IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

DL-CRISPR: a deep learning method for off-target activity prediction in CRISPR/Cas9 with data augmentation

Yu ZHANG[1], Yahui LONG[2], Rui YIN[1], and Chee Keong KWOH[1]
[1]School of Computer Science and Engineering, Nanyang Technological University
[2]College of Computer Science and Electronic Engineering, Hunan University

Corresponding author: Chee Keong Kwoh (e-mail: asckkwoh@ntu.edu.sg).

ABSTRACT Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/CRISPR- associated (Cas) system is a popular and easy to use gene-editing technique, but it has off-target risk. Cutting the off-target sites will harm the cells severely, hence *in silico* methods are needed to help to avoid this. Most existing *in silico* approaches mainly relied on a relatively small positive dataset and the data imbalance issue still exists. Besides, some samples used to be considered as negative are later proved to be positive. Hence, it is essential to refresh the dataset and develop more accurate off-target activity prediction programs. In this work, firstly, we extended the current positive dataset and explored the potential differences between positive and negative data based on the new dataset. Then we adopted a new data augmentation method to solve the data imbalance issue, and used the ensemble idea to take more negative data into consideration to make the model close to the real scenario, but at the same time keeping the model balance. Finally, we developed DL-CRISPR, a deep learning framework to predict off-target activity in CRISPR/Cas9. DL-CRISPR is evaluated and compared with other state-of-the-art methods on three kinds of datasets: 5-fold cross validation test datasets, putative off-targets datasets related to specific single guide RNAs (sgRNAs), and putative off-targets datasets related to unseen sgRNAs. DL-CRISPR realizes the best average accuracy, i.e. 98.57%, on 5-fold cross validation datasets and correctly detects more off-targets on datasets related to both seen and unseen sgRNAs.

INDEX TERMS CRISPR/Cas9, data augmentation, deep learning, off-target.

I. INTRODUCTION

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/CRISPR- associated (Cas) systems [1] [2], has been widely used in gene editing [3] [4] due to its specificity and ease of use. It uses single guide RNA (sgRNA) to guide the Cas9 nuclease when doing DNA cleavage on a specific site, where the sgRNA is composed of a 20-nt protospacer sequence and a 3-nt protospacer adjacent motif (PAM), usually a sequence of NGG. However, the CRISPR/Cas9 system has a potential off-target risk [5], it may cut the unintended sites with several nucleotide mismatches. The occurrence of off-targets will lead to harmful mutations that can impair or even kill the cells. Therefore, it is essential to identify the possible off-target sites for researchers to check for mutations after the genomic cut.

Varieties of biology assays are available to find off-targets. Polymerase Chain Reaction (PCR) is the most reliable way. Wang *et al.* used integrase-defective lentiviral vectors (IDLVs) to measure off-targets activity and detected off-targets whose frequencies are as low as 1% [6], and Ran *et al.* proposed the BLESS which uses small Cas9 enzymes to improve gene editing efficiency [7]. However, one disadvantage of the PCR methods is that they are not practical to be used on a large number of sites only to find the small number of off-targets. Therefore, all sorts of *in vitro* and cell-based techniques have been developed to detect unbiased and genome-wide off-target sites. *In vitro* methods can detect off-target sites with low mutation frequency, for example, Kim *et al.* introduced Digenome-seq to profile off-target sites on whole genome sequencing [8], Tsai *et al.* developed CIRCLE-seq by reducing random reads in Digenome-seq [9], and Cameron *et al.* presented SITE-seq by enriching and tagging Cas9 cleavage sites in Digenome-seq [10]. Additionally, cell-based methods can identify the mutation in certain cell types and conditions, for instances, GUIDE-seq is proposed to find off-target sites by tagging

DNA double-strand breaks (DSBs) with small double-stranded oligonucleotides [11], BLISS is developed to label the DSBs directly [12], and TEG-seq is presented to enrich the target in GUIDE-seq [13].

Besides, economical and effective *in silico* approaches also have been presented. In the early stage, *in silico* methods mainly relied on mathematical statistics by considering mismatches between sgRNA and DNA, for example, MIT score [14] assigned weights according to mismatch position, identity and density between on-targets and off-targets, and CFD score [15] examined the 1-nt mutations of substantial sgRNAs. However, in recent years, the boom of the machine learning breeds a lot of methods in this area, which improves the off-targets prediction accuracy effectively. Abadi *et al.* developed CRISTA to detect the off-targets probability, which uses the Random Forest to learn a regression model [16], Listgarten *et al.* proposed Elevation to predict off-targets activities, which relies on scoring and aggregating scores machine learning models [17], Peng *et al.* presented an SVM ensemble learning method to determine the off-targets propensity of a sgRNA [18], and Chuai *et al.* implemented DeepCRISPR to design optimal sgRNA as well as predicting the off-target profile with deep learning [19].

However, most machine learning based off-target activity prediction programs mainly rely on a previously integrated positive dataset [20], and the data imbalance issue in CRISPR/Cas9 still exists, which may introduce bias into the model and the scenario build on it would be inconsistence with the practical [21].

To solve the above problems, in this work, we first collected data from a series of *in vitro* and cell-based assays to increase positive data quantity as well as the model competency. Then we applied a novel data augmentation method on positive data to increase its training data amount. Ensemble idea is also employed to make the model closing to the real scenario but at the same time keeping the model balance. Finally, combined with Convolutional Neural Network (CNN), we introduced DL-CRISPR, a deep learning method to evaluate off-target activity in CRISPR/Cas9 system. We evaluated Dl-CRISPR and compared it with other state-of-the-art methods from three aspects: our newly constructed 5-fold cross validation test datasets, putative off-target sites related to two specific sgRNAs, and off-target sites for three unseen sgRNAs on mouse gene. DL-CRISPR outperforms other methods on all kinds of datasets, it achieves accuracy (Acc) from 98.40% to 98.72% on 5-fold cross validation datasets, sorts more off-targets into top position according to their probability scores in the dataset related to two specific sgRNAs, and detects 4, 39, and 30 more off-targets on three unseen mouse gene than other methods.

