frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 23 July 2020
doi: 10.3389/fbioe.2020.00858

# Identifying Antioxidant Proteins by Combining Multiple Methods

*Xianhai Li[1,2], Qiang Tang[2], Hua Tang[2] and Wei Chen[1,2,3]**

*[1] School of Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu, China, [2] Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu, China, [3] School of Life Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan, China*

Antioxidant proteins play important roles in preventing free radical oxidation from damaging cells and DNA. They have become ideal candidates of disease prevention and treatment. Therefore, it is urgent to identify antioxidants from natural compounds. Since experimental methods are still cost ineffective, a series of computational methods have been proposed to identify antioxidant proteins. However, the performance of the current methods are still not satisfactory. In this study, a support vector machine based method, called Vote9, was proposed to identify antioxidants, in which the sequences were encoded by using the features generated from 9 optimal individual models. Results from jackknife test demonstrated that Vote9 is comparable with the best one of the existing predictors for this task. We hope that Vote9 will become a useful tool or at least can play a complementary role to the existing methods for identifying antioxidants.

**Keywords: antioxidant, reduced amino acid composition, g-gap dipeptide composition, feature selection, support vector machine**

OPEN ACCESS

***Edited by:***
*Zhibin Lv,*
*University of Electronic Science and Technology of China, China*

***Reviewed by:***
*Shaherin Basith,*
*Ajou University, South Korea*
*Yongchun Zuo,*
*Inner Mongolia University, China*

***Correspondence:**
*Wei Chen*
*chenweiimu@gmail.com*

***Specialty section:***
*This article was submitted to Synthetic Biology, a section of the journal Frontiers in Bioengineering and Biotechnology*

***Received:*** *30 May 2020*
***Accepted:*** *03 July 2020*
***Published:*** *23 July 2020*

***Citation:***
*Li X, Tang Q, Tang H and Chen W (2020) Identifying Antioxidant Proteins by Combining Multiple Methods. Front. Bioeng. Biotechnol. 8:858. doi: 10.3389/fbioe.2020.00858*

## INTRODUCTION

Reactive oxygen species (ROS) are composed of oxygen free radicals and nitrogen free radicals. Free radicals contain unpaired electron molecules or atoms, which are generally unstable and highly reactive. They can trigger lipid peroxidation during metabolism, which leads to DNA strand breaks, and even oxidize biofilms and almost all molecules in tissues indiscriminately. Fortunately, organisms have evolved effective strategies to detect and prevent molecular oxygen metabolites (Finkel and Holbrook, 2000; Mccord, 2000; Klaus and Heribert, 2004; Li et al., 2015). This is called the antioxidant system of organisms, which can effectively resist the damages caused by ROS (Agus et al., 2011).

Owing to their important roles in the antioxidant system, natural antioxidants have received more and more attentions (Yigit et al., 2014). Antioxidant proteins can neutralize free radicals, thereby blocking cell damage or death caused by free radicals. The consumption of antioxidants can be used to reduce the oxidative stress caused by excessive ROS, and reduce the damage to the organism (Yang et al., 2017). Antioxidants have also been applied to prevent diseases such as heart disease, cancer, cardiovascular disease (Gey, 1990; Dreher and Junod, 1996; Diaz et al., 1997). Its unique role in anti-aging was also reported (Ames et al., 1993).

Accordingly, many proteins extracted from rapeseed, ginkgo and other plant seeds are used as natural antioxidants (Nichole et al., 2008; Huang et al., 2009). Some micronutrients such as vitamin C and vitamin E (Lobo et al., 2010) are also considered as antioxidant molecules. However, our body cannot synthesize these nutrients, so we need to ingest them from the diet. Therefore, it has become an urgent task to identify proteins with antioxidant activity from natural compounds.

Although identifying antioxidant proteins through biochemical experiments is an objective and accurate method, they are still labor intensive and expensive. With the massive production of protein sequences, a series of computational methods have been proposed to identify antioxidant proteins. For the first time, Enrique et al. (2013) proposed a random forest model for predicting antioxidant proteins based on star map topological index and achieved satisfactory results. However, their model was trained based on a dataset including redundant sequences that might lead to overestimation problems (Chou, 2011). In 2013, Feng et al. (2013) constructed a high quality dataset with the sequence similarity less than 60%. Based on this dataset, they developed a Naive Bayes method by using the optimal dipeptides and obtained an average accuracy of 66.88%. Based on this dataset, a series of methods have been proposed in recent years. In 2016, Feng et al. (2016) proposed a support vector machine based method, called AodPred, which identifies antioxidant by using the optimal 3-gap dipeptide features and improves the prediction accuracy to 74.79%. Later on, Lei et al. (2018) developed a computational model called SeqSVM by using support vector machine and obtained an overall accuracy of 89.46%. More recently, Meng et al. (2019) proposed another support vector machine model called AOPs-SVM by integrating multiple kinds of features and obtained an overall accuracy of 94.2%. However, the sensitivity of AOPs-SVM is only 68%.

