Plant Molecular Biology (2022) 110:81–92
https://doi.org/10.1007/s11103-022-01288-3

PseAraUbi: predicting arabidopsis ubiquitination sites by incorporating the physico-chemical and structural features

Wei Wang[1,2] · Yu Zhang[1] · Dong Liu[1] · HongJun Zhang[3] · XianFang Wang[4] · Yun Zhou[1]

Received: 5 February 2022 / Accepted: 9 May 2022 / Published online: 1 July 2022

Abstract

Key message **We makes three kinds of important features from *Arabidopsis thaliana*: protein secondary structure based on the Chou-Fasman parameter, amino acids hydrophobicity and polarity information, and analyze their properties.**

Abstract Ubiquitination modification is an important post-translational modification of proteins, which participates in the regulation of many important life activities in cells. At present, ubiquitination proteomics research is mostly concentrated in animals and yeasts, while relatively few studies have been carried out in plants. It can be said that the calculation and prediction of *Arabidopsis thaliana* ubiquitination sites is still in its infancy. Based on this, we describe a calculation method, PseAraUbi (Prediction of *Arabidopsis thaliana* ubiquitination sites using pseudo amino acid composition), that can effectively detect ubiquitination sites on *Arabidopsis thaliana* using support vector machine learning classifiers. Based on protein sequence information, extract features from the Chou-Fasman parameter, amino acids hydrophobicity features, polarity information and selected for classification with the Boruta algorithm. PseAraUbi achieves promising performances with an AUC score of 0.953 with fivefold cross-validation on the training dataset, which are significantly better than that of the pioneer *Arabidopsis thaliana* ubiquitination sites method. We also proved the ability of our proposed method on independent test sets, thus gaining a competitive advantage. In addition, we also in-depth analyzed the physicochemical properties of amino acids in the region adjacent to the ubiquitination site. To facilitate the community, the source code, optimal feature subset, ubiquitination sites dataset in the *Arbidopsis* proteome are available at GitHub (https://github.com/HNUBioinformatics/PseAraUbi.git) for interest users.

Keywords Ubiquitination sites · *Arabidopsis thaliana* · Sequence information · Support vector machine

Introduction

Ubiquitin (UB), a highly conserved small molecule protein composed of 76 amino acids, is used for post-translational modification of substrate proteins (Nobuhiro 2018; Mattern et al. 2019; Mulder et al. 2019). Modifying the proteins by the addition UB into the substrate proteins, can regulate almost all biochemistry within eukaryotic cells. Ubiquitination requires the synergistic activity of three different ubiquitin enzymes: ATP-dependent ubiquitin activating enzyme (E1), ubiquitin-conjugation enzyme (E2) and ubiquitin-protein ligase (E3). The process of transferring ubiquitin to the lysine residues of the target protein molecule involves three important steps. Firstly, a thioester bond formed between the glycine residue at the carboxyl terminal of ubiquitin and the cysteine sulfhydryl group of the active center of E1 enzyme. And then the activated ubiquitin molecule was

✉ Wei Wang
weiwang@htu.edu.cn

✉ Yun Zhou
zy@htu.edu.cn

[1] College of Computer and Information Engineering, Henan Normal University, Xinxiang 453000, China

[2] Key Laboratory of Artificial Intelligence and Personalized Learning in Education of Henan Province, Xinxiang, China

[3] School of Computer Science and Technology, Anyang University, Anyang 455000, China

[4] College of Computer Science and Technology Engineering, Henan Institute of Technology, Xinxiang 453000, China

transferred to the cysteine sulfhydryl group of E2 enzyme. Finally, the ubiquitin was transferred to the lysine residue of the target protein molecule with help of E3 enzyme. As a post-translational modification, ubiquitination is strongly related to various complex biological processes and diseases in plants and animals. It participates widely in physiological processes, such as immune response regulation, DNA damage repair, cell cycle regulation, cell apoptosis and protein degradation, et al., through regulating protein stability, subcellular localization, activity and interaction (Herrmann et al. 2007; Wagner et al.2011; Huizen and Kikkert 2019; Yau et al. 2017).

Protein ubiquitination plays a key role in many stages of plant development. And many experimental studies have already proved that ubiquitination was involved in the growth and development, biotic and abiotic stress and metabolism et al. of plant. Such as, PUB4, as an E3 ubiquitin ligase, which is proved to be a key regulatory component of root meristem maintenance and plays an important role in the downstream of an exogenous CLV3 peptide synthesis (Kinoshita et al. 2015). In addition, SPL2 is one of the only three types of E3 ligases found in the outer membrane of plant chloroplasts, which can bind lanthanide ions (Tracz et al. 2021). Protein ubiquitination mediating Jasmonates and ethylene signaling pathways is very important for regulating plant responses to low-temperature stress (Gong et al. 2020). Posttranslational modifications (PTMs) can effectively break through the limitation of the number of genes. By modifying proteins, the functional diversity of proteins can be significantly expanded, so that cells can better adapt to various environmental stimuli. The interaction between different PTMs further increases the complexity of PTMs, allowing cells to respond more quickly and accurately to various physiological responses, including biotic stresses. PTM interactions in plants have also been shown to exist in various physiological responses of plants (Zhang and Zeng. 2020). Xie summarized the regulatory mechanism of ubiquitination in different plant responses to low phosphorus. For example, plants can regulate primary root growth and lateral root development by relieving endoplasmic reticulum stress at the root tip through the induction of autophagy pathway (Pan et al. 2019).

Finding the accurate location of ubiquitination sites is the basis for studying protein ubiquitination, and provides an effective way to further elucidate the molecular mechanism of ubiquitination modified proteins. Therefore, different experimental methods, such as Mass Spectrometry (MS) techniques (Wang et al. 2019; Xu et al. 2010), ubiquitin antibodies (Brogi et al. 2020), etc., were adopted to identify the ubiquitination sites in ubiquitinated proteins. Unfortunately, identifying the ubiquitination sites with laboratory tests not only is susceptible to the timeliness and reversibility of the ubiquitination processes, but also is expensive and time-consuming. To solve these problems, machine learning methods was used to predict the ubiquitination sites of because it was cheaper and more time-efficient than laboratory tests.

In fact, there have been some methods to predict ubiquitination sites (Yu et al. 2020; Wang and Zhang 2019; He et al. 2018; Chen et al. 2013; Wang et al. 2017). Yu et al. (Yu et al. 2020) used the deep migration learning method to predict the ubiquitination sites of Homo sapiens, Toxoplasma gondii, rice and other species, and achieved better performance on a small sample of Rattus norvegicus datasets. Chen et al. (Chen et al. 2013). respectively used CKSAAP coding, binary amino acid coding, AAindex physicochemical property coding and protein aggregation tendency coding to predict human ubiquitination sites; And, by comparing with previous studies, the yeast ubiquitination site predictor is often unable to accurately predict human ubiquitination sites because of the significant difference in amino acid preference between the sequence neighbors of human ubiquitination sites and yeast counterparts. And, the ubiquitination sites of different species have their own characteristics (Kumar and Vellaichamy 2019; Chen et al. 2014). So specific prediction tools need to be established for different species. At the present, a variety of ubiquitination site prediction models for humans, mice and yeast have been developed. But as for *Arabidopsis thaliana* species, there are only some predictors have been developed (Chen et al. 2019; Mosharaf et al. 2020; Wang et al. 2021). One of the three prediction methods in common is that they use binary coding or CKSAAP coding schemes for feature extraction. We can draw such a conclusion: existing researches have adopted a single method for extracting features of Arabidopsis protein sequences. In response to the challenge, this study attempted to establish a novel feature computational method for identifying ubiquitination sites based on Arabidopsis protein sequences.

