



VacPred: Sequence-based prediction of plant vacuole proteins using machine-learning techniques

ARVIND KUMAR YADAV^{1,†} and DEEPAK SINGLA^{2*,†} 

¹Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Solan, Himachal Pradesh 173 234, India

²School of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana, Punjab 141 004, India

*Corresponding author (Email, deepak@pau.edu)

†Equal contribution.

MS received 12 March 2020; accepted 30 July 2020

Subcellular localization prediction of the proteome is one of major goals of large-scale genome or proteome sequencing projects to define the gene functions that could be possible with the help of computational modeling techniques. Previously, different methods have been developed for this purpose using multi-label classification system and achieved a high level of accuracy. However, during the validation of our blind dataset of plant vacuole proteins, we observed that they have poor performance with accuracy value range from ~1.3% to 48.5%. The results showed that the previously developed methods are not very accurate for the plant vacuole protein prediction and thus emphasize the need to develop a more accurate and reliable algorithm. In this study, we have developed various compositions as well as PSSM-based models and achieved a high accuracy than previously developed methods. We have shown that our best model achieved ~63% accuracy on blind dataset, which is far better than currently available tools. Furthermore, we have implemented our best models in the form of GUI-based free software called ‘VacPred’ which is compatible with both Linux and Window platform. This software is freely available for download at www.deepaklab.com/vacpred.

Keywords. Software; subcellular localization; support vector machine; vacuole

1. Introduction

Vacuoles represent the cellular component of any living cell that varies in size and shape (Zhang *et al.* 2014a). Plant vacuole is represented by a single large structure that is involved in diverse functions such as plant growth and development, maintaining cellular homeostasis, cellular function to retaining turgor and nutrients, ions and secondary metabolites accretion (Pereira *et al.* 2014). Inside the seeds, the vacuole acts as the storage site of proteins and carbohydrates, various kinds of flavonoids for flower and fruit color, and is also associated with cellular response to the environment (Grotewold 2006; Marty 1999; Park *et al.* 2004). Vacuole proteins

function as a transporter to transport diverse class of ions, sugars, amino acids, and other molecules (Zhang *et al.* 2015). Lytic vacuole plays significant role in the degradation of cellular waste, defence, and program cell death (Ibl and Stoger 2014; Shimada *et al.* 2018).

With the availability of the whole genome or proteome of any plant, the ultimate goal is their fast and accurate functional assignment which depends upon the subcellular location of the proteins. The experimental method of subcellular localization is a very tedious and time-consuming process, therefore the focus is on the development of automatic and fast computational tool for accurate prediction. In the past, different multi-class algorithms have been developed for subcellular localization of pro-

Electronic supplementary material: The online version of this article (<https://doi.org/10.1007/s12038-020-00076-9>) contains supplementary material, which is available to authorized users.

teins such as BaCelLo (Pierleoni *et al.* 2006), MultiLoc2 (Blum *et al.* 2009), Plant-mPLoc (Chou and Shen 2010), PProwler 1.2 (Hawkins and Bodén 2006), Predotar v1.03 (Boden and Hawkins 2005), TargetP 1.1 (Emanuelsson *et al.* 2000), WoLF PSORT (Horton *et al.* 2007), YLoc (Briesemeister *et al.* 2010), pLoc-mPlant (Cheng *et al.* 2017), and Plant-mSubP (Sahu *et al.* 2019). However, none of them is specifically designed for the plant vacuole proteins, and thus perform very poorly in predicting the plant vacuole proteins. This emphasizes the need for an accurate computational model specifically trained for the plant vacuole proteins. Therefore, in this study, we developed a SVM-based prediction model for classification of vacuole proteins which is much better than previously developed software.

