·合成生物技术·

基于多尺度卷积神经网络的 CRISPR/Cas9 脱靶预测方法

谢焕增，黄凌泽，罗烨，张桂珊*

汕头大学工学院，广东 汕头 515063

谢焕增, 黄凌泽, 罗烨, 张桂珊. 基于多尺度卷积神经网络的 CRISPR/Cas9 脱靶预测方法[J]. 生物工程学报, 2024, 40(3): 858-876.
XIE Huanzeng, HUANG Lingze, LUO Ye, ZHANG Guishan. Prediction of CRISPR/Cas9 off-target activity using multi-scale convolutional neural network[J]. Chinese Journal of Biotechnology, 2024, 40(3): 858-876.

摘　要：规律成簇的间隔短回文重复序列/CRISPR 相关蛋白 9 (clustered regularly interspaced palindromic repeats/CRISPR-associated protein 9, CRISPR/Cas9)是新一代基因编辑技术，该技术依靠单向导 RNA 识别特定基因位点，并引导 Cas9 核酸酶对特定位点进行编辑。然而，该技术存在脱靶效应限制了其发展。近年来，运用深度学习辅助 CRISPR/Cas9 脱靶预测研究是一个新兴的思路，有助于研究者实现更高效安全的基因编辑和基因治疗。而现有的深度学习模型对脱靶预测的准确性仍有提高空间。为此，本文基于多尺度卷积神经网络提出 CnnCRISPR 模型预测 CRISPR/Cas9 的脱靶情况。首先，将向导 RNA 和 DNA 序列分别进行独热编码，再将两个二值矩阵按位进行或运算。其次，将编码后的序列输入基于 Inception 模块的网络进行训练和验证分析。最后，输出向导 RNA 和 DNA 序列对的脱靶情况。在公开数据集上的实验结果表明，CnnCRISPR 模型的性能优于现有的深度学习脱靶预测模型，为脱靶问题的研究提供了有效且可行的方法。

关键词：CRISPR/Cas9；脱靶效应；多尺度卷积神经网络

资助项目：国家自然科学基金(62103249)；广东省基础与应用基础研究基金(2022A1515011720)；广东省科技专项资金“大专项+任务清单” (STKJ2021183)；汕头大学科研启动基金(NTF20032)
This work was supported by the National Natural Science Foundation of China (62103249), the Guangdong Basic and Applied Basic Research Foundation (2022A1515011720), the Special Fundation for Science and Technology of Guangdong Province (STKJ2021183), and the STU Scientific Research Foundation for Talents (NTF20032).
*Corresponding author. E-mail: gszhang@stu.edu.cn
Received: 2023-05-21; Accepted: 2023-08-08

Prediction of CRISPR/Cas9 off-target activity using multi-scale convolutional neural network

XIE Huanzeng, HUANG Lingze, LUO Ye, ZHANG Guishan*

College of Engineering, Shantou University, Shantou 515063, Guangdong, China

Abstract: Clustered regularly interspaced short palindromic repeat/CRISPR-associated protein 9 (CRISPR/Cas9) is a new generation of gene editing technology, which relies on single guide RNA to identify specific gene sites and guide Cas9 nuclease to edit specific location in the genome. However, the off-target effect of this technology hampers its development. In recent years, several deep learning models have been developed for predicting the CRISPR/Cas9 off-target activity, which contributes to more efficient and safe gene editing and gene therapy. However, the prediction accuracy remains to be improved. In this paper, we proposed a multi-scale convolutional neural network-based method, designated as CnnCRISPR, for CRISPR/Cas9 off-target prediction. First, we used one-hot encoding method to encode the sgRNA-DNA sequence pair, followed by a bitwise or operation on the two binary matrices. Second, the encoded sequence was fed into the Inception-based network for training and evaluating. Third, the well-trained model was applied to evaluate the off-target situation of the sgRNA-DNA sequence pair. Experiments on public datasets showed CnnCRISPR outperforms existing deep learning-based methods, which provides an effective and feasible method for addressing the off-target problems.

