frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 12 September 2022
DOI 10.3389/fgene.2022.1012828

OPEN ACCESS

EDITED BY
Quan Zou,
University of Electronic Science and Technology of China, China

REVIEWED BY
Chunyu Wang,
Harbin Institute of Technology, China
Ping Xuan,
Heilongjiang University, China

*CORRESPONDENCE
Yu Chen,
nefu_chenyu@163.com

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 05 August 2022
ACCEPTED 22 August 2022
PUBLISHED 12 September 2022

CITATION
Chen D, Li S and Chen Y (2022), ISTRF: Identification of sucrose transporter using random forest.
Front. Genet. 13:1012828.
doi: 10.3389/fgene.2022.1012828

ISTRF: Identification of sucrose transporter using random forest

Dong Chen[1], Sai Li[2] and Yu Chen[2]*

[1]College of Electrical and Information Engineering, Qu Zhou University, Quzhou, China, [2]College of Information and Computer Engineering, Northeast Forestry University, Harbin, China

Sucrose transporter (SUT) is a type of transmembrane protein that exists widely in plants and plays a significant role in the transportation of sucrose and the specific signal sensing process of sucrose. Therefore, identifying sucrose transporter is significant to the study of seed development and plant flowering and growth. In this study, a random forest-based model named ISTRF was proposed to identify sucrose transporter. First, a database containing 382 SUT proteins and 911 non-SUT proteins was constructed based on the UniProt and PFAM databases. Second, k-separated-bigrams-PSSM was exploited to represent protein sequence. Third, to overcome the influence of imbalance of samples on identification performance, the Borderline-SMOTE algorithm was used to overcome the shortcoming of imbalance training data. Finally, the random forest algorithm was used to train the identification model. It was proved by 10-fold cross-validation results that k-separated-bigrams-PSSM was the most distinguishable feature for identifying sucrose transporters. The Borderline-SMOTE algorithm can improve the performance of the identification model. Furthermore, random forest was superior to other classifiers on almost all indicators. Compared with other identification models, ISTRF has the best general performance and makes great improvements in identifying sucrose transporter proteins.

KEYWORDS

machine learning, biological sequence analysis, protein identification, sucrose transporter, random forest

1 Introduction

Sucrose is a kind of disaccharide, which is formed by the condensation of fructose and glucose molecules through dehydration and is widely found in various tissues of plants. In the process of plant photosynthesis, carbon transport is mainly in the form of sucrose (Kühn et al., 1999). Therefore, the distribution of sucrose directly affects the growth and yield of plants (Aluko et al., 2021; Mangukia et al., 2021). In terms of physical properties, sucrose is a non-reducing sugar, which can carry a large amount of carbon. In terms of chemical properties, its properties are very stable, and it is not easy to combine with other compounds during transportation, so it has a certain protective effect on carbon. In terms of biological properties, due to the carbon in sucrose having a higher osmotic potential, the transport speed of sucrose is faster in a sieve tube. Sucrose transporters affect the transport of sucrose, which is mainly distributed in parenchyma cells, companion cells, and vacuolar membranes. They are the mediators of sucrose transport from source leaves to the phloem. In addition,

sucrose transporters also exist in sink organs, such as stems, seeds, and fruits. Sucrose transporters can promote sucrose transport under Pi starvation, salinity, and drought stress (Al-Sheikh Ahmed et al., 2018). At present, many experts have carried out a lot of research studies on sucrose transporters and found sucrose transporters in a variety of plant species, such as rice (Aoki et al., 2003), maize (Tran et al., 2017), grapevine, and tobacco (Wang et al., 2019). Endler et al. (2006) discovered a new sucrose transporter on the vacuolar membrane. They used liquid chromatography–tandem mass spectrometry to analyze tonoplast proteins and identified 101 proteins, including sucrose transporters. By studying the sucrose transporter gene RUSUT2 in blackberry, Yan et al. (2021) found that the sucrose content of mature leaves of the transgenic tobacco is enhanced by the overexpression of RUSUT2. At the same time, they found that Rusut2 has transport activity and may participate in sucrose transport during the growth and development of blackberry plants.

With the development of bioinformatics, more and more scholars used machine learning methods to identify sugar transporters. Mishra et al. (2014) developed a new model that incorporated the PSSM profile, amino acid composition, and biochemical composition of transporter proteins. The SVM algorithm was used as a classifier to classify transporters. Based on Mishra's experiments, Alballa et al. (2020) used a series of features including position information, evolutionary information, and amino acid composition to improve the accuracy and MCC of transporter classification. Ho et al. (2019) used word embedding technology to extract effective features from protein sequences and then adopted traditional machine learning methods to classify a variety of transporters (including sugar transporters). It has been proved that machine learning can effectively solve some problems of protein classification. All of the above studies focused on sugar transporters, while Shah et al. proposed to use natural language processing technology BERT to carry out feature extraction of glucose transporters in sugar transporters and classify three glucose transporters through an SVM classifier (Shah et al., 2021). Using machine learning methods to identify special proteins has become a trend, and a machine learning frame has been employed to identify sugar transporters. All these previous works guide us to build a frame for identifying sucrose transporters. In this study, we constructed an identification model named ISTRF to identify sucrose transporters. First, a dataset is built. Second, protein sequences are encoded with k-separated-bigrams-PSSM. Third, the Borderline-SMOTE algorithm is used to augment the positive samples. Finally, the identification model is trained by the random forest algorithm.

2 Materials and methods

2.1 Frame chart of ISTRF

In the study, we proposed a novel identification model called ISTRF, the frame chart of which is shown in Figure 1. First of all, the sucrose and non-sucrose transporter datasets are obtained using sequence homology analysis technology based on the Uniprot and Pfam databases, and then the CD-HIT program was used to remove redundancy and delete the protein sequences with more than 60% similarity. The sucrose transporter samples are construed for the training identification model. Second, we extracted the k-separated-bigrams-PSSM feature to represent samples. Third, we augment the positive samples to balance the training samples by using the Borderline-SMOTE technology. Finally, we built a random forest-based classifier that takes the balancing feature vectors as input. In the following sections, the dataset, feature extraction, sample balancing, and classifiers will be, respectively, introduced in detail.

