REVIEW ARTICLE

Recent Advances on Antioxidant Identification Based on Machine Learning Methods

Pengmian Feng[1,*] and Lijing Feng[2]

[1]School of Basic Medical Sciences, Chengdu University of Traditional Chinese Medicine, Chengdu 611730, China; [2]School of Sciences, North China University of Science and Technology, Tangshan 063000, China

ARTICLE HISTORY

Received: February 01, 2020
Revised: March 17, 2020
Accepted: May 13, 2020

DOI:
10.2174/1389200221666200719001449

Abstract: Antioxidants are molecules that can prevent damages to cells caused by free radicals. Recent studies also demonstrated that antioxidants play roles in preventing diseases. However, the number of known molecules with antioxidant activity is very small. Therefore, it is necessary to identify antioxidants from various resources. In the past several years, a series of computational methods have been proposed to identify antioxidants. In this review, we briefly summarized recent advances in computationally identifying antioxidants. The challenges and future perspectives for identifying antioxidants were also discussed. We hope this review will provide insights into researches on antioxidant identification.

Keywords: Antioxidant, free radical, diseases, sequence encoding scheme, machine learning methods, molecules.

1. INTRODUCTION

Free radicals are by-products of metabolism and their production is increased by exposure to several noxious environmental factors such as ionizing radiation, cigarette smoke, and environmental pollutants [1, 2]. Several diseases like cancer [3], cardiovascular diseases [4], arteriosclerosis [5], neural disorders [6], skin irritations [7], and even the aging process [8] occur due to accumulation of high levels of free radicals in various tissues.

By donating electrons to rampaging free radicals, antioxidants can protect other molecules from oxidation that can damage vital molecules in cells, including DNA and proteins [9]. Antioxidants cover a wide range of molecules and come from many sources. Some are naturally produced in organisms and some naturally occur in foods, such as tomatoes, red peppers, green tea, apples, grapes, blueberries, etc.

It has been reported that high intakes of antioxidants would extend life expectancy by reducing chronic degenerative diseases like cancer and cardiovascular diseases and perhaps by slowing the aging process [8]. Therefore, it is necessary to identify antioxidants from various resources, which will be helpful for treating these diseases.

Although experimental methods are objective to identify antioxidants [10, 11], they are time-consuming and expensive to detect antioxidants from the avalanche of protein sequences generated in the post genomics era. As complements to experimental methods, computational methods are urgently needed to accurately identify antioxidants.

In recent years, several computational methods have been proposed to identify antioxidants. However, it was found that these methods were trained based on different datasets and evaluated by different cross validation method. To give researchers a catching-up view about the development in this area, we summarized recent advances in machine learning methods on antioxidant identification.

2. RESOURCES OF ANTIOXIDANTS

2.1. Database

The Antioxidant Database (AOD) is a manually curated database depositing experimentally validated antioxidants [12]. At present, the AOD is available at http://lin-group.cn/AODdatabase/ and includes 710 antioxidants. Besides the protein sequence, information on taxonomy, source organism, subcellular location, gene ontology, catalytic activity and function of each proteins was also provided in AOD [12].

2.2. Benchmark Dataset

A high quality benchmark dataset is the key point of developing and validating computational methods for antioxidant identification. By searching multiple protein databases, Fernandez-Blanco *et al.* constructed the first dataset for computationally identifying antioxidant, which contains 324 antioxidants and 1657 non-antioxidant [13]. However, some of the sequences in this dataset share 100% sequence identity. As indicated in a recent review [14], a model if trained and tested on such a redundant dataset, will yield over fitting problem and misleading results.

To overcome this drawback, we built a new benchmark dataset to train computational models for antioxidant identification [15]. We firstly searched the UniProt database [16] and harvested 686 proteins with antioxidant activity. In order to include much more data in the new dataset, 686 proteins were merged with the antioxidants in Fernandez-Blanco *et al.*'s dataset. By winnowing proteins with sequence similarity ≥60% using CD-HIT [17], 253 antioxidants and 1552 non-antioxidants were finally retained in the benchmark dataset, which can be accessed at http://lin-group.cn/server/Aod/data.

3. SEQUENCE ENCODING SCHEME

Since antioxidants in the dataset are with different lengths and couldn't be recognized by machine learning methods, several sequence encoding schemes [18, 19] have been proposed to convert antioxidants into discrete vectors that were further used as inputs of machine learning methods.

[*] Address correspondence to this author at the School of Basic Medical Sciences, Chengdu University of Traditional Chinese Medicine, Chengdu 611730, China; E-mail: fengpengmian@gmail.com

1389-2002/20 $65.00+.00

3.1. Star Graph Topological Indices

The main idea of the star graph topological indices (SGTI) is to represent sequences as graphs by topological indices [20]. By using the Sequence to the Star Networks (S2SNet) [21], Fernandez-Blanco *et al.* converted the sequences into embedded/non-embedded SGTI including the trace of connectivity matrices, Harary number, Wiener index, Gutman index, Schultz index, Moreau-Broto indices, Balaban distance connectivity index, Kier-Hall connectivity indices and Randic connectivity index [13]. Based on SGTI, both antioxidants and non-antioxidants in the dataset were transferred into a 42 dimensional vector. More details about SGTI can be referred to Fernandez-Blanco *et al.*'s work [13].

