Article

iPReditor-CMG: improving predictive RNA editor for crop mitochondrial genomes using genomic sequence features and optimal support vector machine

Sidong Qin [1,2,†], Yanjun Fan[1,2,3,†], Shengnan Hu[1,2], Yongqiang Wang[1,2], Ziqi Wang[1,2], Yixiang Cao[1,2], Qiyuan Liu[4], Siqiao Tan[5,*], Zhijun Dai[1,2,*], Wei Zhou [1,2,*]

[1] Hunan Provincial Engineering Technology Research Center for Agricultural Big Data Analysis Decision-Making, Hunan Agricultural University, Changsha, 410128, China; qinsidong@stu.hunau.edu.cn (S.Q.); 514966019@qq.com (Y.F.); hushengnan1996@163.com (S.H.); wyqwyr666888@163.com (Y.W.); wangshiyixiaosheng@gmail.com (Z.W.); caoyixiang0901@gmail.com (Y.C.);
[2] Hunan Provincial Key Laboratory for Biology and Control of Plant Diseases and Insect Pests, Hunan Agricultural University, Changsha, 410128, China;
[3] Shanxi Province Jincheng City Gardening Service Center, Shanxi, 048000, China;
[4] Key Laboratory of Crop Physiology, Ecology and Genetic Breeding, Ministry of Education, College of Agronomy, Jiangxi Agricultural University, Nanchang, China; qiyuanl@126.com;
[5] College of Information Science and Technology, Hunan Agricultural University, Changsha 410128, China;
* Correspondence: tsq@hunau.edu.cn (S.T.); daizhijun@hunau.edu.cn (Z.D.); mengrzhou@163.com (W.Z.);
† These authors contributed equally to this work.

Abstract: Cytosine (C) to uracil (U) RNA editing is one of the most important post-transcriptional processes, however exploring C-to-U editing events efficiently within the crop mitochondrial genome remains a challenge. An improving predictive RNA editor for crop mitochondrial genomes, iPReditor-CMG, was proposed, which was based on SVM, three common crop mitochondrial genomes and self-sequenced tobacco mitochondrial ATPase. After multi-combination feature extracting, high-dimension feature screening and multi-test independent predicting, the results showed that the average accuracy of intraspecific prediction was 0.85, and the highest value even up to 0.91, which outperformed the previous reference models. While the prediction accuracies were 0.78 between dicotyledons and no more than 0.56 between dicotyledons and monocotyledons, implying a possible similarity in C-to-U editing mechanisms among close relatives. The best model was finally identified with an independent test accuracy of 0.91 and an area under the curve of 0.88, and further suggested that five unreported feature sequences TGACA, ACAAC, GTAGA, CCGTT and TAACA were closely associated with the editing phenomenon. Multiple evaluation findings supported that the iPReditor-CMG could be effectively applied to predict crop mitochondrial editing sites, which may contribute to insight into their recognition mechanisms and even other post-transcriptional events in crop mitochondria.

Keywords: iPReditor-CMG; RNA editing site; Mitochondrial genomes; genomic sequence feature; support vector machine

1. Introduction

RNA editing is an important post-transcriptional regulatory event that occurs through nucleotide substitutions, insertions or deletions. RNA editing events occur frequently and may directly affect protein translation and hydrophobicity, among others. Knie *et al* put forward that U-to-C editing appeared in hornworts, some lycophytes, and ferns [1], but the most common editing types were A-to-I editing in vertebrates and C-to-U editing in mitochondria and chloroplasts of higher plants [2].

Mitochondria is ideal for studying important functions in crops [3]. Discovery and recognition of RNA editing sites is the premise for understanding the mitochondrial-related biological functions [4]. Therefore, the recognition of mitochondrial editing sites has received high attention [5-10]. Traditional methods for editing site recognition are time-consuming and labor-intensive [11]. In the era of big data, a large number of attempts have been made to provide more efficient solutions for high-precision theoretical prediction of RNA editing [12-19]. RDDpred distinguished misjudged RNA editing in the human genome, greatly facilitating the application of machine learning in identifying editing sites [12]. Sun *et al*. predicted 203,202 editing sites in the human genome through machine learning, of which only 9% were reported [13]. In addition, Chen *et al*. proposed a predictor called iRNA-AI by incorporating the chemical properties of nucleotides and their sliding occurrence density distribution along the RNA sequence into the general form of pseudonucleotide composition (PseKNC) [14]. Cummings & Myers [15], Mower [16], Thompson & Gopal [17]and Du *et al*. [18] all provided effective prediction models for plant mitochondrial editing sites based on machine learning methods, but their algorithms were relatively complex and profound. Lenz & Knoop [19] provided an online prediction platform containing 19 plant mitochondrial genomes and 13 plant chloroplast genomes, which could effectively predict RNA editing sites through homologous alignment.

All the above results confirm the necessity and feasibility of using appropriate machine learning methods to construct editing site prediction models. SVM is one of the most important learning machines based on statistical learning theory [20] and has the advantage of simplicity and convenience for binary classification problems [21-24]. Based on structural risk minimization rather than empirical risk minimization, SVM not only helps to solve certain problems such as small samples, nonlinear, dimensional disasters, and local minimum problems, but also helps to provide powerful generalization [25]. Mitochondrial genomes can provide the most comprehensive information of editing sites, whereas the benchmark datasets of RNA editing sites in crop mitochondrial genomes are relatively limited and appropriate for modeling with SVM. In addition, the editing sites are mainly determined by their nearby sequences [13], leading to the challenge of capturing the neighboring sequence information efficiently. To develop the simpler and better theoretical prediction models to effectively investigate mitochondrion-related editing sites in crops, this manuscript used several different methods to capture mitochondrial genome information of three commonly used crops, optimized feature combinations and training-test set ratios, and then removed redundant information and screened the optimal model by independent test results of SVM. More importantly, while previous studies mainly performed intraspecific modeling predictions, our experiments further explored the reliability of model interspecific predictions based on benchmark and self-sequenced datasets to provide reliable support for crop mitochondrial editing sites surveys.

2. Materials and methods

2.1. Materials collection

Ara (*Arabidopsis thaliana*, accession number: Y08501), Bra (*Brassica napus*, accession number: AP006444) and Ory (*Oryza sativa* Japonica Group, accession number: BA000029) were widely used for studying mitochondrial RNA editing sites [15,17-18], so their mitochondrial whole genome sequences and their corresponding editing site annotations were downloaded from the NCBI database. To assess the reliability of our model, six tobacco lines SZY90 and MZY90 (sterile line of zhongyan90 and their maintainer line), SYY85 and MYY85 (sterile line of yunyan85 and their maintainer line), SK326 and MK326 (sterile line of K326 and their maintainer line), were further sampled, mitochondrial ATPase DNA and RNA were extracted from flower buds (in triplicate) using CTAB method and Easy-Pure RNA Kit, and then target genes (*atp9*, *atp6*, *orf25*, *orfB* and *atp1*) were amplified and sequenced to identify the editing sites.

2.2 Positive and negative samples composition

For the mitochondrial editing site sequences of the four species, the positive samples were all centered on editing site C, and then 250 bases were taken before and after the center according to the optimization results of Du et al. [18], making the sample sequence length 501 bp (see attachment 1 for extraction code of the positive samples). Since the negative samples in the literature [15] were 41 bp in length and not C-centered, these samples were first mapped to the corresponding mitochondrial whole genomes and then expanded to 501 bp with C as the midpoint (see attachment 2 for extraction code of the negative samples). The ratio of positive to negative samples was close to 1:1.

2.3 Training-test-validation partitions

All samples in the mitochondrial editing site dataset from each species of Ara, Bra and Ory were randomly sorted using MATLAB software, and then the training and test sets were divided in five ratios of 5:5, 6:4, 7:3, 8:2 and 9:1. Meanwhile, all samples derived from tobacco mitochondrial ATPase were used as the validation set.