The above results indicate that the prediction accuracy still needs to be improved. Therefore, in this study, based on the optimal dipeptide composition and the reduced amino acid composition (Chen D. et al., 2012; Chen W. et al., 2012; Feng et al., 2016; Lv et al., 2019), a new model was constructed. The results show that the performance of the proposed method for identifying antioxidant proteins is better than or at least comparable to existing methods.

# MATERIALS AND METHODS

## Training Set and Test Set

The dataset used in the present work is the same as the one used by Feng et al. (2013, 2017),which includes 253 antioxidant protein sequences and 1552 non-antioxidant protein sequences with the sequence identity less than 60%. The dataset is expressed as:

$$S = S_+ \cup S_- \tag{1}$$

where "S" stands for benchmark dataset, "$S_+$" is the positive dataset and contains 253 antioxidant protein sequences, and "$S_-$" is the negative dataset and contains 1552 non-antioxidant protein sequences. The longest and shortest peptides in the dataset are 1463 and 11 amino acids, respectively.

In the following analysis, the dataset S was divided into two parts. One of them is the training set $S_T$ and includes 80% of the sequences in S, and the remaining 20% sequences form the testing set $S_E$, which are expressed as following,

$$S_T = S_+^* 0.8 \cup S_-^* 0.8 \tag{2}$$

$$S_E = S - S_T \tag{3}$$

## Independent Dataset

To objectively evaluate the proposed method and compare with its counterpart, an independent dataset was built in the present work. By searching the Universal Protein Resource (Uniprot) with the keywords "antioxidant" and "reviewed," and setting the date from March 1, 2014 to March 31, 2020, we obtained 22 antioxidant protein sequences that are independent from the sequences in the dataset S.

## Support Vector Machine

Support Vector Machine (SVM) is a method for effectively identifying data according to supervised learning method, which is widely used in bioinformatics and other fields (Feng et al., 2016; Liao et al., 2018; Wang et al., 2019; Liu and Chen, 2020). If the samples are linearly separated, the basic idea of the SVM algorithm is to solve the separation hyperplane that can correctly divide the training dataset and have the largest geometric interval; when the samples are nonlinearly separated, SVM maps the low-dimensional data to the high-dimensional data by the kernel function space. In this work, the LIBSVM package downloaded from https://www.csie.ntu.edu.tw/~cjlin/libsvm/ was used to perform the prediction. The best regularization parameter *C* and kernel width parameter *g* were determined by using the grid search method.

## Sequence Representation

### g-gap Dipeptide Composition

The *g*-gap dipeptide composition was proposed to describe the long-range correlation between two amino acid residues and has been proved to be effective in the field of protein recognition (Ding et al., 2013; Lin et al., 2013; Tan et al., 2019). Accordingly, in the present work, the *g*-gap dipeptide composition was used to encode the sequences in both benchmark dataset and independent test dataset.

The g-gap dipeptide composition is expressed as following,

$$F = [f_1^g \; f_2^g \ldots f_i^g \ldots f_{400}^g]^T \tag{4}$$

$$f_i^g = \frac{n_i^g}{L - g - 1} \tag{5}$$

where $f_i^g$ represents the frequency of the *i*-th ($i$ = 1, 2,..., 400) dipeptide with *g*-gap interval in the protein sequence, and T