Keywords: CRISPR/Cas9; off-target effects; multi-scale convolutional neural network

自 1987 年规律成簇的间隔短回文重复序列(clustered regularly interspaced short palindromic repeat, CRISPR)系统首次在大肠杆菌中被发现后，科学家又陆续在许多细菌与古细菌中发现这一序列，这些细菌和古细菌利用这种免疫抵抗病毒等外源物质入侵[1]，并证实了这种系统是原核生物的一种免疫机制[2]。2012 年，CRISPR/Cas9 系统首次被引入基因编辑中[3]，之后许多研究领域借助该系统取得了突破性的进展，比如在感染了人类免疫缺陷病毒(human immunodeficiency virus, HIV)-1 病毒的细胞中灭活了病毒基因等[4]。

CRISPR/Cas9 由单向导 RNA (singe guide RNA, sgRNA)和 Cas9 蛋白结合而成，它依靠 sgRNA 识别靶点 DNA 序列，随后用 Cas9 核酸酶在 sgRNA 引导下结合前间区序列邻近基序(proto-spacer adjacent motif, PAM)[5]位点靶向切割目标序列，从而实现对基因进行删除、替换、插入等操作。然而，CRISPR 系统在靶向目标基因组的过程中，由于其在组织细胞内的传递不具有特异性，可能会在 sgRNA 与目标 DNA 序列碱基互补配对时出现多个位点的碱基不匹配，从而导致 Cas9 核酸酶在序列的非目标靶点裂解引发非预期的突变，导致发生脱靶效应[6-7]。针对这个问题，Sanger 测序法通过脱靶预测软件获取潜在脱靶位点，并对这些位点进行 PCR 扩增、测序，判断是否脱靶，但是这类技术灵敏度较低[8]。基于经验的方法根据已有脱靶数据的特征信息，使用特定的脱靶分数来分析脱靶效应，如切割频率确定(cutting frequency determination, CFD)[9]、CRISPR/Cas9 脱靶预测和识别工具

(CRISPR/Cas9 off-target prediction and identification tool, CROP-IT)[10]和错配容忍指数(mismatch tolerance index, MIT)[11]等。尽管这些分数在一定程度上可以表明脱靶效应的特征，但是都有自身的局限性，它们都没有建立起未脱靶位点与脱靶位点之间的联系，而且当数据量增加到一定程度后其准确率无法进一步提升。

近年来，机器学习在图像分类、自然语言处理等方面都取得了可观的成果[12]。随着新一代测序技术的发展，研究者积累了大量的CRISPR/Cas9 基因组编辑数据。如何利用这些数据，从中挖掘出有用的信息是当前一项重要的任务。近年来，以深度学习为代表的人工智能技术取得快速的发展，将人工智能与基因编辑数据有效地结合起来，运用深度学习辅助CRISPR/Cas9 脱靶预测研究是一个新兴的思路，是使用该技术成功进行基因编辑的关键步骤之一，将有助于实现更加高效安全的基因编辑和基因治疗。2018 年，Lin 等[13]基于卷积神经网络提出 CNN_std 脱靶预测模型，将 sgRNA 和 DNA 序列分别进行独热编码，再按位进行“或”运算后输入模型进行预测。在开源数据集的实验结果表明，CNN_std 优于传统脱靶分数计算方法。2020 年，Lin 等[14]基于循环神经网络提出了 CRISPR-Net 模型，首次将 sgRNA-DNA 序列对插入缺失和错配进行编码后输入到模型预测脱靶活性，模型在接收器工作特性曲线下面积(area under the receiver operating characteristic curve, AUROC)和精确召回曲线下面积(area under the precision recall curve, AUROC)等指标上表现出良好的结果，在脱靶问题上有良好的预测性能。2021 年，Charlier 等[7]等提出新的 sgRNA-DNA 序列编码作为前馈神经网络(feedforward neural network, FNN)、卷积神经网络(convolutional neural network, CNN)和循环神经网络(recurrent neural network, RNN)模型的输入，用于脱靶预测研究，模型在 AUROC 指标上得到了改善，验证了其模型能准确地预测脱靶结果。然而，现有的脱靶数据集大多数存在数据严重不平衡的问题，现有模型不够有说服力。本文运用数据集成和深度学习进行 CRISPR/Cas9 系统脱靶预测研究。首先，运用重采样法整合不同平台的数据集，构建脱靶基准数据集；其次，运用不同的编码方式对 DNA 与 sgRNA 序列对进行独热编码；最后，运用多尺度卷积神经网络构建 CnnCRISPR 脱靶预测模型，从而更好地提取基因编码特征。在基准数据集上与现有算法比较，CnnCRISPR 模型具有较好的准确性和泛化能力。