2.4 Feature extraction

All the features were shown in Table 1. T-features were extracted using a program written in-house by referring to the design of Du *et al*. [18], whose composition could be denoted as a 64-dimensional vector (see attachment 3 for extraction code). Additionally, P, M and A-features were extracted by using the program previously developed in the laboratory [26]. Where P-features reflected single base difference between positive and negative samples; M-features reflected sequence features based on different scales (generally between 1 and 5), and the maximum scale was chosen in this manuscript to collect more comprehensive feature information; while A-features specifically reflected the correlation information between two different editing sites.

Table 1 15 groups of sequence features.

Features	Description	Features	Description
T	Triplet features	A+M	Combined A and M-features
P	Statistical difference table features	T+M	Combined T and M-features
M	Multiscale component features	P+A+T	Combined P, A and T-features
A	Multiscale correlation features	P+A+M	Combined P, A and M-features
P+A	Combined P and A-features	P+T+M	Combined P, T and M-features
P+T	Combined P and T-features	A+T+M	Combined A, T and M-features
P+M	Combined P and M-features	P+A+T+M	Combined P, A, T and M-features
A+T	Combined A and T-features		

2.5 Algorithm design and assessment

The LIBSVM software is a concrete implementation of SVM, so it (a subroutine of LIBSVM) was employed to build the classifier. In this case, the radial basis function was selected as the kernel function and the parameters were optimized using grid.py in Python. The experiments were designed through the following steps: 1) obtaining, classifying datasets and extracting sequence features; 2) evaluating different feature groups and different training-test set ratios; 3) evaluating the iPReditor-CMG model based on different training-test set ratios and intraspecific test sets after filtering features using the high-dimensional descriptors selection nonlinearly (HDSN) method; 4) evaluating the iPReditor-CMG model by interspecific datasets; 5) evaluating the iPReditor-CMG model based on the validation sets integratedly and determining the optimal models; and 6) establishing an interpretability system for the best models. Sensitivity (*Sn*) (formula 1), specificity

(Sp) (formula 2) and accuracy (ACC) (formula 3) values were applied to assess the predictive power of all models, and the statistical differences of all assessed values were tested using Duncan's multiple range test (DMRT).

$$Sn = \frac{TP}{TP+FN} \times 100\% \qquad \text{(formula 1)}$$

$$Sp = \frac{TN}{TN+FP} \times 100\% \qquad \text{(formula 2)}$$

$$ACC = \frac{TP+T}{TP+TN+FP+F} \times 100\% \qquad \text{(formula 3)}$$

TP: True positive; FN: False negative; TN: True negative; FP: False positive

3. Results and discussion

3.1 Dataset description extraction of four different crops

3.1.1 Sources and composition of the training and test sets

Three mitochondrial genome sequences containing editing site information for Ara, Bra and Ory were downloaded from NCBI, with lengths of 366,924 bp, 221,853 bp and 490,520 bp, respectively. According to the relevant references and the editing sites extraction principles of this manuscript, 454 C-to-U editing sites of Ara, 423 sites of Bra and 485 sites of Ory were extracted as the positive samples by self-scripting procedure in attachment 1. Then, 439 non-editing sites of Ara, 399 sites of Bra and 531 sites of Ory were extracted as negative samples by the self-programmed procedure in attachment 2. The resulting training and test sets for the three datasets of Ara, Bra and Ory were obtained.

3.1.2 Sources and composition of the validation set

ATPase DNA and RNA were extracted from buds of six tobacco lines according to the general method (replicated three times). PCR results for DNA and complementary DNA (cDNA) of *atp9*, *atp6*, *orf25*, *orfB* and *atp1* were showed in attachment 4, and a total of 180 PCR results were sequenced. The five target genes of tobacco were sequenced at 225 bp, 1,188 bp, 597 bp, 471 bp and 1,530 bp, corresponding to cDNA templates with PCR products of 338 bp, 1,300 bp, 837 bp, 604 bp and 1,600 bp in length, which were similar to the above DNA bands. The alignment results showed that both the amplified DNA sequences and cDNA sequences were consistent with the GenBank references, and the result supported that the amplified products were derived from the target genes.

RNA editing sites for the five target genes were confirmed by aligning DNA and their cDNA sequences, and all RNA editing types were found to be from C to U(T) (Table 2). There were ten, six, ten, four and six editing sites for *atp9*, *atp6*, *orf25*, *orfB* and *atp1* genes in tobacco maintainer lines and even fewer in CMS lines, so a maximum of 36 positive samples were collected as the validation set in this study. Then the C-centered negative samples were searched according to attachment 2 and a total of 33 were matched. Thus, the total number of samples from the validation set of tobacco mitochondrial ATPase reached 69.

doi:10.20944/preprints202110.0332.v1

Table 2 Edited sites of the ATPase from tobacco.

Genes	Sites	Edited bases changes	Genes	Sites	Edited bases changes	Genes	Sites	Edited bases changes
		M/S			M/S			M/S
atp9	20,50,212	T$\underline{C}$A→T$\underline{T}$A		658	$\underline{C}$GC→$\underline{T}$GC/-		416	A$\underline{C}$T→A$\underline{T}$T
	81	GT$\underline{C}$→GT$\underline{T}$/-		683	T$\underline{C}$G→T$\underline{T}$G/-	*orfB*	47	T$\underline{C}$A→T$\underline{T}$A
	82	$\underline{C}$TT→$\underline{T}$TT		1100	T$\underline{C}$T→T$\underline{T}$T/-		58	$\underline{C}$TC→$\underline{T}$TC
	90	TC$\underline{C}$→TC$\underline{T}$	*orf25*	59	T$\underline{C}$T→T$\underline{T}$T		76	$\underline{C}$CT→$\underline{T}$CT
	92,182	T$\underline{C}$G→T$\underline{T}$G		71,89,395	T$\underline{C}$A→T$\underline{T}$A		443	C$\underline{C}$A→C$\underline{T}$A
	191	C$\underline{C}$A→C$\underline{T}$A/-		215	T$\underline{C}$G→T$\underline{T}$G	*atp1*	1039	$\underline{C}$CC→$\underline{T}$CC
	223	$\underline{C}$AA→$\underline{T}$AA		227	C$\underline{C}$C→C$\underline{T}$C		1178	T$\underline{C}$A→T$\underline{T}$A
atp6	466	$\underline{C}$CA→$\underline{T}$CA/-		248	C$\underline{C}$T→C$\underline{T}$T		1216	$\underline{C}$TT→$\underline{T}$TT
	545	T$\underline{C}$A→T$\underline{T}$A/-		251	C$\underline{C}$G→C$\underline{T}$G		1292	C$\underline{C}$G→C$\underline{T}$G
	602	C$\underline{C}$G→C$\underline{T}$G/-		407	C$\underline{C}$A→C$\underline{T}$A		1415,1490	C$\underline{C}$A→C$\underline{T}$A/-

Notes: editing sites were underlined, and - implied there were no editing sites in the sterile lines.

3.1.3 Description extraction of four datasets

Description extraction was performed on all samples, of which there were 12 P-features, 64 T-features, 1024 M-features and 25 A-features. It should be especially noted that there were only 8 P-features of Ara affected by the threshold.

3.2 Dataset classification

All positive and negative samples of the mitochondrial genomic datasets of Ara, Bra and Ory were randomly arranged using MATLAB software, and then the training and test sets were divided in five ratios of 5:5, 6:4, 7:3, 8:2 and 9:1. Meanwhile, samples from tobacco mitochondrial ATPase all consituted the validation set. The detailed classification of all datasets was listed in Table 3.