2.2 Dataset

In this study, a self-built dataset is constructed and used. To obtain a reliable experimental result, it is necessary to use a high-quality benchmark data set, and then the initial data must be processed strictly and standardly. UniProt (Consortium, 2019) database is an authoritative protein database, in which we searched by the keyword "sucrose transporter" to obtain the initial positive sample data set. From the protein family database PFAM (Mistry et al., 2021), families containing positive samples were deleted, and the protein sequence with the longest length was extracted from every remaining family as a negative sample to construct the initial negative sample data set. Next, we processed the initial data set. The first step was to delete the protein sequences containing illegal characters; the second step was to delete the protein sequences with a length less than 50; in the third step, the CD-HIT (Fu et al., 2012) program was used to remove redundancy and delete the protein sequences with more than 60% similarity. We eventually obtained 382 SUTs and 9,109 non-SUTs. This data set is extremely unbalanced, so we divided the negative sample data set into ten equally and took one as the experimental data, which is 911 non-SUTs. We divided the data set into an 80% training dataset and a 20% testing dataset and constructed the dataset as shown in Table 1.

2.3 Feature extraction

In the process of protein identification, feature extraction is a crucial step (Yang and Jiao, 2021). To improve the identification performance of the model, we tried to extract features with high identification and good specificity. In this study, we considered this problem from two perspectives, namely, physicochemical properties and evolutionary information. We tried three features and their various combinations. Finally, the k-separated-bigrams-PSSM which has the best performance according to the experimental result is selected as the feature representation method in our model.

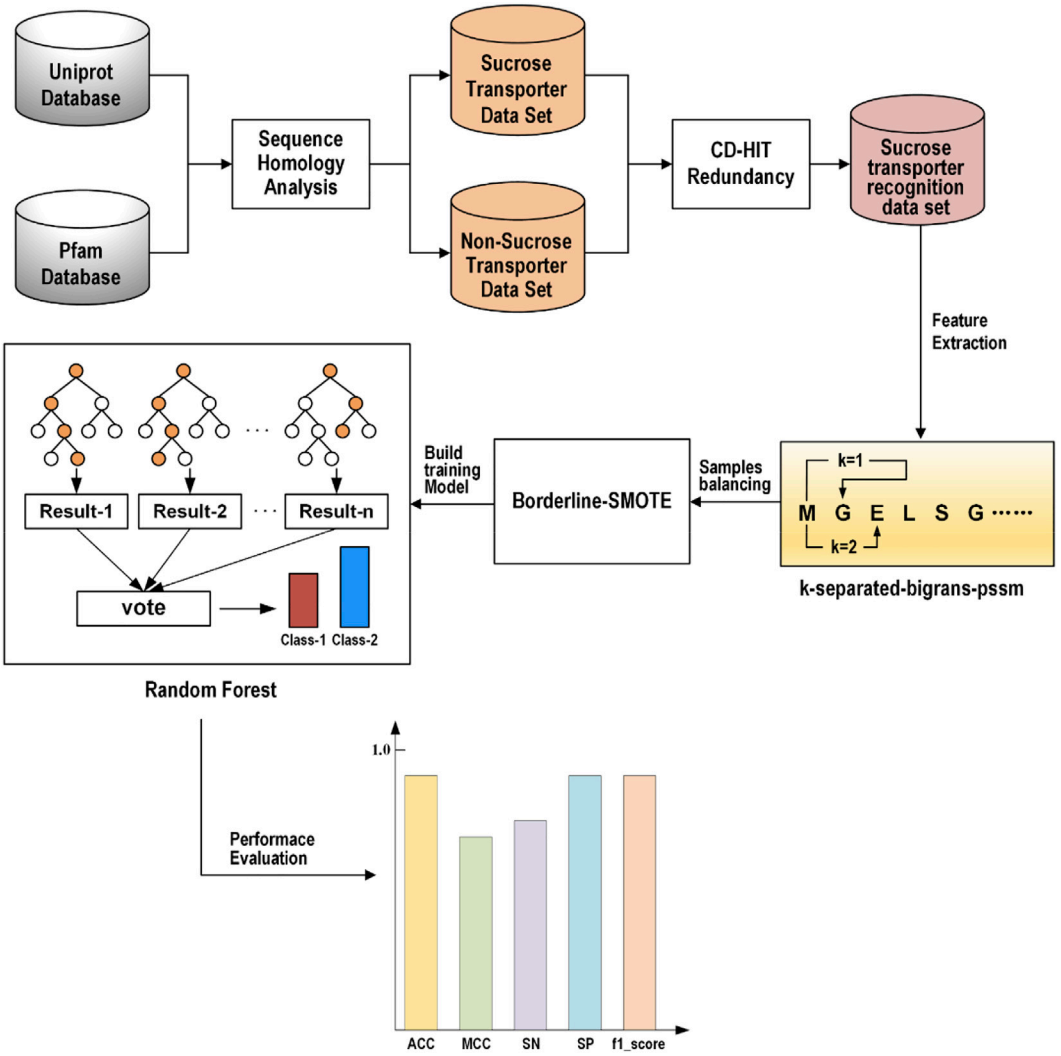

FIGURE 1
Frame chart of ISTRF.

TABLE 1 Self-built dataset.

Dataset	SUT	Non-SUT
Training dataset	306	729
Testing dataset	76	182

2.3.1 188D

188D includes the frequency of 20 amino acids and eight physical and chemical properties (Cai et al., 2003).

The formula for calculating the frequency of 20 amino acids is as follows:

$$F_i = \frac{N_i}{L}, (i = A, C, D, \ldots, Y),$$

where N_i is the number of amino acid type i, and L is the length of a protein sequence.

The composition, transition, and distribution are used to describe eight physicochemical properties of proteins (Xiong et al., 2018; Zou et al., 2019; Masoudi-Sobhanzadeh et al., 2021). Taking the hydrophobicity attribute as an example, "RKEDQN" is polar, "GASTPHY" is neutral, and "CVLIMFW" is hydrophobic. The frequency of each group can be expressed as follows:

$$C_i = \frac{N_i}{L}, i \in \{polar, neutral, hydrophobic\}.$$

The transition from polar group to neutral group is the frequency of polar residue following neutral residue or neutral residue following polar residue. The transition between the neutral group and hydrophobic group and the transition between the hydrophobic group and polar group have similar definitions. It can be expressed by the following formula:

$$T(i_1, i_2) = \frac{N(i_1, i_2) + N(i_2, i_1)}{L - 1}, (i_1, i_2) \in \{(polar, neutral), (neutral, hydrophobic), (hydrophobic, polar)\}.$$

The distribution consists of five values, which are the first, 25, 50, 75, and 100% positions of each group of amino acid in the sequence.

2.3.2 PSSM composition

PSSM composition is a feature that describes the evolutionary information of protein sequences, and it is also used to identify a variety of proteins (Wang et al., 2018; Ali et al., 2020; Qian et al., 2021). First, we run the PSI-BLAST tool (Ding et al., 2014) against the Uniref50 database with the e-value set to 0.001. We can obtain the original PSSM profile. Then, we summed the same amino acid rows together and divided the results by the number of amino acids in the protein sequence. Finally, a 400-dimensional PSSM composition was obtained.