II. MATERIALS AND METHODS

A. DATASETS

There are three kinds of off-target types: nucleic acid mismatch with on-target sequence, nucleic acid deletion from on-target sequence, and nucleic acid insertion from on-target sequence. Here we only focus on the mutation type off-targets, which occupies a large proportion in all off-targets with sequence length 23.

For emerging of new data experimentally verified, firstly, we collected off-target sequences as well as their sgRNAs from *in vitro* [8] - [10] [22] [23] and cell-based genome-wide assays [11] - [13] [24], among these assays, 3 of them contain PCR validation experiments [8] [22] [23]. As we hypothesized that the *in vitro* and cell-based assays may not be the reliable off-target data sources due to the differences in setting the environments between them and PCR assays, we further split the off-targets to two types. We defined the reliable off-targets as the off-target sequences which are either detected by at least two different *in vitro* and cell-based studies, or validated by PCR, and defined the "detected" off-targets as the off-targets which are not validated by PCR and only detected by one of *in vitro* or cell-based studies. With the definition, 957 unique reliable off-targets and 5,435 "detected" off-targets were obtained from *in vitro* and cell-based assays, from which we deleted sgRNA 'TCATCCTCCTGACAATCGATAGG' on gene CCR5_9 with its off-target site, as there is only 1 sample for this sgRNA. Then we downloaded off-targets obtained by PCR assays [18] and got 215 more samples. We integrated the reliable off-targets got above without repetition to construct the final positive dataset. The complete positive dataset contains 1128 unique reliable off-target sequences related to 28 sgRNAs. The process to split reliable and "detected" off-targets and the corresponding data amount is illustrated in Fig. 1, and the overview of the assays where we collected data is implemented in Supplementary Table S1.

For negative data collection, we downloaded the negative dataset from Peng's work [18], which were found by Cas-OFFinder [25] in human gene hg38 related to 29 sgRNAs with no more than 6 mismatches and excluded those already in his positive dataset. It should be noticed that two sgRNAs in the 29 sgRNAs have the same 20-nt protospacer sequences (GGGTGGGGGGAGTTTGCTCC) but only different PAM sequences (AGG and TGG), these two sgRNAs are considered as one with PAM sequence NGG in our positive data, thus the negative data we downloaded are in accordance with our positive data. However, 379 samples in this negative dataset were later proved to be reliable off-targets and were put into our new positive dataset. Besides, 3,485 samples in the original negative dataset belong to "detected" off-targets as we defined. To avoid confusion, we deleted these newly identified reliable and "detected" off-targets

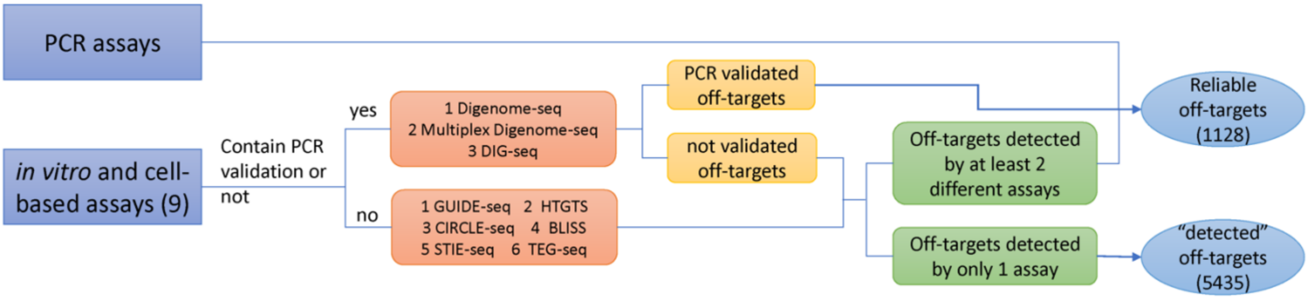
PCR assays
in vitro and cell-based assays (9)
Contain PCR validation or not
yes
no
1 Digenome-seq
2 Multiplex Digenome-seq
3 DIG-seq
1 GUIDE-seq 2 HTGTS
3 CIRCLE-seq 4 BLISS
5 STIE-seq 6 TEG-seq
PCR validated off-targets
not validated off-targets
Off-targets detected by at least 2 different assays
Off-targets detected by only 1 assay
Reliable off-targets (1128)
"detected" off-targets (5435)

FIGURE 1. **The process to split reliable and "detected" off-targets. 9 *in vitro* and cell-based assays are considered here, their names are shown in the orange box. After processing, 1128 reliable off-targets and 5435 "detected" off-target are obtained.**

from the original negative data. Our new negative dataset contains 403,953 samples, from which the DNA sites are called no-editing sites.

B. FEATURES

In this paper, the features are all extracted directly from the sequence, as the previous study reported that the sequence-derived feature is quite dependable for off-target prediction [20]. And also, according to work [14], only PAM NGG and NAG have editing efficiencies, almost all experiments used NGG as PAM because it is much more efficient than NAG, hence the 3-length PAM sequence does not contain too much useful sequence composition information and we would not consider it in feature extraction.