Table 3 Dataset profiles for the four crops

		Positive samples					Negative samples					Total samples				
		5:5	6:4	7:3	8:2	9:1	5:5	6:4	7:3	8:2	9:1	5:5	6:4	7:3	8:2	9:1
Ara	Training sets	226	271	316	352	400	220	264	309	362	403	446	535	625	714	803
	Test sets	228	183	138	102	54	219	175	130	77	36	447	358	268	179	90
Bra	Training sets	201	235	283	331	375	210	258	292	326	364	411	493	575	657	739
	Test sets	222	188	140	92	48	189	141	107	73	35	411	329	247	165	83
Ory	Training sets	235	286	333	386	430	273	323	378	426	484	508	609	711	812	914
	Test sets	250	199	152	99	55	258	208	153	105	47	508	407	305	204	102
Tobacco	Validation sets			36					33					69		

3.3 Evaluation of different feature groups and different training-test set ratios

It should be noted that it was impossible to know which ratio was the best when making a priori statistical difference table, and this manuscript could only choose the common ratio of 6:4 between the training and test sets for feature analysis. To understand the effect of 15 groups of features on editing site recognition, the SVM results of the independent tests were shown in Figure 1A. Since the difference in feature accuracy between Ara and Ory was evident, it is more meaningful to evaluate using the mean value, whereby the most features combination (P+A+T+M) was known to be the best. After determining the best feature group, five different training-test set ratios were assessed, and the results in Figure 1B revealed that the prediction accuracy improved to different degrees as the training set ratio increased, with 9:1 being the best.

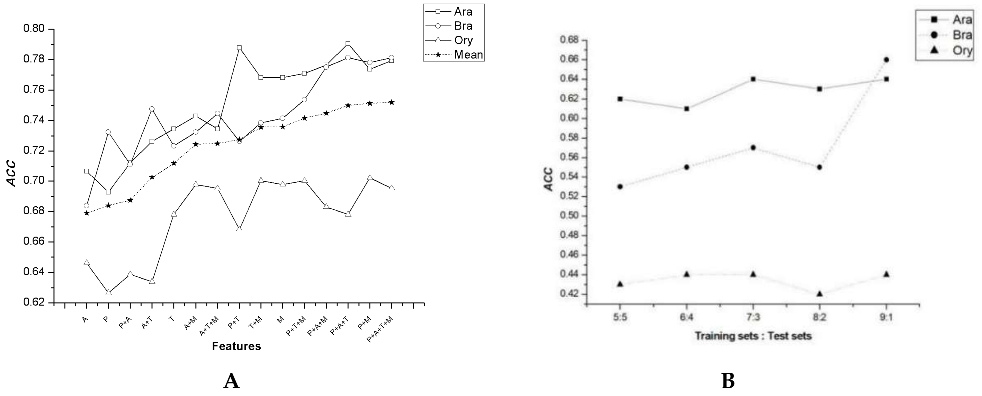


Figure 1 Comparison of classification results of seven groups of features (**A**) and five ratios of training-test sets (**B**)

3.4 Evaluating the models after screening by HDSN method

3.4.1 Evaluating the five ratios training-test sets

The total number of features in the optimal group amounted to 1125 (with 1121 for Ara), but the redundant descriptors needed to be removed for more efficient modeling. High-dimensional feature screening was performed with the HDSN method on all proportions of the training sets for the three species, where each training set was screened 20 times. The results of a total of 300 independent tests of our novel iPReditor-CMG method showed (Figure 2 and Attachment 6) that 1) in the Ara and Ory datasets, the mean values of the three evaluation indexes (*ACC*, *Sn* and *Sp*) were significantly highest and even highly significant when the training set accounted for 90% of the total dataset, and 2) in the Bra dataset, the highest mean values were found when the training set was 80%. It was remarkable that *Sn* (0.72±0.05) and *Sp* (0.82±0.03) were not significantly lower than the corresponding highest values (0.73±0.05 and 0.84±0.04) when the training set was 90%, while *ACC* (0.78±0.02), although significantly lower than the highest value (0.79±0.03), was still the second highest value.

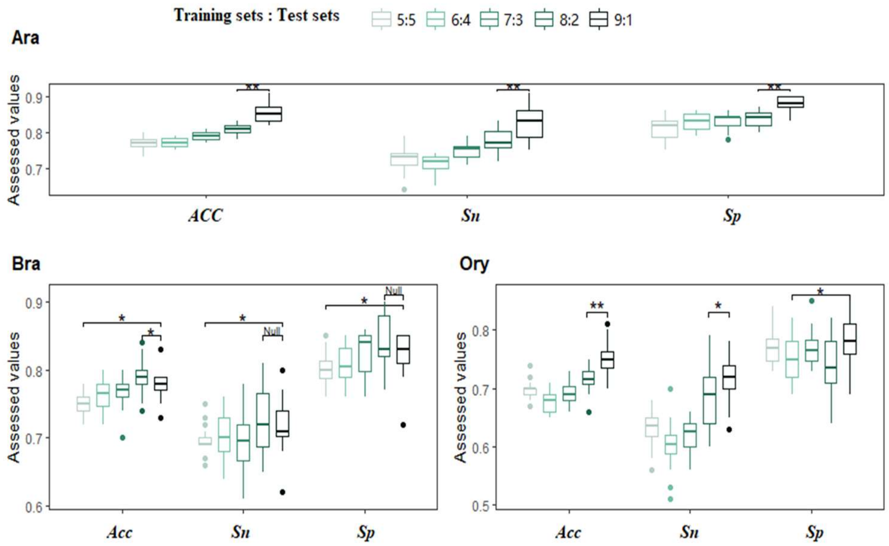


Figure 2 Comparison results of the five proportional training-test sets after screening by HDSN method. ("**" and "*" were indicated by DMRT for highly significant differences (p < 0.01) and significant differences (p < 0.05), respectively. Rules for marking the significant differences in the evaluation values of each group: if the mean value of 9:1 was the maximum, it was compared with those of the other four ratios from high to low, marking only the first significant difference; if the mean value of 9:1 was not the maximum, it was compared with the lower mean values from high to low, and with the higher mean values from low to high, again marking only the first significant difference.)

3.4.2 Validity evaluation of the iPReditor-CMG model based on reference models

To evaluate the validity of our iPReditor-CMG models, five reference models, namely tree-based statistical model (TSM) [15], random forest (RF) [15], predictive RNA editor for plant mitochondrial genes (PREP-Mt) [16], RNA editing site prediction by genetic algorithm learning (REGAL) [17], and Du-SVM [18], were selected for comparison. The two models, TSM and RF, were relatively simple and performed moderately well in predicting which C would be edited to U, providing the first quantitative prediction models for RNA editing sites in plant mitochondria. Previous analyses showed that the identity of the nucleotide -1 to the edited C and the estimated folding free energy of the 41 nt region surrounding the edited C were the most important variables in distinguishing most editing sites, but also found that the information individually was not sufficient to make highly accurate predictions [15]. PREP-Mt could recognize RNA editing sites in plant mitochondrial coding genes [16]. REGAL could optimize genetic algorithm-based variables by setting a weight when C was most likely to be edited in a given sequence, and then using the optimized weights and scoring functions to score each C in the test set and determine whether a threshold for whether it was likely to be edited [17]. Subsequently, Du et al. developed the Du-SVM model based on SVM combined with a triplet feature extraction method to further optimize the SVM results, which was the first time that only nucleic acid sequence features were used to predict the editing sites [18]. In our study, the novel prediction model for crop mitochondrial genome editing sites, iPReditor-CMG, was constructed by combining the nonlinear modeling method SVM, the multiscale feature extraction methods and the nonlinear feature screening method HDSN, and the results of our independent test analysis showed that the new model outperformed any other reference models with an average accuracy of 0.85 (Table 4). The iPReditor-CMG model both minimized redundant features and maintained high accuracy, and the model simplified the prediction process by extracting DNA sequence features. In summary, the novel models we developed could provide more accurate predictions for C-to-U editing sites with better generalization ability .

