International Journal of Biological Macromolecules 228 (2023) 706–714

Contents lists available at ScienceDirect

International Journal of Biological Macromolecules

journal homepage: www.elsevier.com/locate/ijbiomac

AcrPred: A hybrid optimization with enumerated machine learning algorithm to predict Anti-CRISPR proteins

Fu-Ying Dao [a,b], Meng-Lu Liu [a], Wei Su [a], Hao Lv [a,c,d], Zhao-Yue Zhang [a], Hao Lin [a,*], Li Liu [e,*]

[a] Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China
[b] School of Biological Sciences, Nanyang Technological University, Singapore 639798, Singapore
[c] Department of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
[d] SIB Swiss Institute of Bioinformatics, Zurich, Switzerland
[e] Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou 324003, China

ARTICLE INFO

Keywords:
Anti-CRISPR protein
Machine learning
Web-server

ABSTRACT

CRISPR-Cas, as a tool for gene editing, has received extensive attention in recent years. Anti-CRISPR (Acr) proteins can inactivate the CRISPR-Cas defense system during interference phase, and can be used as a potential tool for the regulation of gene editing. In-depth study of Anti-CRISPR proteins is of great significance for the implementation of gene editing. In this study, we developed a high-accuracy prediction model based on two-step model fusion strategy, called AcrPred, which could produce an AUC of 0.952 with independent dataset validation. To further validate the proposed model, we compared with published tools and correctly identified 9 of 10 new Acr proteins, indicating the strong generalization ability of our model. Finally, for the convenience of related wet-experimental researchers, a user-friendly web-server AcrPred (Anti-CRISPR proteins Prediction) was established at http://lin-group.cn/server/AcrPred, by which users can easily identify potential Anti-CRISPR proteins.

1. Introduction

CRISPR-Cas (clustered regularly interspaced short palindromic repeats-CRISPR-associated) is the prokaryotic adaptive immune system across bacteria and archaea, which recognizes and cleaves invading nucleic acids in a sequence-specific manner for defense process [1]. The CRISPR array and multiple cas genes make up its two essential components (Fig. 1). The former contains short repetitive elements (repeats, 25–35 bp long) separated by unique sequences (spacers, 26–72 bp), and the latter is translated into numerous effector proteins that eventually cleave foreign nucleic acid aggressor [2]. In 2012, the CRISPR-Cas engineered system was first presented as a potent gene-editing tool [3], and then was used for a wide variety of purposes. It has been used to treat neurodegenerative diseases, changing the DNA of plants, and particularly cancer therapy [4–7].

To combat the adaptability of CRISPR-Cas surveillance complexes, bacteriophages have evolved natural protein inhibitors for CRISPR-Cas systems, known as Anti-CRISPR (Acr) proteins [8]. To date, the reported Arcs span five CRISPR–Cas subtypes, including class 2 subtypes II, V, and class 1 subtypes I and III. However, no Acr protein of subtype IV has been reported so far [9–11]. The known mechanism of Acr proteins is to achieve the purpose of defense by directly interfering with foreign DNA, which can be roughly divided into three types: crRNA loading interference, DNA binding blockage and DNA cleavage prevention (Fig. 1). Researchers have found that Acr proteins control CRISPR-Cas-based genome editing in temporal, spatial and specific conditions by interfering with CRISPR-Cas activity. Shin et al. have revealed that the timed delivery of AcrIIA4 can allow on-target Cas9-mediated gene editing and significantly reduce off-target editing in human cells [12], indicating that Acr protein has potential clinical application prospects in the future. Moreover, Arc can be a robust "off-switch" for CRISPR-Cas systems. Hammond et al. have found that subtype II-A or II-C Acr proteins can inhibit the Cas9-based gene drive and have applied them to inhibit the transmission of malaria in mosquito [13]. As a natural inhibitor for CRISPR-Cas systems, the research on Acr proteins will help to understand the interaction between phage and host, and promote the development of biological editing technology.

Although Acr proteins of some subtypes of CRISPR-Cas systems have been identified, most CRISPR-Cas inhibitors are unknown. In addition, the inhibitory mechanisms of many Acr proteins have not been

* Corresponding authors.
E-mail addresses: hlin@uestc.edu.cn (H. Lin), liliu2010imu@163.com (L. Liu).

https://doi.org/10.1016/j.ijbiomac.2022.12.250
Received 14 November 2022; Received in revised form 12 December 2022; Accepted 22 December 2022
Available online 28 December 2022

described [14]. With the rapid increase in the number of validated Acr proteins, many computational algorithms for detecting Acr proteins have been developed. Eitzinger et al. designed AcRanker algorithm based on XGBoost and discovered two new Anti-CRISPR proteins: AcrllA20 and AcrIIA21 [15]. Gussow et al. found two new Anti-CRISPR proteins (AcrIC9 and AcrIC10) using random forests (RF)-based model, which was proved by wet experiments [16]. Wang et al. developed an ensemble-learning predictor, named PaCRISPR, by combining support vector machine (SVM) with sequence evolution information to accurately identify and visualize Anti-CRISPR proteins in genome and metagenome sequencing projects [17]. Yi et al. published a powerful tool called AcrFinder to pre-screen genomic data for finding potential Acr proteins by combining homology search, guilt-by-association (GBA), and CRISPR-Cas self-targeting spacers [18]. Wang et al. created an integrative database AcrHub to investigate, predict and map Acr proteins by integrating published state-of-the-art prediction tools [19]. Wandera et al. reported DeepAcr based on a deep learning algorithm for Acr identification and revealed a potent inhibitor of Cas13b nucleases named AcrVIB1 [20]. Recently, Zhu et al. proposed PreAcrs to identify Acr proteins from only sequences, which is a machine learning-based ensemble framework [21].

The above models reported good performances on the identification of Acr proteins, but there are still some issues. (i) With the progress of experimental technology and the discovery of new Acr proteins, the quality of the benchmark dataset of Acr proteins will become higher and more representative, so that high-quality models can be constructed under more favorable conditions. However, the datasets utilized in the earlier models have not been updated. (ii) Despite the considerable performance of published predictors, there is still much room for improvement in prediction accuracy. Establishing a high-precision prediction model is helpful to discover more novel Acr proteins. (iii) Acr proteins are widely present in bacteria, archaea and viruses, and there is an urgent need for a computational model that can accurately predict Acr proteins in three species. In view of these ideas, we collected and constructed a high-quality Acr protein benchmark dataset and obtained a computational model with the pretty prediction performance. Additionally, we established a user-friendly online server, AcrPred, to quickly identify potential Acr proteins.

2. Materials and methods

2.1. Data collection and preprocessing

The most popular and well-known Anti-CRISPR protein database (anti-CRISPRdb) was constructed in 2018 [22], and has been updated to version 2.2, containing 3679 real and predicted Acr proteins [23]. Another database is a unified resource provided by Bondy-Denomy et al. [9], which is mainly used for the registration and tracking of Anti-CRISPR names.