2. Materials and methods

2.1 Dataset preparation

The dataset used in this study was derived from publicly available database UniprotKB/SwissProt (release 3 July 2019). We searched the database with the query: (taxonomy: viridiplantae, location: SL-0272, length: >50, and reviewed: Yes), removed sequences with non-standard amino acids and identified a total of 579 plant vacuole proteins. To develop a supervised machine-learning-based model, the requirement of negative data is must. Thus, in a similar manner, we created the negative dataset with the query (taxonomy: viridiplantae, NOT location: SL-0272, length: >50 and reviewed: Yes) and identified 36,189 non-vacuole proteins from plants. To develop a non-redundant dataset of both vacuole and non-vacuole proteins, we used the CD-HIT program at 40% and 60% sequence identity threshold (Li and Godzik 2006). In case of vacuole proteins, this results in a total of 200 and 274 sequences at 40% and 60% identity cut-off. Similarly, CD-HIT results in 9485 proteins sequences at 40% identity cut-off from the non-vacuole protein dataset. To create a balanced dataset, we randomly selected 200 proteins from negative dataset and used them for developing the prediction model (Wei *et al.* 2018a). Thus, our final training dataset had 200 vacuole and 200 non-vacuole plant proteins. Hereafter, we call them positive and negative datasets.

2.2 Independent dataset

Evaluation of the performance of any machine-learning-based model requires an independent dataset.

Therefore, to create an independent dataset, we considered the difference of proteins at 60% (274) and 40% (200) cut-off and used it as an independent positive dataset. An equal number from the negative dataset which was not present in negative training dataset was used for creating negative independent dataset. Thus, our independent dataset consisted of 74 vacuole and 74 non-vacuole protein sequences.

2.3 Blind dataset

A blind dataset of plant vacuole proteins was created from the cropPAL database (Hooper *et al.* 2016). This database had experimentally determined (FP or MS/MS experiments) and predicted (more than 10 software) subcellular location of proteins from the 12 different plants. We extracted 228 vacuole proteins which were experimentally verified by either FP or MS/MS experiment. Further, one protein of length shorter than 50 amino acids was removed, thus making our final dataset of 227 vacuole proteins.

2.4 Feature calculation

To develop a machine-learning-based predictive model, a fixed-length vector is an essential requirement. In the past, different types of protein features, such as composition-based, physicochemical properties, and position-specific sequence matrix (PSSM), were used to develop robust prediction models. These features can be easily calculated by simple mathematical expressions. In this study, we used 7 types of composition-based features and 21 types of PSSM-based features to develop an efficient and reliable prediction model (supplementary table 1).

1. Composition-based features

- a. Amino acid composition (AAC): This is most widely used in developing protein sequence-based prediction models. In this case, any protein sequence is represented by 20 amino acids of fixed length. Percentage of each amino acid residue in a protein sequence is calculated as:

$$\begin{aligned} & \text{Percentage of amino acid}(i) \\ &= \frac{\text{Total number of amino acid}(i)}{\text{Total number of amino acids in protein}} \times 100 \end{aligned} \quad (1)$$

where i represent one of the 20 standard amino acids.

- b. Dipeptide composition (DPC): In this method, the composition of two consecutive amino acids of a sequence is calculated. This has a total vector of size 400 (20×20) having partial information of the order of amino acids. It can be calculated using the following formula:

$$\begin{aligned} & \text{Dipeptide composition}(i) \\ &= \frac{\text{Total number of Dipeptide } (i)}{\text{Total number of all possible dipeptides}} \times 100 \end{aligned} \quad (2)$$

where i represent one out of 400 dipeptides.

- c. Tripeptide composition (TPC): Tripeptide composition represents the percentage composition of each of the 8000 possible tripeptide form by 20 amino acids and calculated as:

$$\begin{aligned} & \text{Tripeptide composition}(i) \\ &= \frac{\text{Total number of Tripeptide } (i)}{\text{Total number of all possible Tripeptides}} \times 100 \end{aligned} \quad (3)$$

where i represent one out of 8000 tripeptides.

- d. C-terminal Composition: Previously, it was observed that the C-terminal protein region might have any significant roles in biological activity, so this portion could be used for separate sequence composition calculations. In this study, we extracted the 5 and 10 amino acid residues from the C-terminal region of the protein and used for calculation of amino acid composition.
- e. Split and rest amino acid composition: Previous studies reported that some important sequence motifs might be present in a specific protein region and help to improve prediction accuracy (Srinivasan *et al.* 2013). In the case, the whole protein is split into three equal parts and composition of each part is calculated separately. However, in case of rest composition method, amino acid composition of protein is calculated after removing the specified N- and C-terminal residues. In our case, we removed 10 residues from each of N- and C-terminal of protein and calculate the composition of rest region of protein.