2.3.3 The k-separated-bigrams-PSSM

The k-separated-bigrams-PSSM is generated from the original PSSM profile by column transformation. It represents the transition probabilities from one amino acid to another amino acid in a protein sequence (Wang et al., 2020), and the interval of the two amino acids is K. N represents the PSSM matrix, and L is the number of amino acids in the protein sequence and also the number of rows in the PSSM matrix. The transition from the m-th amino acid to the n-th amino acid can be expressed by the following formulas:

$$T_{m,n}(k) = \sum_{i=1}^{L-k} N_{i,m} N_{i+k,n},$$

where $1 \le m \le 20$, $1 \le n \le 20$, and $1 \le k \le K$.

$$T(k) = [T_{1,1}(k), T_{1,2}(k), \ldots, T_{1,20}(k), T_{2,1}(k), \ldots, T_{2,20}(k), \ldots, T_{20,1}(k), \ldots, T_{20,20}(k)].$$

For each k, T(k) is a 400-dimensional feature that represents 400 amino acid transitions. The k ranges from 1 to 11. When k is set to 1, it represents the transition probabilities between neighboring amino acids; when k is set to 2, it represents the transition probabilities between amino acids with one amino acid between them. We can obtain k-separated-bigrams-PSSM (k = 1) and a PSSM-related transformation matrix through POSSUM (Wang et al., 2017). The website is open, and users can easily obtain the required features.

2.4 Sample balancing

The training dataset constructed in Section 2.2 is an imbalance dataset, on which the classifier trained is biased to identify the unseen sample as the majority class (Shabbir et al., 2021). Therefore, we use the Borderline-SMOTE algorithm to balance the feature set. The SMOTE (Chawla et al., 2002) algorithm is an oversampling technique for synthesizing minority classes. It uses the KNN algorithm to calculate the k nearest neighbors of each minority class sample, randomly selects N samples, and performs random linear interpolation on the k nearest neighbors to construct new minority class samples. However, it does not consider the position of the adjacent majority class samples, which usually leads to the phenomenon of sample overlap and affects the classification effect (Chen et al., 2021). Borderline-SMOTE (Han et al., 2005) is an improved oversampling algorithm based on SMOTE. Because the boundary samples are more likely to be misclassified than those far from the boundary, the algorithm only oversamples the boundary samples of the minority class. In the Borderline-SMOTE algorithm, we used the KNN algorithm with k = 5 to balance the feature set of sucrose transporters, so that the 306 SUTs and 729 non-SUTs in the training set were expanded to 729 SUTs and 729 non-SUTs.

2.5 Classifier

In this study, we tried a lot of classification algorithms such as SVM, naive Bayes, SGD, and random forest (Ao et al., 2022). Eventually, we selected the random forest as our classifier based on the experimental results shown in Section 3.3. These machine learning algorithms can be implemented by the WEKA (Holmes et al., 1994; Garner, 1995) software. WEKA is an open data mining platform that can perform data processing such as classification, regression, and clustering. It contains a variety of machine learning algorithms and is simple to operate.

SVM is a supervised learning algorithm and is implemented by the SMO (sequential minimal optimization) algorithm in WEKA (Vapnik, 2006). The classical SVM algorithm has been applied to many problems of bioinformatics, especially in binary classification (Manavalan et al., 2018; Zhang et al., 2019; Ao et al., 2021; Zeng et al., 2021). The main idea is to find an optimal segmentation hyperplane and measure the maximum geometric distance between the nearest sample and the hyperplane so as to divide the data set correctly. The SMO algorithm is an improved support vector machine algorithm that aims to improve the efficiency of the support vector machine. It breaks the large quadratic programming (QP) problem into many smaller QP problems and avoids the problem that the time-consuming numerical QP optimization is used in the inner loop (Platt, 1998).

Naive Bayes is a very classical and simple classification algorithm (Cao et al., 2003). The idea of the algorithm is also

very simple. For a given sample to be classified, the probability that it belongs to the positive sample and the negative sample is solved firstly. And then the sample will be classified into the category with the higher probability. It assumes that each input variable is independent. Although real life cannot meet this assumption, it is still valid for most complex problems.

Stochastic gradient descent (SGD) is often used to learn linear classifiers under convex loss functions such as logistic regression and support vector machines (Bottou, 2010). The SGD algorithm is proposed to solve the problem that batch gradient descent needs to use all the samples for each parameter update, and the speed is slow when the number of samples is large. The characteristic of the SGD algorithm is that in each iteration, a group of samples is randomly chosen for training. After N iterations, it finds out the coefficient which leads to the minimum error of these models.

Random forest is based on the idea of ensemble learning, and it integrates multiple decision trees to obtain classification results (Breiman, 2001). First of all, select k samples repeatedly and randomly from the original training sample set N to generate a new training sample set. Then, n decision trees are generated using the training sample set as input. These decision trees form a random forest. Each decision tree is a classifier. As many decision trees as there are, there are as many classification results. Finally, the random forest integrates the classification results of n decision trees and identifies the class with most votes as the classification result of the sample. Because of this randomness, the random forest has a good anti-noise capability and is very suitable for processing high-dimensional data and avoiding overfitting. In many studies, random forest has shown a good classification effect (Lv et al., 2019; Ru et al., 2019; Ao et al., 2020; Lv et al., 2020; Petry et al., 2020; Zhang et al., 2021a).

2.6 Measurement

We used five indicators to evaluate the performance of our identification model: sensitivity (SN), specificity (SP), accuracy (ACC), Marshall correlation coefficient (MCC), and F-measure (Basith et al., 2020; Zhang et al., 2021b; Lee et al., 2021). These evaluation indicators were the results of the confusion matrix calculation obtained from the experiment, and the calculation formula is as follows:

$$SN = \frac{TP}{TP + FN}$$

$$SP = \frac{TN}{TN + FP}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

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

$$FR = \frac{TP}{TP + FP}$$

$$F - Measure = \frac{2 \times SN \times PR}{SN + PR}$$

where TP represents the number of correctly predicted sucrose transporters, TN represents the number of correctly predicted non-sucrose transporters, FP represents the number of incorrectly predicted sucrose transporters as non-sucrose transporters, and FN represents the number of incorrectly predicted non-sucrose transporters as sucrose transporters.

