ORIGINAL RESEARCH
published: 16 July 2021
doi: 10.3389/fcell.2021.707938

Identification of Disease-Related 2-Oxoglutarate/Fe (II)-Dependent Oxygenase Based on Reduced Amino Acid Cluster Strategy

*Jian Zhou[1]†, Suling Bo[2]†, Hao Wang[1], Lei Zheng[1], Pengfei Liang[1] and Yongchun Zuo[1]**

[1] State Key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, College of Life Sciences, Inner Mongolia University, Hohhot, China, [2] College of Computer and Information, Inner Mongolia Medical University, Hohhot, China

The 2-oxoglutarate/Fe (II)-dependent (2OG) oxygenase superfamily is mainly responsible for protein modification, nucleic acid repair and/or modification, and fatty acid metabolism and plays important roles in cancer, cardiovascular disease, and other diseases. They are likely to become new targets for the treatment of cancer and other diseases, so the accurate identification of 2OG oxygenases is of great significance. Many computational methods have been proposed to predict functional proteins to compensate for the time-consuming and expensive experimental identification. However, machine learning has not been applied to the study of 2OG oxygenases. In this study, we developed OGFE_RAAC, a prediction model to identify whether a protein is a 2OG oxygenase. To improve the performance of OGFE_RAAC, 673 amino acid reduction alphabets were used to determine the optimal feature representation scheme by recoding the protein sequence. The 10-fold cross-validation test showed that the accuracy of the model in identifying 2OG oxygenases is 91.04%. Besides, the independent dataset results also proved that the model has excellent generalization and robustness. It is expected to become an effective tool for the identification of 2OG oxygenases. With further research, we have also found that the function of 2OG oxygenases may be related to their polarity and hydrophobicity, which will help the follow-up study on the catalytic mechanism of 2OG oxygenases and the way they interact with the substrate. Based on the model we built, a user-friendly web server was established and can be friendly accessed at http://bioinfor.imu.edu.cn/ogferaac.

Keywords: 2-oxoglutarate/Fe (II)-dependent oxygenase, reduced amino acid cluster, machine learning, anova, incremental feature selection, 10-fold cross-validation test

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Leyi Wei,
Shandong University, China
Shihua Zhang,
Wuhan University of Science and Technology, China

****Correspondence:***
Yongchun Zuo
yczuo@imu.edu.cn

†These authors share first authorship

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *11 May 2021*
Accepted: *10 June 2021*
Published: *16 July 2021*

Citation:
Zhou J, Bo S, Wang H, Zheng L, Liang P and Zuo Y (2021) Identification of Disease-Related 2-Oxoglutarate/Fe (II)-Dependent Oxygenase Based on Reduced Amino Acid Cluster Strategy. Front. Cell Dev. Biol. 9:707938. doi: 10.3389/fcell.2021.707938

INTRODUCTION

2-Oxoglutarate/Fe (II)-dependent (2OG) oxygenases (EC:1.14.11), generally using nonheme iron as an active-site cofactor, promote oxidative decarboxylation of the substrate to produce carbon dioxide and succinic acid (Hausinger, 2004; Hewitson et al., 2005; Islam et al., 2018). 2OG oxygenases, which can catalyze many different oxidation reactions, are a superfamily with members widely distributed in animals, plants, and microorganisms. In animals, their catalytic range includes hydroxylation and N-demethylation proceeding *via* hydroxylation; in plants and microbes, they affect a wider range, including hydroxylation, ring formations, cleavage, oxidation, rearrangements, desaturations, and halogenations (Farrow and Facchini, 2014; Kawai et al., 2014). The proteins of this superfamily can be divided into 2OG oxygenase domain-containing oxygenases and JmjC domain-containing oxygenases (Jia et al., 2017). **Figure 1** is a schematic diagram of the structure of 2OG oxygenases.

Due to the diversity of 2OG oxygenases and the wide range of binding substrates, these oxygenases play an important role in physiology and have high therapeutic value and therapeutic potential as targets in cancer and many other diseases (Rose et al., 2011). For example, the protein containing the JmjC domain (JMJD6) is located in the nucleus that catalyzes lysine hydroxylation and arginine demethylation of histone and non-histone peptides (Chang et al., 2007; Liu et al., 2013). JMJD6 promotes cell proliferation and migration *in vitro* and accelerates tumor growth *in vivo*, so it may become an attractive target for a new generation of anticancer drugs (Lin et al., 2006; Lee et al., 2012). Prolyl 4-hydroxylase (P4H) plays a vital role in the synthesis of collagen and the regulation of oxygen homeostasis. Collagen P4Hs are considered to be attractive targets for drug inhibitors and involved in the treatment of fibrotic diseases and cancer metastasis (Vasta and Raines, 2018). Hypoxia-inducible transcription factor-prolyl 4-hydroxylase inhibitors are believed to have beneficial effects in the treatment of diseases such as myocardial infarction, stroke, peripheral vascular disease, diabetes, and severe anemias (Myllyharju, 2008; Liao and Zhang, 2020). ALKB homologs (ALKBH) homologs can regulate the physiological and pathological processes of cardiovascular diseases (CVDs), which have great potential in the development of CVD drugs and are expected to become a potential target for the treatment of CVD (Xiao et al., 2020). The change in the catalytic activity or expression level of lysine demethylases (KDMs) is closely related to many diseases, including cancer genesis and progression, neurological disorders, inflammatory and immune disorders, metabolic diseases, and regenerative diseases. Modulators/inhibitors of KDMs may be used as new treatments for cancer and other diseases (Arifuzzaman et al., 2020). Therefore, it is particularly meaningful to predict 2OG oxygenases and find more potential 2OG oxygenases. Since the identification of 2OG oxygenase is time-consuming and expensive, machine learning is an effective and fast method to predict it.

In the past, many machine learning methods for the prediction of metal ion-binding proteins have achieved excellent results. For example, Lin et al. (2006) applied the sequence information used by support vector machine (SVM) to predict the metal ion-binding protein and got a relatively marvelous prediction result. Mohan et al. (2010) used a set of physicochemical parameters of metal ion-binding proteins encoded by the three genes *CzcA*, *CzcB*, and *CzcD* as the training set of the supervised classifier, establishing a model to identify metal ion-binding proteins from unknown proteins. Valasatava et al. (2016) developed MetalPredator, a web server used to predict iron–sulfur cluster-binding proteomes, and it featured an excellent performance in terms of precision and recall. Many studies have also achieved good results in the prediction of metal ion-binding sites, including iron ion-binding sites (Liu and Hu, 2011; Liou et al., 2014), zinc ion-binding sites (Shu et al., 2008; Chen et al., 2013; Yan et al., 2019), copper ion binding sites (Levy et al., 2009; Brylinski and Skolnick, 2011). The above indicate that machine learning is suitable for the application of metal ion-binding proteins (Valasatava et al., 2016). Not only that, studies have shown that using the reduced amino acid cluster (RAAC) strategy to predict the types of proteins can reduce noise and achieve higher accuracy (Zheng et al., 2019). In the prediction of human and nonhuman enzymes (Wang H. et al., 2021), ion channel-targeted conotoxins (Sun et al., 2020), plasmodium secretory protein (Zhang et al., 2020), and defensin peptides (Zuo et al., 2019), the method of reduced amino acid has shown superior performance.

In this study, we established a prediction model for 2OG oxygenases based on SVM, which can effectively identify 2OG oxygenases. A new feature representation scheme (amino acid reduction cluster) was involved in this work. The RAAC strategy can greatly decrease the complexity of protein sequences and extremely reduce the use of computer memory (Zuo et al., 2017; Zheng et al., 2019). The workflow of constructing the OGFE_RAAC is shown in **Figure 2**. Firstly, an objective dataset was established, which contains 734 2OG oxygenases and 385,381 non-2OG oxygenases from the InterPro database. Subsequently, reduced amino acid composition combined with K-mer strategy was used to represent sequence features, and the optimal one was selected from 673 reduction schemes (Zuo et al., 2015). At the same time, we obtained the best feature combination through analysis of variance (ANOVA) combined with incremental feature selection (IFS) and applied SVM to establish the model. The results of 10-fold cross-validation and independent test set showed that OGFE_RAAC could accurately predict 2OG oxygenases.

