iANOP-Enble: a sequence-based ensemble classifier for identifying antioxidant proteins by PseAAC and Random Forests

Xuan Xiao[1, a], Weifeng Ju[1, b], Mengjuan Hui[1, c]

[1]Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, 333403, China

[a]email: jdzxiaoxuan@163.com, [b]email: 1832015504@qq.com, [c]email: huimengjuan@163.com

Keywords: Antioxidant proteins; Voting system; Ensemble classifier

Abstract. Inoxidizability of proteins is one of the most basic function attribute, and shares a sustainable effect for biological process in protein repair and regulate redox-sensitive signaling pathways. In the genome era, however, it is urgent to design an effective computation method to rapidly detect the antioxidant proteins based on sequence information due to the addition of the larger amount of sequence. We designed a novel automations computational algorithm named "iANOP-Enble". In this predictor, the protein sample was formulated by protein similarity scores matrix and amino acid prosperities information into Random Forest. The process of the new predictor algorithm to identify antioxidants protein is designed as a voting system, which consists of eleven sub-classifiers. In order to verify our algorithm availabilities, we adopted a fair comparison method that used the same bench data set. Finally, the result shows that our algorithm is more promising than existing method on the basis of the same standard of comparison

Introduction

Antioxidant proteins are low concentration substance which have certain effect on cells, especially in inhibiting the oxidation of free radicals reaction[1]. The mechanism of the antioxidant have direct effects on free radical, or indirectly consumed material. Research proves that the body's antioxidant system[2] is a perfect and complex function system which is similar to the immune system[3].

In recent decades, researches show that Antioxidants protein can help prevent some diseases such as cancer, or even coronary heart disease in some cases. For an unknown protein sequence, existing bioexperiment can succeed in estimating whether it is antioxidant protein or not. However, in the genome era with the addition of the larger mount of sequence, many researchers have designed a lot of effective computational approaches to identify the antioxidants proteins based on some protein properties, and achieved some success.

In the field of molecular biology, it is pretty common knowledge that for a protein sequence information, it is insufficient to simply consider the appearing frequency information of single amino acid. Therefore, in order to display the sequence order information, previous studies have developed an effective method which doesn't directly use protein amino acid[4], but use protein sequence properties or amino acid properties to represent the whole sequence information. This method can represent a part of sequence order information. Since PseAAC was proposed, it has been successfully used to predict some biological problems.

The aim of this study is to design a voting system to identify antioxidants protein based on the protein similarity scores matrix and amino acid appear frequency information.

Material and Methods

In order to design an effective algorithm, the construction of dataset, which can be used to train and test our algorithm model, is the key step. In this study, we collected the validated dataset from the publication papers [5], which contained 254 (positive examples) antioxidant proteins and 1567 (negative examples) non antioxidant proteins. The dataset is described by

$$\mathbb{S} = \mathbb{S}^+ \cup \mathbb{S}^- \tag{1}$$

In this formula, $\mathbb{S}^+$ means the union of contain antioxidant proteins only, while the $\mathbb{S}^-$ means the union of contain non antioxidant proteins only, and the symbol $\cup$ means union.

For an unknown protein sequence P can been described as

$$P = R_1R_2R_3R_4R_5R_6 \cdots R_L \tag{2}$$

where R_1 means the first amino acid residue, R_2 means the second amino acids residue, and so forth.

Unfortunately, although many existing kinds of straight forward discrete feature models can simply describe samples features, they can not describe an unknown sample **P** which have the sequence evolutionary information that can reflect the characteristics of the sequence. Thus, in order to overcome abovementioned problem, various discrete models were proposed to describe effective features of protein sequence. The purpose of the motivation is to perfect the predictor algorithm performance. For the mount of discrete feature model, the usual description is its amino acid information [6] and protein sequence attribute[7].

Usually, some sequence essence information, especially order information, may be lost if amino acid compostion vector is used to construct a protein sample mathematical modeling. The lost of order information will cause poor performance of algorithm. The problems can be understood as how to construct an effective feature model, and whether the algorithm performance can be improved. This is because the existing classifier algorithm, such as neural network, (SVM) [8], random forest [9], K-nearest neighbor (KNN)[10], can highly effectively deal with biological problems.

Based on the previous study [11], a protein sequence:

$$P = \left[\psi_1 \ \psi_2 \ \cdots \ \psi_u \ \cdots \ \psi_\Omega\right]^T \tag{3}$$

Next, we will introduce how to extract those parameters.