2. PSSM-based features
PSSM profile generated using the PSI-blast search is based on evolutionary information used to identify

the remote homologs. Previously, it has been used in developing various machine-learning-based models for the sequence annotation (McGuffin *et al.* 2000; Saha *et al.* 2006). POSSUM server was used to produce a PSSM profile based on the uniref50 database searched for three iterations at e-value 0.001 and 21 different types of PSSM-based features were calculated and used for modelling (Wang *et al.* 2017). POSSUM divided these features set into four major groups, i.e. generated by transformation of rows, columns, both row and column, combination of all these features (supplementary table 1).

2.5 Support Vector Machine (SVM)

SVM is a powerful machine-learning software that has been extensively used in various bioinformatics analyses (FY *et al.* 2019; Boopathi *et al.* 2019; Manavalan *et al.* 2018a, 2018b; Manavalan and Lee 2017; Wei *et al.* 2018b). This is a very reliable technique for biological sequence analysis due to its capability of handling noise and high-dimensional feature space (Zavaljevski *et al.* 2002). SVM allows the users to tune various parameters available for different kernels such as linear, polynomial, sigmoid, or radial basis function (RBF) (Ramana and Gupta 2009; Ramana 2015; Mishra *et al.* 2014). In this study, we used freely available software SVM^{light} (<http://svmlight.joachims.org>) to train SVM classifiers and develop prediction models.

2.6 Five-fold cross-validation

A five-fold cross-validation technique was used to examine the quality of develop models. In the case, the complete dataset was divided into five equal subsets of which four subsets were combined and used for as a training set and fifth subset was used as test set. The complete process was repeated five times so that each subset was used as a test set at least one time.

2.7 Performance evaluations

To evaluate the quality of developed models, we used confusion matrix metrics with sensitivity, specificity, accuracy, and Matthew correlation coefficient (MCC) as described previously (Dao *et al.* 2019). Area under receiver opening curve (ROC) was also considered to measure the overall prediction performance.

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} \times 100 \quad (4)$$

$$\text{Specificity} = \frac{TN}{(TN + FP)} \times 100 \quad (5)$$

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)} \times 100 \quad (6)$$

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \quad (7)$$

Here TP, FP, TN, and FN are the true positives, false positives, true negatives, and false negatives, respectively.

3. Results

3.1 Composition-based models

The present study reported the various composition-based models developed using amino acid, dipeptide, C-terminal, and rest and split amino acid composition for the annotation of plant vacuole proteins. Firstly, the model developed on AAC resulted in ~79% accuracy with an MCC value 0.58 on the independent dataset (table 1). Similarly, the dipeptide composition-based model showed 82.43% sensitivity and 78.38% specificity with an accuracy of 80.41% on the independent dataset (table 1). However, C-terminal-based C5 and C10 model performed poorly with maximum accuracy ~49% and 55% on independent datasets respectively. Furthermore, the model developed using split and rest amino acid composition achieve 70.95% and 75.68% accuracy on independent datasets. As observed from table 1, DPC-based model performs the best compared with all the other composition-based models.

3.2 PSSM-based models

Based on the PSSM profile, we developed 21 different models for evaluating the significance and performance of each of the PSSM-based features. Among the various row-transform-based features, AAC-PSSM showed 93.24% sensitivity, 78.38% specificity, and 85.81% accuracy with MCC value 0.72 on an independent dataset. From table 2, we observed that F-PSSM and Smooth-PSSM performed poorly with maximum MCC values 0.30 and 0.46 on independent dataset respectively.

Table 1. Performance of composition-based models on training and independent datasets

Feature	Sensitivity	Specificity	Accuracy	MCC	ROC
AAC-Train	76.00	72.00	74.00	0.48	0.79
AAC-Ind	83.78	74.32	79.05	0.58	0.87
DPC-Train	70.00	81.00	75.50	0.51	0.80
DPC-Ind	82.43	78.38	80.41	0.61	0.84
TPC-Train	62.00	83.50	72.75	0.47	0.78
TPC-Ind	74.32	78.38	76.35	0.53	0.86
Split-Train	74.50	68.00	71.25	0.43	0.76
Split-Ind	82.43	59.46	70.95	0.43	0.80
Rest-Train	76.50	70.00	73.25	0.47	0.79
Rest-Ind	82.43	68.92	75.68	0.52	0.85
C5-Train	55.00	61.50	58.25	0.17	0.59
C5-Ind	41.89	56.76	49.32	-0.01	0.53
C10-Train	57.50	59.00	58.25	0.17	0.57
C10-Ind	60.81	50.00	55.41	0.11	0.59