3.2. Amino acid Composition

Amino acid composition (AAC) is the most straightforward sequence encoding method [22]. By using AAC, a protein sequence will be converted into a 20 dimensional discrete vector, in which each element indicates the occurrence frequency of the 20 natural amino acids in the sequence.

3.3. Dipeptide Composition

Compared with AAC, the dipeptide composition (DPC) integrates the local order information in a sequence and is the occurrence frequency of two proximate amino acids in the sequence [23-25]. By using DPC, a protein sequence will be converted into a 400 dimensional vector.

3.4. G-Gap Dipeptide Composition

Although the *k*-tuple amino acid composition contains global correlation information, the dimension of the vector will increase with the increment of *k*. To deal with this problem, the *g*-gap dipeptide composition (GDC) method was proposed to describe the long-range correlations between amino acids in a sequence [26]. By using GDC, a protein sequence will be represented by a 400 dimensional vector. Different from DPC, each element in the vector generated based on GDC represents the occurrence frequency of dipeptide with *g* amino acids interval [27-29].

Generally speaking, *g* is an integral number and ranges from 0 to 10. g=0 indicates the correlation of two proximate amino acids; g=1 is the correlation between two amino acids with the interval of one amino acid; g=2 is the correlation between two amino acids with the interval of two amino acids, and so forth.

3.5. 188D

The 188D feature was proposed by Cai *et al.* to represent protein sequences [30], which include the information derived from both amino acid composition and their physicochemical properties (i.e. secondary structure, solvent accessibility, normalized Van der Waals volume, hydrophobicity, charge, polarity, polarizability, and surface tension) [30]. Detail information about 188D feature can be found in Cai *et al.*'s work [30]. It was widely employed in recent bioinformatics works [31-33].

3.6. Secondary Structure Features (SSF)

Considering the fact that the function of a protein is closely correlated with its structure, the secondary structure features (SSF) was used to encode proteins. The secondary structure of a protein can be obtained by using PSI-PRED [34], and each amino acid will be assigned a secondary structure state, namely helix (H), strand (E) and coil (C), respectively. Therefore, a protein sequence with the length of L can be transferred into a new sequence encoded by H/E/C. The following features derived based on SSI was defined to describe proteins in the realm of computational proteomics [35].

(1) By counting the number of helix (H), strand (E) and coil (C), their content information S_i was defined as shown Eq. 1, which describes the frequency of H, E and C appeared in the sequence.

$$S_i = \frac{n_i}{L} \quad i \in \{H, E, C\} \tag{1}$$

where n_j is the number of secondary structure H/E/C.

(2) Transition information of H/E/C is defined by the following equation.

$$T_{ij} = \frac{n_{ij}}{L-1} \quad i, j \in \{H, E, C\} \tag{2}$$

where n_{ij} is the number of secondary structure state *i* and *j* of the neighboring amino acids in the sequence.

(3) The average length and the normalized maximal length of the secondary structure segment are as follows, respectively.

$$\overline{Seg(i)} = \frac{Len(Seg(i))}{\sum n_{Seg(i)}} \quad i \in \{H, E, C\} \tag{3}$$

$$Max_{Seg(i)} = \frac{Max(Seg(i))}{L} \quad i \in \{H, E, C\} \tag{4}$$

where *Seg*(*i*) is the segment containing the secondary structure *i* (*i*=H, E, C), *Len*(*Seg*(*i*)) is the length of *Seg*(*i*), $n_{Seg(i)}$ is the number of *Seg*(*i*) in the sequence, and *Max*(*Seg*(*i*)) is the maximum length of *Seg*(*i*).

(4) In order to reflect the special arrangements of the secondary structure elements, the position related secondary structure content was defined a following,

$$F_i = \frac{\sum_{j=1}^{n_i} p_{ij}}{L(L-1)} \quad i \in \{H, E, C\} \tag{5}$$

where n_i is the number of secondary structure H/E/C, p_{ij} indicates the position of the *j*-th order secondary structure.

3.7. Relative Solvent Accessibility (RSA)

Since the solvent accessibility of a protein is closely correlated with its antioxidant activity, Zhang *et al.* proposed a new method to represent antioxidants by using relative solvent accessibility (RSA) [36]. The 28 features based on RSA can be calculated by using the software PaleAle 4.0 [37], which are described in Zhang *et al.*'s work.

3.8. Composition, Transition, Distribution (CTD)

In order to globally describe protein sequences, Dubchak *et al.* proposed three global descriptors [38], namely composition (C), transition (T) and distribution (D), based on the amino acid distribution of a specific structural or physicochemical property. The composition descriptor C is the global percent composition of 20 native amino acids. The transition descriptor T is the frequency with amino acids of one type of native amino acids followed by another type. The distribution descriptor D indicates the respective locations of the first, 25%, 50%, 75% and 100% of each type of 20 native amino acids. The CTD descriptors have been included in several recent software [39, 40].