1 方法

1.1 数据集

本文采用 9 个开源数据集(包括 hek、iGWOS、II5、II6、k562、II1、II2、II3 和 II4)进行实验(其中，前面 5 个数据集用于构建基准数据集，后面 4 个数据集作为独立测试集用于测试模型的泛化性能)。表 1 描述了上述 9 个开源数据集的整体信息，总的来说，数据正负样本比例最大的是 II6，达到了 1:6 846.55，而数据正负样本比例最小的是 II1，仅为 1:1.14。另外，数据量达到 10 万以上的数据集占了一半，并且 iGWOS 的样本量最大，有 444 921 个，而不足 10 万的数据集占了 4 个，样本量最小的数据集为 II1，仅有 4 853 个。每个数据样本包含长度为 23 bp 的 sgRNA 序列、DNA 序列及标签。其中，标签为 1 表示脱靶，标签为 0 表示未检测到脱靶。研究表明，与回归分析相比，脱靶预测更适合采用分类法进行分析[15]。本文采用二分类法进行脱靶预测。此外，数据集类别不平衡给预测模型构建带来了极大的挑战，如图 1

所示，原始数据集的正负分布极度不平衡，例如，k562 数据集的正样本只有 120 个，而负样本达到了 20 199 个，正负比例达到了 1:168.33，同样，hek 数据集的正负样本比例为 1:246.97。目前，解决类不平衡问题通常采用数据重采样技术等。对此，本文采用数据重采样技术构建实验的基准数据集。首先，依次提取每个数据集，将正负样本分成 2 个子集。保留所有的正样本，并从负样本集中随机抽取比正样本集多的负样本。然后将抽出的数据进行一系列整合等操作得到一个用于后续模型训练的基准数据集，该基准数据集含有 2 616 个正样本，3 616 个负样本。最后，用独热编码法将碱基符号序列转化为数字特征序列。

1.2 数据编码

由于计算机只能处理数字信号，而 CRISPR/Cas9 序列通常由‘A’、‘G’、‘C’和‘T’这些符号表示，因此需要选择适当的方法将非数值数据处理成数值数据。常用的方法有单字母法 (one-letter code)[16] 和独热编码法 (one-hot coding)[17]。独热编码解决了分类器不好处理属性数据的问题，适用于特征不多的情况。本文运用独热编码将核苷酸‘A’表示为[1,0,0,0]，‘G’表示为[0,1,0,0]，‘C’表示为[0,0,1,0]，‘T’表示为[0,0,0,1]。

1.3 深度学习算法

1.3.1 卷积神经网络

由于图像包含的数据量太大，早期应用人工智能处理图像时效果并不理想。而且，在图像数字化过程中，特征很难保留下来。卷积神经网络(convolutional neural network, CNN)[18]的出现解决了上述难题，近年来已被广泛应用于自然语言处理、场景分类等领域。CNN 能够通过降维减少参数，并用类似视觉的方式从图像中学习特征，极大地扩展了深度学习的应用。与传统神经网络相比，CNN 的创新之处在于利

表 1 实验数据集的正负样本分布情况

Table 1 The distribution of positive and negative samples in the experimental dataset

	hek	iGWOS	II5	II6	k562	II1	II2	II3	II4
Number of positive	536	1 850	54	56	120	2 273	52	3 767	354
Number of negative	132 378	443 071	95 775	383 407	20 199	2 580	10 077	213 966	294 180
Total	132 914	444 921	95 829	383 463	20 319	4 853	10 129	217 733	294 534
IR*	1:246.97	1:239.50	1:1 773.61	1:6 846.55	1:168.33	1:1.14	1:193.79	1:56.80	1:831.02

*: IR, imbalance rate.