Based on the above data resources, we constructed a benchmark dataset according to following steps: (i) A total of 1472 Acr proteins was collected, including 1375 Acrs verified by experiment from anti-CRISPRdb and 97 Acr proteins from unified resource. (ii) By using CD-HIT software with a threshold of 0.4 to remove high similarity sequences, 231 non-redundant Acr proteins were remained and randomly divided into 205 and 26 Acr proteins for training and testing, respectively. (iii) 1162 non-Acr proteins were downloaded directly from PaCRISPR project [17]. Likewise, the non-Acrs were split into training and test datasets, which contain 902 and 260 samples, respectively. The training dataset is obviously out of balance, therefore we used the downsampling approach to address the problem [24]. The 5-fold cross-validation test was applied to evaluate model performance when training model based on training dataset, and test dataset was as independent data to further validate the obtained model.

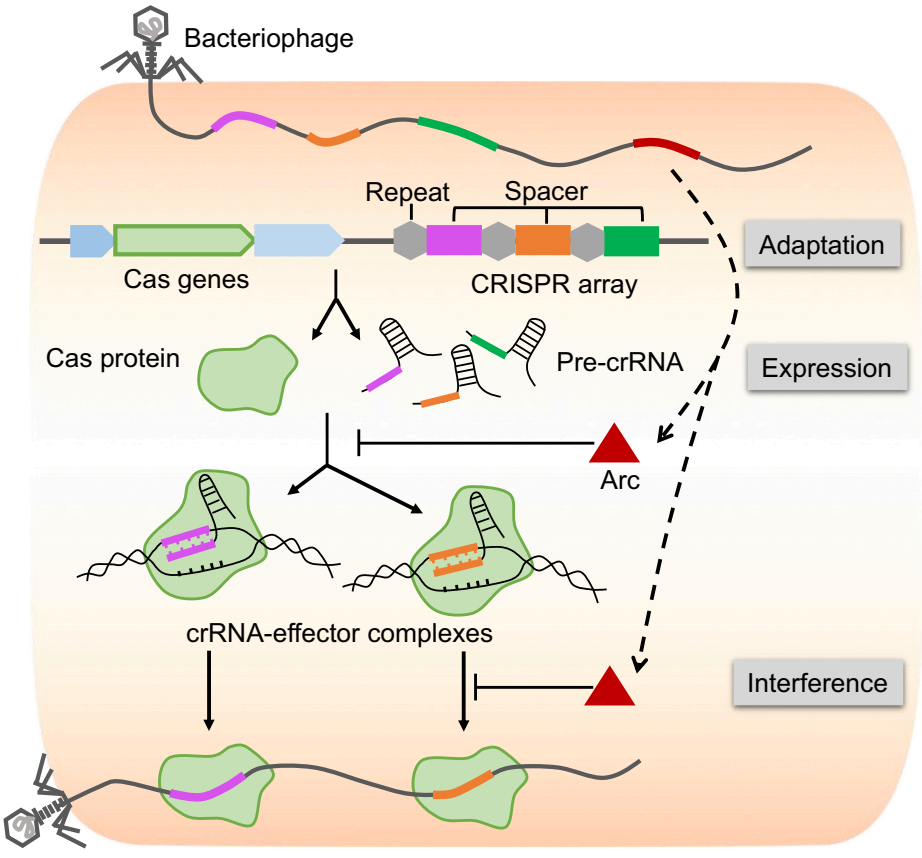


Fig. 1. Stages of CRISPR-Cas immunity and mechanisms of Acr function. The adaptive immunity by CRISPR-Cas systems functions in three stages: adaptation, processing, and interference. In the adaptation stage, invading short DNA fragments integrate into the host CRISPR array. The CRISPR array consists of some invading genetic elements (known as spacers) separated by repeats. During the expression stage, the CRISPR array is transcribed into precursor CRISPR-derived RNAs (crRNAs) and further processed into short mature crRNAs. At the target interference stage, crRNA-effector complexes are responsible for sequence-specific recognition and degradation of the invading nucleic acids. However, bacteriophages have evolved natural protein inhibitors for CRISPR-Cas systems. Anti-CRISPR (Acr) proteins can directly inactivate CRISPR–Cas systems by interfering at different immunity stages.

2.2. Feature extraction methods

To create a high-performance predictor, it is imperative to use an efficient mathematical expression to encode protein samples [25,26]. In this study, the following six feature encoding methods [27,28] were used to convert protein sequences into feature vectors.

2.2.1. DPC

Amino acid sequence is the most critical factor to determine protein function. To extract the short-range correlation information hidden behind residue sequences, the dipeptide composition (DPC) is a kind of *k*-peptide composition algorithm for describing the order between two amino acids. Protein sequences are formulated by DPC as follow:

$$DPC = \left[\frac{dp_1}{l-1}, \frac{dp_2}{l-1}, \ldots, \frac{dp_{400}}{l-1}\right]^{\mathrm{T}} \tag{1}$$

where dp_i is denoted as the occurrences number of *i*-th dipeptide. *T* is transpose symbol. *l* represents the length of a protein sequence, the same is true for the following formulas.

2.2.2. CTD

Composition, Transition, and Distribution (CTD) features, which indicate the amino acid distribution patterns of a certain structural or physicochemical property in a protein sequence [29], have been successfully employed in various functional and structural studies of proteins [26,30]. The CTD scheme groups 20 amino acids into three categories according to their corresponding attributes. Details about the division of the amino acids are provided in Table S1. Specifically, composition (C) refers to the global percent composition of 20 native amino acids, which is defined as:

$$C_j^i = \frac{n_j^i}{l} (i = 1, \ldots, 8; j = 1,2,3) \tag{2}$$

where n_j^i represents the number of the *j*-th class of amino acids in the *i*-th physicochemical properties.

Transition (T) characterizes the percent frequency with amino acids from one type of native amino acid followed by another type, which can be calculated by:

$$T_{j,k}^i = \frac{m_{j,k}^i + m_{k,j}^i}{l-1} \ (i = 1, \ldots, 8; j = 1,2,3; j < k \leq 3) \tag{3}$$

where for *i*-th physicochemical property, $m_{j,k}^i$ denotes number of occurrences of a transition from class *j* to class *k* and $m_{k,j}^i$ represents the number of occurrences of a transition from class *k* to class *j*.

D (distribution) refers to the respective locations of the first, 25 %, 50 %, 75 % and 100 % of amino acids with certain characteristics in a protein sequence. This feature can be computed as follow:

$$D_{j,q}^i = \frac{p_{j,q}^i}{l} \ (i = 1, \ldots, 8; j = 1,2,3; q = 1,25,50,75,100) \tag{4}$$

where $p_{j,q}^i$ denotes the minimum length of the sequence containing *q*% of amino acids given the *j*-th class for the *i*-th physicochemical property.