MATERIALS AND METHODS

Dataset

The 2OG oxygenase superfamily can be classified into 2OG oxygenase domain-containing oxygenases and JmjC domain-containing oxygenases, so we collected all the verified 734 proteins of these two domains in the IPR number (IPR005123 and IPR003347) of the InterPro public database as a positive sample. Concurrently, 385381 protein data verified by SwissProt were gathered as negative samples, which is the manual annotation and review part of UniProt. Then, CD-HIT (Huang et al., 2010) was used to remove sequences with a similarity of more than 50% (Zou et al., 2020), and 480 samples are selected as the training set (Fu et al., 2012). We chose 150 samples from the rest as the test set, and the dataset was named 2OG-SwissProt. For the purpose of getting a better model, we also used iron-binding protein as a negative sample to construct a dataset. We acquired 593 iron-binding proteins (GO:0005506, 2OG oxygenase proteins removed) from the InterPro public database and processed

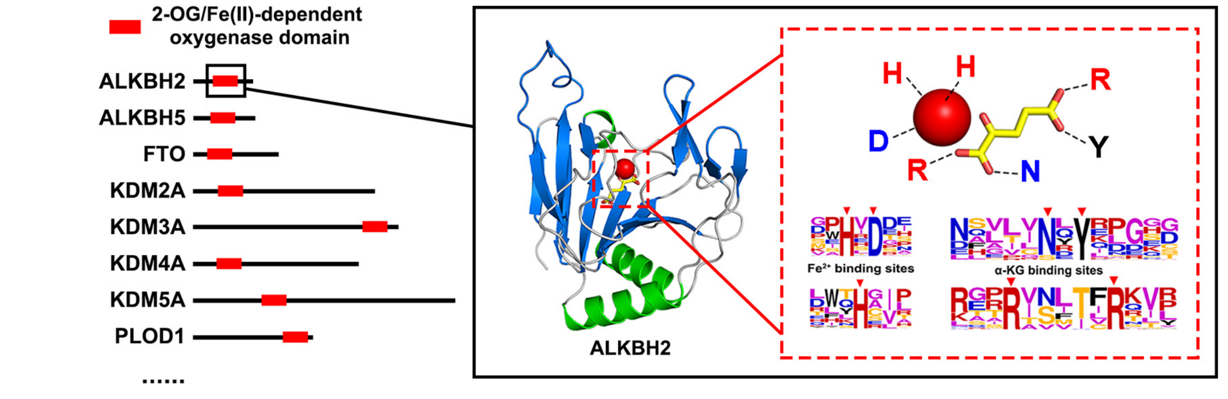


FIGURE 1 | Schematic diagram of the structure of 2-oxoglutarate/Fe (II)-dependent (2OG) oxygenase.

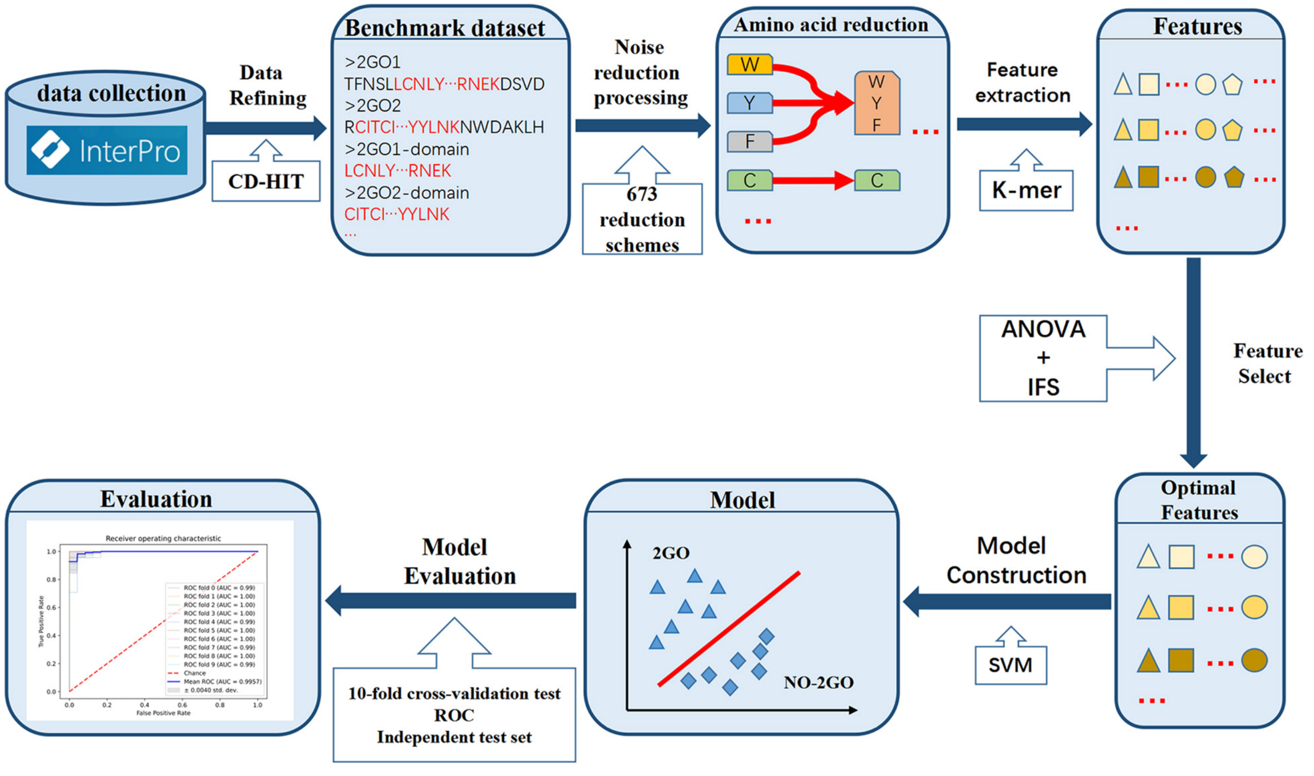


FIGURE 2 | The workflow of OGFE_RAAC predictor.

them in the same way as the 2OG-SwissProt dataset to obtain 471 training set samples and 159 test set samples; the dataset was named 2OG-Fe.

For further research, we manually extracted the domain sequences of 2OG oxygenase and iron-binding proteins. The processing method is the same as the above; in order to better verify the prediction results, we used CD-HIT processing sequence similarity less than 50% as the training set and the rest as the independent test set. Among them, 1,036 samples constitute an independent test set, 621 positive samples and 415 negative samples; 283 samples constitute a training set, 113 positive samples and 170 negative samples. This dataset was named 2OG-domain (**Table 1**).

Reduce Protein Sequence

Under normal circumstances, protein is composed of 20 natural amino acids. We combine amino acids with similar characteristics based on the physicochemical properties and atomic arrangement of amino acids. For instance, using fuzzy clustering technology and matrices cluster amino acids and interpret the sequence in a new encoding method (Georgiou et al., 2009; Zuo and Li, 2009). The strategy of RAACs can

TABLE 1 | Data composition of each dataset.

Dataset	Group	Training set	Test set
2OG-SwissProt	Positive	240	75
	Negative	240	75
2OG-Fe	Positive	240	75
	Negative	231	84
2OG-domain	Positive	113	621
	Negative	170	415

effectively reduce the complexity of the sequence and improve computational efficiency. In the study, we used 673 amino acid reduction schemes generated by 74 types to predict 2OG oxygenases, and each type has a reduced size of 2–19 (Zuo et al., 2019; Zheng et al., 2020).