3.9. Position-Specific Score Matrix (PSSM)

Insertion/deletion (indel) and mutation are accumulated during protein evolution, which indeed reduces the protein sequence similarity. However, homologous proteins still share similar structures and functions. The position specific score matrix (PSSM) was adopted to obtain evolutionary essential signatures of protein sequences [41-43]. The PSSM is a matrix including L*20 elements and is generated by PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool) [44]. The elements in the *i*th row of the matrix denote the probabilities of the *i*th residue of the given protein sequence mutating to 20 native amino acids during the evolution process.

Considering the length variation of different proteins, in order to formulate different proteins into feature vectors with the same dimension, the original PSSM was further transformed into a 20×20 dimensional matrix.

4. EVALUATION METRICS

Both 10-fold cross validation test and jackknife test have been used to examine the quality of a predictor for identifying antioxidants. In the 10-fold cross validation test, the samples in the dataset was divided into 10 sub-sets, 9 of them are used for training, while the left one is used for testing. The process will be repeated 10 times. In the jackknife test, each antioxidant in the training dataset is in turn singled out as an independent test one and all the properties are calculated without including the one being identified. The process was repeated until all the samples were used as the independent test one.

The sensitivity (Sn), specificity (Sp), accuracy (Acc) and Mathew's correlation coefficient (MCC) and F1 score have been introduced to evaluate the performance of the proposed methods for identifying antioxidants, which are defined as follows [45-47].

$$\begin{cases} \text{Sn} = \dfrac{TP}{TP+FN} \\ \text{Sp} = \dfrac{TN}{TN+FP} \\ \text{Acc} = \dfrac{TP+TN}{TN+FP+FN+TP} \\ F1 = 2\times\dfrac{\dfrac{TP}{TP+FP}\times\dfrac{TP}{TP+FN}}{\dfrac{TP}{TP+FP}+\dfrac{TP}{TP+FN}} \\ \text{MCC} = \dfrac{(TP\times TN - FP\times FN)}{\sqrt{(TP+FN)\times(TP+FP)\times(TN+FP)\times(TN+FN)}} \end{cases} \tag{6}$$

where TP, TN, FP and FN represent true positive, true negative, false positive and false negative, respectively. In addition, the area under the receiver operating characteristic curve (AUC) [48, 49] was also used to objectively evaluate the performance of the predictors. The AUC value of 0.5 is a random prediction, while a value of 1 is a perfect prediction.

5. METHODS FOR IDENTIFYING ACP

In the past several years, at least 9 machine learning based methods have been proposed for identifying antioxidants, which are listed in Table **1**.

In the following section, we will briefly introduce these methods in terms of the used machine learning method and the obtained accuracies for identifying antioxidants.

In 2013, Fernandez-Blanco *et al*. [13] reported the first computational method to identify antioxidants, in which the protein sequences were encoded by using the SGTI method. By using the Random Forest as the classification engine, in the 10 fold cross validation test, an accuracy of 94.6% was obtained for identifying antioxidants. However, there are redundant sequences in the dataset used by Fernandez-Blanco *et al*., which makes their results incredible.

Later on, Feng *et al.* [15] developed a Native Bayes (NB) model to identify antioxidants. Different from Fernandez-Blanco *et al.*'s work [13], Feng *et al.* constructed a high quality benchmark dataset with the sequence similarity less than 60%. By encoding the protein sequences using amino acid and dipeptide composition, an accuracy of 55.85% with AUC of 0.68 was obtained for identifying antioxidants in the jackknife test. To further improve the model's accuracy, the Correlation-based Feature Selection combined with Best First Search strategy was used to perform feature selection. Finally, 44 optimal features were selected out from the 420 original features and used as the input of NB. In the jackknife test, the final method obtained an accuracy of 66.89% with the AUC of 0.768 for identifying antioxidants in the benchmark dataset. In addition, the model was also evaluated on an independent dataset containing 20 antioxidants and accurately identified 16 antioxidants.

In 2015, inspired by Feng *et al.*'s work, Zhang and her colleagues proposed a Random Forest (RF) based method to identify antioxidants [50]. In their model, the antioxidants were encoded by using the *g*-gap (*g*=4) dipeptide composition and PSSM. In order to improve the predictive accuracy, the Information Gain combined with Incremental Feature Selection (IG-IFS) was used to select optimal features for building the computational model. In the 10-fold cross validation test, an accuracy of 90% with AUC of 0.94 was obtained for identifying antioxidants. It should be pointed out that the Zhang *et al*'s training dataset is a balanced one and includes 100 antioxidant proteins and 100 non-antioxidant proteins, which is smaller than Feng *et al.*'s one that includes 253 antioxidants and 1552 non-antioxidants.