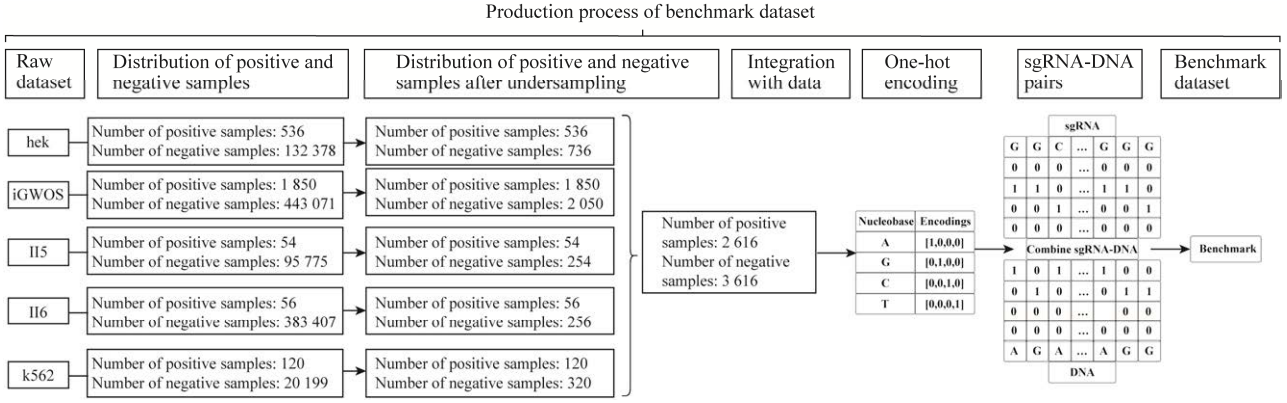

图 1 构建基准数据集流程图

Figure 1 Flowchart of building the benchmark dataset.

用卷积来提取图像中的特征，通过将输入矩阵与一个指定尺寸的卷积核以一定的步长做卷积操作，得到包含原图像信息的特征图，从而简化参数并保留特征。

1.3.2 GoogLeNet 模型

通常，研究者大多通过增加网络的深度、堆叠模型的参数量来提高模型精度，如 Alexnet[19]。但如果训练集有限，不断增加参数会容易产生过拟合；参数越多，计算压力越大；网络越深，容易出现梯度弥散问题。2015 年，Szegedy 等[20]发表了在 2014 年 ImageNet 大型视觉识别挑战赛(ImageNet large scale visual recognition challenge 2014, ILSVRC2014)上提出的 GoogLeNet 模型，创新性地引入 Inception 模块来组装网络，采用并行式网络结构代替以往的串联式结构，在增加网络深度和宽度的同时减少了参数，大大提升了模型的性能。Inception 结构是把不同尺寸的卷积核(1×1, 3×3, 5×5)以及最大池化层堆叠成一个分支网络结构，从而提供不同的感受野，让网络自适应地选择合适的感受野，提升了网络对尺度的适应性。但是，由于用到了 5×5 的卷积，容易造成特征图厚度很大的问题。因此该作者提出了 Inception v1 模块，如图 2 所示，可以结合处于不同通道但相关性很高的特征，降低了特征图厚度，大大减少了参数量。

1.4 算法评价体系

在分类问题中，评判模型的性能好坏通常采用准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1 分数(F1 score)、接收器工作特性(receiver operating characteristic, ROC)曲线以及精确召回曲线(precision recall curve, PRC)等评价指标。

准确率(Accuracy, ACC)指预测正确的样本数与整体的样本数之比，其公式为：

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}} \tag{1}$$

其中，TP 为真阳性，表示正确预测的正样本数量；FP 为假阳性，表示错误预测的负样本数量；TN 为真阴性，表示正确预测的负样本数量；FN 为假阴性，表示错误预测的正样本数量。

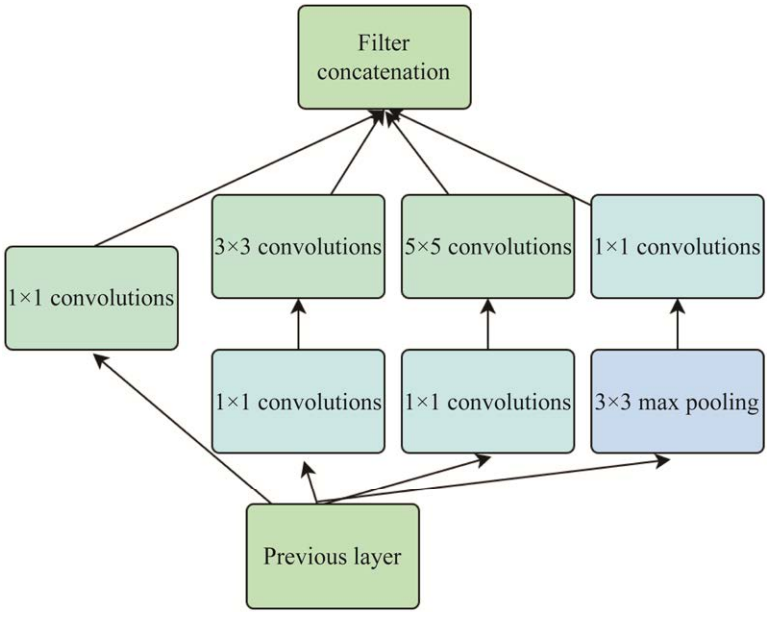

图 2 Inception v1 结构

Figure 2 Structure of inception v1.