To solve this problem, we used an Arabidopsis dataset containing 1607 protein sequences. In addition, we proposed a novel method, PseAraUbi, for predicting ubiquitin sites in Arabidopsis. To further improve the prediction accuracy, we computed an optimal set of features from 30 features selected by the Boruta algorithm from various Chou-Fasman parameters, hydrophobicity and polarity of amino acids. Then, the SVM classifier based on this feature representation was used to predict ubiquitination sites. Furthermore, it demonstrated that the PseAraUbi could significantly improve the overall performance of the cross-validation dataset and independent dataset, and compared with other state-of-the-art predictors, it could predict ubiquitination sites more accurately. The flowchart of PseAraUbi is shown in Fig. 1. In addition, we also analyzed the physical and chemical properties of the amino acids near the ubiquitination site in the Arabidopsis protein sequence, and found that the amino acid properties

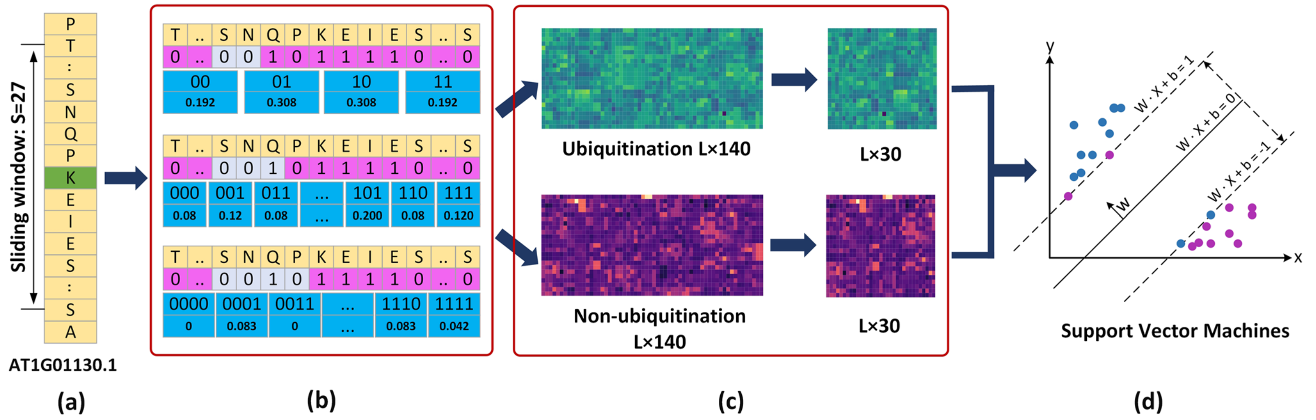


Fig. 1 Flowchart of PseAraUbi. (a) represents the length of the amino acid sequence selected with the sliding window size of 27, which represent the ubiquitinated or non-ubiquitinated sites in green and yellow. (b) represents feature extraction. Yellow represents amino acid sequence, and pink represents encoding amino acid into binary vector, and gray represents grouping according to consecutive 2, 3 and 4 amino acids respectively. The feature extraction part in the figure is an example of the tendency of amino acids to form an alpha helix (Pα), Pα > 100 represented by 1, otherwise represented by 0; blue is the probability of the occurrence of 2, 3 and 4 consecutive amino acids. (c) is feature extraction. Green indicates positive sample features, and amaranth indicates negative sample features, where L is the characteristic dimension, and Boruta algorithm is used to screen the optimal feature of the 30-dimension from the 140-dimension feature. (d) represents the final prediction model constructed by the support vector machine model

of the ubiquitinated sequence and the non-ubiquitinated sequence are different.

Materials and method

Data preparation

In this study, the protein data we used is derived from a previously published paper, in which the ubiquitination site (lysine residues) was verified by the ubiquitin binding diagonal chromatography experiment (Walton et al. 2016). The ubiquitination sites annotations were extracted from the UniProtKB/Swiss-Prot and NCBI protein sequence database (https://www.ncbi.nlm.nih.gov/gene/) regarding the model plant *Arabidopsis thaliana*. In this study, we refer to the experimentally verified lysine ubiquitination sites as positive samples (i.e., ubiquitination sites), and the remaining lysine residues as negative samples (i.e., non-ubiquitination sites). We collected a total of 1,607 Arabidopsis proteins, from which 500 proteins were randomly selected without duplication for model training, and a total of 1,120 ubiquitination sites were obtained from these 500 Arabidopsis proteins. In order to overcome the problem of model overfitting in prediction, CD-HIT server is used to reduce sequence homology, and a 40% sequence identity threshold is used to solve homology redundancy (Li 2006; Fu et al. 2012). In the end, we obtained 472 Arabidopsis protein including 1054 ubiquitination sites. Subsequently, randomly selected non-ubiquitination sites equivalent to the number of ubiquitination sites from the negative samples. In order to verify the performance of the model, cross-validation experiments and independent test data sets were used. The independent dataset is consisted of 300 ubiquitinated protein having 612 ubiquitination sites, which also used CD-HIT for eliminating homologous protein redundancy. Similarly, the independent dataset also contained 1:1 ratio of positive and negative samples.

Performance evaluation

To assess the performance, we adopt several widely used measures, including accuracy (ACC), sensitivity (SEN/Recall), specificity (SPE), precision (PRE), the Matthew's correlation coefficient (MCC) and the area under the ROC curve (AUC). These measurements are defined as:

$$ACC = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{1}$$

$$SEN = \frac{T_P}{T_P + F_N} \tag{2}$$

$$SPE = \frac{T_N}{T_N + F_N} \tag{3}$$

$$PRE = \frac{T_P}{T_P + F_P} \tag{4}$$

$$MCC = \frac{T_P \times T_N - F_P \times F_N}{\sqrt{(T_P + F_P)(T_P + F_N)(T_N + F_P)(T_N + F_N)}} \quad (5)$$

Among them, true positive (TP) represents the number of true ubiquitination sites that are predicted correctly; true negative (TN) represents the number of true non-ubiquitination sites that are correctly predicted; false negative (FN) represents the true ubiquitination sites number of points, and these sites are designated as non-ubiquitination; false positive (FP) represents the number of true non-ubiquitination sites, and these sites are designated as ubiquitination sites.

Features extraction

We initially used a comprehensive set of 140-dimensional features to make further selections. The feature set is composed of the Chou-Fasman parameter (Zhou 1998), hydrophobicity (Xiao and Chou 2007; Matsui et al. 2017) of amino acid and polarity (Maheshwari and Dhathathreyan 2004) of amino acids. Here, in order to encode the Arabidopsis protein into a suitable sequence fragment, and then convert the sequence into a computer-recognizable binary vector. We consider using the optimal protein sequence fragment size of 27, also known as the window size. The window size of both training set and independent test dataset is 27-long sliding window from whole protein sequences. The middle position of each sequence fragment contains a ubiquitination site (lysine) or a non-ubiquitination site. If the ubiquitination site is located at the beginning or end of the protein sequence, resulting in the sequence length being shorter than 27, the missing position is filled with the character X.