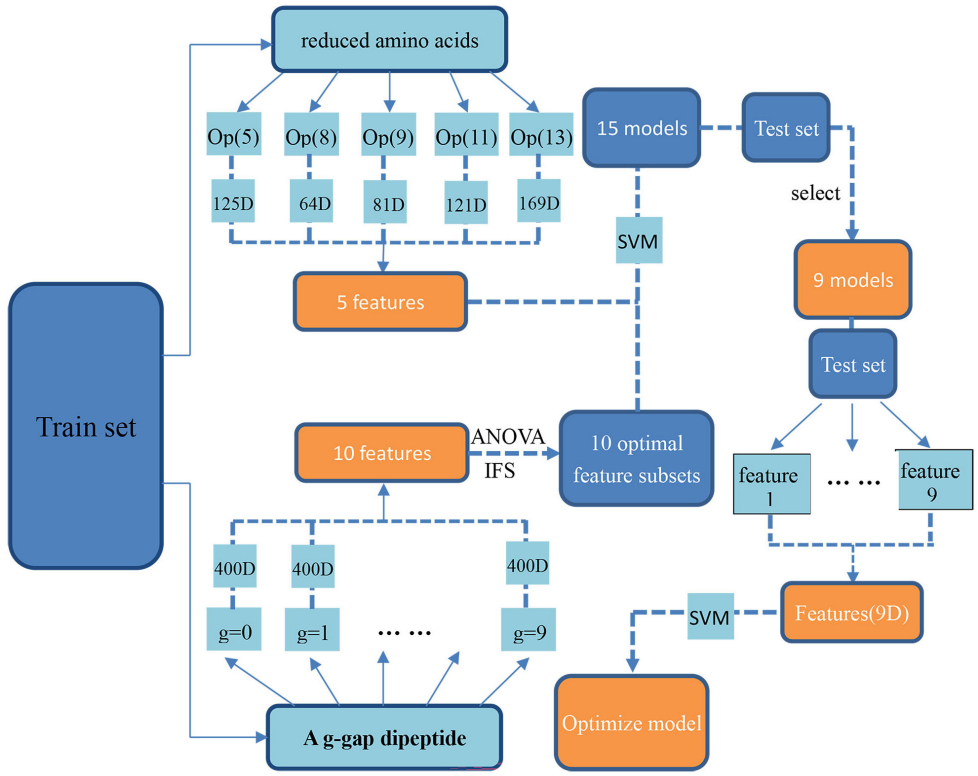
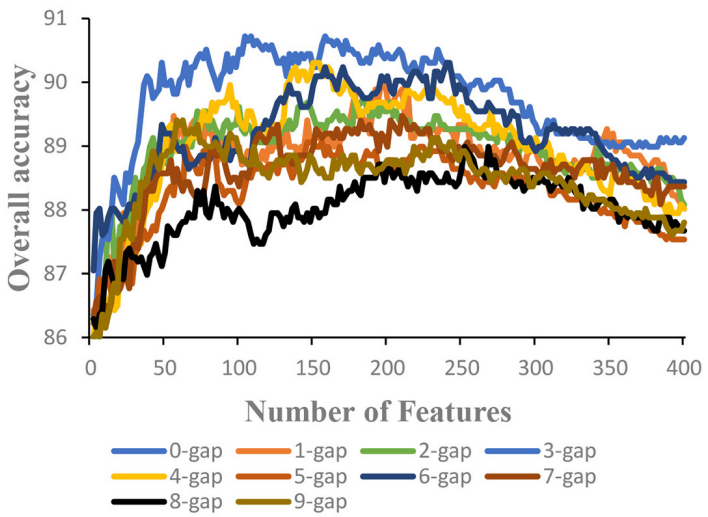



**FIGURE 1 |** The flowchart of building the proposed method. The samples in the training dataset were firstly encoded by using reduced amino acid compositions and the optimal g-gap dipeptide compositions, respectively. Accordingly, 15 SVM models based on these different kinds of features was built. After validating the combinational performance of these models on the test dataset, 9 of the 15 SVM models were selected out as the optimal models. Finally, the SVM outs of these 9 models were used as the new features and used as the inputs of the SVM for building the proposed model.




| g-gap | optimal features | accuracy |
|---|---|---|
| 0-gap | 172 | 0.886 |
| 1-gap | 197 | 0.898 |
| 2-gap | 184 | 0.892 |
| 3-gap | 158 | 0.909 |
| 4-gap | 150 | 0.903 |
| 5-gap | 147 | 0.9 |
| 6-gap | 158 | 0.9 |
| 7-gap | 210 | 0.898 |
| 8-gap | 252 | 0.875 |
| 9-gap | 58 | 0.873 |

**FIGURE 2 |** The IFS curves of different g-gap dipeptides (g = 0, 1, 2,..., 9). The optimal number of features and the accuracy based on the optimal features were shown in the right of the figure.

represents the transposition of the vector. $n_i^g$ represents the number of the $i$-th $g$-gap dipeptide. In the present work, $g$ is an integer in the range of [0, 9]. For example, $g$ = 0 represents the correlation between two adjacent amino acid residues, and $g$ = 1 represents the correlation of two amino acid residues separated by one residue, and so forth.