Extract Features Based on K-mer

The typical K-mer (N-peptide) composition can effectively dig out the detailed information of the amino acid composition of the sequence (Zhu et al., 2019; Jaillard et al., 2020). We use K-mer ($K = 1, 2, 3$) to extract amino acid sequence information. Due to the limited memory, the maximum K value is 3, and a total of 20^K features can be obtained according to the original amino acid composition. The composition of K-mer ($K = 2$) can be expressed as follows:

$$P = R_1R_2R_3 \cdots R_{L-1}R_L \tag{1}$$

$$F = \left[d_1, d_2, \cdots d_{400}\right]^T \tag{2}$$

Here, R_i represents the i-th residue of the 2OG oxygenases. L represents the total length of the amino acid sequence. d_i ($i = 1$, 2,..., 400) is the i-th dipeptide in the 400-amino acid combination, and T means transposition operator. The d_i can be calculated as follows:

$$d_i = {}^{n_i}\Big/_{\sum_{i=1}^{400} n_i} \tag{3}$$

Here, n_i denotes the number of the i-th dipeptide. Combined with RAAC strategy, the feature extraction method can be expressed as follows:

$$F = \left[P^1_{1,1}, P^2_{1,2}, \ldots, P^k_{i,j}, \ldots, P^N_{T,C}\right] \tag{4}$$

where $P^k_{i,j}$ denotes the method of the N-peptide with different RAAC descriptors (N-peptide). N denotes the N-peptide. T denotes the type of different amino acid alphabets, and C denotes the cluster of the reduced amino acid alphabet. The parameters of the above equation can be limited as follows:

$$\begin{cases} 1 \le k \ \le N, N = \ [1, 2, 3] \\ 1 \le \ i \le \ T, T = \ [1, 2, \ldots, 74] \\ 1 \ \le j \ \le C, C = \ [2, 3, \ldots, 19] \end{cases} \tag{5}$$

Support Vector Machine

Support vector machine is a machine learning model that classifies data according to supervised learning methods and has been widely used in bioinformatics (Beer, 2017; Huang et al., 2018; Manavalan et al., 2018; Meng et al., 2020; Tahir and Idris, 2020). There are four types of kernel function, including linear functions, polynomial functions, S-shaped functions, and radial basis functions (RBFs). In the past predictions of proteins, the RBF kernel function had better performance, and we have verified that the RBF kernel function has better performance in our model through the calculation and comparison of the four kernel functions. Accordingly, we used the SVM package with RBF kernel for the classifier, which can be obtained from https://www.csie.ntu.edu.tw/~cjlin/libsvm (Chang and Lin, 2011). The libsvm package provides a grid search program to optimize the parameters C and γ. The kernel parameter γ and the regularization parameter C are used to adjust the SVM model to obtain the best performance. The selection ranges of C and γ are as follows:

$$2^{-5} < C < 2^{15} \tag{6}$$

$$2^{-15} < \gamma < 2^3 \tag{7}$$

Feature Screening

The initial features extracted by K-mer are exclusive features, not the optimal combination of features (Zou et al., 2016; He et al., 2020). ANOVA is a popular feature selection method that can help us measure the weight value of each feature (Saeys et al., 2007; Tang et al., 2018). Then, we used IFS to determine the dimensionality of the best feature set according to the feature weights obtained by the ANOVA. The ANOVA equations are as follows:

$$F = \frac{S_x^2}{S_y^2} \tag{8}$$

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{9}$$

$$S_y^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(y_i - \bar{y}\right)^2 \tag{10}$$

where F is the variance value of the feature. S_x^2 is the sample variance between groups. S_y^2 denotes the sample variance within groups.

Performance Evaluation

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling (K-fold cross-validation) test, and jackknife test. However, among the three cross-validation methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset and hence has been increasingly used and widely recognized by investigators to examine the accuracy of various predictors (Chou and Shen, 2008; Chou, 2011; Chou et al., 2012; Zhang et al., 2021). However, since the current study would involve feature selection as described above, to reduce the computational time, the 10-fold cross-validation test and independent dataset test would be adopted as done by many investigators using SVM as the prediction engine. The performance can be measured in term of Sensitivity (Sn), Specificity (Sp), F1 score, Matthew's correlation coefficient (MCC), and Accuracy (Acc; Li et al., 2020; Shen and Zou, 2020; Yang et al., 2021), which are expressed as follows:

$$\mathrm{Sn} = \frac{TP}{TP + FN} \tag{11}$$

$$\mathrm{Sp} = \frac{TN}{TN + FP} \tag{12}$$

$$F1\ \mathrm{score} = \frac{2TP}{2TP + FP + FN} \tag{13}$$

$$\mathrm{Acc} = \frac{TP + TN}{TP + FN + TN + FP} \tag{14}$$

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}} \tag{15}$$

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

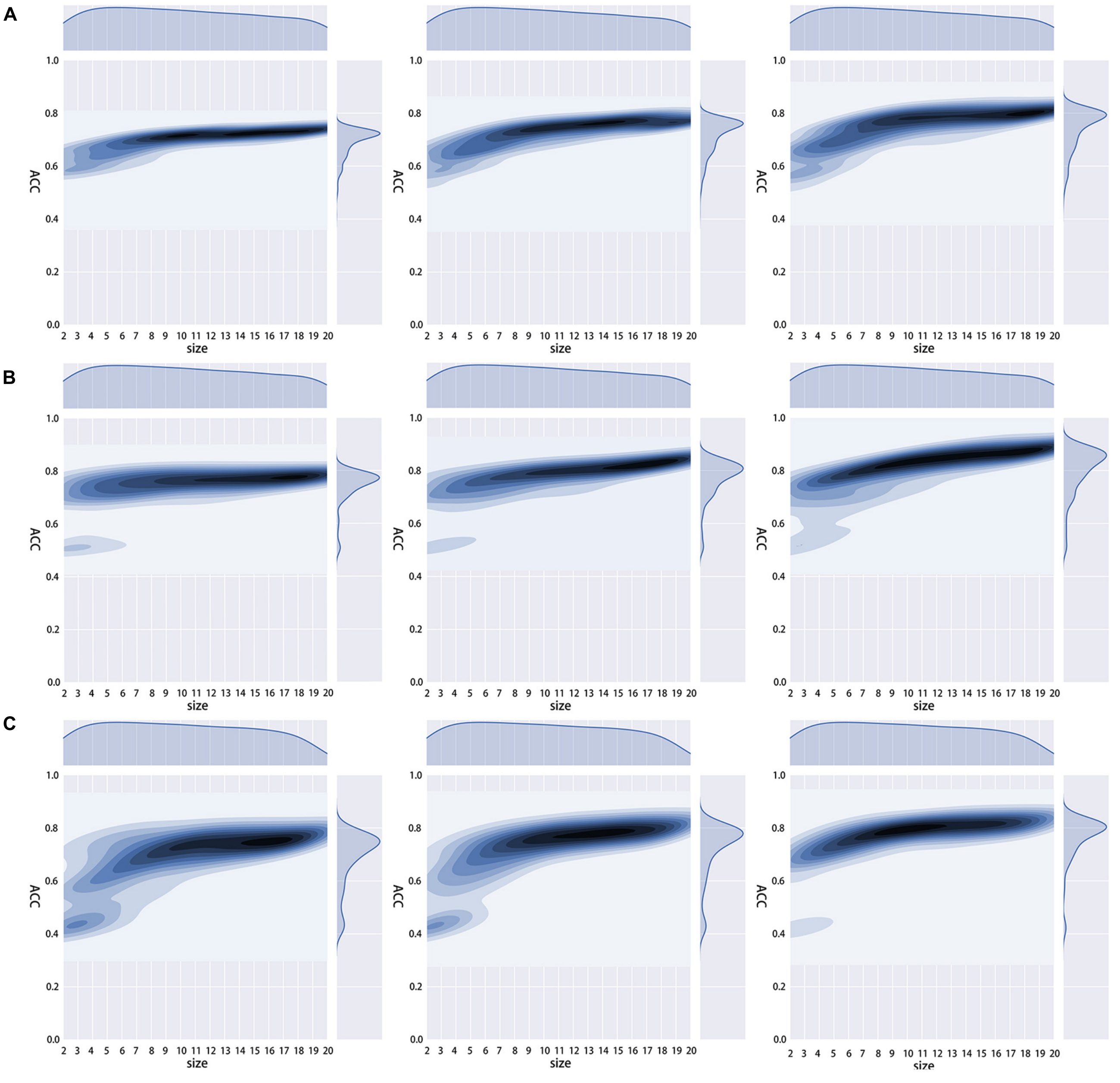


FIGURE 3 | Density distribution diagram of different *K* value accuracy rates. **(A–C)** are the density distribution diagrams of the 2OG-SwissProt set, 2OG-Fe set, and 2OG-domain set at different *K* values in 673 reduction schemes.