Conversely, S-PSSM and RPM-PSSM performed well among the row-transformed features with RPM-PSSM more balanced compared to S-PSSM in terms of sensitivity and specificity values (table 2). Similarly, among the column-transformed features, K-PSSM and TRI-PSSM performed the best with sensitivity 90.54%/93.24%, specificity 82.43%/82.43% with accuracy value 86.49%/87.84% respectively. However, DPC-PSSM and TPC-PSSM were not as good as compared to the others. We observed that the model developed on mixed features performed better compared to individual transformed features (table 2). As evident from table 2, K-PSSM was the best performing balanced model in terms of accuracy and MCC value among all the PSSM-based models.

3.3 Validation on blind dataset

The blind dataset of vacuole proteins constructed in this study was used to evaluate the performance of our best models. We considered our two best model: DPC model and K-PSSM model, and compared the performance with previously developed models. The prediction results of previously developed models showed

Table 2. Performance of the PSSM-based models on training and independent datasets

Feature	Sensitivity	Specificity	Accuracy	MCC	ROC
<i>Row-transformation-based descriptors</i>					
AAC-PSSM-Train	75.00	76.50	75.75	0.52	0.80
AAC-PSSM-Ind	93.24	78.38	85.81	0.72	0.91
S-PSSM-Train	86.50	70.50	78.50	0.58	0.82
S-PSSM-Ind	97.30	63.51	80.41	0.65	0.91
F-PSSM-Train	62.00	65.00	63.50	0.27	0.66
F-PSSM-Ind	90.54	33.78	62.16	0.30	0.72
AB-PSSM-Train	76.50	75.00	75.75	0.52	0.80
AB-PSSM-Ind	93.24	68.92	81.08	0.64	0.90
PSSM-Comp-Train	74.00	76.50	75.25	0.51	0.81
PSSM-Comp-Ind	94.59	83.78	89.19	0.79	0.94
RPM-PSSM-Train	78.50	77.00	77.75	0.56	0.84
RPM-PSSM-Ind	94.59	74.32	84.46	0.70	0.91
Smooth-PSSM-Train	71.50	66.50	69.00	0.38	0.73
Smooth-PSSM-Ind	83.78	60.81	72.30	0.46	0.84
<i>Column-transformation-based descriptors</i>					
DPC-PSSM-Train	64.50	86.00	75.25	0.52	0.81
DPC-PSSM-Ind	91.89	85.14	88.51	0.77	0.96
K-PSSM-Train	76.50	87.00	81.75	0.64	0.86
K-PSSM-Ind	90.54	82.43	86.49	0.73	0.93
TRI-PSSM-Train	74.00	85.50	79.75	0.60	0.87
TRI-PSSM-Ind	93.24	82.43	87.84	0.76	0.93
EEDP-PSSM-Train	74.50	76.00	75.25	0.51	0.80
EEDP-PSSM-Ind	98.65	75.68	87.16	0.76	0.92
TPC-PSSM-Train	88.00	53.00	70.50	0.44	0.79
TPC-PSSM-Ind	95.95	51.35	73.65	0.53	0.88
<i>Mixed row- and column-transformation-based descriptions</i>					
EDP-PSSM-Train	70.00	67.50	68.75	0.38	0.71
EDP-PSSM-Ind	87.84	68.92	78.38	0.58	0.81
PSE-PSSM-Train	80.00	78.00	79.00	0.58	0.85
PSE-PSSM-Ind	97.30	71.62	84.46	0.71	0.91
DP-PSSM-Train	83.00	75.00	79.00	0.58	0.84
DP-PSSM-Ind	98.65	68.92	83.78	0.71	0.91
PSSM-AC-Train	56.50	77.00	66.75	0.34	0.70
PSSM-AC-Ind	70.27	77.03	73.65	0.47	0.78
PSSM-CC-Train	63.50	82.50	73.00	0.47	0.78
PSSM-CC-Ind	77.03	83.78	80.41	0.61	0.89
R-PSSM-Train	77.00	69.00	73.00	0.46	0.77
R-PSSM-Ind	98.65	71.62	85.14	0.73	0.90
<i>Combinative descriptors</i>					
AADP-PSSM-Train	64.00	87.50	75.75	0.53	0.81
AADP-PSSM-Ind	91.89	89.19	90.54	0.81	0.96
AATP-Train	80.50	63.50	72.00	0.45	0.78
AATP-Ind	91.89	68.92	80.41	0.62	0.90
MEDP-Train	74.00	75.50	74.75	0.50	0.80
MEDP-Ind	95.95	75.68	85.81	0.73	0.92