The pseudo amino acid composition (PseAAC) (Ju and Wang 2018; Naseer et al. 2021) that can truly reflect the intrinsic correlation between the protein sample and the attribute to be predicted, the general form of Chou's PseAAC (Chou 2001) can be calculated as:

$$P = \left[P_1, P_2, P_3, \ldots, P_{20}, P_{20+1}, \ldots P_{20+\lambda}\right]^T \quad (6)$$

$$P = \begin{cases} \frac{f_u}{\sum_{u=1}^{20} f_u + \sum_{k=1}^{\lambda} \tau_k} 1 \le u \le 20 \\ \frac{\mathrm{w}\tau_{u-20}}{\sum_{u=1}^{20} f_u + w \sum_{k=1}^{\lambda} \tau_k} 20 + 1 \le u \le 20 + \lambda \end{cases} \quad (7)$$

where W is weight factor,τ_k is the sequential information of the sequence, *fu* is the number of occurrences of amino acid *u*.

Chou and Fasman (Chou and Fasman 1978) provided a statistical method for the prediction protein second structure prediction, named as Chou-Fasman methods. The basic idea is to assign three numbers for every amino acid, which describes the propensity of the amino acid to being part of α-helices, β-sheets and turns, Pα, Pβ and Pτ, respectively. The Pα, Pβ and Pτ parameters of the 20 amino acids are showed in Fig. 2.

Chou and Fasman considered four consecutive amino acids as the core, so this study also gave priority to their proposition to study the information of four consecutive amino acids. Let's take the Pα feature as an example. For *i*-th amino acid in a protein sequence S, if it's Pα > 100, we set P*i* = 1. If Pα < 100, we set P*i* = 0, in this way, protein sequences can be translated into binary sequences. For example, sequence "S = TSPESDYARSNQPK(Ubi)EIESRVSDKETAS" can be translated to "S = 0,001,010,100,010, 111,100,101 11,010". For a total of 16 different situations

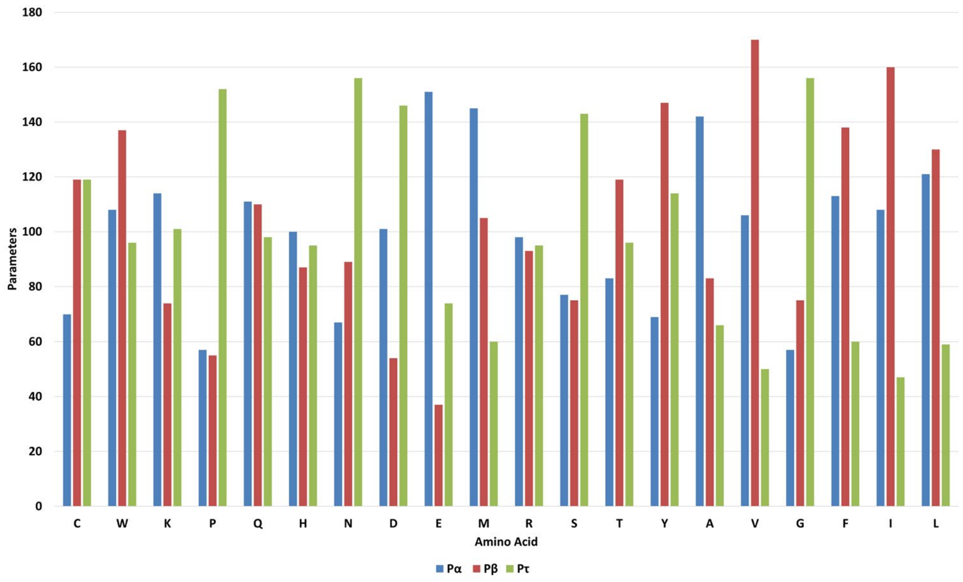


Fig. 2 The Pα, Pβ and Pτ parameters of the 20 Amino Acids. the amino acid to being part of α-helices (Pα) is indicated by blue, the amino acid to being part of β-sheets (Pβ) is indicated by red and the amino acid to being part of turns is indicated by green

for four consecutive amino acids, they are "0000", "0001", "0010"… "1111", and then calculate the probability of each of these sixteen situations in the sequence. Therefore, we can get the 16-dimensional digital features of each protein sequence. By the same way, we can get 32-dimensional features for protein sequence based on the Pβ and Pτ.

The hydrophobicity of amino acids is useful information for many researches relating to proteins. When we consider the hydrophobicity index, if the hydrophobicity index of amino acid Y is greater than 0, we set Y = 1, else we set Y = 0. By repeating the method of 4 consecutive amino acids in the previous paragraph, we can obtain 16-dimensional features for a sequence. The polarity of the amino acid side chains is important for protein stability from a biochemical perspective. So, we also take account of the polarity of amino acids. If the amino acid is polarity, we set 1, otherwise set to 0. Repeating the method in the previous paragraph, we can get 16-dimensional features for one sequence. Furthermore, the character X is replaced by 0.

Protein is formed by a large number of amino acid residues connected to each other. So, it is simply considering the characteristics of a single amino acid without considering the interaction between amino acid molecules may result in insufficient amount of information extracted. Therefore, when extracting the numerical features of amino acids, this paper considers three consecutive amino acid combinations, namely, two consecutive amino acid combinations, three consecutive amino acid combinations, and four consecutive amino acid combinations. Regardless of the case of continuous increase, only the three most basic cases are considered, and the control of computational complexity is also considered. For the combination of two consecutive amino acids, there are four cases of “00, 01, 10, 11”, similarly, there are eight situations for three consecutive amino acid combinations: "000, 001, 010, 011…111", and four consecutive combinations, there are sixteen cases: “0000, 0001, 0010 … 1111”. In this study, we obtained a total of five types of features, and each type of feature can obtain 28-dimensional feature vectors, so 140-dimensional feature vectors can be obtained.

Features selection

Feature selection can easily remove redundant and irrelevant features, which helps to further improve the performance of the classifier (Veredas et al. 2018). Based on the 140-dimensional candidate features obtained above, we use Boruta algorithm (Kursa and Rudnicki 2010; Chen et al. 2021) to further select the optimal feature subset. The Boruta algorithm is a wrapper-base feature selection method, which is constructed based on random forest (RF). Its goal is to find all relevant features useful for prediction, not to find the minimal-optimal feature.

The evaluation criterion *Rc* represents the prediction performance of the classifier with different ranking features. The formula is defined as follows:

$$R_C = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(ACC_i + SEN_i + SPE_i + AUC_i\right) \tag{8}$$

where *n* is the number of repetitions of fivefold cross-validation; *ACCi, SENi, SPEi and AUCi* represent the values of the accuracy, sensitivity, specificity and AUC score of the *i*-th fivefold cross-validation, respectively. We select the top-*k* ranked features with the highest *Rc* score.

Support vector machines

Based on the above characteristics, we constructed an SVM classification model. SVM is a machine learning method based on statistical learning theory (Chang and Lin 2011). The SVM is performed by the Support Vector Machine scikit-learn package for python to evaluate the performance of the model. In this study, we use the linear kernel function of SVM and evaluated the model through 5 cross-validation experiments. Finally, the overall performance is calculated by averaging the performance of the 5 subsets (at the fold level).

Radom Forest

The Random Forest (RF) algorithm (Statistics and Breiman 2001) builds large number of Decision Tree (Breiman et al. 1984) during training, and finally outputs the classification results. A Decision Tree is an attribute classifier that makes decisions based on the structure of the tree. Although the decision tree algorithm is easy to understand, its prediction performance is usually low. The random forest algorithm has the advantages of fast learning speed, high classification accuracy and the ability to evaluate the importance of variables, and is widely used in the fields of image processing and bioinformatics.