Suppose $S = (s_1 s_2 \dots s_j \dots s_n)$ represents the putative off-target DNA sequences from 5' to 3', and $D = (d_1 d_2 \dots d_j \dots d_n)$ demonstrates sgRNA sequences, where $s_j, d_j \in \{A, T, G, C\}$, n denotes the sequence length and equals to 20 here. To represent the sequences, firstly, we used the one-hot vector to encode two raw sequences: sgRNA and its putative off-target sites. Each nucleotide acid (s_j or d_j) will be represented as one of (1,0,0,0), (0,1,0,0), (0,0,1,0) and (0,0,0,1) corresponding to A, C, G and T, respectively. Next, we introduced the mismatch position and type feature to extract the mutation information from the raw sequences. In the on-target and putative off-target sequence pairs, if $d_j \neq s_j$, position j is regarded as mismatch. For each mismatch position j, the mismatch type must belong to one of the 12 mutation types: $\{AT, AG, AC, TA, TG, TC, GA, GT, GC, CA, CT, CG\}$, where the first nucleic acid is from on-target sites and the second is from putative off-target sites. In this way, the mutation position and type information can be incorporated into a 12*20 matrix M, where $M_{(i,j)} = 1$ means at position j, mismatch happens between sgRNA and its putative off-target with type i , and other values in M will be assigned to 0.

Considering the above feature extraction schemes, each sequence can be represented in a 20*20 matrix by concatenating the one-hot features for on-target sequence, one-hot features for putative off-target sequence, and the mismatch position and pair matrix M, along the row.

C. DATA AUGMENTATION

Data imbalance is a severe issue in CRISPR/Cas9 off-target prediction. As mentioned above, the negative data can be as large as several hundred thousand while the positive data is only 1,128. This issue had been addressed by Gao [21] and he called for more efforts to be done in this area. Inspired by the image augmentation methods, in this work, we performed rotation on our built feature matrix to extend the positive dataset. We rotated each original feature matrix to 90 degrees, 180 degrees and 270 degrees, respectively. In this way, the data size can be augmented to four times as before.

D. MODEL CONSTRUCTION

Because our raw sequences are organized in 2D matrices with binary values, and CNN network can learn the hierarchical spatial representations, we believe that the application of multiple layers CNN structure would be a suitable way to learn useful information from the feature matrices we built, and we adopt 4 CNN layers here. To train the model, we first extended the positive training dataset size from 902 to 3608 with our new data augmentation method, then randomly selected the same amount of negative data as positive to ensure the balance of the model. Although some authors stated that the artificial approach to balance dataset may under-estimate the real scenario where CRISPR off-targets can have lots of negative samples in practice, given its whole genomic search range [21], we believe that using unbalance data to train the model will introduce bias and getting undesired results for positive data. Rather, we adopted the ensemble idea to make the model to be close to the real scenario to solve the problem addressed above. We trained the model 10 times, each time we randomly reselected 3608 negative samples to ensure different data can be learned by the network. The final result was got by averaging scores of ten models. The techniques we described above finally composed of DL-CRISPR, a deep learning model for off-target prediction in CRISPR/Cas9. Its working mechanism is illustrated in Fig. 2.

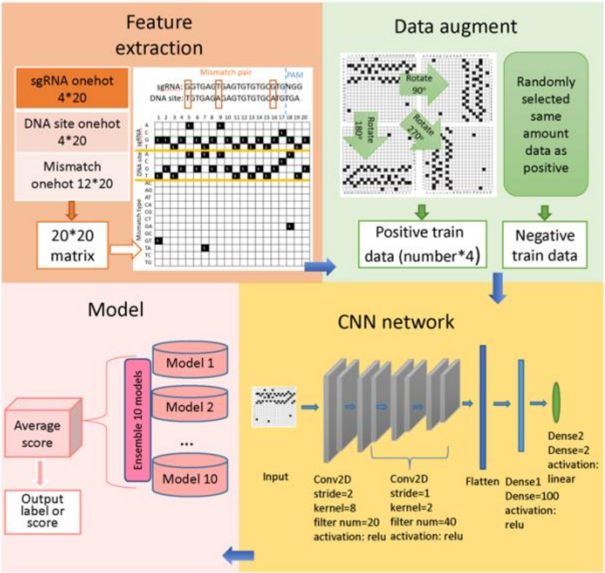

FIGURE 2. **Overview of DL-CRISPR. Each raw sequence data is represented in a 20*20 matrix with binary values, positive train data amount is quadrupled, and 10 deep learning models with 4-layer CNN network are trained with the same quadrupled positive data but the same amount of different negative data. The final result is got as the average score of 10 models.**

III. RESULTS AND DISCUSSIONS

A. DATA RELIABILITY

As mentioned above, most off-target samples were obtained from *in vitro* and cell-based assays, and we hypothesized that the potential differences of experimental settings in these assays from that of PCR assays will induce off-targets that would not happen in practice. To illustrate this, we designed three experiments whose positive and negative training data are 1) reliable off-targets and "detected" off-targets, 2) "detected" off-targets and negative samples, and 3) reliable off-targets and negative samples, from which, all 3 kinds of datasets are randomly selected with sizes of 902 (80% data of reliable off-targets dataset). The models constructed from these three kinds of training data were tested on the remaining 226 reliable off-targets and 226 randomly selected negative samples which are not overlapped with training.

We tested the model 5 times for each group of training data, the results were recorded in Supplementary Table S2. The average Acc of using "detected" off-targets as negative data, positive data, and not use are 77.16%, 82.44%, and 93.72%, respectively. Obviously, differences exist between reliable off-targets and "detected" off-targets due to the over 10% difference in classification accuracies when using "detected" off-targets as positive and not using it, which in turn prove that the off-target samples found by one of *in vitro* or cell-based experiments are not so reliable to be the true off-targets in practice. The results also demonstrate that there are distinctions between "detected" off-targets and negative samples, but the relatively smaller gap between Acc got by the second and the third group of training data, and the first and the third group of training data illustrates that the properties of "detected" off-targets are more consistent with the reliable off-targets than the negative data. Even though, the "detected" off-targets will not be considered as either positive or negative in our training datasets.