3 Results and discussion

3.1 Performance of different features

As shown in the frame chart of ISTRF in Section 2.1, our model extracted the k-separated-bigrams-PSSM feature to encode samples. To prove the effectiveness of our feature extraction method, we conducted experiments to compare the performance of different feature extraction algorithms. Specifically, we selected 188D, PSSM composition, k-separated-bigrams-PSSM, and their combinations. 188D feature reflected the frequency of 20 amino acids and eight physical and chemical properties, while PSSM composition and k-separated-bigrams-PSSM reflected the evolutionary information of protein sequences. We used the random forest as a classifier and did not apply Borderline-SMOTE to the extracted feature, and the experimental results of different features on 10-fold cross-validation are shown in Table 2. Bold values in the table indicate the best results. According to the number of indicators with the highest value, the number of k-separated-bigrams-PSSM is 4, the number of combinational features of 188D and k-separated-bigrams-PSSM is 4, and the number of other features and combinational features is lower or equal to 1. The k-separated-bigrams-PSSM has fewer feature numbers than the combination of 188D and k-separated-bigrams-PSSM; therefore, the former has the best performance according to the number of indicators with the highest value. According to the indicator of ACC and MCC, k-separated-bigrams-PSSM still has the highest value, and it verified that k-separated-bigrams-PSSM has the best general performance. Considering the indicator of SN, our used k-separated-bigrams-PSSM also has the maximum value. It verifies that our feature extraction method has better performance than other methods in predicting sucrose transporter protein from positive examples. Considering the indicator of SP, our feature extraction method is slightly lower than the combinational feature of PSSM composition and k-separated-bigrams- PSSM and is equal to or higher than other methods. However, the indicators of SN, MCC, and ACC of our feature extraction method are obviously larger than the combinational feature of PSSM composition and k-separated-bigrams-PSSM, which verify that

TABLE 2 Result of various feature extraction methods using random forest without Borderline-SMOTE on 10-fold cross-validation.

Feature	SN	SP	ACC	MCC	F-measure
188D	0.895	0.970	0.948	0.874	0.910
PSSM composition	0.876	0.967	0.940	0.855	0.896
k-separated-bigrams-PSSM	**0.925**	0.973	**0.958**	**0.900**	**0.929**
188D + PSSM composition	0.895	0.973	0.950	0.878	0.913
188D + k-separated-bigrams-PSSM	**0.925**	0.973	**0.958**	**0.900**	**0.929**
PSSM composition + k-separated-bigrams-PSSM	0.908	**0.978**	0.957	0.897	0.927
188D + PSSM composition + k-separated-bigrams-PSSM	0.918	0.973	0.957	0.895	0.926

Bold values in the table indicate the best results.

TABLE 3 Result of various feature extraction methods using SGD without Borderline-SMOTE on 10-fold cross-validation.

Feature	SN	SP	ACC	MCC	F-measure
188D	0.866	0.951	0.926	0.821	0.873
PSSM composition	0.873	0.956	0.931	0.834	0.883
k-separated-bigrams-PSSM	**0.964**	0.952	**0.956**	**0.897**	**0.928**
188D + PSSM composition	0.902	0.959	0.942	0.861	0.902
188D + k-separated-bigrams-PSSM	0.912	0.952	0.940	0.857	0.900
PSSM composition + k-separated-bigrams-PSSM	0.905	**0.967**	0.949	0.877	0.913
188D + PSSM composition + k-separated-bigrams-PSSM	0.912	0.960	0.946	0.870	0.909

Bold values in the table indicate the best results.

TABLE 4 Result of various features using random forest with Borderline-SMOTE on 10-fold cross-validation.

Feature	SN	SP	ACC	MCC	F-measure
188D	0.989 + 9.4	0.937–3.3	0.963 + 1.5	0.927 + 5.3	0.964 + 5.4
PSSM composition	0.982 + 10.6	0.952–1.5	0.967 + 2.7	0.935 + 8	0.968 + 7.2
k-separated-bigrams-PSSM	0.986 + 6.1	0.970–0.3	0.978 + 2	0.956 + 5.6	0.978 + 4.9
188D + PSSM composition	0.982 + 8.7	0.952–2.1	0.967 + 1.7	0.935 + 5.7	0.968 + 5.5
188D + k-separated-bigrams-PSSM	0.984 + 5.9	0.945–2.8	0.964 + 0.6	0.929 + 2.9	0.965 + 3.6
PSSM composition + k-separated-bigrams-PSSM	0.984 + 7.6	0.957–2.1	0.971 + 1.4	0.941 + 4.4	0.971 + 4.4
188D + PSSM composition + k-separated-bigrams-PSSM	0.985 + 6.7	0.949–2.4	0.967 + 1	0.935 + 4	0.968 + 4.2

the combinational feature of PSSM composition and k-separated-bigrams-PSSM trends to identify a protein as a non-sucrose transporter protein. Based on the fact that training data are an unbalanced data set in which negative samples are larger than positive ones, our feature extraction method is less affected by unbalanced data. After balancing the training data, the SN of our feature method is larger than the combinational feature of PSSM composition and k-separated-bigrams-PSSM, and the detailed experimental results are shown in Section 3.2. Therefore, from the overall perspective, our method obviously performs better than all other methods.

To further illustrate that our feature extraction method also has better performance using other classifiers, we also conducted experiments on different features using an SGD classifier. Table 3 shows the experimental results. As we can see from Table 3, our feature extraction method has better performance than other methods according to the number of indicators with the highest value or ACC indicator or MCC indicator. All in all, our feature extraction method performs better than other feature extraction methods.

3.2 Experiments on sample balancing

As shown in the frame chart of ISTRF in Section 2.1, the sucrose transporter database built in this study has more negative samples than positive ones, and it is an imbalanced dataset that influences the classification performance of the machine learning

TABLE 5 Result of various features using SGD with Borderline-SMOTE on 10-fold cross-validation.