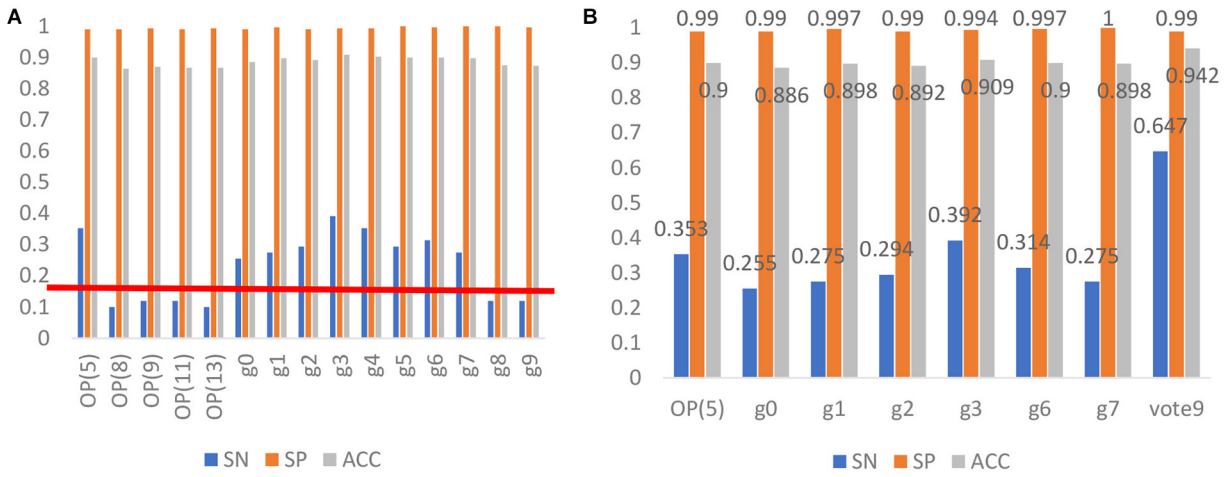



**FIGURE 3 | (A)** The performance of the 15 models for identifying antioxidants. OP (5) stands for the method of optimizing amino acid residues to divide 20 amino acid letters into 5 categories, and then uses LIBSVM to establish a classification model. Using the method of g-gap dipeptide (Feng et al., 2016), we selected the best feature subset of protein sequence steps g = 0, 1... 7 to construct a g0, g1... g7 classification model. Vote9 is a comprehensive classification model that used the prediction results of the above classification models as feature vectors. **(B)** Comparison between Vote9 and single classification model.

## Reduced Amino Acid Composition

With the aim of including structural information, the reduced amino acid composition (RAAC) was applied to encode proteins (Feng et al., 2016). Compared with the classical amino acid composition, the RAACs can reduce protein complexity and eliminate part of the redundant signals without losing sequence information intact (Wang and Wang, 1999; Liu et al., 2018). In order to obtain the RAAC from the sequences, Zuo et al. (2017) established the online webserver and database (Zheng et al., 2019) that can be used to calculate RAAC.

In term of RAAC, based on amino acid sequence and structure information, the 20 natural amino acids can be aggregated into a smaller number of representative amino acid residues (Thomas and Dill, 1996; Mirny and Shakhnovich, 1999; Solis and Rackovsky, 2000). According to the different optimization procedures (Op) for protein sequences proposed by Etchebest et al. (2007), there are 5 different cluster files for the 20 natural amino acids, i.e., Op(5), Op(8), Op(9), Op(11)and Op(13), which are formulated as below:

$$\mathrm{Op}\,(i) = \begin{cases} \mathrm{Op}\,(5):\{\mathrm{G};\ \mathrm{IVFYW};\ \mathrm{ALMEQRK};\ \mathrm{P};\ \mathrm{NDHSTC}\} \\ \mathrm{Op}\,(8):\{\mathrm{G};\ \mathrm{IV};\ \mathrm{FYW};\ \mathrm{ALM};\ \mathrm{EQRK};\ \mathrm{P};\ \mathrm{ND};\ \mathrm{HSTC}\} \\ \mathrm{Op}\,(9):\{\mathrm{G};\ \mathrm{IV};\ \mathrm{FYW};\ \mathrm{ALM};\ \mathrm{EQRK};\ \mathrm{P};\ \mathrm{ND};\ \mathrm{HS};\ \mathrm{TC}\} \\ \mathrm{Op}\,(11):\{\mathrm{G};\ \mathrm{IV};\ \mathrm{FYW};\ \mathrm{A};\ \mathrm{LM};\ \mathrm{EQRK};\ \mathrm{P};\ \mathrm{ND};\ \mathrm{HS};\}\ (6) \\ \mathrm{T};\ \mathrm{C}\} \\ \mathrm{Op}\,(13):\{\mathrm{G};\ \mathrm{IV};\ \mathrm{FYW};\ \mathrm{A};\ \mathrm{L};\ \mathrm{M};\ \mathrm{E};\ \mathrm{QRK};\ \mathrm{P};\ \mathrm{ND}; \\ \mathrm{HS};\ \mathrm{T};\ \mathrm{C}\} \end{cases}$$

where $i$ indicates the different cluster profiles ($i$ = 5, 8, 9, 11, 13), and the letters between the two semicolons belong to the same cluster.