2.2.3. PSSM

Position-specific scoring matrix (PSSM) is defined as a matrix that involves information about the probability of amino acids or nucleotides occurrence in each position, which is derived from a multiple sequence alignment by running PSI-BLAST [31] against the uniref 50 database. The features extracted from PSSM to some extent track the evolutionary history of proteins to learn more informative patterns. It has been proven to improve the predictive performance of computational model both structurally and functionally about the query proteins [32,33]. The PSSM of a protein sequence was denoted as:

$$PSSM = [S_1, S_2, \ldots, S_{20}] \tag{5}$$

where S_i, ($i = 1, 2, \ldots, 20$) is the column vector of amino acid type *i* in the matrix. And each column vector can be represented as:

$$S_j = \left[S_{1,j}, S_{2,j}, \ldots, S_{l,j}\right]^T, (j = 1, \ldots, 20) \tag{6}$$

where $S_{i,j}$ is the score of number *j* residue in position *i* corresponding to the sequence order and *l* is the length of the protein sequence.

2.2.4. PSSM-composition

PSSM-composition can calculate the feature by summing rows that correspond to the same amino acid residues in the PSSM matrix. The sum value was divided by the length of the protein sequence for each type of amino acid (there is a total of 20 types). Thus, a vector of size at 400 (20 × 20) is finally used for representing a protein sequence sample.

2.2.5. DPC-PSSM

The algorithm is related to DPC and originally was proposed for protein structural class prediction [34]. For calculating this descriptor, the elements of two successive rows and two different columns are multiplied in PSSM matrix, which could generate a 400-dimension feature vectors as following:

$$DPC - PSSM = \left[D_{1,1}, \ldots, D_{1,20}, \ldots, D_{20,1}, \ldots, D_{20,20}\right]^{\mathrm{T}} \tag{7}$$

$$D_{i,j} = \frac{1}{l-1}\sum_{k=1}^{l-1} S_{k,i} \times S_{k+1,j} \ (1 \leq i,j \leq 20) \tag{8}$$

where $S_{k,i}$ represents the score of the *k*-th row and the *i*-th column in PSSM matrix.

2.2.6. PSSM-AC

The method stands for auto-covariance transformation [35], which can calculate the correlation between two residues within PSSM matrix. The feature vector is generated by:

$$PSSM\_AC_{j,g} = \frac{1}{l-g}\sum_{i=1}^{l-g}\left(S_{i,j} - \frac{1}{l}\sum_{i=1}^{l} S_{i,j}\right)\left(S_{i+g,j} - \frac{1}{l}\sum_{i=1}^{l} S_{i,j}\right) (j = 1, \ldots, 20) \tag{9}$$

where $S_{i,j}$ is the score of the residue of the *i*-th position mutated to the *j*-th amino acids residue in the protein sequence. High score means highly conserved position. As a result, the protein sequence generates a 20 × *g* (*g* = 10) dimensional feature vector.

2.2.7. RPSSM

RPSSM is a new matrix by merging some columns of the original PSSM profile. For convenience, we denoted the standardized PSSM of the query sequence as S = (S_A, S_R, S_N, S_D, S_C, S_Q, S_E, S_G, S_H, S_I, S_L, S_K, S_M, S_F, S_P, S_S, S_T, S_W, S_Y, S_V), in which S_A, S_R, …, S_V represent the 20 columns in the original PSSM corresponding to the 20 native types of amino acids. The new matrix could be denoted as $P = (p_1, p_2, \ldots, p_{10})$, here,

$$\begin{cases} p_1 = \dfrac{S_F + S_Y + S_W}{3} \\ p_2 = \dfrac{S_M + S_L}{2} \\ p_3 = \dfrac{S_I + S_V}{2} \\ p_4 = \dfrac{S_A + S_T + S_S}{3} \\ p_5 = \dfrac{S_N + S_H}{2} \\ p_6 = \dfrac{S_Q + S_E + S_D}{3} \\ p_7 = \dfrac{S_R + S_K}{2} \\ p_8 = S_C \\ p_9 = S_G \\ p_{10} = S_P \end{cases} \tag{10}$$

Then, RPSSM is transformed into a 10-dimensional feature vector by using the following formula.

$$D_s = \frac{1}{l}\sum_{i=1}^{l}\left(p_{i,s} - \frac{1}{l}\sum_{i=1}^{l} p_{i,s}\right)^2 \quad (s = 1, \ldots, 10) \tag{11}$$

where $p_{i,\ s}$ represents the element in the i-th row and s column of the simplified PSSM. The simplified PSSM can also be further transformed into a 10 × 10 matrix by mining its local sequence order information.

$$D_{s,t} = \frac{1}{l-1}\sum_{i=1}^{l-1} X_{i,i+1} \tag{12}$$

$$X_{i,i+1} = \frac{\left(p_{i,s} - p_{i+1,t}\right)^2}{2} (s, t = 1, \ldots, 10) \tag{13}$$

Finally, RPSSM can be expressed as a 110-dimensional features by integrating $D_{s,\ t}$ and D_s.

2.3. Support vector machine (SVM)

SVM is one of the most widely used supervised learning algorithms in computational biology, especially suitable for the binary classification tasks [36]. The basic idea of SVM is to transform the input vector into a high-dimensional Hilbert space and seek a separating hyperplane in this space. Here, we use the Scikit-learn Python library (sklearn.svm.SVC) to construct predictor based on feature vectors extracted from protein sequence data, and performed a grid search method to optimize the two parameters C and γ in the search spaces [2^{-5}, 2^{13}] and [2–13,23], respectively.

2.4. Performance evaluation

To evaluate the prediction performance of the proposed method, four standard statistical indicators are commonly performed, namely *ACC*, *SN*, *SP*, and *MCC* [37,38], which are expressed as:

$$\begin{cases} SN = \dfrac{TP}{TP+FN}, 0 \le Sn \le 1 \\ SP = \dfrac{TN}{TN+FP}, 0 \le Sp \le 1 \\ ACC = \dfrac{TP+TN}{TP+TN+FP+FN}, 0 \le Acc \le 1 \\ MCC = \dfrac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)\times(TN+FN)\times(TP+FP)\times(TN+FP)}}, -1 \le MCC \le 1 \end{cases} \tag{14}$$

where *TP*, *TN*, *FP* and *FN* represent the number of true positive, true negative, false positive, and false negative, respectively.

Additionally, the AUC (area under the receiver operating characteristic curve) [39], a comprehensive indicator that reflects the *SN* and *SP* of continuous variables, is used to objectively evaluate the proposed model. The AUC ranges from 0 to 1, the larger the better.

3. Results

3.1. The training strategy of AcrPred

Inspired by ensemble learning [40,41], an efficient scheme that minimizes the generalization error rate, we employed a two-step fusion model strategy to solve the issue of imbalance classification and maximize the performance of the model as following steps (Fig. 2).

First, the balanced training subsets were constructed. We down-sampled the negative samples five times to create five balanced training subsets. Among them, there is no overlap among the five negative subsets. This enables us to make fully use of the information provided by the negative samples in the case of the ratio of 1:5 for positive and negative samples.

Second, classification model was generated for each feature extraction method by using SVM with 5-fold cross-validation test. We first encoded five training subsets by a feature extraction algorithm. Then, considering the weakening of model performance and the burden of computational resources caused by redundant features, we applied analysis of variance (ANOVA) [42] and incremental feature selection method (IFS) [43] to score feature values and select the five optimal feature subsets, respectively. Next, five optimal sub-models were built by using SVM on the basis of five optimal feature subsets. Ultimately, the classification model was obtained by integrating the five sub-models, which is the first-step model fusion.