RESULTS

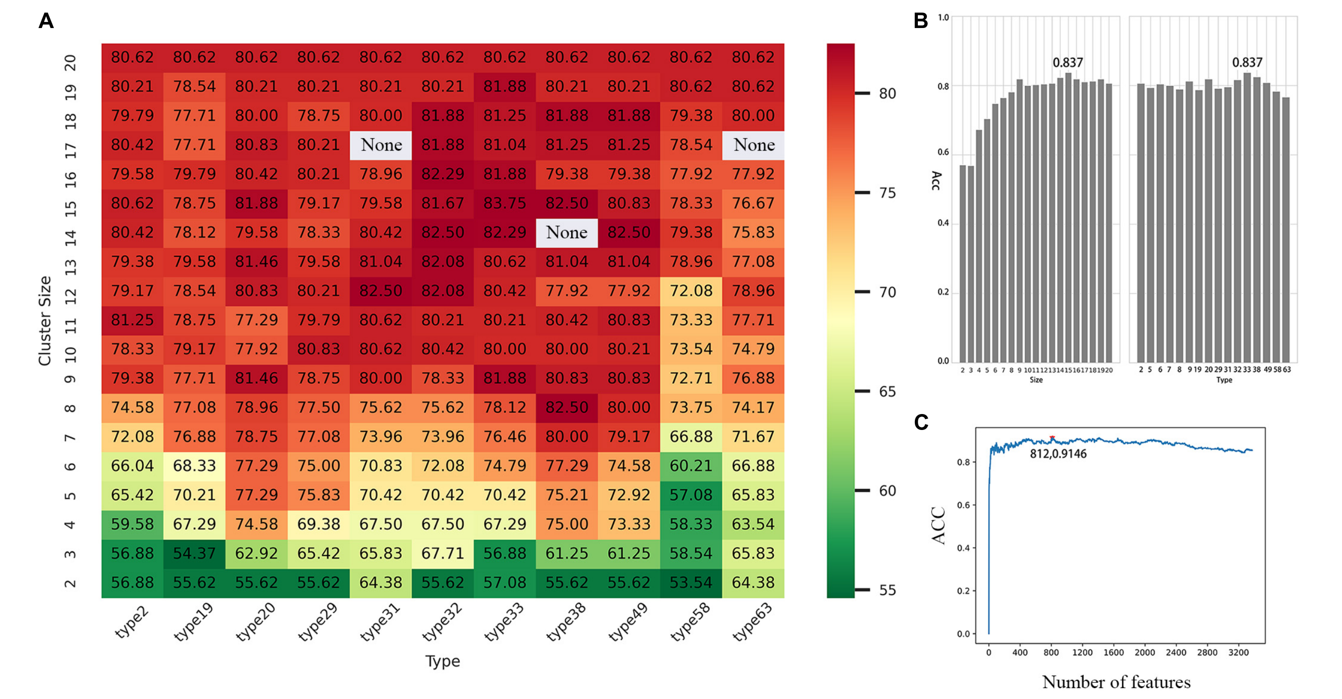
Predictive Performance of Different Reducing Amino Acid Cluster

To obtain the optimal amino acid reduction scheme and the appropriate K value ($K = 1, 2, 3$), we calculated the accuracy of the 673 reduction schemes mentioned in RAACBook (Zheng et al., 2019) with the different K values. We found that all three models showed the best performance at $K = 3$, and most of the reduction schemes had higher accuracy when $K = 3$ (**Figure 3**). We guessed that there would be more features when $K = 3$, and they would better reflect the properties of the protein and get a more accurate model.

After confirming that the model has better performance when $K = 3$, we then selected the best scheme from 673 RAAC schemes to construct the model. In the 2OG-SwissProt model, we tested each size of each reduction type and compared different reduction sizes of different reduction types (**Figure 4A**). We found that when $t = 33$ (**Table 2**), $s = 15$ (t represents the t-th reduction type in RAACBook; s represents the size of the RAAC), the highest accuracy rate is 83.75% (**Figure 4B**). In the prediction of the 2OG-Fe dataset, we were pleasantly surprised to find that the highest accuracy rate also appears in the reduction type 33, and the highest accuracy rate is 90.04% when $s = 16$ (**Supplementary Figure 1B**). There is also a very high accuracy rate at $s = 15$, reaching 88.76% (**Supplementary Figure 1A**). The reduction method of type 33 uses a database of aligned protein structures to propose a new clustering method based on the substitution scores, which aggregates 20 amino acids in two groups, namely, the hydrophobic groups and the polar groups (Li and Wang, 2007). Therefore, we speculated that the function of 2OG oxygenases may be related to its polarity and hydrophobicity.

To further prove that polarity and hydrophobicity may be related to the function of 2OG oxygenases, we manually extracted the 2OG oxygenase domain and JmjC domain sequences and other iron-binding domain sequences for prediction. Protein functions mainly through its domain region, and 2OG oxygenases also bind Fe(II) and 2-oxoglutarate in their domain position to perform their functions. Therefore, the region outside the domain may be noise information for feature extraction, and only using the domain sequence to extract features can better reflect the function of 2OG oxygenases (Shen and Zou, 2020). The result is the same as we expected, when $t = 33$ and $s = 15$, the highest accuracy rate is obtained (**Supplementary Figure 1B**). The same result is obtained with the complete sequence, which further proves that the polarity and hydrophobicity may be related to the function of 2OG oxygenases.

The functional domain of 2OG oxygenases contains Fe^{2+}-binding sites and α-ketoglutarate-binding sites, and their amino acid composition is almost completely conserved. The Fe^{2+}-binding motif (HXD-H) and α-KG-binding motif (N-Y-R-R) of the ALKBH family are entirely conserved in the homologs (Bjornstad et al., 2011; Fedeles et al., 2015; Alemu et al., 2016; Xu et al., 2021), and other 2OG oxygenases have similar structures (Bleijlevens et al., 2008; Islam et al., 2018;