B. DATA ANALYSIS AND FEATURE INVESTIGATION

To show the differences between positive data and negative data as well as the effectiveness of our features, mutation position and type for two kinds of data were visualized with line charts and heatmaps in Fig. 3. From the line chart, the distribution of mutation position and type in negative data are much more uniform than that of positive, whose mutation frequency for each position almost remain at around 5%, whereas in positive data, the mutation position and type have some obviously preferences: some positions, like position 1-4 and 8, show higher mutation frequencies, especially in position 1 and 2, whose mutation frequencies are about as twice as those in negative, whereas positions like 10-16, 19 and 20 seem to less likely to mutate. Besides, the mutation types in the same positions also can be different in two kinds of data. For instance, at position 1, although the heatmaps show that nucleic acid 'G' is more likely to mutate than other nucleic acids in both positive and negative, 'G' is more likely to mutate to 'A' in positive while to 'T' in negative, in addition, 'A' and 'T' in positive nearly show no mutation trends, whereas in negative, 'A', 'C' and 'T' almost have the same probabilities to mutate; at position 2, the mutation from 'G' to 'A' shows much higher frequency in positive, but in negative, it is not so noticeable as mutation 'G-C' and 'G-T' also share very high frequencies; at position 8, 'T' is much more likely to mutate to C in positive, however, this circumstance rare happens in negative samples; at position 19, mutation type 'T-C' happens most often whereas other types rarely happen, however in negative, all mutation types can happen but 'C-G' has the largest probability.

In addition, even for the positions share nearly the same mutation frequencies in positive and negative, such as position 5-7, 9, 17 and 18, the mutation types can be different. For example, at position 8, mutation type 'T-C' is most likely to happen in positive, but in negative, this rarely happens, in contrast, nucleic acids 'C' and 'T' are more likely to mutate to other nucleic acids. And at position 18, mutation type 'G-A' is quite noticeable in positive, whereas in negative, the mutations for all nucleic acids share nearly the same possibilities.

The mutation position trends we found above, although with limited positive data, proved that in 20-bp PAM-proximal seed sequence, the sites which have mismatch nucleic with sgRNA in PAM-distal region and match with sgRNA in PAM-proximal region are more likely to be off-target sites, this finding is consistent with the work reported before [14]. At the same time, from the mutation type plotting, mutation type 'G-A' in positive happens quite often. The findings here give us some insights to distinguish the different types of data and indicate that our feature matrices can contain useful information in predicting off-targets.

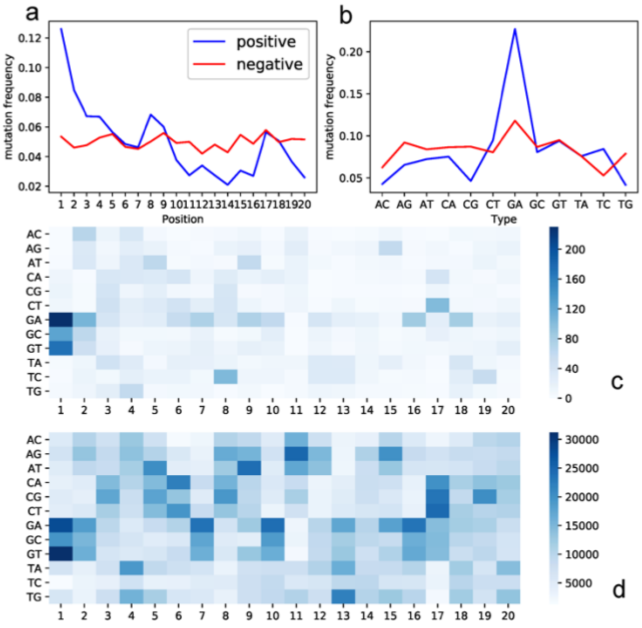

FIGURE 3. **Mutation information in positive and negative samples. (a) Mutation position comparison, (b) mutation type comparison, (c) heatmap of mutation in positive data, and (d) heatmap of mutation in negative data.**

C. 5-FOLD CROSS VALIDATION

To better demonstrate the performance of DL-CRISPR and show its robust, we evaluate it with 5-fold cross validation. The original data were randomly divided into 5 subsets, in each validation, one subset is used as the test and the others are used as training. As described above, we use ensemble learning to incorporate more negative data at the same time to maintain the model balance, hence, in each fold, 10 models are constructed and the results are taken as their average. The off-target prediction for given sgRNAs is similar to the topic of optimal sgRNA design, some sgRNA design tools also can output possible off-targets and therefore can be used for comparison here. However, most optimal sgRNA selection tools only output optimal sgRNA in a gene, for tools which also can output off-targets, other input information may be required, like longer input sequence, e.g. 30-nt before PAM [26], or epigenetic features of the sequence [19]. Considering these, DL-CRISPR is compared with three recently published machine learning methods: ensemble SVM [18], CRISTA [16], and Elevation [17]. Elevation is designed as an optimal sgRNA selection tool, it outputs the optimal sgRNAs for input gene as well as their corresponding off-target sites, as some sgRNAs in our dataset are not outputted as optimal by Elevation, we cannot get the predicted off-targets for these sgRNAs. Therefore, the performance of Elevation is only evaluated on the 13 sgRNAs it found, where the 13 sgRNAs are recorded in Supplementary Table S3. In addition, all these three methods only can output scores of predicted positive data, so it is impossible for us to compare the results via ROC and PRC curves.

The 5-fold cross validation results for all four methods are demonstrated with box and whisker plots in Fig. 4. Box and whisker plots incorporate the information of center, spread and overall range for a group of data, where the skewed distribution and potential unusual samples can be indicated clearly, hence can give us insights about the model robust.