Feature	SN	SP	ACC	MCC	F-measure
188D	0.966 + 10	0.909–4.2	0.938 + 1.2	0.877 + 5.6	0.939 + 6.6
PSSM composition	0.975 + 10.2	0.938–1.8	0.957 + 2.6	0.914 + 8	0.958 + 7.5
k-separated-bigrams-PSSM	0.997 + 3.3	0.942–1	0.970 + 1.4	0.941 + 4.4	0.971 + 4.3
188D + PSSM composition	0.985 + 8.3	0.931–2.8	0.958 + 1.6	0.918 + 5.7	0.959 + 5.7
188D + k-separated-bigrams-PSSM	0.985 + 7.3	0.931–2.1	0.958 + 1.8	0.918 + 6.1	0.959 + 5.9
PSSM composition + k-separated-bigrams-PSSM	0.984 + 7.9	0.951–1.6	0.967 + 1.8	0.945 + 6.8	0.968 + 5.5
188D + PSSM composition + k-separated-bigrams-PSSM	0.988 + 7.6	0.940–2	0.964 + 1.8	0.928 + 5.8	0.965 + 5.6

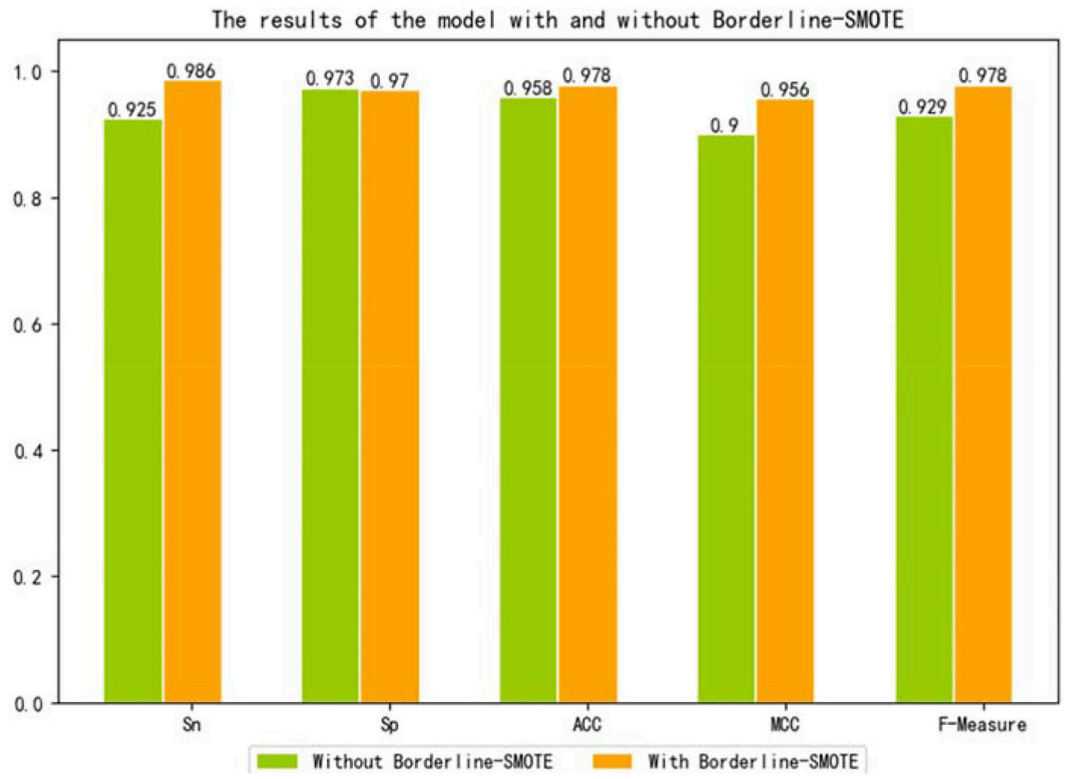

FIGURE 2
Results of the model with or without Borderline-SMOTE on 10-fold cross-validation.

algorithm. We adopted Borderline-SMOTE to augment the positive samples, and finally the number of positive samples is equal to negative samples. To verify that Borderline-SMOTE is effective for our model, we, respectively, conducted experiments using random forest and SGD on the basis of different features with Borderline-SMOTE. Experimental results are shown in Tables 4 and 5. The first number is the experimental result using Borderline-SMOTE, the second number is the percentage of increase or decrease relative to one without Borderline-SMOTE, and the plus sign denotes an increase, while the minus sign denotes a decrease. By comparing Table 4 with Table 2, we can see that the performance of features using Borderline-SMOTE is better than features not using Borderline-SMOTE in all indicators except indicator SP. The same conclusion is also obtained by comparing Table 5 with Table 3. In general, the features of Borderline-SMOTE can improve classification performance.

To further verify that our model can use Borderline-SMOTE to improve the classification performance, that is, Borderline-SMOTE is effective in our model. We compared the performance of our model with Borderline-SMOTE and

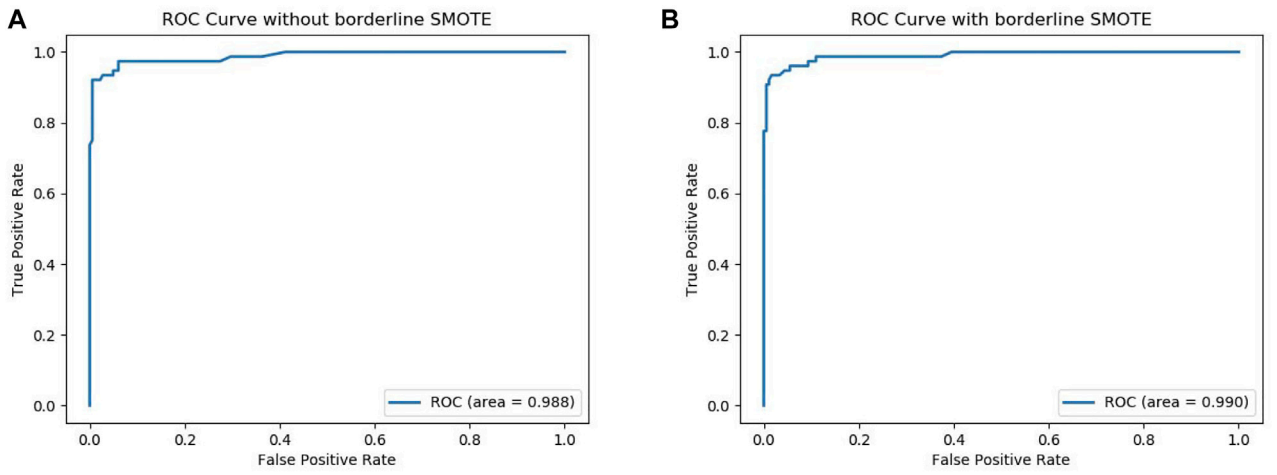

FIGURE 3
ROC curve with or without Borderline-SMOTE. **(A)** ROC curve without Borderline-SMOTE. **(B)** ROC curve with Borderline-SMOTE.