Fully connected neural Layer

Fully connected neural network (FCNN) (Zhang et al. 1998; Hsu et al. 1990) is a widely used artificial neural network, in which the neurons in the fully connected layer are completely connected with the neurons in the two layers before and after, but there is no connection between the neurons in the same fully connected layer. The input layer is mainly the encoding matrix; the hidden layer is to effectively integrate the local features; the output layer is mainly to output the prediction score.

Results and analysis

Optimal selection of features

The 140 candidate characteristics can be grouped into: the propensity of amino acids to become alpha helices (Pα), the propensity of amino acids to become beta sheets (Pβ), The tendency of amino acids to become turns (Pτ), hydrophobicity, polarity. We combine the Pα, Pβ and Pτ features as Combined1 (Pα + Pβ + Pτ), and combine hydrophobicity and polarity features as Combined2 (hydrophobicity + polarity). We compare the predictive performance of different feature categories. As listed in Table 1, the hydrophobicity features obtain the best performance among the five basic feature categories, with the highest ACC, MCC and AUC values of 0.800, 0.602 and 0.832, respectively. We also find that the novel feature combination (Combined2) performs much better than the combination of Chou-Fasman parameter features (Combined1). As expected, the combination of all the features (Pα + Pβ + Pτ + hydrophobicity + polarity) shows the highest performance. The results suggest that the three categories of features may be complementary, and their combination can help predict *Arabidopsis thaliana* ubiquitination sites.

Screening valuable information is an important step in constructing a classifier for predicting ubiquitination sites. In this study, we use the RF-based Bortua algorithm for feature selection. The Rc value is the highest when using the top 30 features, as shown in Fig. 3. The relative importance and rankings of the 30 best features are displayed on GiuHub. We found that the hydrophobicity of amino acid, the amino acid to being part of α-helices and β-sheets features dominate the top-10 list. By carrying out experiments, we found out, that if the features contain the hydrophobicity of “100” “101”, the Pα of “100” “101” and the Pβ of “1101” “1011”, thus the better prediction results can be got. By digging deeper and mapping back to the amino acids fragments, we found that the highest proportion of sequence fragments "100" and "101" based on hydrophobicity parameters are AKR and AEL in protein sequences; The highest proportion of sequence fragments of "100" and "101" based on Pα parameter are KRR and KRK. The highest proportion of sequence fragment "1101" and "1011" based on Pβ parameter is QFPV and TELL. These motifs may be involved in ubiquitination.

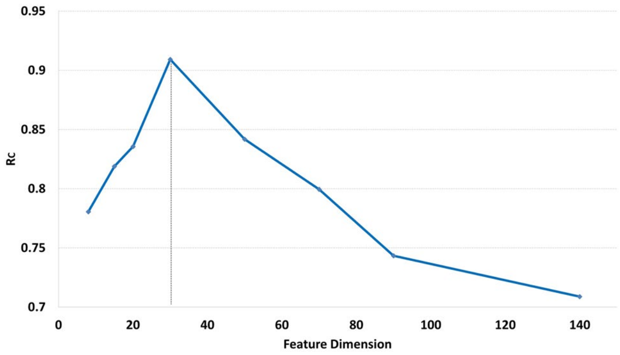


Fig. 3 The *Rc* values of top-*k* feature sets obtained by using Boruta and SVM. The abscissa indicates the number of feature dimensions, and the ordinate indicates the *Rc* value

In addition, we also compare the predictive performance of different feature categories in the optimal feature set. We evaluate the performance based on SVM with fivefold cross-validation on the dataset. As listed in Table 2, the hydrophobicity features obtain the best performance among the five basic feature categories, with the highest SEN and AUC values of 0.824 and 0.832, respectively. The following is Pα and Pβ. At the same time, it is found from Table 2 that the top 30-dimensional features we obtained based on the Boruta feature selection algorithm achieved the highest ACC and AUC values, which were 0.904 and 0.936, respectively. As a result, we select the top 30 features as the optimal feature set.

We also calculated the numbers of features of each feature type in the candidate full feature set and the selected optimal feature set, respectively, and redrawn the pie chart. The candidate complete feature set contains a total of 140-dimensional feature vectors, and the five categories of features contain 28-dimensional feature vectors respectively, which are allocated to the fan chart with a proportion of

Table 1 Performance comparison of different feature combinations of all features

Feature	ACC	PRE	SEN	SPE	MCC	AUC
Pα	0.650	0.660	0.647	0.653	0.300	0.682
Pβ	0.770	0.760	0.775	0.764	0.540	0.802
Pτ	0.680	0.640	0.695	0.667	0.361	0.715
Hydrophobicity	0.800	0.760	0.826	0.778	0.602	0.832
Polarity	0.670	0.612	0.697	0.648	0.344	0.697
Combined1	0.784	0.645	0.801	0.775	0.552	0.813
Combined2	0.792	0.780	0.796	0.788	0.584	0.826
All features	0.860	0.840	0.875	0.846	0.720	0.902

Table 2 Performance comparison of different feature combinations of top 30

Feature	ACC	PRE	SEN	SPE	MCC	AUC
Pα	0.640	0.667	0.632	0.647	0.280	0.654
Pβ	0.713	0.733	0.705	0.722	0.427	0.726
Pτ	0.613	0.620	0.612	0.614	0.227	0.642
Hydrophobicity	0.820	0.813	0.824	0.815	0.640	0.832
Polarity	0.516	0.526	0.516	0.517	0.033	0.554
Top 30	0.908	0.885	0.927	0.891	0.795	0.953

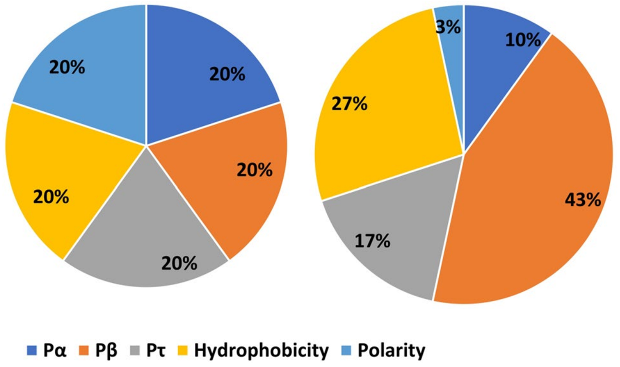


Fig. 4 The proportion of each feature in the feature set. the amino acid to being part of α-helices feature (Pα) is shown in Navy blue, and the amino acid to being part of β-sheets feature (Pβ) is shown in orange, and the amino acid to being part of turns (Pτ) feature is shown in gray, and hydrophilicity feature is shown in yellow and polarity feature is shown in light blue. The picture on the left shows the proportion of each feature in the full feature set, and the picture on the right shows the proportion of each feature in the optimal feature set

20%. The optimal feature collection consists of 30 dimensional feature vectors, and the Pα, Pβ, Pτ, hydrophilic and hydrophobic feature consist of 3-dimensional feature vectors, 13-dimensional feature vectors, 4-dimensional feature vectors, 8-dimensional feature vectors and 1-dimensional feature vectors respectively, accounting for 10%, 43%, 17%, 27% and 1%, respectively. As shown in Fig. 4, the Pα features with the highest proportion, and increased significantly in the optimal feature set (from 20 to 43%). the second is hydrophobicity features in the optimal feature set. All the results suggest that Pα and hydrophobicity features are more predictive than other features in determining *Arabidopsis thaliana* ubiquitination sites.