Accordingly, a sequence can be encoded based on the reduced amino acid composition. As indicated in Eq. 6, for the $n$-peptide composition with various cluster profiles, the components and dimensions of the feature vector will be different.

$$\Psi = [\Psi_1,\ \Psi_2,\ \cdots,\ \Psi_\Omega]^T \tag{7}$$

where $\Omega$ is the dimension of the vector, and is based on the selected $n$ and cluster profiles. For example, for the dipeptide composition with the cluster profile of Op(5), the $\Omega$ will be 25. In the current work, our initial tests demonstrate that the optimal $n$ for different cluster profiles is as following,

$$\Omega = \begin{cases} 5^3 = 125 & for\ Op(5)\ cluster \\ 8^2 = 64 & for\ Op(8)\ cluster \\ 9^2 = 81 & for\ Op(9)\ cluster \\ 11^2 = 121 & for\ Op(11)\ cluster \\ 13^2 = 169 & for\ Op(13)\ cluster \end{cases} \tag{8}$$

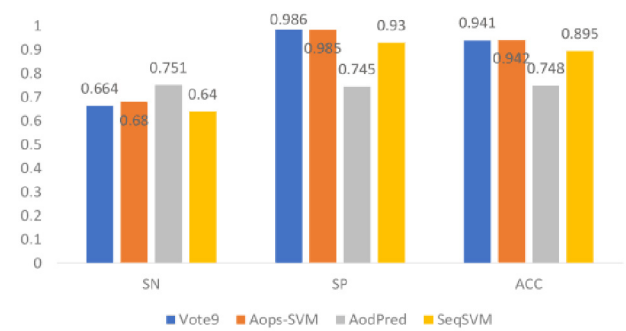



**FIGURE 4 |** Comparison of Vote9 with existing methods.

## Performance Evaluation

There are usually three methods for evaluating the performance of computational models, namely independent dataset test, k-fold

cross-validation test, and jackknife test (Wei et al., 2017; Chen et al., 2019; Manavalan et al., 2019a,b; Yang et al., 2019; Hasan et al., 2020; Lv et al., 2020). Among the three evaluation methods, the most rigorous and least random jackknife test was used to evaluate the proposed method.

The sensitivity (Sn), specificity (Sp), accuracy (Acc) and Mathew's correlation coefficient (MCC) was selected as the evaluation metrics that are defined as following,

$$Sn = \frac{TP}{TP + FN} \tag{9}$$

$$Sp = \frac{TN}{TN + FP} \tag{10}$$

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \tag{11}$$

$$MCC = \frac{TN^{*}TP - FP^{*}FN}{\sqrt{(TP + FP)^{*}(FN + TN)^{*}(TP + FN)^{*}(TN + FP)}} \tag{12}$$

where TP, FP, FN, and TN represent true positive, false positive, false negative and true negative, respectively.

## Feature Selection

The principle of analysis of variance (ANOVA) is to measure the characteristic variance by calculating the ratio (*F*-value) between the characteristics of the groups and the internal characteristics of the groups (Lin and Ding, 2011; Basith et al., 2019). The larger the *F*-value, the greater the probability that each sample comes from a different population. In order to exclude redundant features and enhance the robustness of the proposed model, the ANOVA that widely used in computational proteomics (Ding et al., 2013; Lin et al., 2013; Basith et al., 2020) combined with the incremental feature selection (IFS) strategy was used to select the optimal features.

## Flowchart of the Method

By following the above procedure, we proposed a new computational method for identifying antioxidants. The flowchart of how to build it was shown in **Figure 1**.

# RESULTS AND DISCUSSION

## Prediction Performance

In order to obtain the optimal features, for a given kind of g-gap dipeptide composition, the 400 g-gap dipeptide compositions were ranked based on their F-scores. Each of the 400 dipeptide compositions were added one by one from higher to lower rank. This procedure was repeated 400 times, and for each time a SVM model was built. The accuracies of these models were then used to plot the IFS curve. Accordingly, the 10 IFS curves for $g$ = 0 to 9 were obtained (**Figure 2**), where the abscissa is the number of features and the ordinate is the corresponding accuracy. In each curve, the optimal number of features were obtained when the curve reaches its peak. The optimal number of features and the accuracy based on the optimal features were shown in the right of **Figure 2**. Accordingly, 10 models were obtained based on g-gap dipeptide compositions.

Based on the reduced amino acid composition, another five models were built for identifying antioxidants. Their predictive performances together with that of the 10 models based on g-gap dipeptide composition were indicated in **Figure 3A**.

According to the prediction results of the 15 models, we removed 6 models with the sensitivity less than 20%. Therefore, 9 models were left and were combined to build the final model in the following analysis. To do so, the out of the nine SVM based models (1 or −1) were further used as the input of the SVM. Therefore, each sequence will be re-encoded by a 9-dimension vector with the element of 1 or −1. The model thus obtained is called Vote9. In the jackknife test, Vote9 obtained an accuracy of 0.94 with the sensitivity of 0.65, specificity of 0.99 and MCC of 0.74.