Cluster Size	type2	type19	type20	type29	type31	type32	type33	type38	type49	type58	type63
20	80.62	80.62	80.62	80.62	80.62	80.62	80.62	80.62	80.62	80.62	80.62
19	80.21	78.54	80.21	80.21	80.21	80.21	81.88	80.21	80.21	80.62	80.62
18	79.79	77.71	80.00	78.75	80.00	81.88	81.25	81.88	81.88	79.38	80.00
17	80.42	77.71	80.83	80.21	None	81.88	81.04	81.25	81.25	78.54	None
16	79.58	79.79	80.42	80.21	78.96	82.29	81.88	79.38	79.38	77.92	77.92
15	80.62	78.75	81.88	79.17	79.58	81.67	83.75	82.50	80.83	78.33	76.67
14	80.42	78.12	79.58	78.33	80.42	82.50	82.29	None	82.50	79.38	75.83
13	79.38	79.58	81.46	79.58	81.04	82.08	80.62	81.04	81.04	78.96	77.08
12	79.17	78.54	80.83	80.21	82.50	82.08	80.42	77.92	77.92	72.08	78.96
11	81.25	78.75	77.29	79.79	80.62	80.21	80.21	80.42	80.83	73.33	77.71
10	78.33	79.17	77.92	80.83	80.62	80.42	80.00	80.00	80.21	73.54	74.79
9	79.38	77.71	81.46	78.75	80.00	78.33	81.88	80.83	80.83	72.71	76.88
8	74.58	77.08	78.96	77.50	75.62	75.62	78.12	82.50	80.00	73.75	74.17
7	72.08	76.88	78.75	77.08	73.96	73.96	76.46	80.00	79.17	66.88	71.67
6	66.04	68.33	77.29	75.00	70.83	72.08	74.79	77.29	74.58	60.21	66.88
5	65.42	70.21	77.29	75.83	70.42	70.42	70.42	75.21	72.92	57.08	65.83
4	59.58	67.29	74.58	69.38	67.50	67.50	67.29	75.00	73.33	58.33	63.54
3	56.88	54.37	62.92	65.42	65.83	67.71	56.88	61.25	61.25	58.54	65.83
2	56.88	55.62	55.62	55.62	64.38	55.62	57.08	55.62	55.62	53.54	64.38

FIGURE 4 | Performance evaluation of different reduced amino acid clusters. **(A)** Heat map of accuracy distribution of different reduced amino acid clusters. **(B)** The accuracy rate of the reduced amino acid cluster ($t = 33$, $s = 15$) with the highest accuracy rate reaches 83.75%. **(C)** The incremental feature selection (IFS) curve shows that prediction accuracy is 91.46% when using 812 optimal features based on the tripeptide combination ($t = 33$, $s = 15$).

TABLE 2 | Cluster size of reduced amino acid alphabet of type 33.

Size	Reduced amino acid cluster
2	STANDGRQEKHPIVLMWYF-C
3	STANDGRQEKHP-IVLMWYF-C
4	STANDG-RQEKHP-IVLMWYF-C
5	STAND-G-RQEKHP-IVLMWYF-C
6	STAND-G-RQEK-HP-IVLMWYF-C
7	STA-ND-G-RQEK-HP-IVLMWYF-C
8	STA-ND-G-RQ-EK-HP-IVLMWYF-C
9	STA-ND-G-RQ-EK-HP-IVLM-WYF-C
10	ST-A-ND-G-RQ-EK-HP-IVLM-WYF-C
11	ST-A-ND-G-RQ-EK-H-P-IVLM-WYF-C
12	ST-A-N-D-G-RQ-EK-H-P-IVLM-WYF-C
13	ST-A-N-D-G-RQ-EK-H-P-IV-LM-WYF-C
14	S-T-A-N-D-G-RQ-EK-H-P-IV-LM-WYF-C
15	S-T-A-N-D-G-RQ-EK-H-P-IV-L-M-WYF-C
16	S-T-A-N-D-G-RQ-E-K-H-P-IV-L-M-WYF-C
17	S-T-A-N-D-G-RQ-E-K-H-P-IV-L-M-WY-F-C
18	S-T-A-N-D-G-R-Q-E-K-H-P-IV-L-M-WY-F-C
19	S-T-A-N-D-G-R-Q-E-K-H-P-I-V-L-M-WY-F-C

TABLE 3 | The results of each evaluation index of the three models.

Model	Acc (%)	Sn (%)	SP (%)	MCC (%)	*F*1 score (%)	AUC (%)
2OG-SwissProt	91.04	93.33	88.75	82.34	91.26	97.15
2OG-Fe	97.23	97.92	96.53	94.48	97.31	99.57
2OG-domain	97.87	98.23	97.65	95.60	97.37	99.89

Acc, accuracy; AUC, area under the curve; MCC, Matthew's correlation coefficient; Sn, sensitivity; and Sp, specificity.

Wang et al., 2021). They all combine Fe^{2+} and α-ketoglutarate through conserved polar amino acid regions, which may be the reason why polarity is an essential feature of 2OG oxygenase identification. In addition, in the best reduction scheme, Phenylalanine (F), Tryptophan (W), and Tyrosine (Y) are recombined into a new letter, and these three amino acids are all aromatic amino acids. We speculate that the function of 2OG oxygenases may be related to the hydrophobicity of aromatic amino acids and the unique properties of its benzene ring.

Feature Selection

Although we can get more features when $K = 3$, not every feature can be helpful to the prediction of 2OG oxygenases; some features may even become noise information and affect the final result. Therefore, we used ANOVA combined with IFS to select the best feature combination. Through 10-fold cross-validation, the 2OG-SwissProt model achieves an optimal performance of 91.46% with 812 feature combinations (**Figure 4C**); the 2OG-Fe model achieves an optimal performance of 96.61% with 1,181 feature combinations (**Supplementary Figure 1C**); 2OG-domain

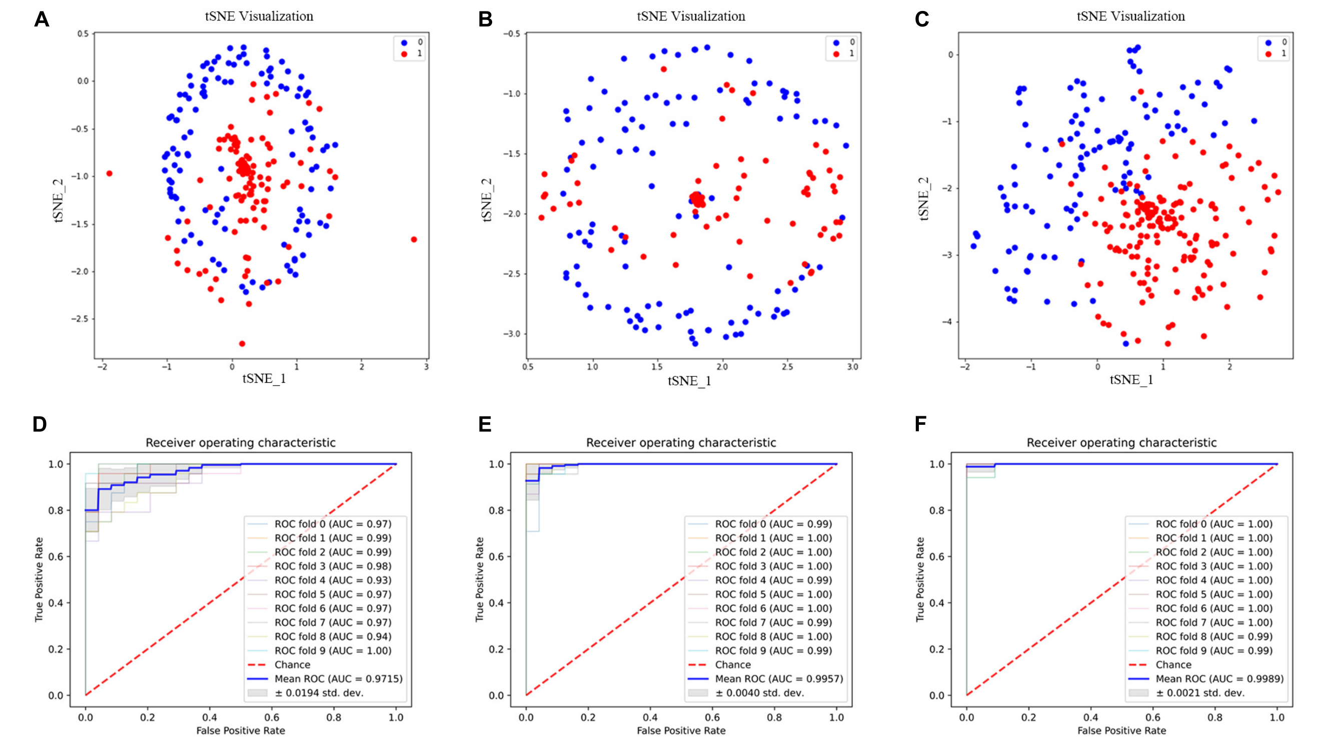
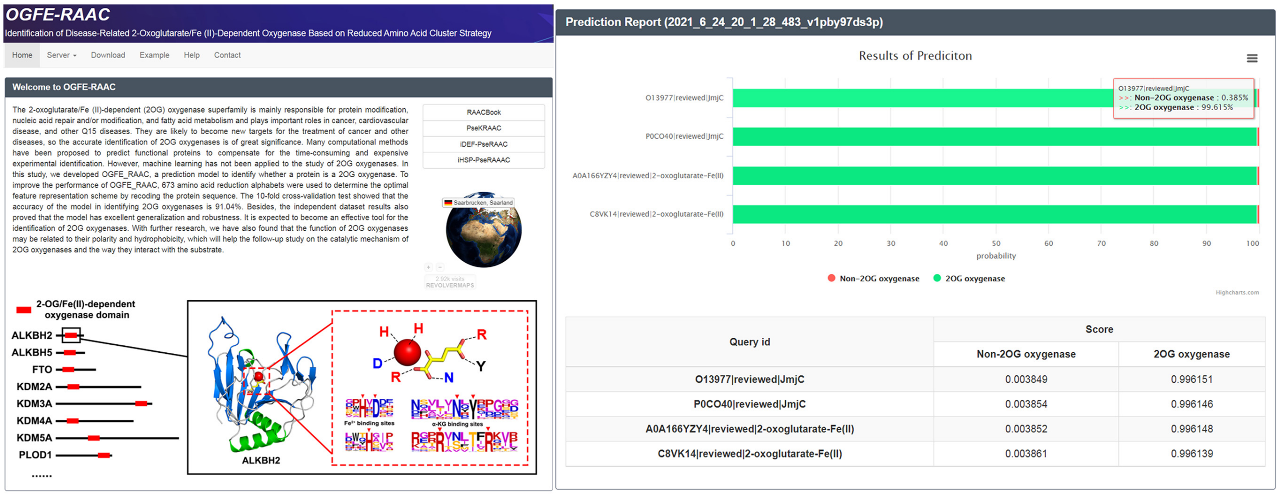