Thirdly, six classification models (M_1-M_6 in Fig. 2) were established based on CTD, DPC, PSSM-composition, DPC-PSSM, RPSSM and PSSM-AC using the same training strategy in the second step. Their prediction performance was evaluated by independent test dataset (Fig. 3b).

Fourthly, the predictive accuracy is further improved by integrating six models. We considered to fuse the six models to produce a total of 63 different combinations, which is the second-step model fusion. The average output probability of models was used as the final prediction score for each fusion model. After processing with the two-step model fusion strategy, the best combination of models will be selected by evaluating on the independent test dataset (Fig. 3c).

3.2. Performance evaluation of six feature extraction methods

According to the second step of training strategy described above, we can obtain six models by using CTD, DPC, PSSM-composition, DPC-PSSM, RPSSM and PSSM-AC, respectively. To emphasize the necessity of feature selection, we took the CTD descriptor as an example to show the IFS process based on ANOVA (Fig. 3a). One may observe that the maximum AUC value of 0.789 is obtained when the top 90-dimension features are used as input. After that, with the increase of the number of features, the model's performance gradually deteriorates. Thus, we could conclude that some redundant information can prevent the classifier from making correct choices.

We also assessed the robustness and generalization of the six models on the test dataset. From Fig. 3b, we found that: (i) PSSM-based features could always obtain better performance in terms of *SP*, *ACC*, *MCC* and AUC values. Although both DPC and CTD achieved acceptable model performance, it is clear that the short-range sequence information and physicochemical properties-based information are insufficient to characterize Acr proteins. Instead, the evolutionary information captured by PSSM-based approaches could track the evolutionary history of proteins and learn more informative patterns from Acr protein sequences. This finding is also consistent with the conclusion of previous study that Acr

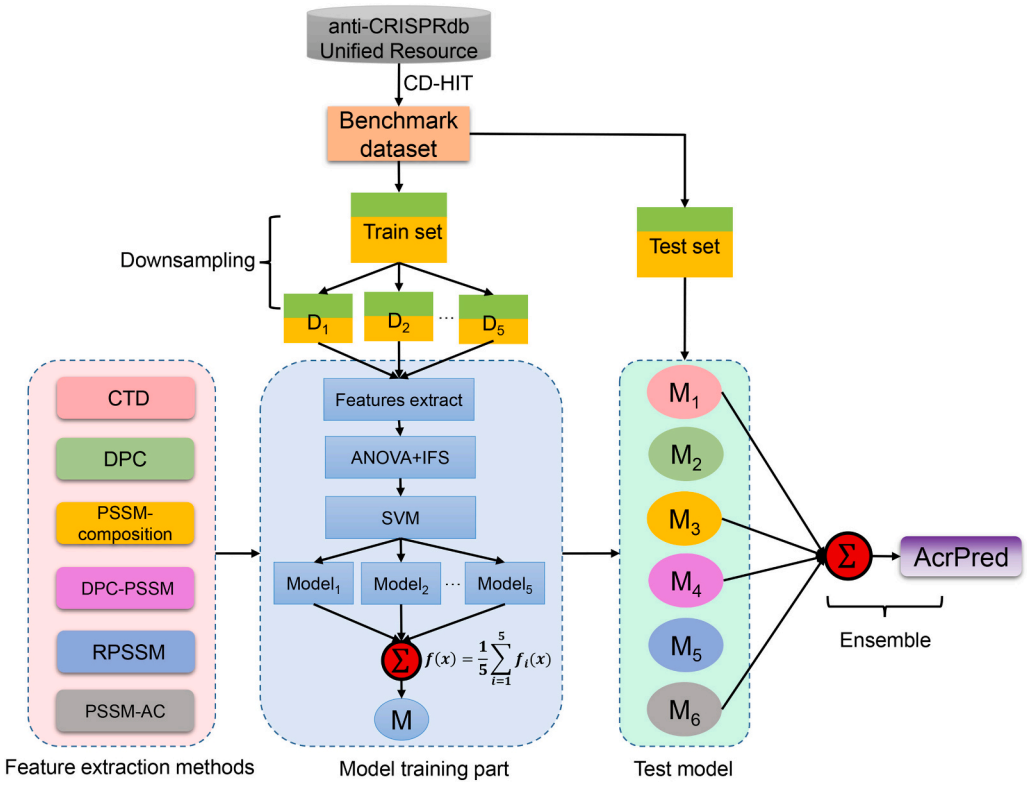


Fig. 2. The methodology of ensemble model construction by using two-step fusion model strategy. For train set, five sub-datasets are generated by 5-time downsampling to solve the data imbalance problem in model training. In the model training part, firstly, each sub-dataset is extracted features by feature extraction methods. Then, ANOVA-based IFS process is used to obtain optimal sub-classifiers. Finally, integrating sub-classifiers to obtain M (represents the final classifier produced by a certain feature extraction method) by averaging their prediction outputs. M_1 to M_6 represent the final classifiers obtained with CTD, DPC, …, PSSM-AC, respectively. Eventually, the best fusion model is based on the CTD, PSSM-composition, DPC-PSSM and PSSM-AC.