We observe that the overall performance of DL-CRISPR on 5-fold cross validation is the best even under the truth that test data may already exist in training for other methods, whose average values of Acc, sensitivity (Sn), specificity (Sp), f-score, and Harmonic mean (Hm) are 98.57%, 95.57%, 98.58%, 0.9928, and 0.9705, respectively. Besides, DL-CRISPR outperforms the other three methods with better performance matrices in all 5 validations. Although the Acc of CRISTA are also very large and within a narrower spread than DL-CRISPR, they are mainly contributed by the negative prediction accuracies due to the heavily imbalanced data. However, in this classification problem, we should pay more attention to the positive prediction accuracy, as it can

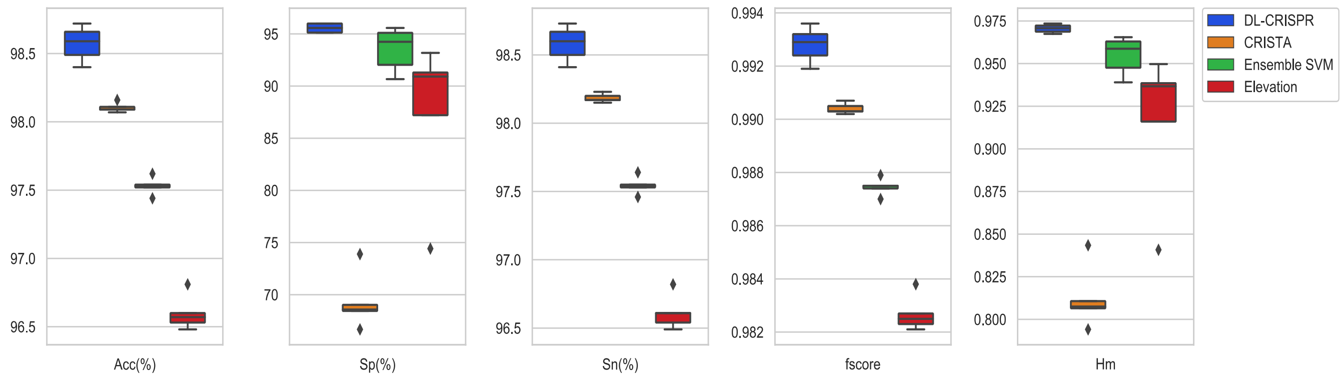

FIGURE 4. **The Acc, Sn, Sp, fscore, and Hm for different methods on 5-fold cross validation.**

provide valuable references for the potential off-target sites. The smallest sensitivity in DL-CRISPR is 95.13% whereas the largest sensitivity in CRISTA is only 73.89%. Additionally, the average sensitivities for CRISTA, Ensemble SVM and Elevation are 69.32%, 93.53%, and 87.40%, respectively, which are all worse than DL-CRISPR's 95.66%.

D. PERFORMANCE ON SPECIFIC SGRNAS

In this part, we want to evaluate the ability of DL-CRISPR in identifying the most possible off-targets from a large number of putative sites. Each input sequence will be assigned a probability score of being off-target by the model, whether the experimentally verified samples can be assigned high probability scores and whether they can be ranked at top positions among all potential sites by the model can give us significant insights about the model effectiveness.

We test DL-CRISPR on putative off-target datasets for specific sgRNAs. We selected two sgRNAs to study, one from gene VEGFA (on-target sequence: GGGTGGGGGGAGTTTGCTCCNGG) and the other from gene EMX1 (on-target sequence: GAGTCCGAGCAGAAGAAGAANGG), we annotate these two sgRNAs as VEGFA and EMX1 in the following description. These two sgRNAs are chosen because they are investigated most times in nine *in vitro* and cell-based techniques (9 times for VEGFA and 6 times for EMX1). The model we used is the ensemble model constructed in the first cross validation from the above part, only Ensemble SVM and CRISTA are compared here, as the Elevation does not output the two sgRNAs as optimal on gene VEGFA and EMX1.

To prepare the data, firstly, we download the putative off-target data in Frock's work [24] for these two sgRNAs from [20], 41,746 and 76,251 samples are obtained for sgRNA VEGFA and EMX1 separately. After deleting those already in training, 99 for VEGFA and 64 for EMX1, we get 29 and 12 reliable off-targets, and 437 and 92 "detected" off-targets for VEGFA and EMX1 separately. To evaluate the methods on new datasets for two sgRNAs, we compare the number of reliable off-targets and the number of "detected" off-targets in the first m highest scores predicted by different methods, where $m = 29, 40, 70$ for VEGFA and $m = 12, 20, 30$ for EMX1. These reliable and "detected" off-target sequences are expected to be ranked in top positions with larger probability scores, especially so for the reliable off-targets. The results were demonstrated in Fig. 5.

From the comparisons, DL-CRISPR sort one more reliable off-target into the first 29 and 70 most possible off-target positions for sgRNA VEGFA and the first 30 most possible off-target position for sgRNA EMX1. And DL-CRISPR also ranks more "detected" off-targets, i.e. no less than 6 for sgRNA VEGFA and no less than 2 for sgRNA EMX1, into the top m positions compared to CRISTA and Ensemble SVM.