Train: training dataset; Ind: individual dataset; MCC: Matthews correlation coefficient; ROC: area under receiver opening curve.

very poor performance with accuracy varies from 1.32% to 41.85% (figure 1, table 3). BaCello has incorporated the amino acid composition, sequence profile, and signal information to develop the SVM-based model. However, MultiLoc2 is based on a six-layered prediction that uses gene ontology and

phylogenetic information along with sequence composition and motif analysis. Similarly, Plant-mPLoc is specially designed for plant proteins classification with protein domain, gene ontology, and evolutionary information along with sequence composition. TargetP is based on N-, C-terminal signal sequences, while

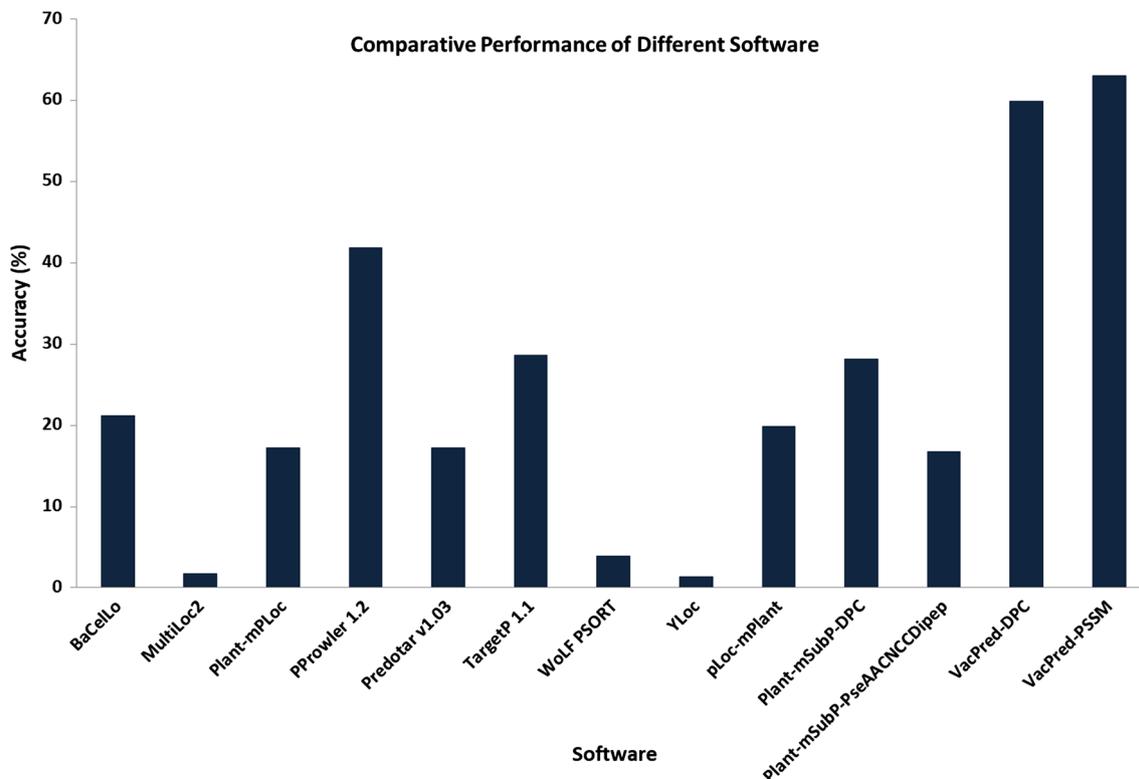


Figure 1. Performance of different software on the blind dataset.