## Comparison With Single Model

In order to demonstrate the better performance of Vote9, we compared its performance with that of the single model for identifying antioxidants in the test dataset. The result is shown in **Figure 3B**. It was found that the sensitivity, specificity and accuracy of Vote9 are all significantly better than those of any

**TABLE 1 |** Comparative results of different methods for identifying antioxidants in independent dataset.

| Sample | Aops-SVM | Aodpred | Vote9 | Sample | Aops-SVM | Aodpred | Vote9 |
|---|---|---|---|---|---|---|---|
| P9WQB7 | Y | Y | N | P9WIS6 | Y | N | N |
| P9WHH9 | Y | N | N | P9WQB6 | Y | Y | N |
| P9WIS7 | Y | N | **Y** | P9WID9 | Y | Y | N |
| P9WG35 | Y | Y | N | O17433 | Y | Y | N |
| P9WGE9 | Y | Y | N | P9WIE0 | Y | N | N |
| P9WQB5 | Y | Y | N | P9WID8 | Y | Y | N |
| P9WIE3 | Y | Y | N | P9WGE8 | Y | Y | N |
| P0CU34 | Y | Y | N | C0HK70 | Y | Y | N |
| Q5ACV9 | N | N | N | P9WQB4 | Y | Y | N |
| P9WHH8 | Y | N | **Y** | P9WG34 | Y | Y | N |
| P9WIE1 | Y | N | **Y** | P9WIE2 | Y | Y | N |

single model, demonstrating that it's necessary to built the model by combining the optimal single models.

## Comparison With Existing Methods

In this section, we compared the performance of Vote9 with the performance of other existing methods (Aops-SVM, AodPred, and SeqSVM) that all trained based on the same dataset. Their performances were shown in **Figure 4**.

It was found that the accuracy of Vote9 is better than that of AodPred and SeqSVM, and is comparable with that of Aops-SVM. Although the sensitivity of Vote9 is lower than that of Aops-SVM and AodPred, its specificity is higher than that of the other three methods (Aops-SVM, AodPred, and SeqSVM). This result indicate that Vote9 might also become a useful tool for identifying antioxidants.

In order to objectively evaluate the performance of different methods for identifying antioxidants, a comparison was performed based on the independent dataset. Since some of the previous methods didn't provide publicly available tool or doesn't work properly, the comparison was also performed among Vote9, Aops-SVM, and AodPred. Their performances for identifying antioxidants in independent dataset were reported in **Table 1**. As shown in **Table 1**, we found that Aops-SVM performs the best, and Vote9 and AodPred can be used as complementary tools.

## Conclusion

The role of antioxidant proteins in neutralizing free radicals and preventing the damage of free radicals to cells is well known. Unfortunately, there are very few molecules with antioxidant properties in nature. Therefore, in order to accelerate researches on antioxidant proteins, there is an urgent need to develop effective methods for identifying them.

In the present work, we proposed a new method, called Vote9, in which the sequences were encoded by using the features generated from 9 optimal individual models. Results from jackknife test demonstrated that Vote9 is comparable with the best of the existing predictors for this task. The results of independent dataset test demonstrate that Vote9 can play a complementary role to the existing methods in this area. We hope that Vote9 will become a useful method for identifying antioxidants.

# DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/supplementary material.

# AUTHOR CONTRIBUTIONS

WC conceived and designed the experiments. XL, QT, and HT performed the experiments. XL and WC wrote the manuscript. All authors read and approved the final manuscript.

# REFERENCES


Agus, S. T., Eka, M., Oh, L. K., and Keizo, H. (2011). Isolation and characterization of antioxidant protein fractions from melinjo (*Gnetum gnemon*) seeds. *J. Agric. Food Chem.* 59, 5648–5656. doi: 10.1021/jf2000647

Ames, B. N., Shigenaga, M. K., and Hagen, T. M. (1993). Oxidants, antioxidants, and the degenerative diseases of aging. *Proc. Natl. Acad. Sci. U.S.A.* 90, 7915–7922. doi: 10.1073/pnas.90.17.7915

Basith, S., Manavalan, B., Hwan Shin, T., and Lee, G. (2020). Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. *Med. Res. Rev.* doi: 10.1002/med.21658 [Epub ahead of print].