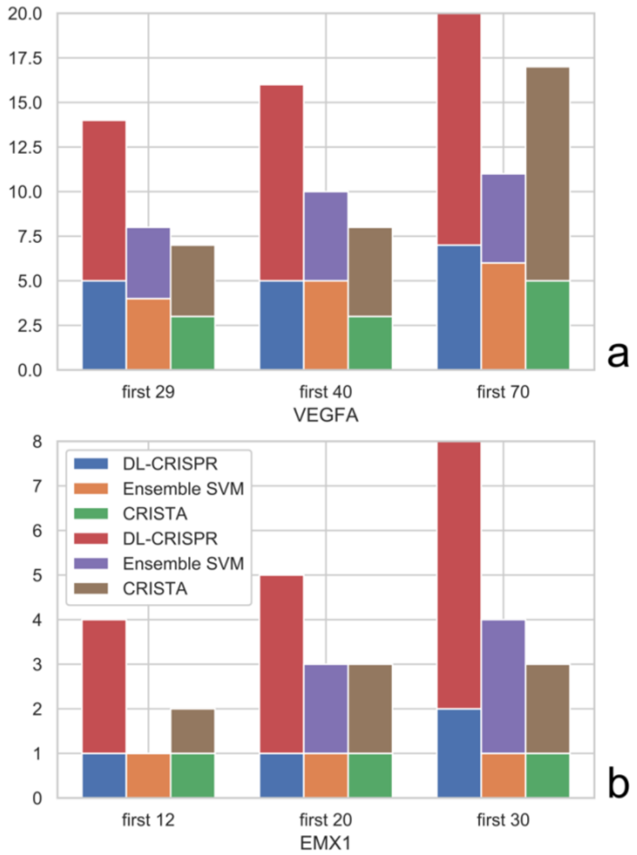

FIGURE 5. Comparisons among DL-CRISPR, Ensemble learning and CRISTA about how many reliable or "detected" off-targets were ranked into the top m largest probabilities on the datasets related to sgRNA (a) VEGFA and (b) EMX1. Bottom bar: the number of reliable off-targets, top bar: the number of "detected" off-targets, whole bar: the number of off-targets found in any biology assays.

E. PERFORMANCE ON UNSEEN SGRNAS

We implement the studies in this part on three new sgRNAs: gM (GGCTGATGAGGCCGCACATGTGG), gMH (CAGGTTCCATGGGATGCTCTGGG) and gp (AGCAGCAGCGGCGGCAACAGCGG), targeted to the mouse Pcsk9 gene, with off-target sequences identified by CIRCLE-seq on WT and KI mouse genomic DNA [27]. Three new test datasets contain 166 off-targets for sgRNA gM, 439 for gMH and 3,381 for gp, respectively. Since all the off-targets are already detected by CIRCLE-seq, these data should have relatively higher scores. As these three sgRNAs are new, we reconstruct the model under the DL-CRISPR working mechanism using all positive data in our dataset.

DL-CRISPR identifies 75, 174 and 3,323 sequences from gM, gMH and gp datasets as off-targets respectively. The prediction result of gp is in high agreement with the *in vitro* experiment while the results of the other two sgRNAs are not so consistent. This phenomenon can be explained from their heatmaps in Fig. 6. The mutations in gp are much more identical, they mainly happen at position 10, 13 and 17 with mutation type 'GC' or 'AG', therefore, the samples in this

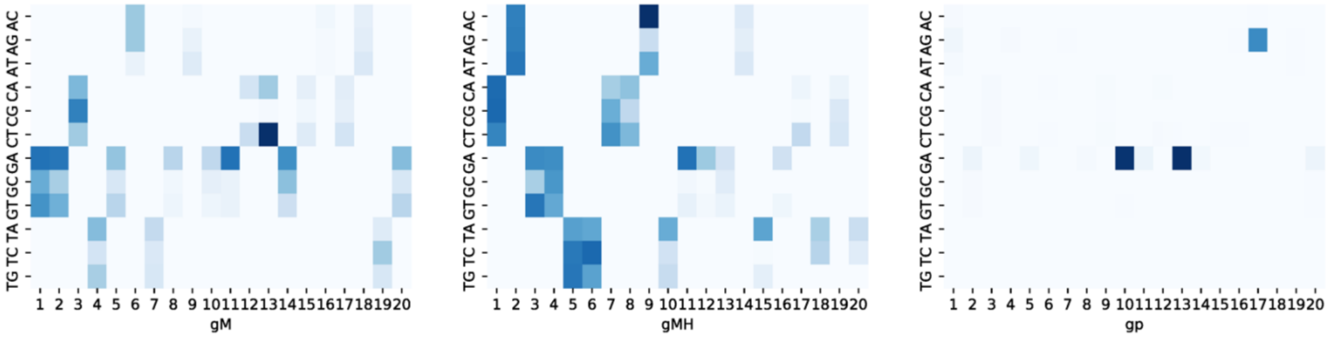

FIGURE 6. **Heatmaps for the off-targets found by CIRCLE-seq of three mouse sgRNA gM, gMH, and gp.**

dataset are very likely to be ascribed into the same class, and obviously, these samples are identified correctly by DL-CRISPR, which illustrates its ability to predict off-targets of unseen sgRNA. However, as to the other two sgRNAs, the mutation positions and types of their off-targets sequences are much variety, hence the scores may vary in scopes and the sequences may be allocated to different classes. Furthermore, as we defined above, the test samples in this part are "detected" off-targets rather than the reliable off-targets, whereas the DL-CRISPR is trained with reliable off-targets, hence it may not favor all "detected" off-targets and assign high probability scores to them.

We also explore the prediction results of Ensemble SVM and CRISTA on the same datasets. Their results together with the results of DL-CRISPR are demonstrated in TABLE I. CRISTA outputs 888, 1,410 and 15,953 off-targets in total for gM, gMH and gp sgRNAs, from which 26, 80 and 2,913 samples are in the off-target datasets identified by CIRCLE-seq. The trends for the prediction results using Ensemble SVM and CRISTA on three sgRNAs are consistent with what we found in DL-CRISPR above, where the off-target sequences related to gp are the easiest to be predicted and the gMH's mutation sites are the hardest to be defined. CRISTA gives the least number of predicted off-targets, and Ensemble SVM is slightly worse than that of DL-CRISPR on these unseen sgRNAs, which predicts 4, 39, and 30 fewer sequences in datasets related to gene gM, gMH, and gp, respectively. Through the comparison, DL-CRISPR is proved to be an effective off-targets activity prediction program for unseen sgRNAs.