Based on the high quality benchmark dataset built in 2013, Feng and her colleagues proposed a support vector machine (SVM) based model, called AodPred, for identifying antioxidant proteins [51]. In order to include the global sequence order information, the optimal *g*-gap (*g*=3) dipeptide composition obtained by using the analysis of variance (ANOVA) feature selection method was used to encode the sequences in the dataset. In the jackknife test, an accuracy of 74.79% was obtained for identifying antioxidants in the dataset. Moreover, a user friendly webserver was provided at http://lin-group.cn/server/AntioxiPred, which was the first online tool for identifying antioxidants at that time.

Later on, based on the same dataset as that used in previous work, Zhang *et al*. proposed an ensemble based method [35]. The Relief combined with Incremental Feature Selection (IFS) method was used to select optimal features from the hybrid features including SSF, PSSM, RSA and CTD. Based on the optimal features, an ensemble classifier including RF, SMO, NNA and J48 obtained the best accuracy of 94% with AUC of 0.978 for identifying antioxidants in the dataset, which is better than that reported in their previous work. A web-server for the proposed method was provided at http://antioxidant.weka.cc. However, it was found that the webserver is out of service at present.

On the basis of Feng *et al.*'s dataset, Xiao *et al*. applied the AAC, PSSM and 60 features obtained by using grey system model to represent the sequences in the dataset [52]. Considering the unbalanced number between positive and negative samples, they designed a voting model which consists of eleven Random Forest based sub-predictors. In the 10-fold cross validation test, they obtained an accuracy of 88.25% with the AUC of 0.935 for identifying antioxidants in the dataset.

In 2018, Xu *et al.* [53] developed the sequence based support vector machine method to identify antioxidants in the dataset constructed by Feng *et al*. They applied the synthetic minority oversampling technique (SMOTE) method to overcome the class imbalance problem. The maximum relevance maximum distance (MRMD) [54, 55] method was used to select optimal features from the 188D feature. Finally, an SVM based model called SeqSVM was proposed and obtained an accuracy of 89.46% in the jackknife test.

With the wide application of deep learning in bioinformatics [56-60], it has also been used to identify antioxidants. In 2018, Shao *et al*. proposed a deep learning based classifier, called IDAod, to identify antioxidants [61]. Different from the above mentioned methods, the deep autoencoder and full connect neural network were used to extract features from the mixed *g*-gap (*g*=0 and 1) dipeptide compositions. In the 10 fold cross validation test, an F1 score of 0.8842 with the accuracy of 97.05% was obtained for iden-

Table 1. Summary of existing tools for identifying antioxidants.

Methods	Algorithm	Webserver	Year
Fernandez-Blanco's	RF	Not provided	2012
Feng's	NB	Not provided	2013
Zhang's	RF	Not provided	2015
AodPred	SVM	http://lin-group.cn/server/AntioxiPred	2015
AOP-Pred	Ensemble Classifiers	http://antioxidant.weka.cc	2016
iANOP-Enble	RF	Not provided	2017
IDAod	Deep Learning	http://bigroup.uestc.edu.cn/IDAod/	2018
SeqSVM	SVM	Not provided	2018
AOPs-SVM	SVM	http://server.malab.cn/AOPs-SVM/	2019

tifying antioxidants in the dataset. For the convenience of scientific community, an online webserver was developed for IDAod, which is available at http://bigroup.uestc.edu.cn/IDAod/.

More recently, Meng *et al*. developed a powerful predictor for identifying antioxidants [62]. They firstly extracted 473 features from AAC, DPC, PSSM, and SSI, and then applied the MRMD method to select optimal features. Finally, 176 optimal features were obtained and were used as the input of SVM. Accordingly, a predictor called AOPs-SVM was proposed and obtained an accuracy of 94.2% with the AUC of 0.832 in the jackknife test. At present, AOPs-SVM performs the best for identifying antioxidants. A webserver was also established and could be freely accessible at http://server.malab.cn/AOPs-SVM/index.jsp.

6. CHALLENGES AND PERSPECTIVES

By neutralizing free radicals, antioxidants can prevent cells from further damage or death. The potential of antioxidants in prevention of diseases has also been reported in recent years. Therefore, accurate identification of antioxidants will pave the ways to speed up the researches on antioxidants.

It is exciting that several computational methods have been proposed to identify antioxidants. Although these works promoted researches on antioxidants and facilitated the identification of antioxidants, the following challenges should be considered in future works.

Although the predictor AOPs-SVM obtained the highest accuracy of 94.2%, its sensitivity is less than 70%. This is due to the following fact. Both AOPs-SVM and the other existing methods were all trained based on an imbalanced dataset with the ratio of positive to negative samples approximately 1:6. To solve this problem, it is necessary to collect much more molecules with antioxidant activity to enlarge the number of antioxidants in the dataset.