In this paper, the protein sequence properties or amino acid properties were selected to extract protein sequence features, which means two different methods of amino acid composition and Grey Position Specific Scoring Matrix are used to represent the elements in **Eq. 3**.

The AAC information can extract a protein sequences into 20-D features vectors The 20D elements, which means 20 different amino acids frequency in protein sequence. So the twenty different frequency elements in Eq.3 can be given as

$$\psi_i = f_i^{(1)} \quad (i = 1, 2, \cdots, 20) \tag{4}$$

However, all the sequence order information would be lost if only AAC information features are used. In order to solve the problem, we need to consider the order information and the next step will describe how to overcome the shortcoming.

In order to make our feature extraction method meaningful and reserve the original sequence features, in the next process, we will take sequence evolutionary information into consideration in modeling the sequence features. By using the BLAST [12] software, we can obtain a query protein sample evolution information, PSSM matrix, which is a 20×L matrix, it can be written as

$$P_{PSSM}^{(0)} = \begin{bmatrix} \Xi^0_{1\to1} & \Xi^0_{1\to2} & \cdots & \Xi^0_{1\to20} \\ \Xi^0_{2\to1} & \Xi^0_{2\to2} & \cdots & \Xi^0_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \Xi^0_{L\to1} & \Xi^0_{L\to2} & \cdots & \Xi^0_{L\to20} \end{bmatrix} \tag{5}$$

where $\Xi^0_{i\to j}$ means evolution score, that is for each of the sequence position that amino acids turned into the j-th (j = 1, 2,…, 20) types amino acid. The scores matrix was in accordance with using PSI-BLASR software to research the similar protein sequence in the UniProtKB/Swiss-Prot database. In order to keep the score in the same ranges, we use the standard sigmoid function to normalize the original score, given by[13]

$$P_{PSSM}^{(1)} = \begin{bmatrix} \Xi^1_{1\to1} & \Xi^1_{1\to2} & \cdots & \Xi^1_{1\to20} \\ \Xi^1_{2\to1} & \Xi^1_{2\to2} & \cdots & \Xi^1_{2\to20} \\ \vdots & \vdots & \vdots & \vdots \\ \Xi^1_{L\to1} & \Xi^1_{L\to2} & \cdots & \Xi^1_{L\to20} \end{bmatrix} \tag{6}$$

where

$$\Xi^1_{i\to j} = \frac{1}{1+e^{-\Xi^0_{i\to j}}} \quad (1 \le i \le L, 1 \le j \le 20) \tag{7}$$

Next, we use following step to analyze the key information from **Eq.6**, in the Eq.3 in the front 40 elements except the amino acid components, so that the other elements can be defined as

$$\psi_{j+20} = \omega_j \quad (j = 1, 2, \cdots, 20) \tag{8}$$

where

$$\omega_j = \frac{1}{L} \times \sum\nolimits_{K=1}^{L} \Xi^1_{K\to j} \quad (j = 1, 2, \cdots, 20) \tag{9}$$

And, in order to obtain another 60D vectors of **Eq.3** , a named grey system model, method [14] was used in this study

$$\psi_{j+40} = \mu_j \quad (j = 1, 2, \cdots, 60) \tag{10}$$

where

$$\begin{cases} \mu_{3j-2} = w_1 f_j^{(1)} a_1^j \\ \mu_{3j-1} = w_2 f_j^{(1)} a_2^j \quad (j = 1, 2, \cdots, 20) \\ \mu_{3j} \quad = w_3 f_j^{(1)} b^j \end{cases} \tag{11}$$

In Eq.11, w_1, w_2, w_3 means the weight factors, $a_1^j \quad a_2^j \quad b^j$ can be formulated as

$$\begin{bmatrix} a_1^j \\ a_2^j \\ b^j \end{bmatrix} = (H_j^T H_j)^{-1} H_j^T U_j \quad (j = 1, 2, \cdots, 20) \tag{12}$$

where

$$H_j = \begin{bmatrix} -\Xi^1_{2\to j} & -(\Xi^1_{1\to j} + 0.5\Xi^1_{2\to j}) & 1 \\ -\Xi^1_{3\to j} & -(\sum_{i=1}^{2} \Xi^1_{i\to j} + 0.5\Xi^1_{3\to j}) & 1 \\ \vdots & \vdots & \vdots \\ -\Xi^1_{L\to j} & -(\sum_{i=1}^{L-1} \Xi^1_{i\to j} + 0.5\Xi^1_{L\to j}) & 1 \end{bmatrix} \tag{13}$$

and

$$U_j = \begin{bmatrix} \Xi^1_{2\to j} - \Xi^1_{1\to j} \\ \Xi^1_{3\to j} - \Xi^1_{2\to j} \\ \vdots \\ \Xi^1_{L\to j} - \Xi^1_{L-1\to j} \end{bmatrix} \tag{14}$$