Table 3. Benchmarking of different software on blind datasets

Software	Vacuole predicted	Accuracy (%)
BaCelLo	48	21.15
MultiLoc2	4	1.76
Plant-mPLoc	39	17.18
PProWler 1.2	95	41.85
Predotar v1.03	39	17.18
TargetP 1.1	65	28.63
WoLF PSORT	9	3.96
YLoc	3	1.32
pLoc-mPlant	45	19.83
Plant-mSubP-DPC	64	28.19
Plant-mSubP-PseAACNCCDipep	38	16.74
VacPred-DPC	136	59.91
VacPred-PSSM	143	62.99

WoLF PSORT, in addition to signal, is based on composition and functional motifs. Recently, Plant-mSubP for classification of single as well as dual-label plant protein has been developed which shows the best performance on a hybrid model (PseAAC-NCC-DIPEP) composed of pseudo-amino acid composition,

N-terminal signal, and dipeptide composition. Furthermore, most of the good performing models have predicted multiple locations rather than one location, as evident from the cropPAL database. This clearly shows that these models have not captured sufficient features for plant vacuole proteins. However, our DPC and K-PSSM model has correctly classified 136 and 143 proteins with accuracy 59.91%, and 62.99% respectively (figure 1). The high accuracy clearly indicates the applicability of our models.

3.4 Software

Based on our study, we have developed GUI-based software ‘VacPred’ that is compatible with the different operating systems (figure 2). We have incorporated our two best algorithms – DPC model and K-PSSM model – for the prediction of plant vacuole proteins. To execute the DPC-based prediction, users only need a protein sequence fasta file without any limitation on file size or number of sequences. Our K-PSSM-based model is based on the features calculated using POSSUM software; thus, users need to first calculate the K-PSSM features using POSSUM software and the output file of this software is directly given as input for