Basith, S., Manavalan, B., Shin, T. H., and Lee, G. (2019). SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. *Mol. Ther. Nucleic Acids* 18, 131–141. doi: 10.1016/j.omtn.2019.08.011

Chen, D., Lu-Feng, Y., Shou-Hui, G., Hao, L., and Wei, C. (2012). Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions. *J. Proteom.* 77, 321–328. doi: 10.1016/j.jprot.2012.09.006

Chen, W., Feng, P., and Lin, H. (2012). Prediction of replication origins by calculating DNA structural properties. *FEBS Lett.* 586, 934–938. doi: 10.1016/j.febslet.2012.02.034

Chen, W., Feng, P., Liu, T., and Jin, D. (2019). Recent advances in machine learning methods for predicting heat shock proteins. *Curr. Drug Metab.* 20, 224–228. doi: 10.2174/1389200219666181031105916

Chou, K.-C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* 273, 236–247. doi: 10.1016/j.jtbi.2010.12.024

Diaz, M. N., Frei, B., Vita, J. A., and Keaney, J. F. (1997). Antioxidants and atherosclerotic heart disease. *N. Engl. J. Med.* 337, 408–416. doi: 10.1056/nejm199708073370607

Ding, H., Guo, S.-H., Deng, E.-Z., Yuan, L.-F., Guo, F.-B., Huang, J., et al. (2013). Prediction of Golgi-resident protein types by using feature selection technique. *Chemometr. Intellig. Lab. Syst.* 124, 9–13. doi: 10.1016/j.chemolab.2013.03.005

Dreher, D., and Junod, A. F. (1996). Role of oxygen free radicals in cancer development. *Eur. J. Cancer* 32, 30–38. doi: 10.1016/0959-8049(95)00531-5

Enrique, F.-B., Vanessa, A.-P., Robert, M. C., and Julian, D. (2013). Random forest classification based on star graph topological indices for antioxidant proteins. *J. Theor. Biol.* 317, 331–337. doi: 10.1016/j.jtbi.2012.10.006

Etchebest, C., Benros, C., Bornot, A., Camproux, A.-C., and Brevern, A. G. (2007). A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. *Eur. Biophys. J.* 36, 1059–1069. doi: 10.1007/s00249-007-0188-5

Feng, P., Ding, H., Lin, H., and Chen, W. (2017). AOD: the antioxidant protein database. *Sci. Rep.* 7:7449.

Feng, P., Chen, W., and Lin, H. (2016). Identifying antioxidant proteins by using optimal dipeptide compositions. *Interdiscipl. Sci. Comput. Life Sci.* 8, 186–191. doi: 10.1007/s12539-015-0124-9

Feng, P.-M., Hao, L., and Wei, C. (2013). Identification of antioxidants from sequence information using naïve Bayes. *Comput. Math. Mat. Methods Med.* 2013:567529.

Finkel, T., and Holbrook, N. J. (2000). Oxidants, oxidative stress and the biology of ageing. *Nature* 408, 239–247. doi: 10.1038/35041687

Gey, K. F. (1990). The antioxidant hypothesis of cardiovascular disease: epidemiology and mechanisms. *Biochem. Soc. Trans.* 18, 1041–1045. doi: 10.1042/bst0181041

Hasan, M. M., Schaduangrat, N., Basith, S., Lee, G., Shoombuatong, W., and Manavalan, B. (2020). HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. *Bioinformatics* 36, 3350–3356. doi: 10.1093/bioinformatics/btaa160

Huang, W., Deng, Q., Xie, B., Shi, J., Huang, F., Tian, B., et al. (2009). Purification and characterization of an antioxidant protein from *Ginkgo biloba* seeds. *Food Res. Intern.* 43, 86–94. doi: 10.1016/j.foodres.2009.08.015

Klaus, A., and Heribert, H. (2004). Reactive oxygen species: metabolism, oxidative stress, and signal transduction. *Annu. Rev. Plant Biol.* 55, 373–399. doi: 10.1146/annurev.arplant.55.031903.141701

Lei, X., Guangmin, L., Shuhua, S., and Changrui, L. (2018). SeqSVM: a sequence-based support vector machine method for identifying antioxidant proteins. *Intern. J. Mol. Sci.* 19:1773. doi: 10.3390/ijms19061773

Li, S., Tan, H.-Y., Wang, N., Zhang, Z.-J., Lao, L., Wong, C.-W., et al. (2015). The role of oxidative stress and antioxidants in liver diseases. *Intern. J. Mol. Sci.* 16, 26087–26124.

Liao, Z. J., Li, D. P., Wang, X. R., Li, L. S., and Zou, Q. (2018). Cancer diagnosis through isomir expression with machine learning method. *Curr. Bioinform.* 13, 57–63. doi: 10.2174/1574893611666160609081155

Lin, H., Chen, W., and Ding, H. (2013). AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes. *PLoS One* 8:75726. doi: 10.1371/journal.pone.0075726

Lin, H., and Ding, H. (2011). Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. *J. Theor. Biol.* 269, 64–69. doi: 10.1016/j.jtbi.2010.10.019

Liu, D., Li, G., and Zuo, Y. (2018). Function determinants of TET proteins: the arrangements of sequence motifs with specific codes. *Brief. Bioinform.* 20, 1826–1835.