Finally, according to the abovementioned process, the parameters Ω can be given by

$$\Omega = 20 + 20 + 60 = 100 \tag{15}$$

In detail, each of the part in Eq.15 can be expressed as

$$\psi_j = \begin{cases} f_j^{(1)} \quad (1 \le j \le 20) \\ \omega_j \quad (21 \le j \le 40) \\ \mu_j \quad (41 \le j \le 100) \end{cases} \tag{16}$$

Random Forest algorithm

Classifier is an essential part of the predictor, the Random Forest (RF) classifier is one of the frequently-used classifiers, which has a very powerful effect on the biological problems. In this study, according to a lot of experiments, we select 50 as the RF classifier trees parameters when the computational cost and overfitting problems are considered. However, in our study, the dataset is unbalanced, which means the number of the negative dataset is larger than the positive, but almost all of the classifier are often used to train the balanced data. For our study, dataset is unbalanced so that the performance of the RF is not satisfing. Therefore, we proposed a two-layer ensemble learning to overcome this problem. And the process of the predictor iANOP-Enble was exhibited in **Fig.1.**

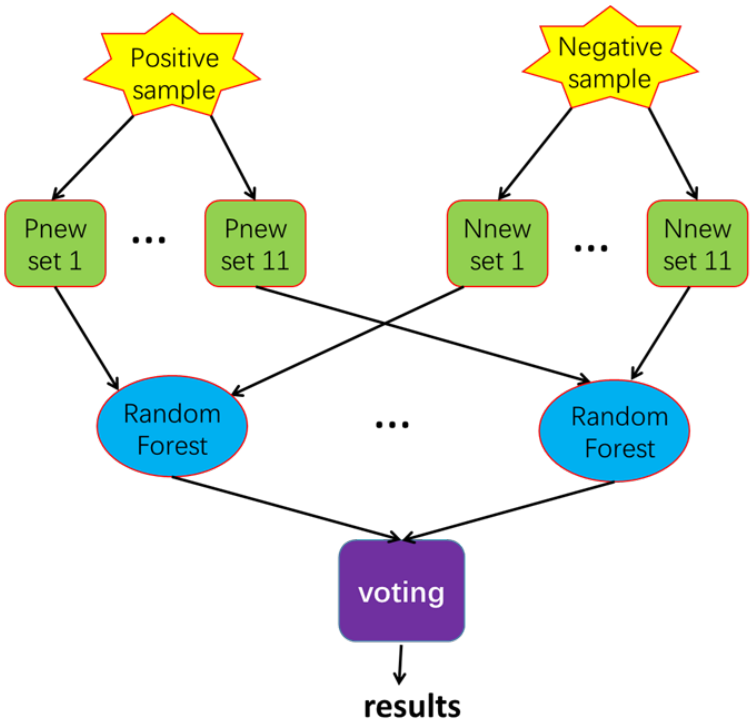

Fig.1. the figure shows the ensemble learning process.

According to the abovementioned process, the predictor called iANOP-Enble was obtained. An image shows how the antioxidant proteins are identified in the process of the predictor, as shown in **Fig.2**

Prediction of Antioxidant Proteins

In this paper, we adopted the cross validation to test our predictor. In order to avoid the deviation, we adopt 10-fold cross validation and random sampling 10 times. For each time the results were shown in Table 1.

According to the test, the result of our predictor is satisfying (see Table 1), and the final result is about average ten times result, in which the total of rate is 0.8825, the Matthew's correlation coefficient is 0.5725, the sensitivity is 0.7278 and specificity is 0.9075.