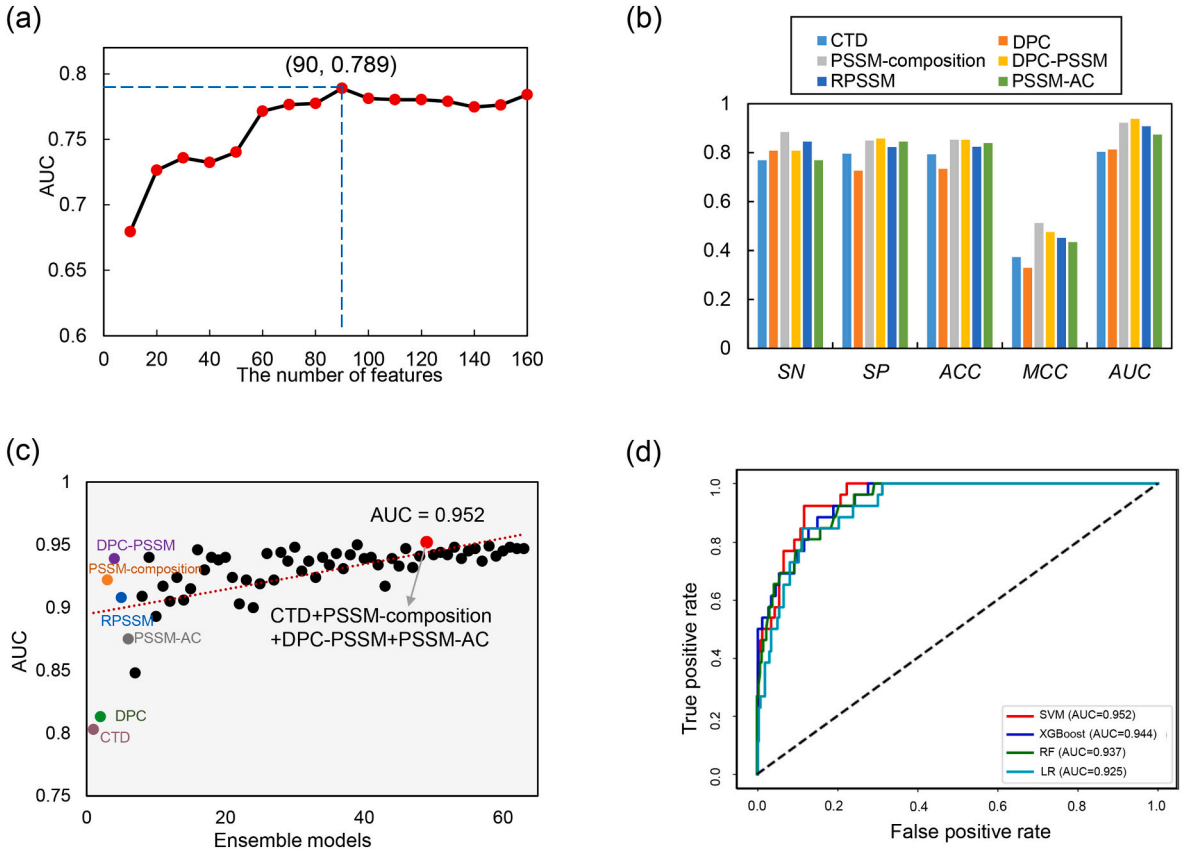


Fig. 3. Model training. (a) ANOVA-based IFS process in each downsampling when using CDT to extract feature vectors. (b) The histogram to show the performances of six feature encoding schemes in test datasets. (c) Figure shows the AUC values of 63 fusion models, in which the best fusion model is based on the CTD, PSSM-composition, DPC-PSSM and PSSM-AC. (d) Comparison of the performance on different classifiers based on best fusion model.

proteins may initially originate from bacteriophages and then be transferred to in bacteria and archaea as a mobile genetic elements (MGEs) [44]. (ii) As expected, the DPC-PSSM method which combines the advantages of DPC and PSSM showed the best performance with the AUC of 0.939. Therefore, we speculate ensemble model can achieve a more stable and accurate prediction probability by combining short-range sequence information and evolutionary features of proteins.

3.3. Prediction performance on different fusion models

Generally, the combination of models can lead to better prediction accuracy. We obtained 63 different fusion models by combining the above-mentioned six classifiers (M_1 to M_6) and evaluated their performance on the test dataset (Fig. 3c and Table S2). We observed that most of the fusion models achieved better performance when compared with their corresponding single models. For example, the voting fusion model based on CTD, PSSM-composition, DPC-PSSM and PSSM-AC significantly could produce the best prediction performance with highest average AUC values of 0.952. However, the performance of the fusion model based on the combination of CTD and PSSM-composition is worse than single model constructed by PSSM-composition, which indicates that fusing model may bring noise to reduce the robustness of the model in a few cases. Nevertheless, we can still conclude that the model fusion strategy is effective in the detection of Acr proteins and could produce significant improvement of model performance.

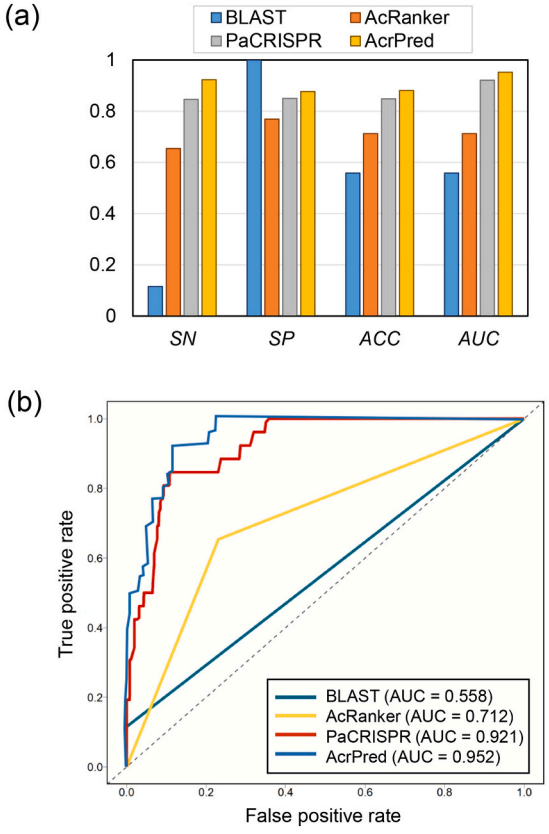
There are also some classification algorithms that perform well on small sample datasets, such as RF, XGBoost and LR, etc. Therefore, we compared four models trained by the four different classification algorithms (Fig. 3d). It can be seen that SVM exhibits a superior performance (AUC = 0.952) in identifying Acr proteins compared with other classification algorithms. Eventually, the final Acr proteins prediction model, named AcrPred, was generated through training using SVM. This model is an ensemble classifier based on four single models, namely CTD, PSSM-composition, DPC-PSSM and PSSM-AC.

In short, we established AcrPred through two-step model fusion strategy, in which the first step is to take the average of the sub-models generated by multiple downsampling for each single feature set, the second step is to further integrate the models constructed by multiple feature sets (Fig. 2). The strategy not only overcomes the problem of imbalance classification, but also picks out the optimal model with the maximum accuracy.

3.4. Comparison with published tools

To further prove the superiority of our proposed method, we need to compare AcrPred with other published tools based on machine learning methods. Here, three most popular tools were selected: BLAST [45], AcRanker [15] and PaCRISPR [17]. The same test dataset was utilized to maintain fairness and accuracy of the comparison, and the generated prediction results were recorded in Fig. 4a–b.

Obviously, AcrPred is significantly superior to other existed models for identifying Acr proteins as shown in *SN*, *ACC*, AUC and ROC curves, and is lower than BLAST only in SP. The prediction principle of BLAST is sequence alignment to find similar sequences. When two peptide sequences are quite similar, they most likely belong to the same protein. However, sequence similarity between Acrs and non-Acrs is generally poor, and BLAST can only accurately predict a small percentage of Acrs (high *SP* and low *SN*). Moreover, AUC is the most important metrics in binary classification. We found that BLAST produced the lowest AUC value. Overall, AcrPred produced stable performance on *SN*, *SP*, *ACC* and AUC, which are 0.923, 0.877, 0.881 and 0.952, respectively, indicating that our proposed AcrPred provides excellent predictive ability



(c)

Name	Predicted results: (Correct Incorrect)		
	AcrPred	PaCRISPR	AcRanker
AcrIA1	0.299	0.407	-5.51
AcrIIA24	0.724	0.783	-5.54
AcrIIA25	0.538	0.632	-5.37
AcrIIA26	0.75	0.896	-4.04
AcrIIA27	0.663	0.647	-3.58
AcrIIA28	0.786	0.886	-4.56
AcrIIA29	0.718	0.783	-5.62
AcrIIA30	0.598	0.589	-3.61
AcrIIA31	0.611	0.603	-1.43
AcrIIA32	0.502	0.608	-4.04

Fig. 4. Model evaluation and comparison. (a) Compared AcrPred with other published existing tools in *SN*, *SP*, *ACC* and AUC. (b) ROC curves of AcrPred, PaCRISPR, AcRanker and the BLAST-based baseline predictor based on the same independent dataset. (c) Prediction comparison of the case study Anti-CRISPR proteins. All models adopt default cut-off thresholds of 0.5 (for AcrPred and PaCRISPR) or −5 (for AcRanker) to determine prediction results.

compared with existing tools.