Model prediction performance

By predicting performance results on the optimal feature set, we found that support vector machines (SVM) can achieve better performance than random forest (RF) classifiers and deep-learning fully connected neural network (FCNN). So, PseAraUbi uses SVM to build the final model that contains the 30-dimensional optimal features. At the same time, five-fold cross-validation and independent testing are used to evaluate different models. In addition, AraUbiSite (Chen et al. 2019), ArabidopsisUb (Mosharaf et al. 2020) and CNN_Binary (Wang et al. 2021) are prediction tools for Arabidopsis ubiquitination sites. AraUbiSite uses two amino acid coding schemes based on protein sequence fragments, namely CKSAAP coding and binary coding methods, and uses machine learning methods to build models. Through comparative research, the predictor based on random forest is selected as the best predictor under the CKSAAP coding scheme. ArabidopsisUb also chose the predictor based on random forest as the best predictor under the CKSAAP coding scheme. Wang used the convolutional neural network to train the model, and used the binary code and the physical and chemical properties of the amino acid to code the amino acid in the sequence, and CNN_Binary model achieved better results. The above three tools for predicting Arabidopsis ubiquitination sites have a common trait that uses binary coding and K-spaced amino acid pair composition as input features to build models. while, in this work, we obtained new characteristics based on the Chou-Fasman parameter, combined with the amino acid hydrophobicity characteristics and polarity information. And, our model achieves better performance. The detail results are as follows.

Since AraUbiSite, ArabidopsisUb and CNN_Binary provide five-fold cross-validation and independent test prediction scores on the web server, and our method is test by using the same data set, which is feasible to directly compare with other methods. The other metrics for performance of the five methods in five-fold cross-validation and independent dataset are listed in Table 3 and 4, respectively. The "N/A" in the table indicates that the algorithm does not provide this performance value. We can see that SVM_PseAraUbi perform much better than others both in train dataset and independent dataset. The sensitivity values of SVM_PseAraUbi, RF_PseAraUbi and FCNN_PseAraUbi are larger than that of AraUbiSite and ArabidopsisUb, but their specificity values are smaller than that of AraUbiSite and ArabidopsisUb. The possible reason is that SVM_PseAraUbi, RF_PseAraUbi and FCNN_PseAraUbi predict more ubiquitination sites, which also increases the number of true positives and false positives. By comparing AUC values, we find that our method exhibits a competitive advantage.

Table 3 Prediction performance in five-fold cross-validation

Method	ACC	SEN	SPE	PRE	MCC	AUC
AraUbiSite	0.818	0.533	0.913	N/A	0.485	0.877
ArabidopsisUb	0.838	0.772	**0.915**	0.763	0.680	0.910
CNN_Binary	0.854	0.881	0.827	N/A	N/A	0.924
RF_ PseAraUbi	0.874	0.882	0.867	0.898	0.704	0.931
FCNN_PseAraUbi	0.883	0.891	0.877	0.915	0.710	0.940
SVM_PseAraUbi	**0.908**	**0.927**	0.891	**0.922**	**0.725**	**0.953**

In the Method column, the words in bold represent the experimental method we used. In addition, the other data in bold represent the best predicted performance values

Table 4 Prediction performance in independent test

Method	ACC	SEN	SPE	PRE	MCC	AUC
AraUbiSite	0.814	0.513	**0.914**	N/A	0.468	0.868
ArabidopsisUb	0.802	0.801	0.782	N/A	0.580	0.861
CNN_Binary	0.854	0.892	0.817	N/A	N/A	0.921
RF_ PseAraUbi	0.857	0.861	0.853	0.872	0.693	0.923
FCNN_PseAraUbi	0.872	0.881	0.863	0.895	0.706	0.930
SVM_PseAraUbi	**0.887**	**0.894**	0.879	**0.913**	**0.722**	**0.942**

In the Method column, the words in bold represent the experimental method we used. In addition, the other data in bold represent the best predicted performance values

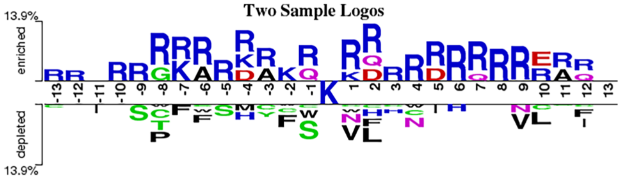


Fig. 5 The engagement of amino acids residues around the ubiquitination sites compared to non-ubiquitination sites is represented by Two-Sample Logos software (statistical t-test, p-value < 0.05)

Amino acid properties analysis

In this paper, the position-specific residue composition surrounding the ubiquitination sites was analyzed using the two-sample logo. The height of the residues in the X-axis was in proportion to the percentage of corresponding residue in the positive and negative samples. Form the Fig. 5 we can see Arg and Lys are enriched near ubiquitination sites, while Ser, Leu and Phe are depleted. Form the Fig. 5 we can see Arg and Lys are enriched near ubiquitination sites, while Ser, Leu and Phe are depleted. In the above sequence fragment analysis, we found that Lys and Arg also appeared more frequently than other amino acids. Mand studies have shown that although the replacement of positively charged amino acids with negatively charged or neutral amino acids maintains the binding and synergistic interaction of mutant proteins, these proteins will also have defects in other aspects (Kang et al. 2007; Hiller et al. 2020). And, positively charged amino acids are commonly used excipients for stabilizing therapeutic proteins in biopharmaceutical formulations. The positively charged side chains of amino acids play an important role in the mechanism of controlling their influence on protein stability (Platts et al. 2016). The activity of p53 as an inducible transcription factor depends on its rapid nuclear stabilization after stress and is regulated by ubiquitination. The author also found that direct binding of p53 to importin-alpha3 depends on the positive charge contributed by lysine residues 319–321 within nuclear localization signals I (Marchenko et al. 2010). Based on this, we hypothesized that positive charge enrichment at ubiquitin sites plays an important role in the physiological and biochemical processes of ubiquitin proteins in eukaryotic cells.

As a supplement, we analyzed the Chou-Fasman parameter, hydrophobicity and polar properties of residues around the Arabidopsis ubiquitination site, and compared the average values of ubiquitinated and non-ubiquitinated peptides, and Pα denotes alpha helix propensities, and Pβ denotes beta-strand propensities in Fig. 6. In general, residues in the ubiquitination peptides tend to form alpha helices, while residues in the non-ubiquitination peptides tend to form beta strands. And these ubiquitination peptides are less hydrophobic than non-ubiquitination peptides. Polarity represents amino acid polarity while non-ubiquitination peptides are less polarity than Ubiquitination peptides.