Table 4 Comparison of ACC values of different prediction methods

	TSM [15]	RF [15]	REGAL [17]	PREP-Mt [16]	Du-SVM [18]	SVC- HDSN
Ara	0.71	0.74	0.81	0.82	0.85	0.91
Bra	0.69	0.77	0.77	0.87	0.85	0.84
Ory	0.71	0.72	0.75	0.83	0.83	0.81
Mean	0.70	0.74	0.78	0.84	0.84	0.85

3.4.3 Evaluation of the iPReditor-CMG model based on interspecific validation

The generalization ability of a model refers to the predictive ability of the model for unseen data, therefore a better prediction model should have a high generalization ability [27]. To further investigate the effectiveness of our iPReditor-CMG model in predicting mitochondrial genome editing sites in different crops, we performed interspecific prediction for three species, Ara, Bra and Ory. After screening by the previous HDSN method, 100 training sets had been obtained for each species. Since species variability was present in P-feature positions, all training sets could only retain A, T and M features for interspecific validation; while the test set samples consisted of all samples of each species with features referenced to their corresponding training set features. A total of 600 interspecific prediction results revealed (Table 5) that Ara and Bra were predicted independently of each other with an accuracy of up to 0.78, while the accuracy involving Ory did not exceed 0.56.

Table 5 Comparison of interspecific independent prediction results of iPReditor-CMG

Training set		Test set		ACC_{max}	ACC_{mean}	Training set		Test set		ACC_{max}	ACC_{mean}
Species	Number	Species	Number			Species	Number	Species	Number		
Ara	100	Bra	100	0.78	0.72±0.05	Ara+Bra	5	Ory	5	0.58	0.55±0.02
		Ory	100	0.54	0.53±0.01	Ara+Ory	5	Bra	5	0.75	0.71±0.03
Bra	100	Ara	100	0.78	0.72±0.06	Bra+Ory	5	Ara	5	0.77	0.74±0.02
		Ory	100	0.55	0.53±0.01						
Ory	100	Ara	100	0.56	0.50±0.02						
		Bra	100	0.55	0.50±0.02						

The original dataset after feature extraction was further combined in pairs according to species, and the training set was obtained after removing redundant features using HDSN method (repeated five times), and the corresponding test set was identified from the remaining another species based on the screening results. The results dispalyed (Table 5) that the accuracy of independent predictions was poor when Ory alone was used as the test set, but only slightly decreased when Ory was combined with other species. The overall trend was consistent with the interspecific results between individual species described above, which may be attributed to the fact that Ara and Bra originated from the same dicotyledons, whereas Ory being a monocotyledon. It is hypothesized that the closer the origin of the species, the more similar the sequence coding mechanism adjacent to the editing site seems likely to be. Therefore, the iPReditor-CMG model should have better generalization ability in predicting the editing site of dicotyledons.

doi:10.20944/preprints202110.0332.v1

3.4.4 Validity evaluation of the iPReditor-CMG model based on the validation sets

To further verify the generalization ability of the iPReditor-CMG model, the self-sequenced RNA editing sites of tobacco mitochondrial ATPase were also evaluated as the validation set. Based on the 300 sets of features reserved in the training sets, 300 validation sets from tobacco were also obtained accordingly. After independent testing, a total of 600 *ACC* values from the test sets of Ara, Bra, and Ory and the validation sets of tobacco were analyzed synthetically to obtain 31 high-precision models (i.e., the models with the highest or double-high *ACC* values in the test or validation sets were regarded as high-precision models) (Attachment 5) and 2 optimal candidate models (Ara_9_1 and Ara_9_4, where 9_1 and 9_4 respectively denoted the first and fourth results of 20 repetitions at a training-test set ratio of 9:1) (Table 6). The analysis of the two best candidates was generated by the same proportion of the same species, with Ara_9_1 having more features and lower accuracy in the validation set but highest accuracy in the test set, while Ara_9_4 had fewer features and fairly high accuracy in both the test and validation sets, although not the highest accuracy. *Sn* and *Sp* represented the prediction accuracy for positive and negative samples, and the results in Table 6 supported the stability of the two optimal candidate models without overfitting.

Table 6 Evaluation of the optimal models

		Screening times	Preserved features	*Sn*	*Sp*	*ACC*
Ara_9_1	Test set	3	400	0.91	0.90	0.91
	Validation set			0.81	0.77	0.79
Ara_9_4	Test set	10	44	0.83	0.90	0.87
	Validation set			0.81	0.91	0.86

Overall, the Ara data set exhibited very high accuracy (Appendix 5) with 2% of the models having *ACC* values above 0.90 and 29% above 0.80, whereas there were no models with *ACC* values above 0.90 in the Bra and Ory datasets, but 6.5% and 2% of the models were above 0.80, respectively. It was speculated that the reasons for this phenomenon might be the following: 1) the homology between Arabidopsis thaliana and tobacco was higher, with more than 70% sequence similarity for the five genes in the validation set; 2) the proportion of positive and negative samples was balanced in the Ara dataset, followed by the Bra dataset and least in the Ory dataset, so the imbalance between positive and negative samples might be a more important reason for the slightly lower prediction accuracy of the Bra and Ory datasets. To effectively utilize the information from the existing dataset, all the positive samples were corrected in this investigation, which resulted in a change in the number of positive samples compared to the reference, while the negative samples remained uncorrected. Even with the imbalance of positive and negative samples in the newly modified datasets, the prediction accuracy of the iPReditor-CMG model was still significantly improved over the reference methods in the Ara dataset and similarly in the Bra and Ory datasets. Therefore, the prediction accuracy of our novel model might still be further improved if the sample sequence homology could be increased and the sample imbalance could be decreased.

3.5 Establishment of an interpretability system for the best model

The Ara_9_1 and Ara_9_4 models were considered as candidates for the best because they had the highest *ACC* values and fewer features. Many previous researches have indicated that SVM has better generalization but weaker interpretability in some nonlinear fields, for which we analyzed two best candidate models according to the previously established SVM interpretability system [28]. Their ROC curve plots were presented in figure 3, and since the closer the AUC value (area under the ROC curve) was to 1, the better the accuracy of the model prediction was [29], Ara_9_1 (AUC = 0.88) was judged to be better than Ara_9_4 (AUC = 0.79). Therefore, Ara_9_1 should be the best model and an interpretable analysis was then performed on it after a comprehensive comparison.

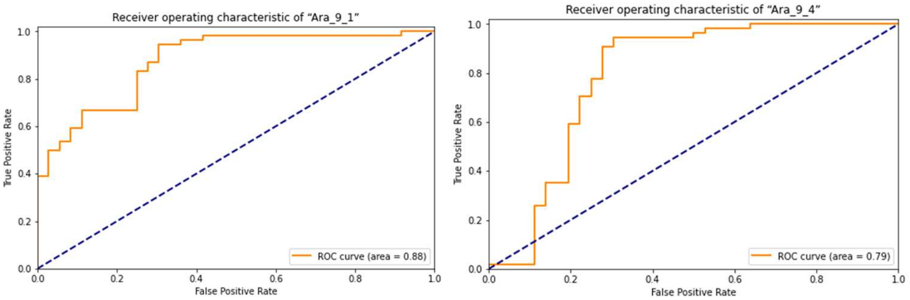


Figure 3 The ROC graph of two best candidate models

In the Ara_9_1 model, P-, A-, T- and M-features were all involved. All feature variables of the Ara_9_1 model were evaluated using the SVM-based t-test, and the top ten features were listed in Table 7, of which the top five features were all P-features and the features ranked 6th to 10th were all M-features. In our P-feature extraction procedure, the presence of a difference threshold (= 0.1) might lead to poor local interpretation, but the P-features were still ranked so high that overall it indicated that the frequencies of the four bases of the sequences with editing sites were extremely different from those of the normal sequences. Then, the M featured sequences ranked 6th to 10th were TGACA, ACAAC, GTAGA, CCGTT and TAACA, but no similar was previously reported in the literature. These variables were closely related to the C-to-U editing phenomenon and further experimental validation was needed to understand their possible functions.