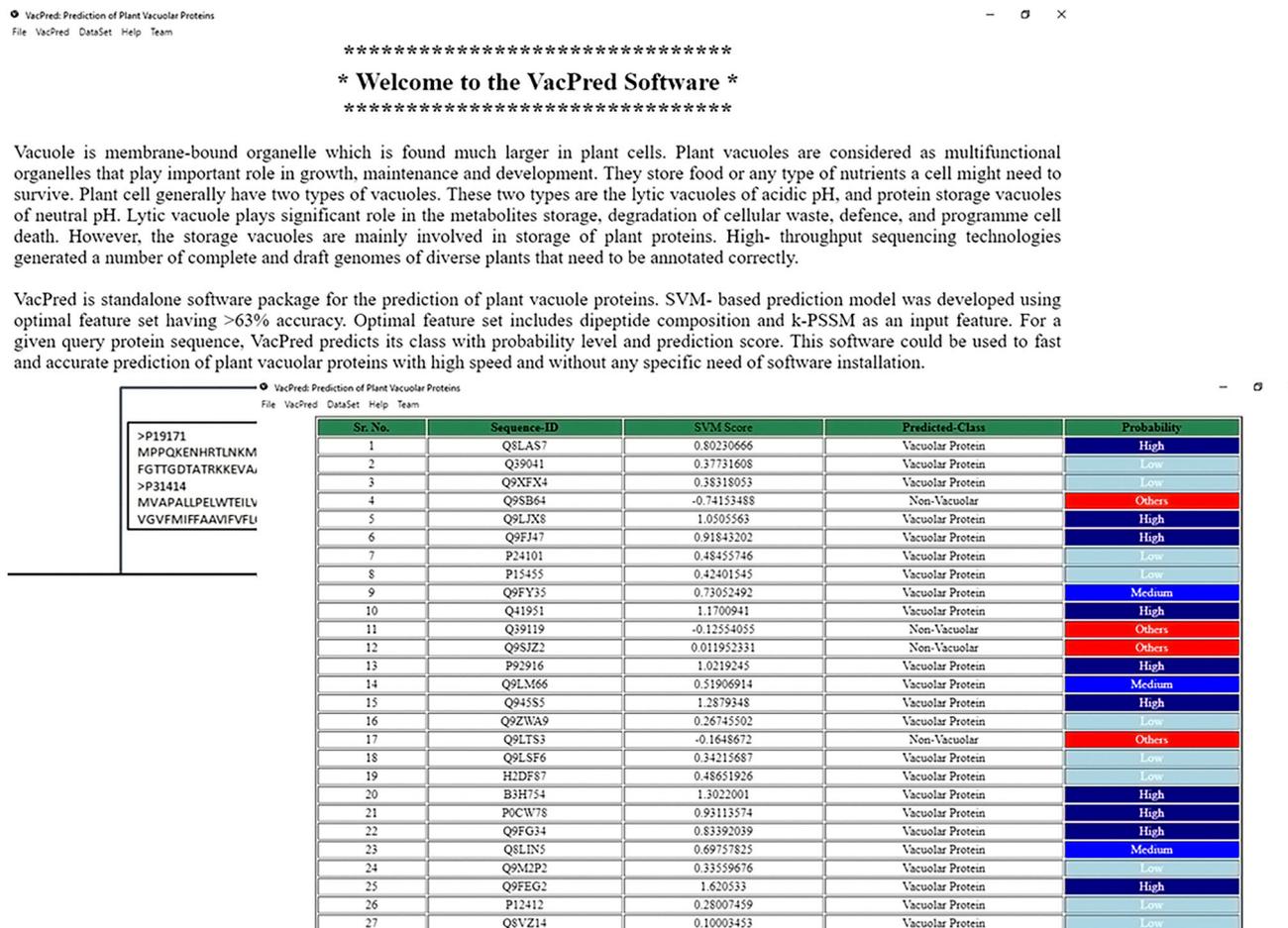


Figure 2. The homepage and prediction result of the VacPred software.

the prediction of plant vacuole proteins. VacPred is developed using nodejs-based framework from electron software that used javascript and PERL in the backend. This is accessible at www.deepaklab.com/vacpred.

4. Discussion

Next-generation sequencing has completed various genome or transcriptome projects and many more are in progress. Genome annotation including subcellular localization is the most crucial and important steps of any genome sequencing projects that shed the light on protein structure and functions. Among the various cellular organelles, plant vacuole is one of the most important components of plant cells that perform diverse functions (Zhang et al. 2014a, b; Pereira et al. 2014). The experimental identification of plant vacuole protein is a time-consuming and costly affair that requires sophisticated instruments and manpower.

To overcome this, machine-learning-based computational methods evolved as highly efficient and less

expensive way of sequence annotation. Furthermore, our analysis confirmed that all the previously developed models were not able to predict plant vacuole proteins with high accuracy. We applied machine-learning-based techniques and developed more than 30 different types of models. In the end, we had selected two best performing models including one dipeptide composition-based and one PSSM-based model. Both models showed similar performance on a blind dataset with ~60% and ~63% accuracy on DPC and KPSSM-based model. Based on this analysis, we developed a standalone GUI software ‘VacPred’ that will be useful for large-scale annotation projects for the plant vacuole protein prediction.

Acknowledgements

The authors are thankful to the DBT-BTISNET for providing the bioinformatics facility at the School of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana.