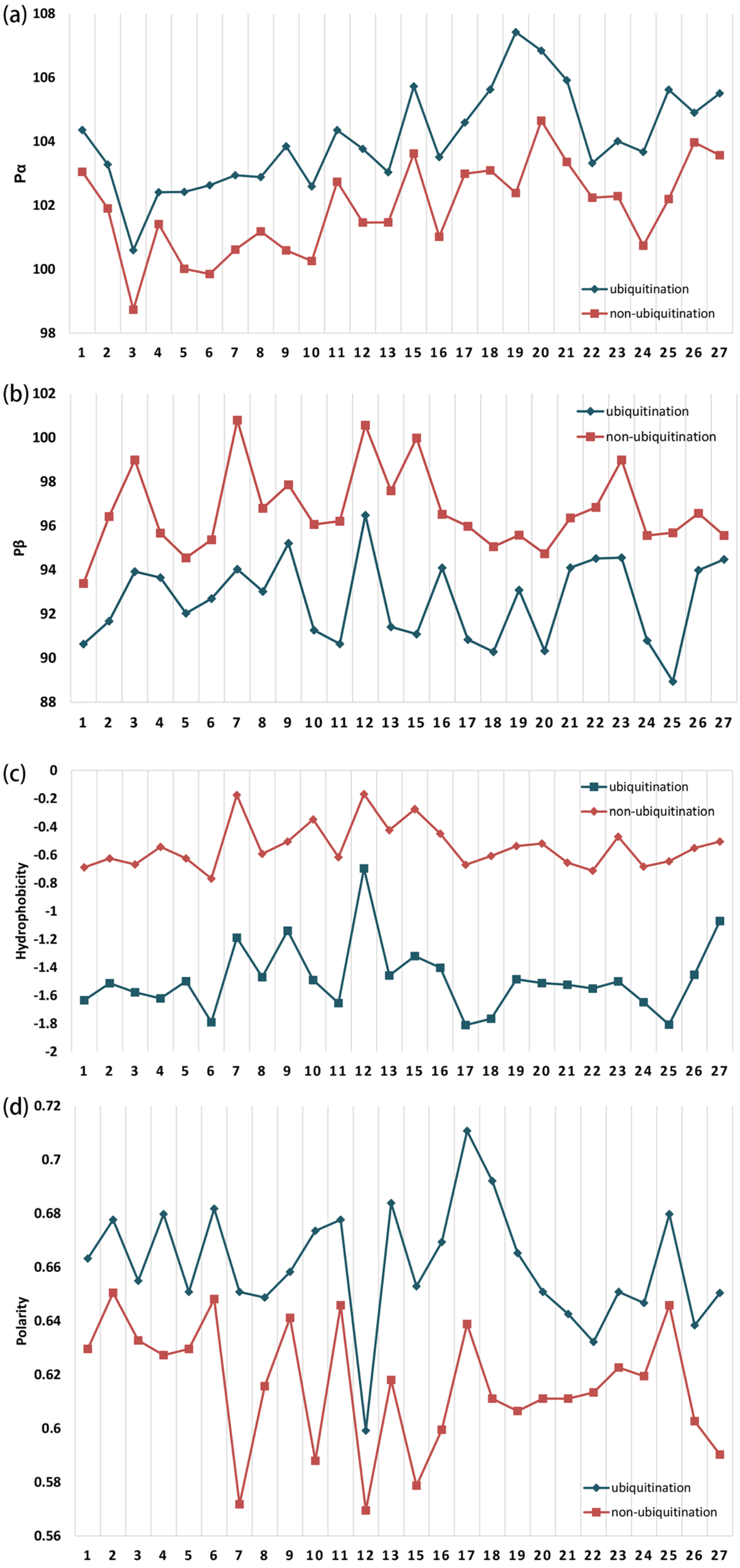
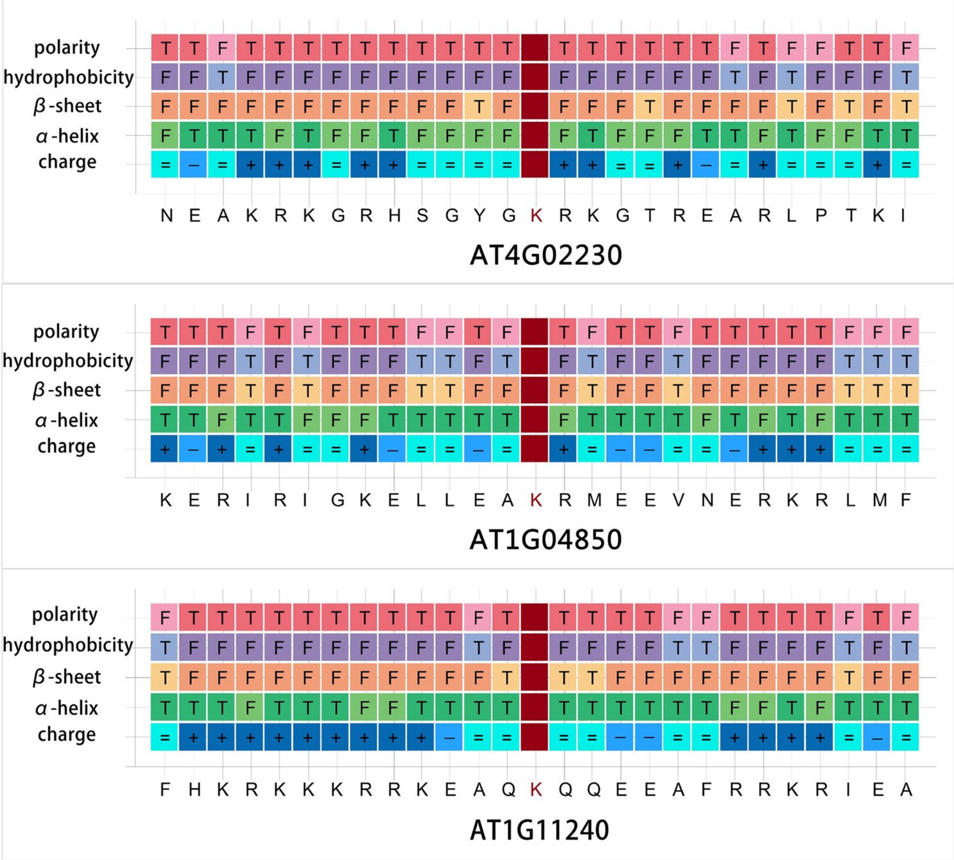


Fig. 6 Comparison of a) alpha helix, b) beta-strand propensities, c) hydrophobicity and d) polarity between ubiquitination and non-ubiquitination peptides. Red lines represent non-ubiquitination peptides, while blue lines denote ubiquitination peptides. Positions from 1 to 13 are the left positions of the ubiquitination site, and positions from 15 to 26 are the right positions of the ubiquitination sites



	N	E	A	K	R	K	G	R	H	S	G	Y	G	K	R	K	G	T	R	E	A	R	L	P	T	K	I
polarity	T	T	F	T	T	T	T	T	T	T	T	T	T		T	T	T	T	T	T	F	T	F	F	T	T	F
hydrophobicity	F	F	T	F	F	F	F	F	F	F	F	F	F		F	F	F	F	F	F	T	F	T	F	F	F	T
β-sheet	F	F	F	F	F	F	F	F	F	F	F	T	F		F	F	F	T	F	F	F	F	T	F	T	F	T
α-helix	F	T	T	T	F	T	F	F	T	F	F	F	F		F	T	F	F	F	T	T	F	T	F	F	T	T
charge	=	−	=	+	+	+	=	+	+	=	=	=	=		+	+	=	=	+	−	=	+	=	=	=	+	=

AT4G02230

	K	E	R	I	R	I	G	K	E	L	L	E	A	K	R	M	E	E	V	N	E	R	K	R	L	M	F
polarity	T	T	T	F	T	F	T	T	T	F	F	T	F		T	F	T	T	F	T	T	T	T	T	F	F	F
hydrophobicity	F	F	F	T	F	T	F	F	F	T	T	F	T		F	T	F	F	T	F	F	F	F	F	T	T	T
β-sheet	F	F	F	T	F	T	F	F	F	T	T	F	F		F	T	F	F	T	F	F	F	F	F	T	T	T
α-helix	T	T	F	T	T	F	F	F	T	T	T	T	T		F	T	T	T	T	F	T	F	T	F	T	T	T
charge	+	−	+	=	+	=	=	+	−	=	=	−	=		+	=	−	−	=	=	−	+	+	+	=	=	=