The function of a protein is correlated with its structure. Although the secondary structure information has been used to encode proteins in antioxidant identification, the structure status for each amino acid was predicted by using PSI-PRED. If the secondary structure is not correctly predicted by PSI-PRED, it will provide wrong information for further analysis in the models that encode antioxidants by using secondary structure information. The reduced amino acid alphabet (RAAA) [63, 64] can effectively extract information of structurally conserved regions and structural similarity of proteins. Therefore, it is necessary to integrate RAAA features when building models in future work.

Antioxidants can be classified into two main classes according to the mechanisms they interrupt the overall oxidation process, namely chain breaking and preventive mechanisms. However, the existing computational methods couldn't distinguish these two kinds of antioxidants. To address such a challenge, more efforts should be made to develop new methods for classifying these two kinds of antioxidants.

CONCLUSION

In this paper, we reviewed recent advances in the application of machine learning methods for identifying antioxidants. The challenges and future perspectives were also discussed. We hope this review will provide insights into researches on antioxidants.

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Lobo, V.; Patil, A.; Phatak, A.; Chandra, N. Free radicals, antioxidants and functional foods: Impact on human health. *Pharmacogn. Rev.,* **2010**, *4*(8), 118-126. http://dx.doi.org/10.4103/0973-7847.70902 PMID: 22228951

[2] Hajhashemi, V.; Vaseghi, G.; Pourfarzam, M.; Abdollahi, A. Are antioxidants helpful for disease prevention? *Res. Pharm. Sci.,* **2010**, *5*(1), 1-8. PMID: 21589762

[3] Pham-Huy, L.A.; He, H.; Pham-Huy, C. Free radicals, antioxidants in disease and health. *Int. J. Biomed. Sci.,* **2008**, *4*(2), 89-96. PMID: 23675073

[4] Jain, A.K.; Mehra, N.K.; Swarnakar, N.K. Role of Antioxidants for the Treatment of Cardiovascular Diseases: Challenges and Opportunities. *Curr. Pharm. Des.,* **2015**, *21*(30), 4441-4455. http://dx.doi.org/10.2174/1381612821666150803151758 PMID: 26234792

[5] Toledo-Ibelles, P.; Mas-Oliva, J. Antioxidants in the Fight Against Atherosclerosis: Is This a Dead End? *Curr. Atheroscler. Rep.,* **2018**, *20*(7), 36.

http://dx.doi.org/10.1007/s11883-018-0737-7 PMID: 29781062

[6] Carvalho, A.N.; Firuzi, O.; Gama, M.J.; Horssen, J.V.; Saso, L. Oxidative Stress and Antioxidants in Neurological Diseases: Is There Still Hope? *Curr. Drug Targets,* **2017**, *18*(6), 705-718. http://dx.doi.org/10.2174/1389450117666160401120514 PMID: 27033198

[7] Pai, V.V.; Shukla, P.; Kikkeri, N.N. Antioxidants in dermatology. *Indian Dermatol. Online J.,* **2014**, *5*(2), 210-214. http://dx.doi.org/10.4103/2229-5178.131127 PMID: 24860765

[8] Fusco, D.; Colloca, G.; Lo Monaco, M.R.; Cesari, M. Effects of antioxidant supplementation on the aging process. *Clin. Interv. Aging,* **2007**, *2*(3), 377-387. PMID: 18044188

[9] Sogut, S.; Zoroglu, S. S.; Ozyurt, H.; Yilmaz, H. R.; Ozugurlu, F.; Sivasli, E.; Yetkin, O.; Yanik, M.; Tutkun, H.; Savas, H. A.; Tarakcioglu, M.; Akyol, O. Changes in nitric oxide levels and antioxidant enzyme activities may have a role in the pathophysiological mechanisms involved in autism *Clinica chimica acta; international journal of clinical chemistry,* **2003**, *331*(1-2), 111-7.

[10] Huang, W.; Deng, Q.C.; Xie, B.J.; Shi, J.; Huang, F.H.; Tian, B.Q.; Huang, Q.D.; Xue, S. Purification and characterization of an antioxidant protein from Ginkgo biloba seeds. *Food Res. Int.,* **2010**, *43*(1), 86-94. http://dx.doi.org/10.1016/j.foodres.2009.08.015

[11] Fu, J.; Tang, J.; Wang, Y.; Cui, X.; Yang, Q.; Hong, J.; Li, X.; Li, S.; Chen, Y.; Xue, W.; Zhu, F. Discovery of the Consistently Well-Performed Analysis Chain for SWATH-MS Based Pharmacoproteomic Quantification. *Front. Pharmacol.,* **2018**, *9*, 681. http://dx.doi.org/10.3389/fphar.2018.00681 PMID: 29997509

[12] Feng, P.; Ding, H.; Lin, H.; Chen, W. AOD: the antioxidant protein database. *Sci. Rep.,* **2017**, *7*(1), 7449. http://dx.doi.org/10.1038/s41598-017-08115-6 PMID: 28784999