3.5. Case studies

Recently, 10 new Acr proteins have been discovered. Among them, 9 Acr proteins (AcrIIA24 to AcrIIA32) were found on the MGEs of *Streptococcus*, which can prevent Cas9 complex of type II-A CRISPR–Cas systems from binding or shearing DNA [46]. AcrIA1 is the first discovered type I-A Acr in *Sulfolobus*, which could effectively inhibit the acquisition of spacers [47]. The sequence similarity between Acr proteins in our benchmark dataset and these 10 new Acr proteins is <40 %. This difference provides the perfect case study to compare generalized prediction ability of our proposed method with its peers.

Here, we compared AcrPred with two other online tools (PaCRISPR and AcRanker). As shown in Fig. 4c, both AcrPred and PaCRISPR successfully identify 9 out of 10 Acrs. Whereas, AcRanker can only correctly pick out 6 Acr proteins. These observations suggest that AcrPred and PaCRISPR can detect new Acr proteins with low false-positive rates, which also indicate that the ability of AcRanker to recognize new Acr proteins needs to be further enhanced. Further analysis, from Fig. 4a–b, we can find AcrPred is significantly superior to PaCRISPR for identifying Acr proteins in *SN*, *SP*, *ACC* and AUC based on same test dataset, as well as the ROC curves. In addition, compared with PaCRISPR, AcrPred applied feature selection technique to obtain features with fewer dimensions but more informative features, which can save more computing resources. Therefore, we could reasonably speculate that as the available case increases, the advantage of AcrPred would be more pronounced.

Notably, none of these three tools could recognize AcrIA1. We speculate that AcrIA1 is specific in both evolution and sequence information compared with other Acr types. This leads to the inability of our model to identify novel features of I-A type. However, it is predictable that the predictive ability of our model for type I-A Acrs can be improved when more type I-A Acrs are discovered and used for model training.

3.6. The AcrPred online server

To provide convenience, we developed a free web server at http://lin-group.cn/server/AcrPred for easy access to AcrPred. Users could access and apply our classification model without the need to solve difficult mathematical equations.

The online server includes *Home page*, *Web Server*, *Download*, *User Guide*, *Citation*, *About Us* and *Contact* modules (Fig. 5). The *Web Server* page is the most important component of this online tool. If users want to know whether a protein sequence is Acr protein, first copy/paste the FASTA sequences of query proteins into the input box, and upload corresponding PSSM files obtained by PSI-BLAST. In addition, we provided a step-by-step tutorial to generate the PSSM file to maximize user's experience. Then, click the *submit* button to wait for the outcomes (Fig. 5a). Finally, the prediction results and the predict probability for each sequence will be displayed in the output panel (Fig. 5b). *Download* page gives the train and test datasets used in this study, as well as the 10 new Acr proteins used for model evaluation in the section of case study analysis (Fig. 5c).

To ensure that the AcrPred platform remains competitive and up-to-date, we will regularly maintain and timely upload newly discovered Anti-CRISPR proteins. And we also will expand the server's memory in the future to enable direct generation of PSSM matrices locally. AcrPred is expected to be a useful preliminary screening toolkit to identify potential Anti-CRISPR proteins, and therefore expediting the discovery of novel Acr proteins for subsequent experimental validation.

4. Discussion

CRISPR-Cas is a promising innovative technology for genomic editing, which offers scientists a chance to edit DNA structure and change gene function. Anti-CRISPR (Acr) proteins have great potential to serve as regulators of CRISPR-Cas genome editing tools for safer and more controllable genome engineering. However, the limitations of the relatively small number of known Acr proteins hinder the further

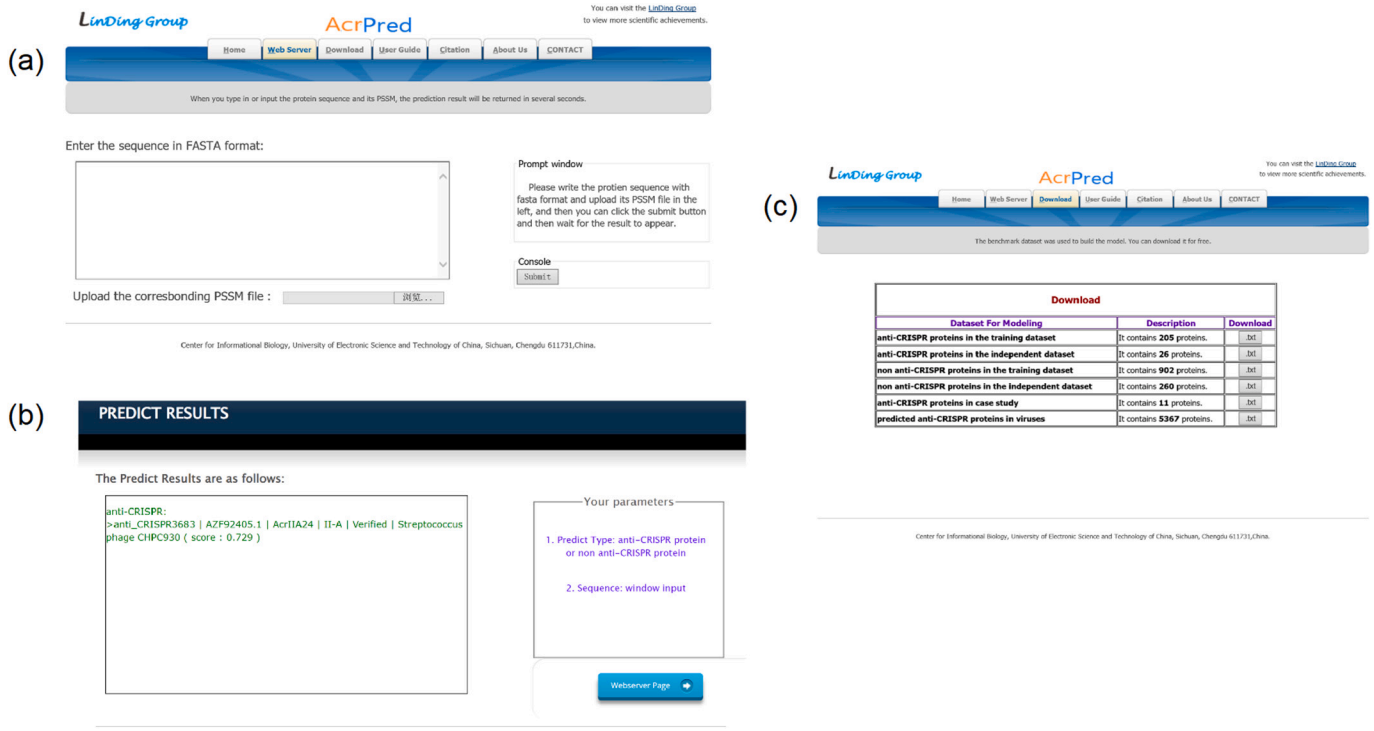


Fig. 5. Semi-screenshots to show the pages of the AcrPred platform. Its website address is at http://lin-group.cn/server/AcrPred/. (a) is the sequence submission page, where users can upload queried protein sequence and the corresponding PSSM matrix generated by PSI-BLAST. (b) is prediction result page, in which we give the identification result according to the prediction probability. (c) is download page with a variety of information including the Acrs in case study.