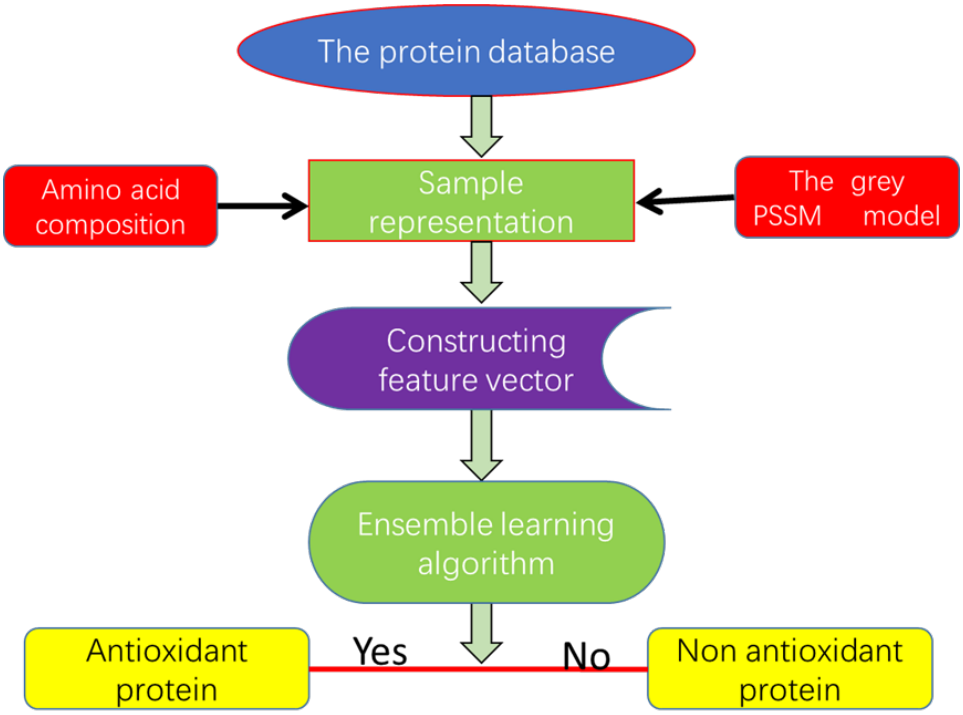

Fig.2. The image shows that the predictor how to identify the antioxidant proteins

Table 1 10-fold cross-validation results

	Evaluating Methods			
	Accuracy	Sensitivity	Specificity	MCC
1	0.9281±0.0081	0.6800±0.0184	0.9360±0.0200	0.5964±0.0153
2	0.9175±0.0046	0.7600±0.0129	0.9358±0.0113	0.6545±0.0089
3	0.8186±0.0056	0.5200±0.0043	0.9299±0.0118	0.4577±0.0101
4	0.8791±0.0050	0.5384±0.0077	0.9108±0.0106	0.4357±0.0093
5	0.8907±0.0026	0.8461±0.0127	0.8726±0.0085	0.5967±0.0058
6	0.9065±0.0067	0.8000±0.0129	0.8917±0.0094	0.5916±0.0136
7	0.8306±0.0031	0.8076±0.0100	0.9044±0.0058	0.6254±0.0068
8	0.8618±0.0054	0.7600±0.0068	0.8846±0.0072	0.5515±0.0106
9	0.8956±0.0084	0.7200±0.0181	0.9235±0.0127	0.5970±0.0170
10	0.8907±0.0065	0.8461±0.0127	0.8853±0.0061	0.6179±0.0135
Mean	0.8825±0.0015	0.7278±0.0025	0.9075±0.0034	0.5725±0.0029

Table 2 the predictor iANOP-Enble comparison the existing predictors

Predictor	Acc(%)	MCC	Sn(%)	Sp(%)	AUC
Randon Forest[a]	87.97	---	28.35	97.64	0.797
Na Ï ve Bayes[b]	66.88	---	72.04	66.05	0.855
iANOP-Enble	88.25	0.5725	72.78	90.75	0.935

[a] Results obtained by Fernández-Blanco[15].
[b]Results obtained by the optimized features for the predictor by Feng[5] and coworkers

Comparison with existing Machine Learning Methods

Our predictor iANOP-Enble has obtained the four criterion index result (see Table 2) via 10-fold cross-validation. In order to make a convenient and accurate comparison, we also listed the existing predictors results. According to the Table 2, it shows that the performance of our predictor iANOP-Enble is much better than those of existing classifier algorithm model [5]. The MCC of the current predictor is 0.5725, however, both Fernández and Feng don't list the value. Secondly, it is also true that the other three metrics are much better. The result of AUC is especially accurate,

which shows that our predictor iANOP-Enble is very effective in identifying antioxidant proteins, as shown in the sixth column of Table 2. But above all, our new predictor can achieve very high accuracy and have stronger robustness. And the ROC curve is shown in Fig 3 about iANOP-Enble.