Table 7 top 10 variables evaluation for Ara_9_1 model

Ranking	Features	t	Ranking	Features	t
1st	P_1	16.09	6th	$M_{900(TGACA)}$	15.56
2nd	P_4	15.93	7th	$M_{65(ACAAC)}$	15.56
3rd	P_3	15.64	8th	$M_{712(GTAGA)}$	15.54
4th	P_2	15.61	9th	$M_{367(CCGTT)}$	15.54
5th	P_7	15.58	10th	$M_{772(TAACA)}$	15.53

Notes: In the column of features, P1 indicated the 1st feature among P-features, M900(TGACA) denoted the 900th feature among M-features with a feature sequence of TGACA, and so on. (t0.05/900 = 1.96, t0.01/900 =2.58).

4. Conclusions

To effectively analyze the important biological issue of crop mitochondrial editing, fifteen groups of DNA sequence features were extracted from four crops in this manuscript, and a large amount of redundant information was removed using the feature nonlinear screening method HDSN developed in our laboratory. Subsequently, the iPReditor-CMG prediction models were constructed based on nonlinear SVM, and the results of independent tests from multiple perspectives showed that the advantages of our models. The model greatly simplified feature sources and reduced the computational effort, while still maintaining a high level of accuracy. The research will play an important role in identifying crop mitochondrial editing sites and even analyzing major bioscience concerns related to editing-based fertility in crops.

Author's contributions

Sidong Qin and Yanjun Fan are mainly responsible for writing papers; Shengnan Hu, Yongqiang Wang, Ziqi Wang, Yixiang Cao, Siqiao Tan and Qiyuan Liu are responsible for collecting and collating data; corresponding author Wei Zhou and Zhijun Dai are responsible for the conception and revision of the articles, and participates in the paper writing guidance. All authors read and agree to the final text.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No. 31960418, 31701164 and 31301388), Hunan Provincial Natural Science Foundation of China (No. 2021JJ30355 and 2018JJ3238) and the Research Foundation of Hunan Provincial Education Bureau of China (No. 19A214). The authors also thank Hunan Agricultural University, Changsha, China for their help during the manuscript preparation.

Attachment 1

```
function    [ pos_sample ] = extract_position( seq,pos_index )
n = length(pos_index);
nucleotide = seq(pos_index);
for i = 1:n
    if strcmp(nucleotide(i),'C')
        front = pos_index(i)-250;
        behind = pos_index(i)+250;
pos_sample{i,1} = seq(front:behind);
else
       front = pos_index(i)-250;
       behind = pos_index(i)+250;
       temp_sample = seq(front:behind);
       temp2 = seqcomplement(seqreverse(temp_sample));
       pos_sample{i,1} = temp2;
    end
end
end
```

Attachment 2

```
function    [ seq_500 ] = get_neg( ara_seq, ara_neg )
%GEN_NEG Summary of this function goes here
% From 41 windows to 501 windows
% ara_seq: Original sequence, ara_neg: Sequence with window 41
%     Detailed explanation goes here
 [m,~]=size(ara_neg);
for i = 1:m
    temp = ara_neg{i};
    index = strfind(ara_seq, temp);
    if ~isempty(index)
       seq_500{i} = ara_seq((index-230): (index+270));
    else
       temp_comre = seqcomplement(seqreverse(temp));
       index_re = strfind(ara_seq, temp_comre);
       temp2 = ara_seq((index_re-230): (index_re+270));
```

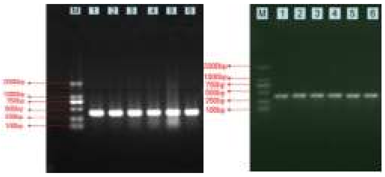
```
            seq_500{i} = seqcomplement(seqreverse(temp2));
        end
    end
end
```

Attachment 3

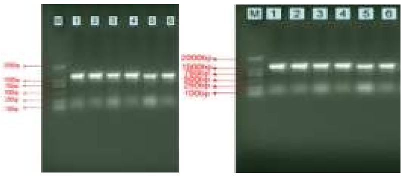
```
function   [ tri_fea ] = triplet_fea( seq )
%TRIPLET_FEA Summary of this function goes here
%     Detailed explanation goes here
nu = {'A','T','C','G'};
table = get_table(nu);
m= length(seq);
fea= [];
for ii = 1:m
    temp = seq{ii};
    for jj = 1:64
    index = strfind(temp, table{jj});
    number(jj) = length(index)/499;
    end
    fea =    [fea;number];
    fprintf('finished %d\n sequeces', ii);
end
tri_fea = fea;
end

function table = get_table(nu)
table = {};
n=0;
for i = 1:4
        temp1 = nu{i};
        for j = 1:4
                temp2 = nu{j};
                for k = 1:4
                        temp3 = nu{k};
                        n = n+1;
                        table{n}= [temp1, temp2, temp3];
                end
        end
end
end
```

Attachment 4

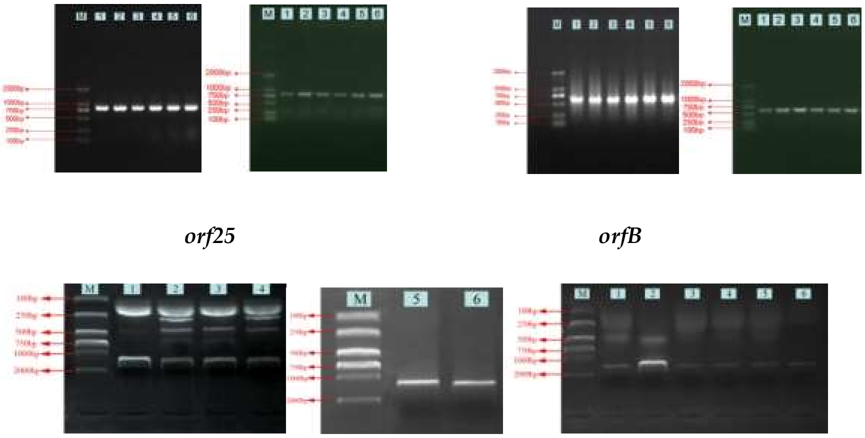


atp9



atp6

doi:10.20944/preprints202110.0332.v1



orf25

orfB

atp1

Figure. Electrophoretogram of PCR products of DNA (left) and cDNA (right) of ATPase genes in tobacco mitochondria. (M: DL2000 DNA Marker; 1: SZY90; 2:MZY90; 3: SYY85; 4: MYY85; 5: SK326; 6: MK326)

Attachment 5

Table Twenty evaluated results using HDSN method of five different proportions datasets in four crop mitochondria