development of gene editing technology to a certain extent. To overcome this shortcoming, we proposed a machine learning model, called AcrPred, based on two-step model fusion strategy. The 5-fold cross-validation test, test dataset validation and case study demonstrated that AcrPred is robustness and has excellent generalization ability. In addition, AcrPred also shows significant advantage over other published tools. We expect that AcrPred will continue to discover more novel Acr proteins for yet unforeseen future biotechnology applications.

Additionally, there are still many issues to be addressed in this field. We plan to explore more information about Acr proteins from the following directions in the future. (i) We are going to identify Acrs using genomic contexts information of Anti-CRISPR genes. Most Anti-CRISPR genes are always found upstream of Anti-CRISPR-associated (ACA) genes encoding proteins containing a helix-turn-helix (HTH) DNA-binding domain. The conservation of ACA genes has served as a signpost for the identification of Acr genes [48]. (ii) A model for the identification of Acr proteins' types should be constructed. The classification of Acr types is based on the type of CRISPR-Cas system they inhibit, because each Acr inhibits a specific CRISPR-Cas system. Therefore, as the number of Acrs increases, it is meaningful to further determine the type after judging whether an unknown protein belongs to Acr. (iii) Potential Acrs in public databases should be identified. Encouraged by the low false positive rate of AcrPred in case study, we can perform AcrPred to identify new Acr proteins from public database (such as GenBank). That will help researchers analyze known and potential Acr proteins.

CRediT authorship contribution statement

Conceptualization: H. Lin, L. Liu, and M.-L.L. Investigation: F.-Y.D., H. Lv. and Z.-Y.Zhang. Coding: M.-L.L. and W. S. Writing - Original Draft: F.-Y. D and M.-L.L. Writing - Review and Editing: H. Lin and F.-Y. D. Funding acquisition: H. Lin. and L. Liu

Conflict of interest

The authors declare that they have no competing interests.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by a grant from the Sichuan Provincial Science Fund for Distinguished Young Scholars (2020JDJQ0012) and the National Natural Science Foundation of China (62272085). Fu-Ying Dao is supported by China Scholarship Council to visit Nanyang Technological University.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijbiomac.2022.12.250.

References

[1] A. Butiuc-Keul, A. Farkas, R. Carpa, D. Iordache, CRISPR-cas system: the powerful modulator of accessory genomes in prokaryotes, Microb. Physiol. 32 (1–2) (2022) 2–17.
[2] R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero, P. Horvath, CRISPR provides acquired resistance against viruses in prokaryotes, Science 315 (5819) (2007) 1709–1712.
[3] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J.A. Doudna, E. Charpentier, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity, Science 337 (6096) (2012) 816–821.
[4] L. Cong, F.A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P.D. Hsu, X. Wu, W. Jiang, L.A. Marraffini, F. Zhang, Multiplex genome engineering using CRISPR/Cas systems, Science 339 (6121) (2013) 819–823.
[5] M. Adli, The CRISPR tool kit for genome editing and beyond, Nat. Commun. 9 (1) (2018) 1911.
[6] L.S. Qi, M.H. Larson, L.A. Gilbert, J.A. Doudna, J.S. Weissman, A.P. Arkin, W. A. Lim, Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression, Cell 152 (5) (2013) 1173–1183.
[7] P.D. Hsu, E.S. Lander, F. Zhang, Development and applications of CRISPR-Cas9 for genome engineering, Cell 157 (6) (2014) 1262–1278.
[8] J. Bondy-Denomy, A. Pawluk, K.L. Maxwell, A.R. Davidson, Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system, Nature 493 (7432) (2013) 429–432.
[9] J. Bondy-Denomy, A.R. Davidson, J.A. Doudna, P.C. Fineran, K.L. Maxwell, S. Moineau, X. Peng, E.J. Sontheimer, B. Wiedenheft, A unified resource for tracking anti-CRISPR names, CRISPR J. 1 (2018) 304–305.
[10] N. Jia, D.J. Patel, Structure-based functional mechanisms and biotechnology applications of anti-CRISPR proteins, Nat. Rev. Mol. Cell Biol. 22 (8) (2021) 563–579.
[11] D. Trasanidou, A.S. Geros, P. Mohanraju, A.C. Nieuwenweg, F.L. Nobrega, R.H. J. Staals, Keeping crispr in check: diverse mechanisms of phage-encoded anti-crisprs, FEMS Microbiol. Lett. 366 (9) (2019) fnz098.
[12] J. Shin, F. Jiang, J.J. Liu, N.L. Bray, B.J. Rauch, S.H. Baik, E. Nogales, J. Bondy-Denomy, J.E. Corn, J.A. Doudna, Disabling Cas9 by an anti-CRISPR DNA mimic, Sci. Adv. 3 (7) (2017), e1701620.
[13] A. Hammond, R. Galizi, K. Kyrou, A. Simoni, C. Siniscalchi, D. Katsanos, M. Gribble, D. Baker, E. Marois, S. Russell, A. Burt, N. Windbichler, A. Crisanti, T. Nolan, A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector Anopheles gambiae, Nat. Biotechnol. 34 (1) (2016) 78–83.
[14] F. Zhang, G. Song, Y. Tian, Anti-CRISPRs: the natural inhibitors for CRISPR-cas systems, Anim. Model. Exp. Med. 2 (2) (2019) 69–75.
[15] S. Eitzinger, A. Asif, K.E. Watters, A.T. Iavarone, G.J. Knott, J.A. Doudna, F. Minhas, Machine learning predicts new anti-CRISPR proteins, Nucleic Acids Res. 48 (9) (2020) 4698–4708.
[16] A.B. Gussow, A.E. Park, A.L. Borges, S.A. Shmakov, K.S. Makarova, Y.I. Wolf, J. Bondy-Denomy, E.V. Koonin, Machine-learning approach expands the repertoire of anti-CRISPR protein families, Nat. Commun. 11 (1) (2020) 3784.
[17] J. Wang, W. Dai, J. Li, R. Xie, R.A. Dunstan, C. Stubenrauch, Y. Zhang, T. Lithgow, PaCRISPR: a server for predicting and visualizing anti-CRISPR proteins, Nucleic Acids Res. 48 (W1) (2020) W348–W357.
[18] H. Yi, L. Huang, B. Yang, J. Gomez, H. Zhang, Y. Yin, AcrFinder: genome mining anti-CRISPR operons in prokaryotes and their viruses, Nucleic Acids Res. 48 (W1) (2020) W358–W365.
[19] J. Wang, W. Dai, J. Li, Q. Li, R. Xie, Y. Zhang, C. Stubenrauch, T. Lithgow, AcrHub: an integrative hub for investigating, predicting and mapping anti-CRISPR proteins, Nucleic Acids Res. 49 (D1) (2021) D630–D638.
[20] K.G. Wandera, O.S. Alkhnbashi, H.V.I. Bassett, A. Mitrofanov, S. Hauns, A. Migur, R. Backofen, C.L. Beisel, Anti-CRISPR prediction using deep learning reveals an inhibitor of Cas13b nucleases, Mol. Cell 82 (14) (2022) 2714–2726, e4.
[21] L. Zhu, X. Wang, F. Li, J. Song, PreAcrs: a machine learning framework for identifying anti-CRISPR proteins, BMC Bioinformatics 23 (1) (2022) 444.
[22] C. Dong, G.F. Hao, H.L. Hua, S. Liu, A.A. Labena, G. Chai, J. Huang, N. Rao, F. B. Guo, Anti-CRISPRdb: a comprehensive online resource for anti-CRISPR proteins, Nucleic Acids Res. 46 (D1) (2018) D393–D398.
[23] C. Dong, X. Wang, C. Ma, Z. Zeng, D.K. Pu, S. Liu, C.S. Wu, S. Chen, Z. Deng, F. B. Guo, Anti-CRISPRdb v2.2: an online repository of anti-CRISPR proteins including information on inhibitory mechanisms, activities and neighbors of curated anti-CRISPR proteins, Database (Oxford) 2022 (2022), baac010.
[24] T. Muhammad Atif, J. Kittler, F. Yan, Inverse random under sampling for class imbalance problem and its application to multi-label classification, Pattern Recogn. 45 (10) (2012) 3738–3750.
[25] F.Y. Dao, H. Lv, H. Zulfiqar, H. Yang, W. Su, H. Gao, H. Ding, H. Lin, A computational platform to identify origins of replication sites in eukaryotes, Brief. Bioinform. 22 (2) (2021) 1940–1950.
[26] H. Lv, F.Y. Dao, Z.X. Guan, H. Yang, Y.W. Li, H. Lin, Deep-kcr: accurate detection of lysine crotonylation sites using deep learning method, Brief. Bioinform. 22 (4) (2021) bbaa255.
[27] Z. Chen, P. Zhao, F. Li, T.T. Marquez-Lago, A. Leier, J. Revote, Y. Zhu, D.R. Powell, T. Akutsu, G.I. Webb, K.C. Chou, A.I. Smith, R.J. Daly, J. Li, J. Song, iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data, Brief Bioinform. 21 (3) (2020) 1047–1057.
[28] Z. Chen, P. Zhao, C. Li, F. Li, D. Xiang, Y.-Z. Chen, T. Akutsu, Roger J. Daly, Geoffrey I. Webb, Q. Zhao, L. Kurgan, J. Song, iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization, Nucleic Acids Res. 49 (10) (2021) e60-e60.
[29] I. Dubchak, I. Muchnik, S.R. Holbrook, S.H. Kim, Prediction of protein folding class using global description of amino acid sequence, Proc. Natl. Acad. Sci. U. S. A. 92 (19) (1995) 8700–8704.
[30] P. Feng, L. Feng, Sequence based prediction of pattern recognition receptors by using feature selection technique, Int. J. Biol. Macromol. 162 (2020) 931–934.
[31] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (17) (1997) 3389–3402.

