Cheng X, Xiao X, Chou K-C. pLoc-mPlant: predict subcellular localization of multi-location plant proteins by incorporating the optimal GO information into general PseAAC. Mol BioSyst 2017;13(9):1722–7.
[11] Sahu SS, Loaiza CD, Kaundal R. Plant-mSubP: a computational framework for the prediction of single- and multi-target protein subcellular localization using integrated machine-learning approaches. Aob Plants 2019;12(3).
[12] Tahir M, Idris A. MD-LBP: An efficient computational model for protein subcellular localization from HeLa cell lines using SVM. Curr Bioinform 2020;15(3):204–11.
[13] Yadav AK, Singla D. VacPred: Sequence-based prediction of plant vacuole proteins using machine-learning techniques. J Biosci 2020;45(1).
[14] Lv Z et al. Anticancer peptides prediction with deep representation learning features. Briefings Bioinf 2021;22(5).
[15] Lv Z et al. Identification of sub-Golgi protein localization by use of deep representation learning features. Bioinformatics 2020;36(24):5600–9.
[16] Anteghini M, dos Santos VM, Saccenti E, In-Pero. Exploiting Deep Learning Embeddings of Protein Sequences to Predict the Localisation of Peroxisomal Proteins. Int J Mol Sci 2021;22(12).
[17] Cui F, Zhang Z, Zou Q. Sequence representation approaches for sequence-based protein prediction tasks that use deep learning. Brief Funct Genom 2021;20 (1):61–73.
[18] Long H et al. Predicting protein phosphorylation sites based on deep learning. Curr Bioinform 2020;15(4):300–8.
[19] Zhang Y et al. Review of the applications of deep learning in bioinformatics. Curr Bioinform 2020;15(8):898–911.
[20] Mcinnes L, Healy J. UMAP: uniform manifold approximation and projection for dimension reduction. J Open Source Software 2018;3(29):861.
[21] Huang Y et al. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 2010;26(5):680–2.
[22] Zou Q et al. Sequence clustering in bioinformatics: an empirical study. Briefings Bioinf 2020;21(1):1–10.
[23] Hooper CM et al. Finding the Subcellular Location of Barley, Wheat, Rice and Maize Proteins: The Compendium of Crop Proteins with Annotated Locations (cropPAL). Plant Cell Physiol 2016;57(1).
[24] Liu B, Gao X, Zhang H. BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. Nucleic Acids Res 2019;47(20).
[25] Chen Z et al. iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics 2018;34 (14):2499–502.
[26] Chen Z et al. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Briefings Bioinf 2020;21(3):1047–57.
[27] Zhu Y et al. Computational identification of eukaryotic promoters based on cascaded deep capsule neural networks. Brief Bioinform 2021;22(4).
[28] Alley EC et al. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods 2019;16(12):p. 1315-+.
[29] Bepler, T. and B. Berger, Learning protein sequence embeddings using information from structure. 2019.
[30] Rao, R., et al., Evaluating Protein Transfer Learning with TAPE. 2019.
[31] Nambiar A et al. Transforming the language of life: transformer neural networks for protein prediction tasks. BCB '20: 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, 2020.
[32] Ke, G., et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. in 31st Annual Conference on Neural Information Processing Systems (NIPS). 2017. Long Beach, CA.
[33] Zou Q et al. A novel features ranking metric with application to scalable visual and bioinformatics data classification. Neurocomputing 2016;173:346–54.
[34] Govern ADM. A new and simpler approximation for ANOVA under variance heterogeneity. J Educat Behav Stat 1994;19(2):91–101.
[35] Pedregosa F et al. Scikit-learn: machine learning in python. J Machine Learn Res 2011;12:2825–30.
[36] He S et al. MRMD2.0: A python tool for machine learning with feature ranking and reduction. Curr Bioinform 2020;15(10):1213–21.
[37] Zhang R et al. Feature selection with multi-view data: A survey. Inform Fusion 2019;50:158–67.
[38] Manayalan B et al. mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. Bioinformatics 2019;35(16):2757–65.
[39] Wei L et al. Improved and Promising Identification of Human MicroRNAs by Incorporating a High-Quality Negative Set. IEEE/ACM Trans Comput Biol Bioinf 2014;11(1):192–201.
[40] Ke, G., et al., LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in Advances in Neural Information Processing Systems 30, I. Guyon, et al., Editors. 2017.
[41] Li, J.P., Yuqian; Tang, Jijun; Zou, Quan; Guo, Fei, DeepATT: a hybrid category attention neural network for identifying functional effects of DNA sequences. Briefings in Bioinformatics, 2020: p. 1-1.
[42] Li J et al. DeepAVP: a dual-channel deep neural network for identifying variable-length antiviral peptides. IEEE J Biomed Health Inf 2020;24 (10):3012–9.