TABLE I
NUMBER OF OFF-TARGETS DEFINED BY DIFFERENT METHODS ON DATASETS OF THREE MOUSE SGRNAS.

	gM	gMH	gp
DL-CRISPR	79	174	3323
Ensemble	75	135	3293
CRISTA	26	80	2913

IV. CONCLUSIONS

Data imbalance is a severe issue in CRISPR/Cas9 system when applying machine learning. To solve the data imbalance problem, we extended the positive dataset size and adopted data augmentation to increase positive training data amount, and we employed ensemble idea to take more negative data into consideration to make the model closing to the real scenario, but at the same time keeping the model balance. Based on the above strategies, we proposed DL-CRISPR, a deep learning model for off-target activity prediction in CRISPR/Cas9. We first explored the off-target data reliability and the differences between positive and negative data on our newly extended dataset. Experiments show that off-targets detected by only one of *in vitro* or cell-based assays have some differences with the reliable off-targets, and positive data have obvious preferences for the mutation positions and types compared to negative data. Then we tested and compared DL-CRISPR with three state-of-the-art methods on different types of datasets. DL-CRISPR achieved the best performance on 5-fold cross validation test datasets with over 98.40% of Acc and over 95.13% of Sn, the general high values and narrow spread of all evaluation matrices illustrate the robust of DL-CRISPR. In addition, DL-CRISPR ranked more reliable and "detected" off-targets in top positions according to the probability scores in datasets related to two specific sgRNAs than other methods. Furthermore, DL-CRISPR also worked well on off-targets prediction for unseen sgRNAs by identifying more off-targets detected by CIRCLE-seq than other methods. In a nutshell, the experimental results in this work fully demonstrated that DL-CRISPR is an effective and robust off-target activity prediction method in CRISPR/Cas9.

APPENDIX

Supplementary Materials, the data used in this work, and the code for DL-CRISPR are available at https://github.com/yuuuuzhang/DL-CRISPR_offtarget_prediction.

ACKNOWLEDGMENT

We thank Dr. Cangzhi Jia for providing valuable suggestions to this work.

REFERENCES

[1] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. Doudna and E. Charpentier, "A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity," *Science,* Vols. 337(6096), pp. pp.816-821, 2012.

[2] P. Mali, L. Yang, K. Esvelt, J. Aach, M. Guell, J. DiCarlo, J. Norville and G. Church, "RNA-guided human genome engineering via Cas9," *Science,* pp. 339(6121), pp.823-826, 2013.

[3] W. Hwang, F. Y. R. D. M. Maeder, S. Tsai, J. Sander, R. Peterson, Y. J. and J. Joung, "Efficient genome editing in zebrafish using a CRISPR-Cas system," *Nature biotechnology,* pp. 31(3), p.227, 2013.

[4] R. Bak, D. Dever and M. Porteus, "CRISPR/Cas9 genome editing in human hematopoietic stem cells," *Nature protocols,* pp. 13(2), p.358, 2018.

[5] Y. Lin, T. Cradick, M. Brown, H. Deshmukh, P. Ranjan, N. Sarode, B. Wile, P. Vertino, F. Stewart and G. Bao, "CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences," *Nucleic acids research,* pp. 42(11), pp.7473-7485, 2014.

[6] X. Wang, Y. Wang, X. Wu, J. Wang, Y. Wang, Z. Qiu, T. Chang, H. Huang, R. Lin and J. Yee, "Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors," *Nature biotechnology,* pp. 33(2), p.175, 2015.

[7] F. Ran, L. Cong, W. Yan, D. Scott, J. Gootenberg, A. Kriz, B. Zetsche, O. Shalem, X. Wu, K. Makarova and E. Koonin, "In vivo genome editing using Staphylococcus aureus Cas9," *Nature,* pp. 520(7546), p.186, 2015.

[8] D. Kim, S. Bae, J. Park, E. Kim, S. Kim, H. Yu, J. Hwang, J. Kim and J. Kim, "Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells," *Nature methods,* pp. 12(3), p.237, 2015.

[9] S. Tsai, N. Nguyen, J. Malagon-Lopez, V. Topkar, M. Aryee and J. Joung, "CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets," *Nature methods,* pp. 14(6), p.607, 2017.

[10] P. Cameron, C. Fuller, P. Donohoue, B. Jones, M. Thompson, M. Carter, S. Gradia, B. Vidal, E. Garner, E. Slorach and E. Lau, "Mapping the genomic landscape of CRISPR–Cas9 cleavage," *Nature methods,* pp. 14(6), p.600, 2017.

[11] S. Tsai, Z. Zheng, N. Nguyen, M. Liebers, V. Topkar, V. Thapar, N. Wyvekens, C. Khayter, A. Iafrate, L. Le and M. Aryee, "GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases," *Nature biotechnology,* pp. 33(2),p.187, 2015.

[12] W. Yan, R. Mirzazadeh, S. Garnerone, D. Scott, M. Schneider, T. Kallas, J. Custodio, E. Wernersson, Y. Li, L. Gao and Y. Federova, "BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks," *Nature communications,* pp. 8,p.15058, 2017.

[13] P. Tang, B. Ding, L. Peng, V. Mozhayskiy, J. Potter and J. Chesnut, "TEG-seq: an ion torrent-adapted NGS workflow for in cellulo mapping of CRISPR specificity," *BioTechniques,* pp. 65(5), pp.259-267, 2018.