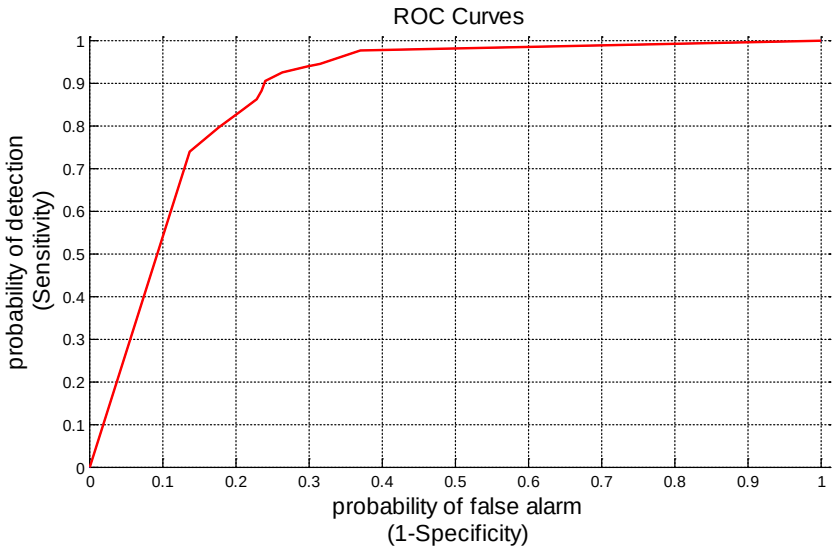

Fig.3. The ROC curves of iANOP-Enble

Conclusion

The new classifier system to identify antioxidant proteins is voting system, which consists of eleven sub-classifiers that can be summarized as an ensemble learning system. To predict antioxidant proteins on the basis of the ensemble learning system is very effective and robustness. Besides, the predictor takes into account the protein sequence evolutionary information, which can reflect the characteristic of the sequence. Because of this, it can achieve higher quality performance than other predictors. Though the result of our predictor is satisfying, it only used the sequence information, not the protein structural information. As a result, some important information may be lost in the extraction protein sequence features. Therefore, we are going to further study this problem by incorporating protein sequence and other properties. In conclusion, we look forward that iANOP-Enble predictor can play a significant role in identifying antioxidant proteins.

Acknowledgement

This work was partially supported by the National Nature Science Foundation of China (No. 31560316).

References

[1] Holley AE, Cheeseman KH. Measuring free radical reactions in vivo. British Medical Bulletin. 1993;49(3):494-505.

[2] Bounous G, Molson JH. The antioxidant system. Anticancer Research. 2003;23(2B):1411-5.

[3] Jones JDG, Dangl JL. The plant immune system. Nature. 2006;444(7117):323-9.

[4] Chou KC. Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology. Current Proteomics. 2009;6(4):262-74(13).

[5] Feng PM, Lin H, Chen W. Identification of antioxidants from sequence information using naïve Bayes. Computational & Mathematical Methods in Medicine. 2013;2013(2):567529-.

[6] Nakashima H, ., Nishikawa K, ., Ooi T, . The folding type of a protein is relevant to the amino acid composition. Journal of Biochemistry. 1986;99(1):153-62.

[7] Chou KC, Zhang CT. Predicting protein folding types by distance functions that make allowances for amino acid interactions. Journal of Biological Chemistry. 1994;269(35):22014-20.

[8] Cheng X, Xiao X, Wu Zc, Wang P, Lin Wz. Swfoldrate: Predicting protein folding rates from amino acid sequence with sliding window method. Proteins: Structure, Function, and Bioinformatics. 2013;81(1):140-8.

[9] Kandaswamy KK, Chou KC, Martinetz T, Möller S, Suganthan PN, Sridharan S, et al. AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. Journal of Theoretical Biology. 2011;270(1):56–62.

[10] Chou KC, Shen HB. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-Nearest Neighbor classifiers. Journal of Proteome Research. 2006;5(8):1888-97.

[11] Kuo-Chen C. Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of Theoretical Biology. 2011;273(1):236-47.

[12] Altschul SF, Madden TL, Sch?Ffer AA, Zhang J, ., Zhang Z, ., Miller W, ., et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997;25(8):3389--402.

[13] Min JL, Xiao X, Chou KC. iEzy-drug: a web server for identifying the interaction between enzymes and drugs in cellular networking. Biomed Res Int. 2013;2013(7):1805-12.

[14] Xiao X, Hui MJ, Liu Z, Qiu WR. iCataly-PseAAC: Identification of Enzymes Catalytic Sites Using Sequence Evolution Information with Grey Model GM (2,1). Journal of Membrane Biology. 2015;248(6):1033-41.

[15] Fernández-Blanco E, Aguiar-Pulido V, Munteanu CR, Dorado J. Random Forest Classification based on Star Graph Topological Indices for Antioxidant Proteins. Journal of Theoretical Biology. 2012;317(1):331–7.