Species	Ratios		Test set/Validation set																			
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Ara	5:5	*ACC*	**0.80/0.71**	0.80/0.56	0.79/0.75	**0.78/0.76**	0.78/0.73	0.78/0.72	0.78/0.69	0.78/0.68	0.78/0.66	0.77/0.73	0.77/0.69	0.77/0.66	0.77/0.62	0.76/0.75	0.76/0.65	0.76/0.63	0.75/0.75	0.74/0.59	0.73/0.63	0.73/0.55
		Sn	**0.74/0.45**	0.76/0.48	0.75/0.63	**0.71/0.75**	0.72/0.66	0.73/0.63	0.70/0.60	0.74/0.63	0.79/0.84	0.72/0.51	0.76/0.66	0.72/0.57	0.76/0.63	0.69/0.57	0.73/0.69	0.73/0.60	0.64/0.54	0.73/0.63	0.69/0.60	0.67/0.78
		Sp	**0.85/0.94**	0.83/0.63	0.83/0.86	**0.85/0.77**	0.83/0.80	0.83/0.80	0.85/0.77	0.82/0.72	0.77/0.50	0.82/0.94	0.78/0.72	0.82/0.75	0.78/0.61	0.82/0.91	0.80/0.61	0.79/0.66	0.86/0.94	0.75/0.55	0.77/0.66	0.80/0.33
	6:4	*ACC*	**0.79/0.85**	0.79/0.76	0.79/0.75	0.79/0.73	0.79/0.65	0.78/0.76	0.78/0.75	0.78/0.73	0.77/0.79	0.77/0.71	0.77/0.71	0.77/0.68	0.77/0.66	0.76/0.81	0.76/0.78	0.76/0.76	0.76/0.75	0.76/0.56	0.75/0.66	0.75/0.65
		Sn	**0.74/0.81**	0.74/0.63	0.73/0.60	0.72/0.69	0.73/0.54	0.72/0.60	0.73/0.60	0.72/0.69	0.73/0.69	0.74/0.51	0.73/0.57	0.69/0.63	0.72/0.54	0.72/0.66	0.70/0.63	0.69/0.66	0.68/0.69	0.72/0.66	0.65/0.63	0.68/0.54
		Sp	**0.84/0.88**	0.84/0.88	0.86/0.88	0.85/0.77	0.85/0.75	0.84/0.91	0.83/0.88	0.85/0.77	0.81/0.88	0.80/0.88	0.81/0.83	0.85/0.72	0.81/0.77	0.79/0.94	0.82/0.91	0.83/0.86	0.83/0.80	0.80/0.47	0.85/0.69	0.83/0.75
	7:3	*ACC*	**0.81/0.71**	**0.81/0.71**	0.80/0.78	0.80/0.76	0.80/0.75	0.80/0.73	0.80/0.73	0.80/0.72	0.80/0.63	**0.79/0.91**	0.79/0.82	0.79/0.76	0.78/0.78	0.78/0.72	0.78/0.68	0.78/0.57	**0.77/0.85**	0.77/0.73	0.77/0.68	0.77/0.65
		Sn	**0.79/0.72**	**0.77/0.57**	0.76/0.60	0.77/0.60	0.76/0.57	0.76/0.54	0.73/0.66	0.75/0.57	0.76/0.63	**0.75/0.84**	0.72/0.75	0.73/0.63	0.74/0.63	0.76/0.63	0.73/0.54	0.73/0.57	**0.76/0.75**	0.71/0.57	0.73/0.63	0.77/0.60
		Sp	**0.84/0.69**	**0.84/0.83**	0.85/0.94	0.82/0.91	0.84/0.91	0.83/0.91	0.86/0.80	0.85/0.86	0.84/0.63	**0.83/0.97**	0.86/0.88	0.84/0.88	0.81/0.91	0.80/0.80	0.83/0.80	0.84/0.58	**0.79/0.94**	0.84/0.88	0.81/0.72	0.78/0.69
	8:2	*ACC*	**0.83/0.79**	0.83/0.72	0.83/0.66	0.83/0.65	0.82/0.81	0.82/0.79	0.82/0.73	0.82/0.71	0.81/0.72	**0.81/0.82**	0.81/0.81	0.81/0.76	0.81/0.75	0.81/0.68	0.80/0.71	0.80/0.63	0.79/0.76	0.79/0.76	0.79/0.69	0.78/0.68
		Sn	**0.80/0.78**	0.80/0.78	0.81/0.57	0.77/0.54	0.76/0.72	0.80/0.78	0.83/0.69	0.81/0.81	0.81/0.72	**0.75/0.84**	0.77/0.66	0.75/0.66	0.77/0.75	0.81/0.72	0.79/0.72	0.76/0.63	0.74/0.63	0.72/0.63	0.76/0.51	0.74/0.66
		Sp	**0.86/0.80**	0.86/0.66	0.85/0.75	0.87/0.75	0.87/0.88	0.84/0.80	0.82/0.77	0.82/0.61	0.80/0.72	**0.86/0.80**	0.84/0.94	0.85/0.86	0.83/0.75	0.80/0.63	0.81/0.69	0.83/0.63	0.84/0.88	0.84/0.88	0.81/0.86	0.81/0.69
	9:1	*ACC*	**0.91/0.79**	**0.88/0.72**	0.88/0.76	**0.87/0.86**	**0.87/0.81**	0.87/0.75	0.87/0.75	0.86/0.60	0.86/0.78	0.85/0.66	0.85/0.63	0.84/0.76	0.84/0.76	0.84/0.63	0.83/0.84	0.83/0.81	0.83/0.78	0.83/0.78	0.83/0.69	0.82/0.63
		Sn	**0.91/0.81**	**0.86/0.78**	0.86/0.69	**0.83/0.81**	**0.83/0.81**	0.86/0.66	0.83/0.66	0.88/0.66	0.86/0.75	0.80/0.69	0.80/0.57	0.83/0.81	0.80/0.57	0.86/0.87	0.75/0.81	0.80/0.78	0.75/0.72	0.75/0.66	0.75/0.66	0.75/0.51
		Sp	**0.90/0.77**	**0.90/0.66**	0.90/0.83	**0.90/0.91**	**0.90/0.80**	0.88/0.83	0.90/0.83	0.85/0.55	0.87/0.80	0.88/0.63	0.88/0.69	0.85/0.72	0.87/0.94	0.83/0.41	0.88/0.86	0.85/0.83	0.88/0.83	0.88/0.88	0.88/0.72	0.87/0.75
	5:5	*ACC*	**0.78/0.62**	0.78/0.46	0.77/0.59	**0.77/0.79**	0.76/0.57	0.76/0.55	0.76/0.56	0.76/0.50	0.75/0.68	0.75/0.66	0.75/0.56	0.75/0.55	0.75/0.55	0.75/0.44	0.74/0.68	0.74/0.68	0.74/0.55	0.73/0.69	0.73/0.59	0.72/0.52
		Sn	**0.69/0.66**	0.70/0.48	0.75/0.63	**0.71/0.75**	0.73/0.60	0.72/0.69	0.70/0.78	0.70/0.54	0.69/0.69	0.69/0.78	0.69/0.51	0.70/0.75	0.66/0.60	0.66/0.48	0.69/0.66	0.69/0.75	0.66/0.63	0.69/0.81	0.69/0.60	0.67/0.69
		Sp	**0.85/0.58**	0.84/0.44	0.78/0.55	**0.82/0.83**	0.80/0.55	0.80/0.41	0.81/0.36	0.81/0.47	0.80/0.66	0.81/0.55	0.80/0.61	0.79/0.36	0.82/0.50	0.82/0.41	0.79/0.69	0.78/0.61	0.81/0.47	0.76/0.58	0.77/0.58	0.77/0.36
	6:4	*ACC*	**0.80/0.73**	0.79/0.57	0.79/0.66	0.79/0.63	0.78/0.69	0.78/0.56	0.78/0.59	0.77/0.62	0.77/0.63	0.77/0.62	0.76/0.60	0.76/0.49	0.75/0.60	0.75/0.60	0.75/0.59	0.74/0.69	0.74/0.57	0.74/0.57	0.72/0.63	0.72/0.60