[14] P. Hsu, D. Scott, J. Weinstein, F. Ran, S. Konermann, V. Agarwala, Y. Li, E. Fine, X. Wu, O. Shalem and T. Cradick, "DNA targeting specificity of RNA-guided Cas9 nucleases," *Nature biotechnology,* pp. 31(9), p.827, 2013.

[15] J. Doench, N. Fusi, M. Sullender, M. Hegde, E. Vaimberg, K. Donovan, I. Smith, Z. Tothova, C. Wilen, R. Orchard and H. Virgin, "Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9," *Nature biotechonology,* pp. 34(2),p.184, 2016.

[16] S. Abadi, W. Yan, D. Amar and I. Mayrose, "A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action," *PloS computational biology,* pp. 13(10), p.e1005807, 2017.

[17] J. Listgarten, M. Weinstein, B. Kleinstiver, A. Sousa, J. Joung, J. Crawford, K. Gao, L. Hoang, M. Elibol, J. Doench and N. Fusi, "Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs," *Nature biomedical engineering,* pp. 2(1),p.38, 2018.

[18] H. Peng, Y. Zheng, Z. Zhao, T. Liu and J. Li, "Recognition of CRISPR/Cas9 off-target sites through ensemble learning of uneven mismatch distributions," *Bioinformatics,* pp. 34(17), pp.i757-i765, 2018.

[19] G. Chuai, H. Y. J. Ma, M. Chen, N. Hong, D. Xue, C. Zhou, C. Zhu, K. Chen, B. Duan and F. Gu, "DeepCRISPR: optimized CRISPR guide RNA design by deep learning," *Genome biology,* pp. 19(1), p.80, 2018.

[20] M. Haeussler, K. Schönig, H. Eckert, A. Eschstruth, J. Mianné, J. Renaud, S. Schneider-Maunoury, A. Shkumatava, L. Teboul, J. Kent and J. Joly, "Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR," *Genome biology,* pp. 17(1),p.148, 2016.

[21] Y. Gao, G. Chuai, W. Yu, S. Qu and Q. Liu, "Data imbalance in CRISPR off-target prediction.," *Briefings in bioinformatics,* 2019.

[22] D. Kim, S. Kim, S. Kim, J. Park and J. Kim, "Genome-wide target specificities of CRISPR-Cas9 nucleases revealed by multiplex Digenome-seq," *Genome research,* pp. 26(3), pp.406-415, 2016.

[23] D. Kim and J. Kim, "DIG-seq: a genome-wide CRISPR off-target profiling method using chromatin DNA," *Genome research,* pp. 28(12), pp.1894-1900, 2018.

[24] R. Frock, J. Hu, R. Meyers, Y. Ho, E. Kii and F. Alt, "Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases," *Nature biotechnology,* pp. 33(2), p.179, 2015.

[25] S. Bae, J. Park and J. Kim, "Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases," *Bioinformatics,* pp. 30(10), pp.1473-1475, 2014.

[26] J. Wang, X. Xiang, L. Cheng, X. Zhang and Y. Luo, "CRISPR-GNL: an improved model for predicting CRISPR activity by machine learning and featurization," *bioRxiv,* p. p.605790, 2019.

[27] P. Akcakaya, M. Bobbin, J. Guo, J. Malagon-Lopez, K. Clement, S. Garcia, M. Fellows, M. Porritt, M. Firth, A. Carreras and T. Baccega, "In vivo CRISPR editing with no detectable genome-wide off-target mutations," *Nature,* pp. 561(7723), p.416, 2018.

Yu ZHANG received her BEng degree from Shandong University, China, and the MSc degree (distinction degree) from Imperial College London, UK, in 2017 and in 2018, respectively. She is currently a Ph.D. candidate in Nanyang Technological University, Singapore. Her research interests include bioinformatics and deep learning.

Yahui LONG received the master's degree in software engineering from Hunan University in 2017, where he is currently pursuing the Ph.D. degree with the College of Computer Science and Electronic Engineering, Hunan University. He is currentlya joint PhD student in School of Computer Science and Engineering, Nanyang Technological University. His research interests include deep learning, bioinformatic and computational biology.

Rui YIN received his B.S. degree in Automation in 2013 from Shandong University, China. He received his M.Sc. degree in Control Engineering in 2016 from Central South University, China. Now, he is a Ph.D. candidate in the school of Computer Science and Engineering in Nanyang Technological University, Singapore. His research interests include data mining and pattern recognition to make sense of big heterogeneous data for real applications in engineering and biomedical science.

Chee Keong KWOH received the bachelor's degree in electrical engineering (first class) and the master's degree in industrial system engineering from the National University of Singapore, Singapore, in 1987 and 1991, respectively. He received the Ph.D. degree from the Imperial College of Science, Technology and Medicine, University of London, in 1995. He has been with the School of Computer Engineering, Nanyang Technological University (NTU), since 1993.

His research interests include data mining, soft computing and graph-based inference; applications areas include bioinformatics and biomedical engineering. He has done significant research work in his research areas and has published many quality international conferences and journal papers. He is an editorial board member of the International Journal of Data Mining and Bioinformatics, the Scientific World Journal, Network Modeling and Analysis in Health Informatics and Bioinformatics, Theoretical Biology Insights, and Bioinformation. He has been a guest editor for many journals such as the Journal of Mechanics in Medicine and Biology, the International Journal on Biomedical and Pharmaceutical Engineering, and others. He has often been invited as organizing member or referee and reviewer for a number of premier conferences and journals including GIW, IEEE, BIBM, RECOMB, PRIB, BIBE, ICDM, and iCBBE. He is a member of the Association for Medical and Bioinformatics, Imperial College Alumni Association of Singapore. He has provided many services to professional bodies in Singapore and was conferred the Public Service Medal by the president of Singapore in 2008. His research interests include data mining, soft computing and graph-based inference; applications areas include bioinformatics and biomedical engineering.