[13] Fernández-Blanco, E.; Aguiar-Pulido, V.; Munteanu, C.R.; Dorado, J. Random Forest classification based on star graph topological indices for antioxidant proteins. *J. Theor. Biol.,* **2013**, *317*, 331-337. http://dx.doi.org/10.1016/j.jtbi.2012.10.006 PMID: 23116665

[14] Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.,* **2011**, *273*(1), 236-247. http://dx.doi.org/10.1016/j.jtbi.2010.12.024 PMID: 21168420

[15] Feng, P.M.; Lin, H.; Chen, W. Identification of antioxidants from sequence information using naïve Bayes. *Comput. Math. Methods Med.,* **2013**, *2013*567529 http://dx.doi.org/10.1155/2013/567529 PMID: 24062796

[16] UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.,* **2018**, *46*(5), 2699. http://dx.doi.org/10.1093/nar/gky092 PMID: 29425356

[17] Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics,* **2012**, *28*(23), 3150-3152. http://dx.doi.org/10.1093/bioinformatics/bts565 PMID: 23060610

[18] Zhu, X.J.; Feng, C.Q.; Lai, H.Y.; Chen, W.; Lin, H. Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowl. Base. Syst.,* **2019**, *163*, 787-793. http://dx.doi.org/10.1016/j.knosys.2018.10.007

[19] Tan, J.X.; Li, S.H.; Zhang, Z.M.; Chen, C.X.; Chen, W.; Tang, H.; Lin, H. Identification of hormone binding proteins based on machine learning methods. *Math. Biosci. Eng.,* **2019**, *16*(4), 2466-2480. http://dx.doi.org/10.3934/mbe.2019123 PMID: 31137222

[20] Randić, M.; Zupan, J.; Vikić-Topić, D. On representation of proteins by star-like graphs. *J. Mol. Graph. Model.,* **2007**, *26*(1), 290-305. http://dx.doi.org/10.1016/j.jmgm.2006.12.006 PMID: 17223597

[21] Munteanu, C.R.; Magalhães, A.L.; Uriarte, E.; González-Díaz, H. Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices. *J. Theor. Biol.,* **2009**, *257*(2), 303-311. http://dx.doi.org/10.1016/j.jtbi.2008.11.017 PMID: 19111559

[22] Chen, W.; Feng, P.; Liu, T.; Jin, D. Recent advances in machine learning methods for predicting heat shock proteins. *Curr. Drug Metab.,* **2018**. PMID: 30378494

[23] Chen, W.; Feng, P.; Nie, F. iATP: A sequence based method for identifying anti-tubercular peptides. *Med. Chem.,* **2019**. http://dx.doi.org/10.2174/1573406415666191002152441 PMID: 31339073

[24] Ding, H.; Deng, E.Z.; Yuan, L.F.; Liu, L.; Lin, H.; Chen, W.; Chou, K.C. iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. *BioMed Res. Int.,* **2014**, *2014*286419 http://dx.doi.org/10.1155/2014/286419 PMID: 24991545

[25] Ding, H.; Li, D. Identification of mitochondrial proteins of malaria parasite using analysis of variance. *Amino Acids,* **2015**, *47*(2), 329-333. http://dx.doi.org/10.1007/s00726-014-1862-4 PMID: 25385313

[26] Chen, W.; Nie, F.; Ding, H. Recent advances of computational methods for identifying bacteriophage virion proteins. *Protein Pept. Lett.,* **2019**. PMID: 30968770

[27] Feng, P.M.; Ding, H.; Chen, W.; Lin, H. Naïve Bayes classifier with feature selection to identify phage virion proteins. *Comput. Math. Methods Med.,* **2013**, *2013*530696 http://dx.doi.org/10.1155/2013/530696 PMID: 23762187

[28] Ding, H.; Feng, P.M.; Chen, W.; Lin, H. Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. *Mol. Biosyst.,* **2014**, *10*(8), 2229-2235. http://dx.doi.org/10.1039/C4MB00316K PMID: 24931825

[29] Yang, W.; Zhu, X.J.; Huang, J.; Ding, H.; Lin, H. A brief survey of machine learning methods in protein sub-Golgi localization. *Curr. Bioinform.,* **2019**, *14*, 234-240. http://dx.doi.org/10.2174/1574893613666181113131415

[30] Cai, C.Z.; Han, L.Y.; Ji, Z.L.; Chen, X.; Chen, Y.Z. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.,* **2003**, *31*(13), 3692-3697. http://dx.doi.org/10.1093/nar/gkg600 PMID: 12824396

[31] Lv, Z.; Jin, S.; Ding, H.; Zou, Q. A random forest sub-Golgi protein classifier optimized via dipeptide and amino acid composition features. *Front. Bioeng. Biotechnol.,* **2019**, *7*, 215. http://dx.doi.org/10.3389/fbioe.2019.00215 PMID: 31552241

[32] Chao, L.; Wei, L.; Zou, Q. SecProMTB: A SVM-based Classifier for Secretory Proteins of Mycobacterium tuberculosis with Imbalanced Data Set. *Proteomics,* **2019**, *19*e1900007 http://dx.doi.org/10.1002/pmic.201900007