Liu, K., and Chen, W. (2020). iMRM:a platform for simultaneously identifying multiple kinds of RNA modifications. *Bioinformatics* 36, 3336–3342. doi: 10.1093/bioinformatics/btaa155

Lobo, V., Patil, A., Phatak, A., and Chandra, N. (2010). Free radicals, antioxidants and functional foods: impact on human health. *Pharm. Rev.* 4, 118–126.

Lv, H., Dao, F., Zhang, D., Guan, Z., Yang, H., Su, W., et al. (2020). iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes. *Science* 23:100991. doi: 10.1016/j.isci.2020.100991

Lv, Z., Jin, S., Ding, H., and Zou, Q. (2019). A random forest sub-Golgi protein classifier optimized via dipeptide and amino acid composition features. *Front. Bioeng. Biotechnol.* 7:215. doi: 10.3389/fbioe.2019.00215

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019a). Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Mol. Ther. Nucleic Acids* 16, 733–744. doi: 10.1016/j.omtn.2019.04.019

Manavalan, B., Shaherin, B., Hwan, S. T., Yeon, L. D., Leyi, W., and Gwang, L. (2019b). 4mCpred-EL: an ensemble learning framework for identification of DNA N4-methylcytosine sites in the mouse genome. *Cells* 8:1332. doi: 10.3390/cells8111332

Mccord, J. M. (2000). The evolution of free radicals and oxidative stress. *Am. J. Med.* 108, 652–659.

Meng, C., Jin, S., Wang, L., Guo, F., and Zou, Q. (2019). AOPs-SVM: a sequence-based classifier of antioxidant proteins using a support vector machine. *Front. Bioeng. Biotechnol.* 7:224. doi: 10.3389/fbioe.2019.00224

Mirny, L. A., and Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* 291, 177–196. doi: 10.1006/jmbi.1999.2911

Nichole, C., Ying, Z., Marian, N., and Fereidoon, S. (2008). Antioxidant activity and water-holding capacity of canola protein hydrolysates. *Food Chem.* 109, 144–148. doi: 10.1016/j.foodchem.2007.12.039

Solis, A. D., and Rackovsky, S. (2000). Optimized representations and maximal information in proteins. *Proteins* 38, 149–164. doi: 10.1002/(sici)1097-0134(20000201)38:2<149::aid-prot4>3.0.co;2-#

Tan, J. X., Li, S. H., Zhang, Z. M., Chen, C. X., Chen, W., Tang, H., et al. (2019). Identification of hormone binding proteins based on machine learning methods. *Math. Biosci. Eng.* 16, 2466–2480. doi: 10.3934/mbe.2019123

Thomas, P. D., and Dill, K. A. (1996). An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl. Acad. Sci. U.S.A.* 93, 11628–11633. doi: 10.1073/pnas.93.21.11628

Wang, J., and Wang, W. (1999). A computational approach to simplifying the protein folding alphabet. *Nat. Struct. Biol.* 6, 1033–1038.

Wang, Y., Shi, F., Cao, L., Dey, N., Wu, Q., Ashour, A. S., et al. (2019). Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images. *Curr. Bioinform.* 14, 282–294. doi: 10.2174/1574893614666190304125221

Wei, L., Tang, J., and Zou, Q. (2017). Local-DPP: an improved DNA-binding protein prediction method by exploring local evolutionary information. *Inform. Sci.* 384, 135–144. doi: 10.1016/j.ins.2016.06.026

Yang, S., Lulu, W., Ying, W., Xiaoqian, O., Zhaoyuan, S., Chongchong, L., et al. (2017). Purification and identification of a natural antioxidant protein from fertilized eggs. *Korea. J. Food Sci. Anim. Resourc.* 37, 764–772. doi: 10.5851/kosfa.2017.37.5.764

Yang, W., Zhu, X., Huang, J., Ding, H., and Lin, H. (2019). A brief survey of machine learning methods in protein sub-golgi localization. *Curr. Bioinform.* 13, 234–240. doi: 10.2174/1574893613666181113131415

Yigit, A. A., Panda, A. K., and Cherian, G. (2014). The avian embryo and its antioxidant defence system. *Worlds Poul. Sci. J.* 70, 563–574. doi: 10.1017/s0043933914000610

Zheng, L., Huang, S., Mu, N., Zhang, H., Zhang, J., Chang, Y., et al. (2019). RAACBook: a web server of reduced amino acid alphabet for sequence-dependent inference by using Chou's five-step rule. *Database* 2019:baz131.

Zuo, Y., Li, Y., Chen, Y., Li, G., Yan, Z., and Yang, L. (2017). PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition. *Bioinformatics* 33, 122–124. doi: 10.1093/bioinformatics/btw564

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.