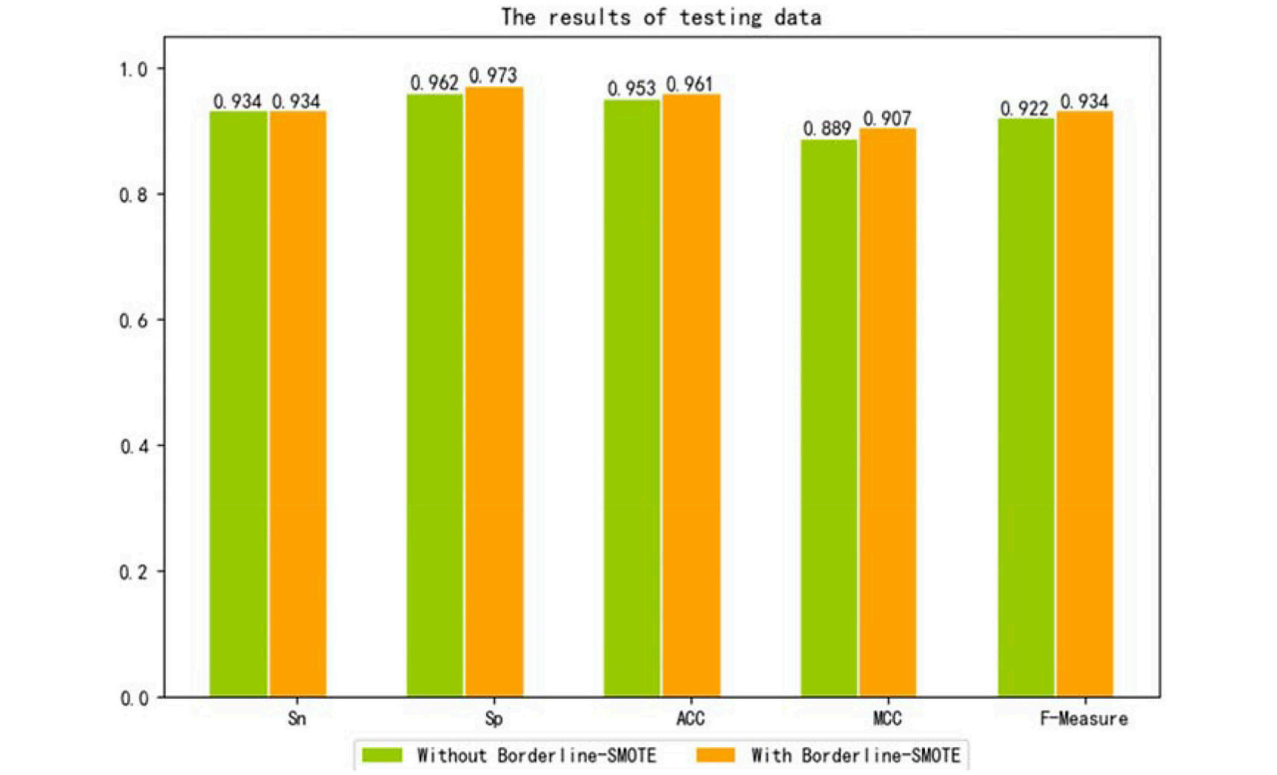

FIGURE 4
Results of the model with or without Borderline-SMOTE on the test dataset.

without Borderline-SMOTE on 10-fold cross-validation, and the experimental result is shown in Figure 2. Except for a slight decrease in SP, all other indicators improved by 2.0–6.1% in Figure 2, especially the indicator SN, which improved to its maximum. The decrease of SP verified that Borderline-SMOTE avoids our model being biased to classifying samples into

TABLE 6 Result of various classifiers using k-separated-bigrams-PSSM feature without Borderline-SMOTE on 10-fold cross-validation.

Classifier	SN	SP	ACC	MCC	F-measure
SVM	0.948	0.944	0.945	0.872	0.911
NB	**0.984**	0.782	0.842	0.703	0.786
SGD	0.964	0.952	0.956	0.897	0.928
RF	0.925	**0.973**	**0.958**	**0.900**	**0.929**

Bold values in the table indicate the best results.

TABLE 7 Result of various classifiers using k-separated-bigrams-PSSM feature with Borderline-SMOTE on 10-fold cross-validation.

Classifier	SN	SP	ACC	MCC	F-measure
SVM	**0.997**	0.877	0.937	0.880	0.940
NB	0.989	0.774	0.881	0.781	0.893
SGD	0.997	0.942	0.970	0.941	0.971
RF	0.986	**0.970**	**0.978**	**0.956**	**0.978**

Bold values in the table indicate the best results.

negative samples. The increase of SN verified that Borderline-SMOTE improves our model's identification ability of positive samples. The improvement of indicators of ACC, MCC, and F-measure verified that Borderline-SMOTE improved our model performance from a general perspective. Furthermore, the ROC curves of our model are plotted in Figure 3, and it can be seen that ISTRF with Borderline-SMOTE is superior to ISTRF without Borderline-SMOTE in the prediction of sucrose transporter protein.

To further evaluate the performance of Borderline-SMOTE in an unseen data set, we conducted experiments on the unseen data set. We used the testing set containing 76 sucrose transporters and 182 non-sucrose transporters to verify the model, and the experimental result is shown in Figure 4. By comparing the two models without and with Borderline-SMOTE, it was found that the latter performs better, which proves once again that Borderline-SMOTE improves the performance of our model.

3.3 Performance of various classifiers

As shown in the frame chart of ISTRF, we adopt random forest as a classifier to train the identification model. To verify that random forest has a better performance than other classifiers, we compared random forest with SVM, NB, and SGD. Table 6 showed the experimental result of 10-fold cross-validation using the k-separated-bigrams-PSSM feature without the Borderline-SMOTE as input. Table 7 showed the experimental result of 10-fold cross-validation using the k-separated-bigrams-PSSM feature with the Borderline-SMOTE as input.

TABLE 8 Experimental result of using different methods.

Model	ACC	MCC	SN	SP
ISTRF	**0.961**	**0.907**	**0.934**	0.973
BioSeq-SVM	0.9457	0.8694	0.9079	0.9615
BioSeq-RF	0.938	0.8505	0.8026	**0.9945**

Bold values in the table indicate the best results.

In Table 6, although the random forest classifier is slightly lower than BN on the SN indicator, it is obviously superior to the other four indicators. According to the number of indicators with the highest value, random forest obtained the four highest values and performs better than the compared classifiers. It is seen in Table 7 that random forest also performs better than the compared classifiers. All in all, random forest is effective in identifying sucrose transporter proteins.

3.4 Comparison with existing methods

To further evaluate the performance of ISTRF, our model is compared with the existing prediction method BioSeq-Analysis (Liu et al., 2019). The online address for this method is http://bioinformatics.hitsz.edu.cn/BioSeq-Analysis/PROTEIN/Kmer/. The SVM and random forest algorithms are used in the BioSeq-Analysis prediction method. We compared them separately. The prediction results are shown in Table 8. It can be seen from Table 8 that our identification model outperforms the compared models on the indicators of ACC, MCC, and SN. It verified that our identification model performs better in general.