[33] Song, L.; Li, D.; Zeng, X.; Wu, Y.; Guo, L.; Zou, Q. nDNA-Prot: identification of DNA-binding proteins based on unbalanced classification. *BMC Bioinformatics,* **2014**, *15*, 298. http://dx.doi.org/10.1186/1471-2105-15-298 PMID: 25196432

[34] Buchan, D.W.A.; Jones, D.T. The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.,* **2019**, *47*(W1), W402-W407. http://dx.doi.org/10.1093/nar/gkz297 PMID: 31251384

[35] Zhang, L.; Zhang, C.; Gao, R.; Yang, R.; Song, Q. Sequence Based Prediction of Antioxidant Proteins Using a Classifier Selection Strategy. *PLoS One,* **2016**, *11*(9)e0163274 http://dx.doi.org/10.1371/journal.pone.0163274 PMID: 27662651

[36] Ehrlich, L.; Reczko, M.; Bohr, H.; Wade, R.C. Prediction of protein hydration sites from sequence by modular neural networks. *Protein Eng.,* **1998**, *11*(1), 11-19. http://dx.doi.org/10.1093/protein/11.1.11 PMID: 9579655

[37] Mirabello, C.; Pollastri, G. Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. *Bioinformatics,* **2013**, *29*(16), 2056-2058. http://dx.doi.org/10.1093/bioinformatics/btt344 PMID: 23772049

[38] Dubchak, I.; Muchnik, I.; Holbrook, S.R.; Kim, S.H. Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. USA,* **1995**, *92*(19), 8700-8704. http://dx.doi.org/10.1073/pnas.92.19.8700 PMID: 7568000

[39] Du, P.F.; Zhao, W.; Miao, Y.Y.; Wei, L.Y.; Wang, L. UltraPse: A Universal and Extensible Software Platform for Representing Biological Sequences. *Int. J. Mol. Sci.,* **2017**, *18*(11)E2400 http://dx.doi.org/10.3390/ijms18112400 PMID: 29135934

[40] Wang, J.; Du, P.F.; Xue, X.Y.; Li, G.P.; Zhou, Y.K.; Zhao, W.; Lin, H.; Chen, W. VisFeature: a stand-alone program for visualizing and analyzing statistical features of biological sequences. *Bioinformatics,* **2019**. http://dx.doi.org/10.1093/bioinformatics/btz689 PMID: 31504195

[41] Xiong, Y.; Wang, Q.; Yang, J.; Zhu, X.; Wei, D.Q. PredT4SE-Stack: Prediction of Bacterial Type IV Secreted Effectors From Protein Sequences Using a Stacked Ensemble Method. *Front. Microbiol.,* **2018**, *9*, 2571.

http://dx.doi.org/10.3389/fmicb.2018.02571 PMID: 30416498

[42] Jiao, Y.S.; Du, P.F. Predicting protein submitochondrial locations by incorporating the positional-specific physicochemical properties into Chou's general pseudo-amino acid compositions. *J. Theor. Biol.,* **2017**, *416*, 81-87.
http://dx.doi.org/10.1016/j.jtbi.2016.12.026 PMID: 28077336

[43] Zhao, W.; Li, G.P.; Wang, J.; Zhou, Y.K.; Gao, Y.; Du, P.F. Predicting protein sub-Golgi locations by combining functional domain enrichment scores with pseudo-amino acid compositions. *J. Theor. Biol.,* **2019**, *473*, 38-43.
http://dx.doi.org/10.1016/j.jtbi.2019.04.025 PMID: 31051179

[44] Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.,* **1997**, *25*(17), 3389-3402.
http://dx.doi.org/10.1093/nar/25.17.3389 PMID: 9254694

[45] Chen, W.; Feng, P.; Song, X.; Lv, H.; Lin, H. iRNA-m7G: Identifying N^7-methylguanosine Sites by Fusing Multiple Features. *Mol. Ther. Nucleic Acids,* **2019**, *18*, 269-274.
http://dx.doi.org/10.1016/j.omtn.2019.08.022 PMID: 31581051

[46] Chen, W.; Lv, H.; Nie, F.; Lin, H. i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics,* **2019**, *35*(16), 2796-2800.
http://dx.doi.org/10.1093/bioinformatics/btz015 PMID: 30624619

[47] Jiao, Y.; Du, P. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quant. Biol.,* **2016**, *4*(4), 320-330.
http://dx.doi.org/10.1007/s40484-016-0081-2

[48] Chen, W.; Yang, H.; Feng, P.; Ding, H.; Lin, H. iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics,* **2017**, *33*(22), 3518-3523.
http://dx.doi.org/10.1093/bioinformatics/btx479 PMID: 28961687

[49] Zhou, Y.-K.; Shen, Z.-A.; Yu, H.; Luo, T.; Gao, Y.; Du, P.-F. *Predicting lncRNA–Protein Interactions With miRNAs as Mediators in a Heterogeneous Network Model.,* **2020**, *10*(1341)
http://dx.doi.org/10.3389/fgene.2019.01341