		Sn	**0.75/0.60**	0.74/0.51	0.73/0.57	0.73/0.66	0.76/0.60	0.70/0.39	0.68/0.57	0.73/0.48	0.72/0.51	0.72/0.72	0.71/0.54	0.65/0.54	0.70/0.57	0.68/0.51	0.68/0.51	0.70/0.60	0.70/0.60	0.64/0.45	0.66/0.66	0.64/0.51
		Sp	**0.84/0.86**	0.83/0.63	0.84/0.75	0.82/0.61	0.79/0.77	0.84/0.72	0.85/0.61	0.79/0.75	0.81/0.75	0.80/0.52	0.79/0.66	0.84/0.44	0.79/0.63	0.81/0.69	0.80/0.66	0.78/0.77	0.78/0.55	0.81/0.69	0.76/0.61	0.78/0.69
		ACC	**0.80/0.66**	0.79/0.56	0.79/0.69	0.78/0.56	0.78/0.63	0.78/0.57	0.78/0.69	0.77/0.73	0.77/0.60	0.77/0.65	0.77/0.71	0.77/0.71	0.77/0.63	0.76/0.62	0.76/0.63	**0.76/0.79**	0.76/0.65	0.75/0.69	0.74/0.62	0.70/0.71
Bra	7:3	*Sn*	**0.72/0.75**	0.72/0.66	0.71/0.69	0.70/0.45	0.69/0.63	0.69/0.48	0.68/0.48	0.75/0.90	0.78/0.81	0.72/0.63	0.71/0.69	0.67/0.60	0.65/0.48	0.74/0.63	0.72/0.51	**0.66/0.75**	0.66/0.57	0.65/0.63	0.67/0.63	0.61/0.66
		Sp	**0.86/0.58**	0.84/0.47	0.85/0.69	0.85/0.66	0.85/0.63	0.85/0.66	0.86/0.88	0.79/0.58	0.76/0.41	0.81/0.66	0.82/0.72	0.85/0.80	0.86/0.77	0.77/0.61	0.79/0.75	**0.85/0.83**	0.84/0.72	0.83/0.75	0.80/0.61	0.77/0.75
		ACC	**0.84/0.75**	0.84/0.66	0.83/0.59	0.82/0.76	0.80/0.60	0.80/0.69	0.80/0.65	0.80/0.63	0.79/0.72	0.79/0.71	0.79/0.65	0.78/0.72	0.78/0.79	**0.78/0.81**	0.78/0.65	0.78/0.72	0.77/0.78	0.76/0.66	0.75/0.66	0.74/0.73
	8:2	*Sn*	**0.79/0.69**	0.78/0.60	0.76/0.63	0.80/0.81	0.78/0.48	0.76/0.57	0.75/0.63	0.68/0.48	0.73/0.69	0.69/0.66	0.69/0.57	0.78/0.69	0.71/0.78	**0.68/0.69**	0.67/0.36	0.65/0.75	0.75/0.63	0.71/0.72	0.65/0.63	0.71/0.78
		Sp	**0.88/0.80**	0.89/0.72	0.88/0.55	0.83/0.72	0.82/0.72	0.83/0.80	0.83/0.66	0.90/0.77	0.83/0.75	0.86/0.75	0.86/0.72	0.79/0.75	0.83/0.80	**0.86/0.91**	0.88/0.91	0.88/0.69	0.79/0.91	0.81/0.71	0.82/0.69	0.77/0.69
		ACC	**0.83/0.65**	0.79/0.66	0.79/0.71	0.79/0.57	0.79/0.71	0.79/0.63	0.79/0.57	**0.78/0.84**	0.78/0.60	0.78/0.59	0.78/0.68	0.78/0.65	0.78/0.71	0.77/0.73	0.77/0.72	0.77/0.71	0.75/0.63	0.75/0.60	0.75/0.73	0.73/0.65
	9:1	*Sn*	**0.80/0.63**	0.77/0.60	0.74/0.69	0.74/0.60	0.71/0.63	0.71/0.51	0.71/0.45	**0.77/0.78**	0.77/0.66	0.74/0.57	0.71/0.63	0.71/0.75	0.68/0.63	0.71/0.63	0.68/0.69	0.68/0.72	0.71/0.72	0.62/0.54	0.62/0.66	0.74/0.63
		Sp	**0.85/0.66**	0.81/0.72	0.83/0.72	0.83/0.55	0.85/0.77	0.85/0.75	0.85/0.69	**0.79/0.88**	0.79/0.55	0.81/0.61	0.83/0.72	0.83/0.55	0.85/0.77	0.81/0.83	0.83/0.75	0.83/0.69	0.79/0.55	0.85/0.66	0.85/0.80	0.72/0.66
		ACC	**0.74/0.78**	0.72/0.71	0.71/0.66	0.71/0.66	0.70/0.71	0.70/0.72	0.70/0.68	0.70/0.68	0.70/0.59	0.70/0.56	0.70/0.68	0.70/0.66	0.70/0.63	**0.69/0.79**	0.69/0.66	0.69/0.62	0.69/0.66	0.68/0.69	0.68/0.71	0.67/0.60
	5:5	*Sn*	**0.68/0.72**	0.64/0.48	0.65/0.60	0.65/0.57	0.66/0.57	0.64/0.66	0.64/0.57	0.63/0.51	0.63/0.54	0.63/0.39	0.60/0.51	0.60/0.48	0.56/0.54	**0.66/0.66**	0.65/0.51	0.65/0.51	0.61/0.69	0.63/0.66	0.58/0.72	0.62/0.54
		Sp	**0.80/0.83**	0.80/0.91	0.77/0.72	0.76/0.75	0.75/0.83	0.77/0.77	0.76/0.77	0.78/0.83	0.78/0.63	0.78/0.72	0.81/0.83	0.81/0.83	0.84/0.72	**0.73/0.91**	0.74/0.80	0.73/0.72	0.77/0.63	0.73/0.72	0.78/0.69	0.73/0.66
		ACC	**0.71/0.62**	0.70/0.73	0.70/0.68	0.70/0.62	0.69/0.55	0.69/0.68	0.69/0.60	0.68/0.76	**0.68/0.81**	0.68/0.71	0.68/0.68	0.67/0.65	0.66/0.66	0.66/0.63	0.66/0.66	0.66/0.75	0.66/0.71	0.66/0.60	0.65/0.63	0.65/0.62
	6:4	*Sn*	**0.70/0.63**	0.64/0.66	0.62/0.57	0.61/0.45	0.62/0.57	0.62/0.48	0.58/0.51	0.65/0.66	**0.64/0.84**	0.59/0.54	0.59/0.48	0.58/0.45	0.62/0.39	0.60/0.45	0.59/0.54	0.59/0.60	0.53/0.54	0.51/0.42	0.61/0.48	0.56/0.54
		Sp	**0.72/0.61**	0.75/0.80	0.78/0.77	0.78/0.77	0.77/0.52	0.75/0.86	0.80/0.69	0.71/0.86	**0.71/0.87**	0.78/0.86	0.77/0.86	0.77/0.83	0.69/0.91	0.73/0.80	0.74/0.77	0.72/0.88	0.78/0.86	0.82/0.77	0.69/0.77	0.73/0.69
		ACC	**0.73/0.62**	0.71/0.68	0.71/0.66	0.71/0.56	0.71/0.63	0.70/0.52	0.70/0.76	0.70/0.69	0.70/0.72	0.69/0.59	0.69/0.55	0.69/0.76	0.69/0.59	0.68/0.57	0.68/0.49	0.68/0.50	**0.68/0.79**	0.67/0.66	0.67/0.60	0.66/0.57
Ory	7:3	*Sn*	**0.66/0.39**	0.64/0.63	0.64/0.51	0.64/0.48	0.63/0.51	0.66/0.45	0.63/0.78	0.59/0.45	0.56/0.63	0.66/0.36	0.64/0.63	0.62/0.69	0.60/0.45	0.63/0.51	0.60/0.30	0.60/0.36	**0.60/0.81**	0.62/0.39	0.61/0.51	0.57/0.45
		Sp	**0.80/0.83**	0.78/0.72	0.78/0.80	0.78/0.63	0.79/0.75	0.75/0.58	0.77/0.75	0.81/0.91	0.85/0.80	0.73/0.80	0.74/0.47	0.76/0.83	0.79/0.72	0.73/0.63	0.77/0.66	0.76/0.63	**0.75/0.77**	0.73/0.91	0.73/0.69	0.76/0.69
		ACC	0.75/0.68	**0.75/0.84**	0.75/0.66	0.74/0.62	0.73/0.68	0.73/0.68	0.73/0.62	0.73/0.56	0.72/0.71	0.72/0.59	0.71/0.66	0.71/0.49	0.71/0.63	0.71/0.59	0.71/0.75	0.70/0.65	0.70/0.59	0.70/0.60	0.69/0.69	0.66/0.69
	8:2	*Sn*	0.79/0.54	**0.72/0.84**	0.68/0.51	0.69/0.48	0.74/0.54	0.72/0.60	0.72/0.57	0.63/0.45	0.73/0.51	0.68/0.60	0.70/0.66	0.69/0.60	0.68/0.48	0.64/0.45	0.60/0.63	0.71/0.72	0.63/0.39	0.60/0.48	0.73/0.63	0.64/0.66
		Sp	0.71/0.80	**0.78/0.83**	0.82/0.80	0.78/0.75	0.72/0.80	0.74/0.75	0.73/0.66	0.82/0.66	0.71/0.88	0.75/0.58	0.71/0.66	0.73/0.38	0.73/0.77	0.78/0.72	0.82/0.86	0.68/0.58	0.77/0.77	0.79/0.72	0.64/0.75	0.68/0.72
		ACC	**0.81/0.66**	**0.80/0.79**	0.78/0.69	0.77/0.59	0.77/0.59	0.76/0.63	0.76/0.62	0.76/0.65	0.76/0.68	0.75/0.66	0.75/0.63	0.74/0.65	0.74/0.76	0.74/0.68	0.74/0.62	0.72/0.72	0.72/0.71	0.71/0.66	0.71/0.72	0.70/0.62
	9:1	*Sn*	**0.78/0.57**	**0.74/0.87**	0.72/0.66	0.74/0.63	0.72/0.45	0.74/0.72	0.72/0.66	0.70/0.51	0.65/0.63	0.72/0.51	0.72/0.54	0.72/0.66	0.70/0.54	0.70/0.60	0.68/0.66	0.76/0.63	0.70/0.57	0.74/0.60	0.70/0.60	0.63/0.42
		Sp	**0.83/0.75**	**0.85/0.72**	0.83/0.72	0.80/0.55	0.81/0.72	0.78/0.55	0.80/0.58	0.81/0.77	0.85/0.72	0.78/0.80	0.78/0.72	0.76/0.63	0.78/0.97	0.78/0.75	0.80/0.58	0.69/0.80	0.74/0.83	0.69/0.72	0.72/0.83	0.76/0.80