4 Conclusion

A large number of experiments have proved that sucrose transporters play an important role in plant growth and crop yield. Therefore, the identification of sucrose transporters has become particularly important. With the rapid development of high-throughput sequencing technology, protein sequences can be easily obtained. In contrast, traditional biochemical technology needs a lot of human, material, and financial resources, and the identification of proteins through bioinformatics methods has become a popular trend. In this study, we introduced k-separated-bigrams-PSSM as the input feature, random forest as the classifier, and the Borderline-SMOTE algorithm to balance the training set. We achieved 0.978 accuracy, 0.986 SN, 0.970 SP, 0.956 MCC, and 0.978 F-measure on the training set. In order to verify the effectiveness of the model, the testing set was used for experiments, and the accuracy was 0.961. In the future, we will continue to find breakthroughs, optimize the experimental model, and strive to obtain better results.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

Author contributions

Conceptualization, DC and YC; data curation, SL; formal analysis, YC and DC; project administration, DC; writing—original draft, SL; and writing—review and editing, DC and YC.

Funding

This work is supported by the Research Start-up Funding Project of Qu Zhou University (BSYJ202112, BSYJ202109), the National Natural Science Foundation of China (61901103, 61671189), and the Natural Science Foundation of Heilongjiang Province (LH2019F002).

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Al-Sheikh Ahmed, S., Zhang, J., Ma, W., and Dell, B. (2018). Contributions of TaSUTs to grain weight in wheat under drought. *Plant Mol. Biol.* 98 (4), 333–347. doi:10.1007/s11103-018-0782-1

Alballa, M., Aplop, F., and Butler, G. (2020). TranCEP: Predicting the substrate class of transmembrane transport proteins using compositional, evolutionary, and positional information. *PloS one* 15 (1), e0227683. doi:10.1371/journal.pone.0227683

Ali, F., Arif, M., Khan, Z. U., Kabir, M., Ahmed, S., and Yu, D. J. (2020). SDBP-Pred: Prediction of single-stranded and double-stranded DNA-binding proteins by extending consensus sequence and K-segmentation strategies into PSSM. *Anal. Biochem.* 589, 113494. doi:10.1016/j.ab.2019.113494

Aluko, O. O., Li, C., Wang, Q., and Liu, H. (2021). Sucrose utilization for improved crop yields: A review article. *Int. J. Mol. Sci.* 22 (9), 4704. doi:10.3390/ijms22094704

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

Ao, C., Zhou, W., Gao, L., Dong, B., and Yu, L. (2020). Prediction of antioxidant proteins using hybrid feature representation method and random forest. *Genomics* 112 (6), 4666–4674. doi:10.1016/j.ygeno.2020.08.016

Ao, C., Zou, Q., and Yu, L. (2022). NmRF: Identification of multispecies RNA 2'-O-methylation modification sites from RNA sequences. *Brief. Bioinform.* 23 (1), bbab480. doi:10.1093/bib/bbab480

Aoki, N., Hirose, T., Scofield, G. N., Whitfeld, P. R., and Furbank, R. T. (2003). The sucrose transporter gene family in rice. *Plant Cell. Physiol.* 44 (3), 223–232. doi:10.1093/pcp/pcg030

Basith, S., Manavalan, B., Hwan Shin, T., and Lee, G. (2020). Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening. *Med. Res. Rev.* 40 (4), 1276–1314. doi:10.1002/med.21658

Bottou, L. (2010). "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*, Hamburg (Physica-Verlag HD), 177–186.

Breiman, L. (2001). Random forests. *Mach. Learn.* 45 (1), 5–32. doi:10.1023/a:1010933404324

Cai, C. Z., Han, L. Y., Ji, Z. L., Chen, X., and Chen, Y. Z. (2003). SVM-prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.* 31 (13), 3692–3697. doi:10.1093/nar/gkg600

Cao, J., Panetta, R., Yue, S., Steyaert, A., Young-Bellido, M., and Ahmad, S. (2003). A naive Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins. *Bioinformatics* 19 (2), 234–240. doi:10.1093/bioinformatics/19.2.234

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi:10.1613/jair.953

Chen, Y., Chang, R., and Guo, J. (2021). Effects of data augmentation method borderline-SMOTE on emotion recognition of EEG signals based on convolutional neural network. *IEEE Access* 9, 47491–47502. doi:10.1109/access.2021.3068316

Consortium, UniProt (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* 47 (D1), D506–D515. doi:10.1093/nar/gky1049

Ding, S., Yan, S., Qi, S., Li, Y., and Yao, Y. (2014). A protein structural classes prediction method based on PSI-BLAST profile. *J. Theor. Biol.* 353, 19–23. doi:10.1016/j.jtbi.2014.02.034

Endler, A., Meyer, S., Schelbert, S., Schneider, T., Weschke, W., Peters, S. W., et al. (2006). Identification of a vacuolar sucrose transporter in barley and Arabidopsis mesophyll cells by a tonoplast proteomic approach. *Plant Physiol.* 141 (1), 196–207. doi:10.1104/pp.106.079533

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28 (23), 3150–3152. doi:10.1093/bioinformatics/bts565

Garner, S. R. (1995). Weka: The waikato environment for knowledge analysis. *Proc. N. Z. Comput. Sci. Res. students Conf.* 1995, 57–64.

Han, H., Wang, W. Y., and Mao, B. H. (2005). "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *International conference on intelligent computing* (Berlin, Heidelberg: Springer), 878–887.

Ho, Q. T., Phan, D. V., and Ou, Y. Y. (2019). Using word embedding technique to efficiently represent protein sequences for identifying substrate specificities of transporters. *Anal. Biochem.* 577, 73–81. doi:10.1016/j.ab.2019.04.011

Holmes, G., Donkin, A., and Witten, I. H. (1994). "Weka: A machine learning workbench," in Proceedings of ANZIIS'94-Australian New Zealnd Intelligent Information Systems Conference, Brisbane, QLD, Australia, 29 November 1994 - 02 December 1994, 357–361.