AT1G04850

	F	H	K	R	K	K	K	R	R	K	E	A	Q	K	Q	Q	E	E	A	F	R	R	K	R	I	E	A
polarity	F	T	T	T	T	T	T	T	T	T	T	F	T		T	T	T	T	F	F	T	T	T	T	F	T	F
hydrophobicity	T	F	F	F	F	F	F	F	F	F	F	T	F		F	F	F	F	T	T	F	F	F	F	T	F	T
β-sheet	T	F	F	F	F	F	F	F	F	F	F	F	T		T	T	F	F	F	F	F	F	F	F	T	F	F
α-helix	T	T	T	F	T	T	T	F	F	T	T	T	T		T	T	T	T	T	T	F	F	T	F	T	T	T
charge	=	+	+	+	+	+	+	+	+	+	−	=	=		=	=	−	−	=	=	+	+	+	+	=	−	=

AT1G11240

Fig. 7 Feature visualization of amino acids residues around the ubiquitination sites for gene AT1G11240, AT1G04850, AT4G02230. "+" indicates that the amino acid is positively charged, "-" indicates that the amino acid is negatively charged, and "=" indicates that the amino acid is neutral or nonpolar; "T" indicates that the amino acid belongs to this characteristic property, and "F" indicates the opposite. At the same time, we used unlike colors to represent different features

Case study

To further illustrate the effectiveness of our feature information for predicting ubiquitination site, we present one example that is analyzed by physic-chemical and structural features. As shown in the Fig. 7, the ubiquitination site residue K is located in the center and is indicated in red. It can be seen from the figure that the amino acids around the ubiquitination site residues have more positively charged residues than negatively charged residues. There are more residues with α-helix tendency, less residues with β-sheet tendency, less hydrophobic amino acids, and more polar amino acids. This is consistent with our previous analysis.

Conclusion

Accurate prediction of Arabidopsis ubiquitination sites is of great significance for understanding the mechanism of ubiquitination-related biological processes. In this work, we described a computational identification method, the PseAraUbi method, to predict ubiquitination sites. We integrate a variety of features, including the features based on the Chou-Fasman parameters, amino acid's hydrophobicity and polarity information. We also utilized the Boruta algorithm to select an optimal feature set, which is proved to be able to improve the prediction accuracy and reduce the risk of overfitting. At the same time, combined with support vector machine method to build the mode. The experiments results showed that our method significantly outperformed the other state-of-the-art approach on both benchmark and independent test dataset. Through the analysis of the physicochemical properties of residues surrounding the ubiquitination site, we found that compared with residues near non-ubiquitination sites, and the residues near ubiquitination sites are less hydrophobic, and prefer to form alpha helices and more polarity. We believe that PseAraUbi can be a useful tool for accurately identifying ubiquitination sites with the increasing of experimentally determined ubiquitination sites.

Acknowledgements This work was supported by the Natural Science Foundation of Henan province (Nos. 212300410367, 202300410102); the Science and Technology Research Key Project of Educational Department of Henan Province (Nos. 21A520023, 22A520030); the National Natural Science Foundation of China (Nos. 62072160, 62076089, 62072157); the Key Project of Science and Technology Department of Henan Province (No. 212102310381); the Educational Science Research Foundation of Henan Normal University (No. 2018JK19); the Teaching Reform Research and Practice Project of Henan Normal University (No. 201936); the Production and Learning Cooperation and Cooperative Education Project of Ministry of Education of China (No. 202102633006); the Key Project of Science and Technology Department of Xinxiang city (No .GG2021004) and National Project Cultivation Fund Project of Henan Normal University (No. 2020PL12).

Author contributions WW and YZ studies the ubiquitination features. WW also provided the experimental methods. WW, YZ, DL, HJZ, XFW and YZ performed additional analysis of the results. All authors read and approved the final manuscript.

Data availability The datasets list reported in this paper have been deposited in the supplementary information that are publicly accessible at https://www.ncbi.nlm.nih.gov/gene/.

References

Breiman LI, Friedman JH, Olshen RA, Stone C (1984) Classification and regression trees. Sta Probab Ser 40:358

Chang CC, Lin CJ (2011) Libsvm: a library for support vector machines. ACM Trans Intell Syst Technol 2:27. https://doi.org/10.1145/1961189.1961199

Chen Z, Zhou Y, Song JN, Zhang ZD (2013) Hcksaap_ubsite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. BBA Proteins Proteom 1834:1461–1467. https://doi.org/10.1016/j.bbapap.2013.04.006

Chen Z, Zhou Y, Zhang ZD, Song JN (2014) Towards more accurate prediction of ubiquitination sites: acomprehensive review of current methods, tools and features. Brief Bioinform 16:640–657. https://doi.org/10.1093/bib/bbu031

Chen JJ, Zhao JN, Yang SP, Chen Z, Zhang ZD (2019) Prediction of protein ubiquitination sites in *arabidopsis thaliana*. Curr Bioinform 14:614–620. https://doi.org/10.2174/1574893614666190311141647

Chen L, Li Z, Zeng T, Zhang YH, Li H, Huang T, Cai YD (2021) Predicting gene phenotype by multi-label multi-class model based on essential functional features. Mol Genet Genomics 296:905–918. https://doi.org/10.1007/s00438-021-01789-8

Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43:246–255. https://doi.org/10.1002/prot.1035

Chou PY, Fasman GD (1978) Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol 47:45–148

Fu LM, Niu BF, Zhu ZW, Wu S, Li WZ (2012) Cd-hit: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Gong M, Li ZP, Wan JN, Chen MJ, Wang H, Shang JJ, Zhou SC, Tan Q, Wang Y, Bao DP (2020) Chilling stress reduced protein translation by the ubiquitination of ribosomal proteins in volvariella volvacea. J Proteom 215:103668. https://doi.org/10.1016/j.jprot.2020.103668

He WY, Wei LY, Zou Q (2018) Research progress in protein post-translational modification site prediction. Brief Funct Genomics 18:220–229. https://doi.org/10.1093/bfgp/ely039

Herrmann J, Lerman LO, Lerman A (2007) Ubiquitin and ubiquitinlike proteins in protein regulation. Circ Res 100:1276–1291. https://doi.org/10.1161/01.res.0000264500.11888.f0

Hiller DA, Dunican BF, Nallur S, Li NS, Piccirilli JA, Strobel SA (2020) The positively charged active site of the bacterial toxin rele causes a large shift in the general base pka. Biochemistry 59:1665–1671. https://doi.org/10.1021/acs.biochem.9b01047

Hsu KY, Li HY, Psaltis D (1990) Holographic implementation of a fully connected neural network. Proc IEEE 78:1637–1645. https://doi.org/10.1109/5.58357

Huizen MV, Kikkert M (2019) The role of atypical ubiquitin chains in the regulation of the antiviral innate immune response. Front Cell Dev Biol 7:392. https://doi.org/10.3389/fcell.2019.00392

Ju Z, Wang SY (2018) Prediction of citrullination sites by incorporating k-spaced amino acid pairs into chou's general pseudo amino acid composition. Gene 664:78–83. https://doi.org/10.1016/j.gene.2018.04.055