[32] S.A. Chen, Y.Y. Ou, T.Y. Lee, M.M. Gromiha, Prediction of transporter targets using efficient RBF networks with PSSM profiles and biochemical properties, Bioinformatics 27 (15) (2011) 2062–2067.
[33] T. Liu, X. Zheng, J. Wang, Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile, Biochimie 92 (10) (2010) 1330–1334.
[34] A. Mohammadi, J. Zahiri, S. Mohammadi, M. Khodarahmi, S.S. Arab, PSSMCOOL: a comprehensive R package for generating evolutionary-based descriptors of protein sequences from PSSM profiles, Biol. Methods Protoc. 7 (1) (2022), bpac008.
[35] L. Zou, C. Nan, F. Hu, Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles, Bioinformatics 29 (24) (2013) 3135–3142.
[36] Z.R. Yang, Biological applications of support vector machines, Brief. Bioinform. 5 (4) (2004) 328–338.
[37] H. Lv, Z.-M. Zhang, S.-H. Li, J.-X. Tan, W. Chen, H. Lin, Evaluation of different computational methods on 5-methylcytosine sites identification, Brief. Bioinform. 21 (3) (2020) 982–995.
[38] F.-Y. Dao, H. Lv, Y.-H. Yang, H. Zulfiqar, H. Gao, H. Lin, Computational identification of N6-methyladenosine sites in multiple tissues of mammals, Comput. Struct. Biotechnol. J. 18 (2020) 1084–1091.
[39] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) (1982) 29–36.
[40] S. Wan, Y. Duan, Q. Zou, HPSLPred: an ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source, Proteomics 17 (17–18) (2017), 1700262.
[41] J. Wang, Q. Zou, M.Z. Guo, Mining SNPs from EST sequences using filters and ensemble classifiers, Genet. Mol. Res. 9 (2) (2010) 820–834.
[42] H. Ding, P.M. Feng, W. Chen, H. Lin, Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis, Mol. BioSyst. 10 (8) (2014) 2229–2235.
[43] F.Y. Dao, H. Lv, F. Wang, C.Q. Feng, H. Ding, W. Chen, H. Lin, Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique, Bioinformatics 35 (12) (2019) 2075–2083.
[44] A.L. Borges, A.R. Davidson, J. Bondy-Denomy, The discovery, mechanisms, and evolutionary impact of anti-CRISPRs, Annu. Rev. Virol. 4 (1) (2017) 37–59.
[45] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (3) (1990) 403–410.
[46] G. Song, F. Zhang, C. Tian, X. Gao, X. Zhu, D. Fan, Y. Tian, Discovery of potent and versatile CRISPR-Cas9 inhibitors engineered for chemically controllable genome editing, Nucleic Acids Res. 50 (5) (2022) 2836–2853.
[47] Z. Zhang, S. Pan, T. Liu, Y. Li, N. Peng, Cas4 nucleases can effect specific integration of CRISPR spacers, J. Bacteriol. 201 (12) (2019), e00747-18.
[48] S.Y. Stanley, A.L. Borges, K.H. Chen, D.L. Swaney, N.J. Krogan, J. Bondy-Denomy, A.R. Davidson, Anti-CRISPR-associated proteins are crucial repressors of anti-CRISPR transcription, Cell 178 (6) (2019) 1452–1464, e13.