Kühn, C., Barker, L., Bürkle, L., and Frommer, W. B. (1999). Update on sucrose transport in higher plants. *J. Exp. Bot.* 50, 935–953. doi:10.1093/jexbot/50.suppl_1.935

Lee, Y. W., Choi, J. W., and Shin, E. H. (2021). Machine learning model for predicting malaria using clinical information. *Comput. Biol. Med.* 129, 104151. doi:10.1016/j.compbiomed.2020.104151

Liu, B., Gao, X., and Zhang, H. (2019). BioSeq-Analysis2. 0: An updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level

based on machine learning approaches. *Nucleic Acids Res.* 47 (20), e127. doi:10.1093/nar/gkz740

Lv, H., Dao, F. Y., Zhang, D., Guan, Z. X., Yang, H., Su, W., et al. (2020). iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes. *Iscience* 23 (4), 100991. doi:10.1016/j.isci.2020.100991

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

Manavalan, B., Shin, T. H., and Lee, G. (2018). DHSpred: Support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget* 9, 1944–1956. doi:10.18632/oncotarget.23099

Mangukia, N., Rao, P., Patel, K., Pandya, H., and Rawal, R. M. (2021). Identifying potential human and medicinal plant microRNAs against SARS-CoV-2 3' utr region: A computational genomics assessment. *Comput. Biol. Med.* 136, 104662. doi:10.1016/j.compbiomed.2021.104662

Masoudi-Sobhanzadeh, Y., Jafari, B., Parvizpour, S., Pourseif, M. M., and Omidi, Y. (2021). A novel multi-objective metaheuristic algorithm for protein-peptide docking and benchmarking on the LEADS-PEP dataset. *Comput. Biol. Med.* 138, 104896. doi:10.1016/j.compbiomed.2021.104896

Mishra, N. K., Chang, J., and Zhao, P. X. (2014). Prediction of membrane transport proteins and their substrate specificities using primary sequence information. *PloS one* 9 (6), e100278. doi:10.1371/journal.pone.0100278

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L., et al. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49 (D1), D412–D419. doi:10.1093/nar/gkaa913

Petry, D., de Godoy Marques, C. M., and Marques, J. L. B. (2020). Baroreflex sensitivity with different lags and random forests for staging cardiovascular autonomic neuropathy in subjects with diabetes. *Comput. Biol. Med.* 127, 104098. doi:10.1016/j.compbiomed.2020.104098

Platt, J. (1998). *Sequential minimal optimization: A fast algorithm for training support vector machines*. Redmond, Washington, United States: Microsoft Research.

Qian, L., Jiang, Y., Xuan, Y. Y., Yuan, C., and SiQiao, T. (2021). PsePSSM-based prediction for the protein-ATP binding sites. *Curr. Bioinform.* 16 (4), 576–582. doi:10.2174/1574893615999200918183543

Ru, X., Li, L., and Zou, Q. (2019). Incorporating distance-based top-n-gram and random forest to identify electron transport proteins. *J. Proteome Res.* 18 (7), 2931–2939. doi:10.1021/acs.jproteome.9b00250

Shabbir, S., Asif, M. S., Alam, T. M., and Ramzan, Z. (2021). Early prediction of malignant mesothelioma: An approach towards non-invasive method. *Curr. Bioinform.* 16 (10), 1257–1277. doi:10.2174/1574893616666210616121023

Shah, S. M. A., Taju, S. W., Ho, Q. T., and Ou, Y. Y. (2021). GT-Finder: Classify the family of glucose transporters with pre-trained BERT language models. *Comput. Biol. Med.* 131, 104259. doi:10.1016/j.compbiomed.2021.104259

Tran, T. M., Hampton, C. S., Brossard, T. W., Harmata, M., Robertson, J. D., Jurisson, S. S., et al. (2017). *In vivo* transport of three radioactive [18F]-fluorinated deoxysucrose analogs by the maize sucrose transporter ZmSUT1. *Plant Physiol. biochem.* 115, 1–11. doi:10.1016/j.plaphy.2017.03.006

Vapnik, V. (2006). *Estimation of dependences based on empirical data*. Berlin, Heidelberg: Springer Science Business Media.

Wang, C., Li, J., Zhang, Y., and Guo, M. (2020). Identification of Type VI effector proteins using a novel ensemble classifier. *IEEE Access* 8, 75085–75093. doi:10.1109/access.2020.2985111

Wang, J., Yang, B., Revote, J., Leier, A., Marquez-Lago, T. T., Webb, G., et al. (2017). Possum: A bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. *Bioinformatics* 33 (17), 2756–2758. doi:10.1093/bioinformatics/btx302

Wang, S., Yang, J., Xie, X., Li, F., Wu, M., Lin, F., et al. (2019). Genome-wide identification, phylogeny, and expression profile of the sucrose transporter multigene family in tobacco. *Can. J. Plant Sci.* 99 (3), 312–323. doi:10.1139/cjps-2018-0187

Wang, Y. B., You, Z. H., Li, L. P., Huang, D. S., Zhou, F. F., and Yang, S. (2018). Improving prediction of self-interacting proteins using stacked sparse auto-encoder with PSSM profiles. *Int. J. Biol. Sci.* 14 (8), 983–991. doi:10.7150/ijbs.23817

Xiong, Y., Wang, Q., Yang, J., Zhu, X., and Wei, D. Q. (2018). PredT4SE-stack: Prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method. *Front. Microbiol.* 9, 2571. doi:10.3389/fmicb.2018.02571

Yan, Z. X., Yang, H. Y., Zhang, C. H., Wu, W. L., and Li, W. L. (2021). Functional analysis of the blackberry sucrose transporter gene RuSUT2. *Russ. J. Plant Physiol.* 68 (2), 246–253. doi:10.1134/s1021443721020217

Yang, L., and Jiao, X. (2021). Distinguishing enzymes and non-enzymes based on structural information with an alignment free approach. *Curr. Bioinform.* 16 (1), 44–52. doi:10.2174/1574893615666200324134037

Zeng, R., Lu, Y., Long, S., Wang, C., and Bai, J. (2021). Cardiotocography signal abnormality classification using time-frequency features and Ensemble Cost-sensitive SVM classifier. *Comput. Biol. Med.* 130, 104218. doi:10.1016/j.compbiomed.2021.104218

Zhang, H., Zhang, Q., Jiang, W., and Lun, X. (2021). Clinical significance of the long non-coding RNA NEAT1/miR-129-5p axis in the diagnosis and prognosis for patients with chronic heart failure. *Exp. Ther. Med.* 16 (4), 512–523. doi:10.3892/etm.2021.9943

Zhang, L., Huang, Z., and Kong, L. (2021). CSBPI_Site: Multi-information sources of features to RNA binding sites prediction. *Curr. Bioinform.* 16 (5), 691–699. doi:10.2174/1574893615666210108093950

Zhang, M., Li, F., Marquez-Lago, T. T., Leier, A., Fan, C., Kwoh, C. K., et al. (2019). MULTiPly: A novel multi-layer predictor for discovering general and specific types of promoters. *Bioinformatics* 35 (17), 2957–2965. doi:10.1093/bioinformatics/btz016

Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: Gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. *Rna* 25 (2), 205–218. doi:10.1261/rna.069112.118