FIGURE 5 | Feature set t-SNE clustering scatter diagram and receiver operating characteristic (ROC) curve diagram. **(A–C)** are the t-SNE clustering analysis diagrams of the feature set after unreduced, reduced, and feature screening, respectively. 0 and 1 represent positive samples and negative samples, respectively. **(D–F)** are the ROC curves of the three models 2OG-SwissProt, 2OG-Fe, and 2OG-domain, respectively.



The 2-oxoglutarate/Fe (II)-dependent (2OG) oxygenase superfamily is mainly responsible for protein modification, nucleic acid repair and/or modification, and fatty acid metabolism and plays important roles in cancer, cardiovascular disease, and other Q15 diseases. They are likely to become new targets for the treatment of cancer and other diseases, so the accurate identification of 2OG oxygenases is of great significance. Many computational methods have been proposed to predict functional proteins to compensate for the time-consuming and expensive experimental identification. However, machine learning has not been applied to the study of 2OG oxygenases. In this study, we developed OGFE_RAAC, a prediction model to identify whether a protein is a 2OG oxygenase. To improve the performance of OGFE_RAAC, 673 amino acid reduction alphabets were used to determine the optimal feature representation scheme by recoding the protein sequence. The 10-fold cross-validation test showed that the accuracy of the model in identifying 2OG oxygenases is 91.04%. Besides, the independent dataset results also proved that the model has excellent generalization and robustness. It is expected to become an effective tool for the identification of 2OG oxygenases. With further research, we have also found that the function of 2OG oxygenases may be related to their polarity and hydrophobicity, which will help the follow-up study on the catalytic mechanism of 2OG oxygenases and the way they interact with the substrate.

Query id	Score	
	Non-2OG oxygenase	2OG oxygenase
O13977|reviewed|JmjC	0.003849	0.996151
P0CO40|reviewed|JmjC	0.003854	0.996146
A0A166YZY4|reviewed|2-oxoglutarate-Fe(II)	0.003852	0.996148
C8VK14|reviewed|2-oxoglutarate-Fe(II)	0.003861	0.996139

FIGURE 6 | Home page and results page of OGFE-RAAC web server.

model also achieves an optimal performance of 96.07% with 350 feature combinations (**Supplementary Figure 1C**). For more clearly showing that the filtered features can better reflect the nature of 2OG oxygenases, we used t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the feature sets after unreduced, reduced, and feature screening in a 2D feature space (**Figures 5A–C**). Obviously, the results show that the feature set clustering effect after feature screening is better, and it can effectively separate 2OG oxygenases from non-2OG oxygenases.

Performance Evaluation

We evaluated our model by 10-fold cross-validation to verify that our model is effective (**Table 3**). At the same time, we drew the receiver operating characteristic (ROC) curve through the 10-fold cross-validation (**Figures 5D–F**).

In order to further evaluate our predictor, we used an independent test set to test 2OG-SwissProt, 2OG-Fe, and 2OG-domain models. The 2OG-SwissProt model accurately predicts 143 samples out of 150 test set samples, and the accuracy rate is 95.33%. The 2OG-Fe model accurately predicts 149 samples out of 159 test set samples, with an accuracy rate of 93.71%. The 2OG-domain model accurately predicts 963 samples out of 1,036 test set samples, with an accuracy rate of 92.95%. These show that our predictor is effective and robust.

Web Server Guidance

For the purpose of other researchers to use our model more conveniently, an easy-to-use web server was established to implement our predictor, which can be freely accessed at http://bioinfor.imu.edu.cn/ogferaac. When you want to use our tool, you need to click the "Service" module and then import the FASTA protein sequence into the input box or upload the button to upload your protein data. Meanwhile, according to the different sequences you provide, you can also choose different modules (2OG-SwissProt, 2OG-Fe, and 2OG-domain) for prediction. After submitting the task, the website will provide the corresponding forecast report, which will display the forecast results and probability of each sequence in the form of tables and flowcharts (**Figure 6**).

DISCUSSION

At present, the research on 2OG oxygenases is more in-depth, and its many functions (such as demethylation) occupy an important position in the research of diseases (Liu et al., 2019; Ao et al., 2021). Based on RAAC strategy and SVM, the prediction model of 2OG oxygenases is constructed. t-SNE results show that RAAC can effectively reduce protein complexity, extract conservative features hidden in noise information, and improve prediction accuracy. OGFE_RAAC has strong robustness and generalization to accurately predict 2OG oxygenases. We anticipate that OGFE_RAAC can accurately and rapidly identify 2OG oxygenases based on peptide sequence and promote the development of related drug research. Not only that, we also found that the function of 2OG oxygenases may be related to its hydrophobicity and polarity during the prediction process, which also provides a new research idea for the future study of 2OG oxygenases.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://bioinfor.imu.edu.cn/ogferaac/public/Download.

AUTHOR CONTRIBUTIONS

YZ conceived and designed the study. JZ and PL organized and collected the data and carried out the computation.

LZ designed and developed the web server. JZ and HW wrote the manuscript. SB participated in all subsequent revisions of the manuscript. YZ planned overall and revised the manuscript. All authors read and approved the manuscript.