Kang S, Han JS, Kim SH, Park JH, Hwang DS (2007) Aggregation of seqa protein requires positively charged amino acids in the hinge region. Biochem Biophys Res Commun 360:63–69. https://doi.org/10.1016/j.bbrc.2007.05.225

Kinoshita A, Seo M, Kamiya Y, Sawa S (2015) Mystery in genetics: pub4 gives a clue to the complex mechanism of clv signaling pathway in the shoot apical meristem. Plant Signal Behav 10:e1028707. https://doi.org/10.1080/15592324.2015.1028707

Kruijsbergen IV, Mulder MPC, Uckelmann M, Welsem TV, Widt JD, Spanjaard A, Jacobs H, El Oualid F, Ovaa H, Leeuwen FV (2020) Strategy for development of site-specific ubiquitin antibodies. Front Chem. https://doi.org/10.3389/fchem.2020.00111

Kumar VS, Vellaichamy A (2019) Sequence and structure based characterization of ubiquitination sites in human and yeast proteins using Chou's sample formulation. Proteins Struct Funct Bioinform 87:646–657. https://doi.org/10.1002/prot.25689

Kursa MB, Rudnicki WR (2010) Feature selection with the boruta package. J Stat Softw 36:1–13. https://doi.org/10.18637/jss.v036.i11

Li WZ, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Liu Y, Li A, Zhao XM, Wang MH (2020) Deeptl-ubi: a novel deep transfer learning method for effectively predicting ubiquitination sites of multiple species. Methods 192:103–111. https://doi.org/10.1016/j.ymeth.2020.08.003

Maheshwari R, Dhathathreyan A (2004) Investigation of surface properties of amino acids: polarity scale for amino acids as a means to predict surface exposed residues in films of proteins. J Colloid Interface Sci 277:79–83. https://doi.org/10.1016/j.jcis.2004.04.023

Marchenko ND, Hanel W, Li D, Becker K, Reich N, Moll UM (2010) Stress-mediated nuclear stabilization of p53 is regulated by ubiquitination and importin-alpha3 binding. Cell Death Differ 17:255–267. https://doi.org/10.1038/cdd.2009.173

Matsui D, Nakano S, Dadashipour M, Asano Y (2017) Rational identification of aggregation hotspots based on secondary structure and amino acid hydrophobicity. Sci Rep 7:9558. https://doi.org/10.1038/s41598-017-09749-2

Mattern M, Sutherland J, Kadimisetty K, Barrio R, Rodriguez MS (2019) Using ubiquitin binders to decipher the ubiquitin code. Trends Biochem Sci 44:559–615. https://doi.org/10.1016/j.tibs.2019.01.011

Mosharaf MP, Hassan MM, Ahmed FF, Khatun MS, Moni MA, Mollah MNH (2020) Computational prediction of protein ubiquitination sites mapping on *arabidopsis thaliana*. Comput Biol Chem 85:107238. https://doi.org/10.1016/j.compbiolchem.2020.107238

Mulder MPC, Witting KF, Ovaa H (2019) Cracking the ubiquitin code: the ubiquitin toolbox. Curr Issues Mol Biol 37:1–20. https://doi.org/10.1021/acs.biochem.9b01047

Naseer S, Ali RF, Muneer A, Fati SM (2021) Iamidev-deep: valine amidation site prediction in proteins using deep learning and pseudo amino acid compositions. Symmetry 13:560. https://doi.org/10.3390/sym13040560

Nobuhiro N (2018) Ubiquitin system. Int J Mol Sci 19:1080. https://doi.org/10.3390/ijms19041080

Pan WB, Wu YR, Xie Q (2019) Regulation of ubiquitination is central to the phosphate starvation response. Trends Plant Sci 24:755–769. https://doi.org/10.1016/j.tplants.2019.05.002

Platts L, Darby SJ, Falconer RJ (2016) Control of globular protein thermal stability in aqueous formulations by the positively charged amino acid excipients. J Pharm Sci 105:3532–3536. https://doi.org/10.1016/j.xphs.2016.09.013

Statistics LB, Breiman L (2001) Random forests. Mach Learn 45:5–32

Tracz M, Górniak L, Szczepaniak A, Bialek W (2021) E3 ubiquitin ligase SPL2 is a lanthanide-binding protein. Int J Mol Sci 22:5712. https://doi.org/10.3390/ijms22115712

Veredas FJ, Urda D, Subirats JL, Cantón FR, Aledo JC (2018) Combining feature engineering and feature selection to improve the prediction of methionine oxidation sites in proteins. Neural Comput Appl 32:323–334. https://doi.org/10.1007/s00521-018-3655-2

Wagner SA, Beli P, Weinert BT, Nielsen ML, Choudhary C (2011) A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles. Mol Cell Proteom 10(M111):013284. https://doi.org/10.1074/mcp.M111.013284

Walton A, Stes E, Cybulski N, Bel MV, Iñigo S, Durand AN, Timmerman V, Heyman J, Pauwels L, Veylder LD, Goossens A, Smet ID, Coppens F, Goormachtig S, Gevaert K (2016) It's Time for Some "Site"-Seeing: Novel Tools to Monitor the Ubiquitin Landscape in *Arabidopsis thaliana*. Plant Cell 28:6–16. https://doi.org/10.1105/tpc.15.00878

Wang L, Zhang R (2019) Towards computational models of identifying protein ubiquitination sites. Curr Drug Targets 20:565–578. https://doi.org/10.2174/1389450119666180924150202

Wang JR, Huang WL, Tsai MJ, Hsu KT, Huang HL, Ho SY (2017) Esaubisite: accurate prediction of human ubiquitination sites by identifying a set of effective negatives. Bioinformatics 33:661–668. https://doi.org/10.1093/bioinformatics/btw701

Wang K, Lu Q, Li XO, Li SM, Wang YM, Xu XS, He CY, Fang L (2019) Profiling of ubiquitination modification sites in talin in colorectal carcinoma by mass spectrometry. Chem Res Chin 35:377–381. https://doi.org/10.1007/s40242-019-8377-1

Wang XF, Yan RX, Chen YZ, Wang YJ (2021) Computational identification of ubiquitination sites in *arabidopsis thaliana* using convolutional neural networks. Plant Mol Biol 105:601–610. https://doi.org/10.1007/s11103-020-01112-w

Xiao X, Chou KC (2007) Digital coding of amino acids based on hydrophobic index. Protein Pept Lett 14:871–875. https://doi.org/10.2174/092986607782110293

Xu G, Paige JS, Jaffrey SR (2010) Global analysis of lysine ubiquitination by ubiquitin remnant immune affinity profiling. Nat Biotechnol 28:868–873. https://doi.org/10.1038/nbt.1654

Yau RG, Doerner K, Castellanos ER, Haakonsen DL, Werner A, Wang N, Yang XW, Martin NM, Mastumoto ML, Dixit VM, Rape M (2017) Assembly and function of heterotypic ubiquitin chains in cell-cycle and protein quality control. Cell 171:918–933. https://doi.org/10.1016/j.cell.2017.09.040

Zhang Y, Zeng LR (2020) Crosstalk between ubiquitination and other post-translational protein modifications in plant immunity. Plant Commun 1:13. https://doi.org/10.1016/j.xplc.2020.100041

Zhang GQ, Patuwo BE, Hu MY (1998) Forecasting with artificial neural networks: the state of the art. Int J Forecast 14:35–62

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17:729–738. https://doi.org/10.1023/a:1020713915365

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.