Note: the better models were bolded.

References

1. Knie N, Grewe F, Fischer S, *et al*. Reverse U-to-C editing exceeds C-to-U RNA editing in some ferns - a monilophyte-wide comparison of chloroplast and mitochondrial RNA editing suggests independent evolution of the two processes in both organelles. *BMC Evol Biol***2016**,*16*,134.
2. Simpson L, Emeson. RB. RNA editing. *Annu Rev Neurosci* **1996**,*19*,27-52.
3. Wang R, Cai X, Hu S, *et al*. Comparative analysis of the mitochondrial genomes of Nicotiana tabacum: hints toward the key factors closely related to the cytoplasmic male sterility mechanism. *Front Genet* **2020**,*11*,257.
4. Gutmann B, Millman M, Vincis Pereira Sanglard L, *et al*. The pentatricopeptide repeat protein MEF100 is required for the editing of four mitochondrial editing sites in Arabidopsis.*Cells* **2021**,*10*,468.
5. Hiesel R, Wissinger B, Schuster W, *et al*. RNA editing in plant mitochondria. *Science* **1989**,*246*,1632-1634.

6. He P, Xiao G, Liu H, *et al*. Two pivotal RNA editing sites in the mitochondrial atp1 mRNA are required for ATP synthase to produce sufficient ATP for cotton fiber cell elongation. *New Phytol* **2018**,*218*,167-182.
7. Small ID, Schallenberg-Rüdinger M, Takenaka M, *et al*. Plant organellar RNA editing: what 30 years of research has revealed. *Plant J***2020**,*101*,1040-1056.
8. Zheng P, Wang D, Huang Y, *et al*. Detection and analysis of C-to-U RNA editing in rice mitochondria-encoded ORFs.*Plants (Basel)* **2020**,*9*,1277.
9. Edera AA, Small I, Milone DH, *et al*. Deepred-Mt: Deep representation learning for predicting C-to-U RNA editing in plant mitochondria. *Comput Biol Med* **2021**,*136*,104682.
10. Wang Y, Liu XY, Huang ZQ, *et al*. PPR-DYW Protein EMP17 Is required for mitochondrial RNA editing, complex III biogenesis, and seed development in maize. *Front Plant Sci* **2021**,*12*,693272.
11. Grohmann L, Keilwagen J, Duensing N, *et al*. Detection and identification of genome editing in plants: challenges and opportunities. *Front Plant Sci* **2019**,*10*,236.
12. Kim MS, Hur B, Kim S. RDDpred: a condition-specific RNA-editing prediction model from RNA-seq data. *BMC Genomics*
2016,*17*,5.
13. Sun JM, Yang DM, OSMARK P, et al. Discriminative prediction of A-To-I RNA editing events from DNA sequence. *PLoS ONE*
2016, *11*,e0164962.
14. Chen W, Feng P, Yang H, *et al*. iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. *Oncotarget*
2017,*8*,4208-4217.
15. Cummings MP, Myers DS. Simple statistical models predict C-to-U edited sites in plant mitochondrial RNA. *BMC Bioinformatics* **2004**,*5*,132.
16. Mower JP. PREP-Mt: predictive RNA editor for plant mitochondrial genes. BMC Bioinformatics, 2005, 6:96.
17. Thompson J, Gopal S. Correction: genetic algorithm learning as a robust approach to RNA editing site prediction. *BMC*
Bioinformatics **2006**,*7*,145.
18. Du P, He T, Li Y. Prediction of C-to-U RNA editing sites in higher plant mitochondria using only nucleotide sequence features. *Biochem Biophys Res Commun* **2007**,*358*,336-341.
19. Lenz H, Knoop V. PREPACT 2.0: predicting C-to-U and U-to-C RNA editing in organelle genome sequences with multiple references and curated RNA editing annotation. *Bioinform Biol Insights* **2013**,*7*,1-19.
20. Liu Q, Deng J, Liu M. Classification models for predicting the antimalarial activity against Plasmodium falciparum. *SAR QSAR Environ Res***2020**,*31*,313-324.
21. Zhang Y, Chu C , Chen Y, *et al* . Splice site prediction using support vector machines with a Bayes kernel. *Expert Syst App*
2006,*30*,73-81.
22. Zhang H, Wang H, Dai Z, *et al*. Improving accuracy for cancer classification with a new algorithm for genes selection. *BMC*
Bioinformatics **2012**,*13*,298.

doi:10.20944/preprints202110.0332.v1

23. Sun C, Dai Z, Zhang H, *et al*. Binary matrix shuffling filter for feature selection in neuronal morphology classification. *Comput Math Methods Med* **2015**,*2015*,626975.
24. Dai Z, Zhou H, Ba Q, *et al*. Improving depression prediction using a novel feature selection algorithm coupled with context-aware analysis. *J Affect Disord* **2021**,*295*,1040-1048.
25. Cortes C. and Vapnik V. Support-vector networks. *Machine Learning* **1995,*20:273-297.***
26. Li JL, Wang LF, Wang HY, *et al*. High-accuracy splice site prediction based on sequence, component and position features. *Genet Mol Res* **2012**,11,3432-3451.
27. Chen Q, Xue B, Zhang M. Rademacher complexity for enhancing the generalization of genetic programming for symbolic regression. IEEE Trans Cybern **2020**,99,1-14.
28. Zhang H, Wang H, Dai Z, *et al*. Improving accuracy for cancer classification with a new algorithm for genes selection. BMC Bioinformatics **2012**,13,298.
29. Ruisánchez I, Jiménez-Carvelo AM, Callao MP. ROC curves for the optimization of one-class model parameters. A case study: authenticating extra virgin olive oil from a Catalan protected designation of origin. *Talanta* **2021**,222,121564.