FUNDING

This work was supported by the National Natural Scientific Foundation of China (Nos. 62061034 and 61861036), Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (NJYT-18-B01), the Fund for Excellent Young Scholars of Inner Mongolia (2017JQ04), and the Science and Technology Major Project of Inner Mongolia Autonomous Region of China to the State Key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock (2019ZD031).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.707938/full#supplementary-material

REFERENCES

Alemu, E. A., He, C., and Klungland, A. (2016). ALKBHs-facilitated RNA modifications and de-modifications. *DNA Repair* 44, 87–91. doi: 10.1016/j.dnarep.2016.05.026

Ao, C., Yu, L., and Zou, Q. (2021). Prediction of bio-sequence modifications and the associations with diseases. *Brief. Funct. Genomics* 20, 1–18. doi: 10.1093/bfgp/elaa023

Arifuzzaman, S., Khatun, M. R., and Khatun, R. (2020). Emerging of lysine demethylases (KDMs): from pathophysiological insights to novel therapeutic opportunities. *Biomed. Pharmacother.* 129:110392. doi: 10.1016/j.biopha.2020.110392

Beer, M. A. (2017). Predicting enhancer activity and variant impact using gkm-SVM. *Hum. Mutat.* 38, 1251–1258. doi: 10.1002/humu.23185

Bjornstad, L. G., Zoppellaro, G., Tomter, A. B., Falnes, P. O., and Andersson, K. K. (2011). Spectroscopic and magnetic studies of wild-type and mutant forms of the Fe(II)- and 2-oxoglutarate-dependent decarboxylase ALKBH4. *Biochem. J.* 434, 391–398. doi: 10.1042/bj20101667

Bleijlevens, B., Shivarattan, T., Flashman, E., Yang, Y., Simpson, P. J., Koivisto, P., et al. (2008). Dynamic states of the DNA repair enzyme AlkB regulate product release. *EMBO Rep.* 9, 872–877. doi: 10.1038/embor.2008.120

Brylinski, M., and Skolnick, J. (2011). FINDSITE-metal: integrating evolutionary information and machine learning for structure-based metal-binding site prediction at the proteome level. *Proteins* 79, 735–751. doi: 10.1002/prot.22913

Chang, B. S., Chen, Y., Zhao, Y. M., and Bruick, R. K. (2007). JMJD6 is a histone arginine demethylase. *Science* 318, 444–447. doi: 10.1126/science.1145801

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 1–27. doi: 10.1145/1961189.1961199

Chen, Z., Wang, Y., Zhai, Y. F., Song, J., and Zhang, Z. (2013). ZincExplorer: an accurate hybrid method to improve the prediction of zinc-binding sites from protein sequences. *Mol. Biosyst.* 9, 2213–2222. doi: 10.1039/c3mb70100j

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

Chou, K. C., and Shen, H. B. (2008). Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* 3, 153–162. doi: 10.1038/nprot.2007.494

Chou, K. C., Wu, Z. C., and Xiao, X. (2012). iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst.* 8, 629–641. doi: 10.1039/c1mb05420a

Farrow, S. C., and Facchini, P. J. (2014). Functional diversity of 2-oxoglutarate/Fe(II)-dependent dioxygenases in plant metabolism. *Front. Plant Sci.* 5:524. doi: 10.3389/fpls.2014.00524

Fedeles, B. I., Singh, V., Delaney, J. C., Li, D. Y., and Essigmann, J. M. (2015). The AlkB Family of Fe(II)/alpha-ketoglutarate-dependent dioxygenases: repairing nucleic acid alkylation damage and beyond. *J. Biol. Chem.* 290, 20734–20742. doi: 10.1074/jbc.r115.656462

Fu, L. M., Niu, B. F., Zhu, Z. W., Wu, S. T., and Li, W. Z. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Georgiou, D. N., Karakasidis, T. E., Nieto, J. J., and Torres, A. (2009). Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. *J. Theor. Biol.* 257, 17–26. doi: 10.1016/j.jtbi.2008.11.003

Hausinger, R. P. (2004). FeII/alpha-ketoglutarate-dependent hydroxylases and related enzymes. *Crit. Rev. Biochem. Mol. Biol.* 39, 21–68. doi: 10.1080/10409230490440541

He, S., Guo, F., Zou, Q., and Ding, H. (2020). MRMD2.0: a python tool for machine learning with feature ranking and reduction. *Curr. Bioinform.* 15, 1213–1221. doi: 10.2174/1574893615999200503030350

Hewitson, K. S., Granatino, N., Welford, R. W., Mcdonough, M. A., and Schofield, C. J. (2005). Oxidation by 2-oxoglutarate oxygenases: non-haem iron systems in catalysis and signalling. *Philos. Trans. A Math. Phys. Eng. Sci.* 363, 807–828. discussion 1035-1040., doi: 10.1098/rsta.2004.1540

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., and Xu, W. (2018). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics Proteomics* 15, 41–51.

Huang, Y., Niu, B. F., Gao, Y., Fu, L. M., and Li, W. Z. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26, 680–682. doi: 10.1093/bioinformatics/btq003

Islam, M. S., Leissing, T. M., Chowdhury, R., Hopkinson, R. J., and Schofield, C. J. (2018). 2-oxoglutarate-dependent oxygenases. *Annu. Rev. Biochem.* 87, 585–620.

Jaillard, M., Palmieri, M., Van Belkum, A., and Mahe, P. (2020). Interpreting k-mer-based signatures for antibiotic resistance prediction. *Gigascience* 9:giaa110. doi: 10.1093/gigascience/giaa110

Jia, B., Tang, K., Chun, B. H., and Jeon, C. O. (2017). Large-scale examination of functional and sequence diversity of 2-oxoglutarate/Fe(II)-dependent oxygenases in Metazoa. *Biochim. Biophys. Acta Gen. Sub.* 1861, 2922–2933. doi: 10.1016/j.bbagen.2017.08.019

Kawai, Y., Ono, E., and Mizutani, M. (2014). Evolution and diversity of the 2-oxoglutarate-dependent dioxygenase superfamily in plants. *Plant J.* 78, 328–343. doi: 10.1111/tpj.12479

Lee, Y. F., Miller, L. D., Chan, X. B., Black, M. A., Pang, B., Ong, C. W., et al. (2012). JMJD6 is a driver of cellular proliferation and motility and a marker of poor prognosis in breast cancer. *Breast Cancer Res.* 14:R85.

Levy, R., Edelman, M., and Sobolev, V. (2009). Prediction of 3D metal binding sites from translated gene sequences based on remote-homology templates. *Proteins* 76, 365–374. doi: 10.1002/prot.22352

Li, F. Y., Leier, A., Liu, Q. Z., Wang, Y. A., Xiang, D. X., Akutsu, T., et al. (2020). Procleave: predicting protease-specific substrate cleavage sites by combining sequence and structural information. *Genomics Proteomics Bioinformatics* 18, 52–64. doi: 10.1016/j.gpb.2019.08.002

Li, J., and Wang, W. (2007). Grouping of amino acids and recognition of protein structurally conserved regions by reduced alphabets of amino acids. *Sci. China Series C Life Sci.* 50, 392–402. doi: 10.1007/s11427-007-0023-3

Liao, C. H., and Zhang, Q. (2020). ASIP COTRAN EARLY CAREER INVESTIGATOR AWARD LECTURE Understanding the oxygen-sensing pathway and its therapeutic implications in diseases. *Am. J. Pathol.* 190, 1584–1595. doi: 10.1016/j.ajpath.2020.04.003

Lin, H. H., Han, L. Y., Zhang, H. L., Zheng, C. J., Xie, B., Cao, Z. W., et al. (2006). Prediction of the functional class of metal-binding proteins from sequence derived physicochemical properties by support vector machine approach. *BMC Bioinform.* 7(Suppl. 5):S13. doi: 10.1186/1471-2105-7-S5-S13

Liou, Y. F., Charoenkwan, P., Srinivasulu, Y., Vasylenko, T., Lai, S. C., Lee, H. C., et al. (2014). SCMHBP: prediction and analysis of heme binding proteins using propensity scores of dipeptides. *BMC Bioinform.* 15(Suppl. 16):S4. doi: 10.1186/1471-2105-15-S16-S4