References

- Boden M and Hawkins J 2005 Prediction of subcellular localization using sequence-biased recurrent networks. *Bioinformatics* **21** 2279–2286
- Boopathi V, Subramaniyam S, Malik A *et al.* 2019 mACPpred: A Support Vector Machine-Based Meta-Predictor for Identification of Anticancer Peptides. *Int. J. Mol. Sci.* **20** 1964
- Briesemeister S, Rahnenfuhrer J and Kohlbacher O 2010 YLoc—an interpretable web server for predicting subcellular localization. *Nucleic Acids Res.* **38** W497–W502
- Blum T, Briesemeister S and Kohlbacher O 2009 MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics* **10** 274
- Cheng X, Xiao X, and Chou KC 2017 pLoc-mPlant: predict subcellular localization of multi-location plant proteins by incorporating the optimal GO information into general PseAAC. *Mol. Biosyst.* **13** 1722–1727
- Chou KC and Shen HB 2010 Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. *PLoS One* **5** e11335
- Dao FY, Lv H, Wang F, *et al.* 2019 Identify origin of replication in *saccharomyces cerevisiae* using two-step feature selection technique. *Bioinformatics* **35** 2075–2083
- Emanuelsson O, Nielsen H, Brunak S, *et al.* 2000 Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300** 1005–1016
- Grotewold E 2006 The genetics and biochemistry of floral pigments. *Annu. Rev. Plant Biol.* **57** 761–780
- Hawkins J and Bodén M 2006 Detecting and sorting targeting peptides with neural networks and support vector machines. *J. Bioinform. Comput. Biol.* **4** 1–18
- Hooper CM, Castleden IR, Aryamanesh N, *et al.* 2016 Finding the subcellular location of barley, wheat, rice and maize proteins: the compendium of crop proteins with annotated locations (cropPAL). *Plant Cell Physiol.* **57** e9–e9
- Horton P, Park KJ, Obayashi T, *et al.* 2007 WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* **35** W585–W587
- Ibl V and Stoger E 2014 Live Cell imaging during germination reveals dynamic tubular structures derived from protein storage vacuoles of barley aleurone cells. *Plants* **3** 442–457
- Li W and Godzik A 2006 Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22** 1658–1659
- Manavalan B, Subramaniyam S, Shin TH, *et al.* 2018 Machine-Learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. *J. Proteome Res.* **17** 2715–2726
- Manavalan B, Shin TH and Lee G 2018 PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine. *Front. Microbiol.* **9** 476
- Manavalan B and Lee J 2017 SVMQA: support-vector-machine-based protein single-model quality assessment. *Bioinformatics* **33** 2496–2503
- Marty F 1999 Plant vacuoles. *Plant Cell* **11** 587–599
- McGuffin LJ, Bryson K and Jones DT 2000 The PSIPRED protein structure prediction server. *Bioinformatics* **16** 404–5
- Mishra NK, Chang J, and Zhao PX 2014 Prediction of membrane transport proteins and their substrate specificities using primary sequence information. *PLoS One* **9** e100278
- Park M, Kim SJ, Vitale A, *et al.* 2004 Identification of the protein storage vacuole and protein targeting to the vacuole in leaf cells of three plant species. *Plant Physiol.* **134** 625–639
- Pereira C, Pereira S and Pissarra J 2014 Delivering of proteins to the plant vacuole—an update. *Int. J. Mol. Sci.* **15** 7611–7623
- Pierleoni A, Martelli PL, Fariselli P, *et al.* 2006 BaCelLo: a balanced subcellular localization predictor. *Bioinformatics* **22** e408–e416
- Ramana J and Gupta D 2009 LipocalinPred: A SVM-based method for prediction of lipocalins. *BMC Bioinformatics* **10** 445
- Saha S, Zack J, Singh B, *et al.* 2006 VGChan: prediction and classification of voltage-gated ion channels. *Genomics. Proteomics Bioinformatics* **4** 253–8
- Sahu SS, Loaiza CD, Kaundal R, *et al.* 2019 Plant-mSubP: a computational framework for the prediction of single- and multi-target protein subcellular localization using integrated machine-learning approaches. *AoB Plants* **12** 068
- Shimada T, Takagi J, Ichino T, *et al.* 2018 Plant Vacuoles. *Annu. Rev. Plant Biol.* **69** 123–145
- Srinivasan SM, Vural S, King BR, *et al.* 2013 Mining for class-specific motifs in protein sequence classification. *BMC Bioinformatics* **14** 96
- Tamanna and Ramana j 2015 MATEPRED-A-SVM-Based prediction method for multidrug and toxin extrusion (MATE) proteins. *Comput. Biol. Chem.* **58** 199–204
- Wang J, Yang B, Revote J, *et al.* 2017 POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. *Bioinformatics* **33** 2756–2758
- Wei L, Zhou C, Chen H, *et al.* 2018 ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* **34** 4007–4016
- Wei L, Chen H and Su R 2018 M6APred-EL: A sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. *Mol. Ther. Nucleic Acid* **12** 635–644
- Zavaljevski N, Stevens FJ and Reifman J 2002 Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. *Bioinformatics* **18** 689–696

- Zhang C, Hicks G and Raikhel N 2015 Molecular composition of plant vacuoles: important but less understood regulations and roles of tonoplast lipids. *Plants* **4** 320–333
- Zhang C, Hicks GR and Raikhel NV 2014 Plant vacuole morphology and vacuolar trafficking. *Front. Plant Sci.* **5** 476
- Zhang L, Zhao X and Kong L 2014 Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou[U+05F3]s pseudo amino acid composition. *J. Theor. Biol.* **355** 105–110

Corresponding editor: SREENIVAS CHAVALI