[50] Zhang, L.N.; Zhang, C.J.; Gao, R.; Yang, R.T. Incorporating g-Gap Dipeptide Composition and Position Specific Scoring Matrix for Identifying Antioxidant Proteins. *Proceeding of the IEEE 28th Canadian Conference on Electrical and Computer Engineering,* Halifax, Canada**2015**.
http://dx.doi.org/10.1109/CCECE.2015.7129155

[51] Feng, P.; Chen, W.; Lin, H. Identifying Antioxidant Proteins by Using Optimal Dipeptide Compositions. *Interdiscip. Sci.,* **2016**, *8*(2), 186-191.
http://dx.doi.org/10.1007/s12539-015-0124-9 PMID: 26345449

[52] Xiao, X.; Ju, W.F.; Hui, M.J. In iANOP-Enble: a sequence-based ensemble classifier for identifying antioxidant proteins by PseAAC and Random Forests *2nd International Conference on Automation, Mechanical Control and Computational Engineering (AMCCE 2017),* **2017**, pp. 587-593.
http://dx.doi.org/10.2991/amcce-17.2017.103

[53] Xu, L.; Liang, G.; Shi, S.; Liao, C. SeqSVM: A Sequence-Based Support Vector Machine Method for Identifying Antioxidant Proteins. *Int. J. Mol. Sci.,* **2018**, *19*(6)E1773
http://dx.doi.org/10.3390/ijms19061773 PMID: 29914044

[54] Zou, Q.; Zeng, J.C.; Cao, L.J.; Ji, R.R. A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing,* **2016**, *173*, 346-354.
http://dx.doi.org/10.1016/j.neucom.2014.12.123

[55] Zou, Q.; Wan, S.; Ju, Y.; Tang, J.; Zeng, X. Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. *BMC Syst. Biol.,* **2016**, *10*(4)(Suppl. 4), 114.
http://dx.doi.org/10.1186/s12918-016-0353-5 PMID: 28155714

[56] Wei, L.; Ding, Y.; Su, R.; Tang, J.; Zou, Q. Prediction of human protein subcellular localization using deep learning. *J. Parallel Distrib. Comput.,* **2018**, *117*, 212-217.
http://dx.doi.org/10.1016/j.jpdc.2017.08.009

[57] Yu, L.; Sun, X.; Tian, S.W.; Shi, X.Y.; Yan, Y.L. Drug and Nondrug Classification Based on Deep Learning with Various Feature Selection Strategies. *Curr. Bioinform.,* **2018**, *13*(3), 253-259.
http://dx.doi.org/10.2174/1574893612666170125124538

[58] Peng, L.; Peng, M.M.; Liao, B.; Huang, G.H.; Li, W.B.; Xie, D.F. The Advances and Challenges of Deep Learning Application in Biological Big Data Processing. *Curr. Bioinform.,* **2018**, *13*(4), 352-359.
http://dx.doi.org/10.2174/1574893612666170707095707

[59] Nie, L.L.; Deng, L.; Fan, C.; Zhan, W.H.; Tang, Y.J. Prediction of Protein S-Sulfenylation Sites Using a Deep Belief Network. *Curr. Bioinform.,* **2018**, *13*(5), 461-467.
http://dx.doi.org/10.2174/1574893612666171122152208

[60] Lv, Z.; Ao, C.; Zou, Q. Protein Function Prediction: From Traditional Classifier to Deep Learning. *Proteomics,* **2019**, *19*(14)e1900119
http://dx.doi.org/10.1002/pmic.201900119 PMID: 31187588

[61] Shao, L.; Gao, H.; Liu, Z.; Feng, J.; Tang, L.; Lin, H. Identification of Antioxidant Proteins With Deep Learning From Sequence Information. *Front. Pharmacol.,* **2018**, *9*, 1036.
http://dx.doi.org/10.3389/fphar.2018.01036 PMID: 30294271

[62] Meng, C.; Jin, S.; Wang, L.; Guo, F.; Zou, Q. AOPs-SVM: A Sequence-Based Classifier of Antioxidant Proteins Using a Support Vector Machine. *Front. Bioeng. Biotechnol.,* **2019**, *7*, 224.
http://dx.doi.org/10.3389/fbioe.2019.00224 PMID: 31620433

[63] Feng, P.M.; Chen, W.; Lin, H.; Chou, K.C. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem.,* **2013**, *442*(1), 118-125.
http://dx.doi.org/10.1016/j.ab.2013.05.024 PMID: 23756733

[64] Chen, W.; Feng, P.; Liu, T.; Jin, D. Recent Advances in Machine Learning Methods for Predicting Heat Shock Proteins. *Curr. Drug Metab.,* **2019**, *20*(3), 224-228.
http://dx.doi.org/10.2174/1389200219666181031105916 PMID: 30378494