Liu, D. Y., Li, G. P., and Zuo, Y. C. (2019). Function determinants of TET proteins: the arrangements of sequence motifs with specific codes. *Brief. Bioinform.* 20, 1826–1835. doi: 10.1093/bib/bby053

Liu, R., and Hu, J. (2011). HemeBIND: a novel method for heme binding residue prediction by combining structural and sequence information. *BMC Bioinform.* 12:207. doi: 10.1186/1471-2105-12-207

Liu, W., Ma, Q., Wong, K., Li, W. B., Ohgi, K., Zhang, J., et al. (2013). Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. *Cell* 155, 1581–1595. doi: 10.1016/j.cell.2013.10.056

Manavalan, B., Shin, T. H., and Lee, G. (2018). PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine. *Front. Microbiol.* 9:476. doi: 10.3389/fmicb.2018.00476

Meng, C., Guo, F., and Zou, Q. (2020). CWLy-SVM: a support vector machine-based tool for identifying cell wall lytic enzymes. *Comput. Biol. Chem.* 87:107304. doi: 10.1016/j.compbiolchem.2020.107304

Mohan, A., Anishetty, S., and Gautam, P. (2010). Global metal-ion binding protein fingerprint: a method to identify motif-less metal-ion binding proteins. *J. Bioinform. Comput. Biol.* 8, 717–726. doi: 10.1142/s0219720010004884

Myllyharju, J. (2008). Prolyl 4-hydroxylases, key enzymes in the synthesis of collagens and regulation of the response to hypoxia, and their roles as treatment targets. *Ann. Med.* 40, 402–417. doi: 10.1080/07853890801986594

Rose, N. R., Mcdonough, M. A., King, O. N. F., Kawamura, A., and Schofield, C. J. (2011). Inhibition of 2-oxoglutarate dependent oxygenases. *Chem. Soc. Rev.* 40, 4364–4397.

Saeys, Y., Inza, I., and Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics* 23, 2507–2517. doi: 10.1093/bioinformatics/btm344

Shen, Z. J., and Zou, Q. (2020). Basic polar and hydrophobic properties are the main characteristics that affect the binding of transcription factors to methylation sites. *Bioinformatics* 36, 4263–4268. doi: 10.1093/bioinformatics/btaa492

Shu, N., Zhou, T., and Hovmöller, S. (2008). Prediction of zinc-binding sites in proteins from sequence. *Bioinformatics* 24, 775–782. doi: 10.1093/bioinformatics/btm618

Sun, Z. J., Huang, S. H., Zheng, L., Liang, P. F., Yang, W. R. T., and Zuo, Y. C. (2020). ICTC-RAAC: an improved web predictor for identifying the types of ion channel-targeted conotoxins by using reduced amino acid cluster descriptors. *Comput. Biol. Chem.* 89:107371. doi: 10.1016/j.compbiolchem.2020.107371

Tahir, M., and Idris, A. (2020). MD-LBP: an efficient computational model for protein subcellular localization from HeLa cell lines using SVM. *Curr. Bioinform.* 15, 204–211. doi: 10.2174/1574893614666190723120716

Tang, H., Zhao, Y. W., Zou, P., Zhang, C. M., Chen, R., Huang, P., et al. (2018). HBPred: a tool to identify growth hormone-binding proteins. *Int. J. Biol. Sci.* 14, 957–964. doi: 10.7150/ijbs.24174

Valasatava, Y., Rosato, A., Banci, L., and Andreini, C. (2016). MetalPredator: a web server to predict iron-sulfur cluster binding proteomes. *Bioinformatics* 32, 2850–2852. doi: 10.1093/bioinformatics/btw238

Vasta, J. D., and Raines, R. T. (2018). Collagen Prolyl 4-Hydroxylase as a therapeutic target. *J. Med. Chem.* 61, 10403–10411. doi: 10.1021/acs.jmedchem.8b00822

Wang, H., Xi, Q. L. M. G., Liang, P. F., Zheng, L., Hong, Y., and Zuo, Y. C. (2021). IHEC_RAAC: a online platform for identifying human enzyme classes via reduced amino acid cluster strategy. *Amino Acids* 53, 239–251. doi: 10.1007/s00726-021-02941-9

Wang, Z., Liu, D., Xu, B., Tian, R., and Zuo, Y. (2021). Modular arrangements of sequence motifs determine the functional diversity of KDM proteins. *Brief. Bioinform.* 22:bbaa215. doi: 10.1093/bib/bbaa215

Xiao, M. Z., Liu, J. M., Xian, C. L., Chen, K. Y., Liu, Z. Q., and Cheng, Y. Y. (2020). Therapeutic potential of ALKB homologs for cardiovascular disease. *Biomed. Pharmacother.* 131:110645. doi: 10.1016/j.biopha.2020.110645

Xu, B., Liu, D., Wang, Z., Tian, R., and Zuo, Y. (2021). Multi-substrate selectivity based on key loops and non-homologous domains: new insight into ALKBH family. *Cell. Mol. Life Sci.* 78, 129–141. doi: 10.1007/s00018-020-03594-9

Yan, R., Wang, X., Tian, Y., Xu, J., Xu, X., and Lin, J. (2019). Prediction of zinc-binding sites using multiple sequence profiles and machine learning methods. *Mol. Omics* 15, 205–215. doi: 10.1039/c9mo00043g

Yang, H., Luo, Y., Ren, X., Wu, M., He, X., Peng, B., et al. (2021). Risk Prediction of Diabetes: Big data mining with fusion of multifarious physical examination indicators. *Inf. Fusion* 75, 140–149. doi: 10.1016/j.inffus.2021.02.015

Zhang, D., Chen, H.-D., Zulfiqar, H., Yuan, S.-S., Huang, Q.-L., Zhang, Z.-Y., et al. (2021). iBLP: an XGBoost-based predictor for identifying bioluminescent proteins. *Comput. Math. Methods Med.* 2021:6664362.

Zhang, H. Y., Xi, Q., Huang, S. H., Zheng, L., Yang, W., and Zuo, Y. C. (2020). iSP-RAAC: identify secretory proteins of malaria parasite using reduced amino acid composition. *Comb. Chem. High Throughput Screen.* 23, 536–545. doi: 10.2174/1386207323666200402084518

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 (Oxford)* 2019:baz131.

Zheng, L., Liu, D., Yang, W., Yang, L., and Zuo, Y. (2020). RaacLogo: a new sequence logo generator by using reduced amino acid clusters. *Brief. Bioinform.* 22:bbaa096. doi: 10.1093/bib/bbaa096

Zhu, X. J., Feng, C. Q., Lai, H. Y., Chen, W., and Lin, H. (2019). Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowl. Based Syst.* 163, 787–793. doi: 10.1016/j.knosys.2018.10.007

Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2020). Sequence clustering in bioinformatics: an empirical study. *Brief. Bioinform.* 21, 1–10.

Zou, Q., Zeng, J. C., Cao, L. J., and Ji, R. R. (2016). A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 173, 346–354. doi: 10.1016/j.neucom.2014.12.123

Zuo, Y., Lv, Y., Wei, Z., Yang, L., Li, G., and Fan, G. (2015). iDPF-PseRAAAC: a web-Server for identifying the defensin peptide family and subfamily using pseudo reduced amino acid alphabet composition. *PLoS One* 10:e0145541. doi: 10.1371/journal.pone.0145541

Zuo, Y. C., Chang, Y., Huang, S. H., Zheng, L., Yang, L., and Cao, G. F. (2019). iDEF-PseRAAC: identifying the defensin peptide by using reduced amino acid composition descriptor. *Evol. Bioinform.* 15:1176934319867088.

Zuo, Y. C., and Li, Q. Z. (2009). Using reduced amino acid composition to predict defensin family and subfamily: integrating similarity measure and structural alphabet. *Peptides* 30, 1788–1793. doi: 10.1016/j.peptides.2009.06.032

Zuo, Y. C., Li, Y., Chen, Y. L., Li, G. P., Yan, Z. H., 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.