精确率(Precision)指预测为正的样本中预测正确的比例，其公式为：

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{2}$$

召回率(Recall)指真实值为正的样本中预测正确的比例，其公式为：

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{3}$$

F1 分数指精确率和召回率的调和平均值。其公式为：

$$\text{F1}=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{4}$$

接收器工作特性(receiver operating characteristic, ROC)[21]曲线可以直观地评定模型预测结果的图像，其公式为：

$$\begin{cases}\text{TPR}=\dfrac{\text{TP}}{\text{TP}+\text{FN}}\\ \text{FPR}=\dfrac{\text{FP}}{\text{FP}+\text{TN}}\end{cases} \tag{5}$$

ROC 曲线的横轴为假阳性率(false positive rate, FPR)，纵轴为真阳性率(true positive rate, TPR)。ROC 曲线越靠近左上角，分类的效果越好。而曲线下面积(area under the curve, AUC)则是 ROC 曲线下方面积的大小，如果曲线的趋势接近，可通过计算 AUC 值来比较分类器。精确召回曲线(precision recall curve, PRC)的横轴为召回率(Recall)，纵轴为精确率(Precision)，PRC 曲线下的面积称为精确召回曲线下面积(area under the precision recall curve, AUPRC)值，模型的性能好坏与 AUPRC 值呈正相关，值越高，性能越好。

1.5 模型训练与模型选择

1.5.1 CnnCRISPR 模型

本文基于 Inception v1 模块构建 CnnCRISPR 模型，通过改进 Inception v1 模块内卷积层和池化层的尺寸大小来优化模型的性能，如图 3 所示。首先，接入 Inception v1 模块的变体(并行设计 1×1、4×1、8×1 的卷积层和 4×1 的池化层)。接着，添加批量标准化(batch normalization, BN)层，经过展平(Flatten)层展平后接多级全连接(Dense)层与随机失活(Dropout)层(其中，Dense 层的单元数依次为 128、92 和 23，并且均用“ReLu”激活函数；Dropout 层的比率依次为 0.25 和 0.15)来进一步提取数据特征并防止过拟合。最后，输出模型的预测结果。

1.5.2 模型训练

CnnCRISPR 模型使用 Python3.7 编写，由深度学习库 Tensorflow-gpu 2.5.0 作为后端实现的。模型的训练和测试过程是在 Intel Xeon 4210R CPU@2.4GHz、Ubuntu 20.04 LTS、32GB RAM 和 NVIDIA 12GB RTX3080Ti GPU 的服务器上执行的。CnnCRISPR 使用学习率为 0.000 1 的“Adam”优化器自动调整优化参数；损失函数设置为“binary_crossentropy”，评价指标设置为“accuracy”；将“batch_size”设置为 128 来避免发生过拟合。为了获得性能良好的模型，在训练过程中加入回调函数“ReduceLROnPlateau”和“EarlyStopping”，其中“ReduceLROnPlateau”的因子(factor)设置为 0.2，训练最多进行 150 个 epoch，“patience”分别为 8 和 20。若验证集的性能没有提升，训练在 8 个 epoch 后缩小学习率，而在 20 个 epoch 停止训练。其中，先以 8.5:1.5 的比例将数据集划分为训练验证集和测试集，再将划分得到的训练验证集以 7:3 的比例划分训练集和验证集，本文的所有实验均以该划分方式划分数据集。

1.5.3 CnnCRISPR 模型选择

首先，在基准数据集上评估 CnnCRISPR 标准结构及其 4 种变体的预测性能，从中选择最优的模型，包括：(1) Origin_Inception：将 CnnCRISPR 的 Inception 模块改为原始的 Inception 模块；(2) CnnCRISPR_ap：将 CnnCRISPR 的 BN 层后加入尺寸大小为 2×2 的最大池化层；

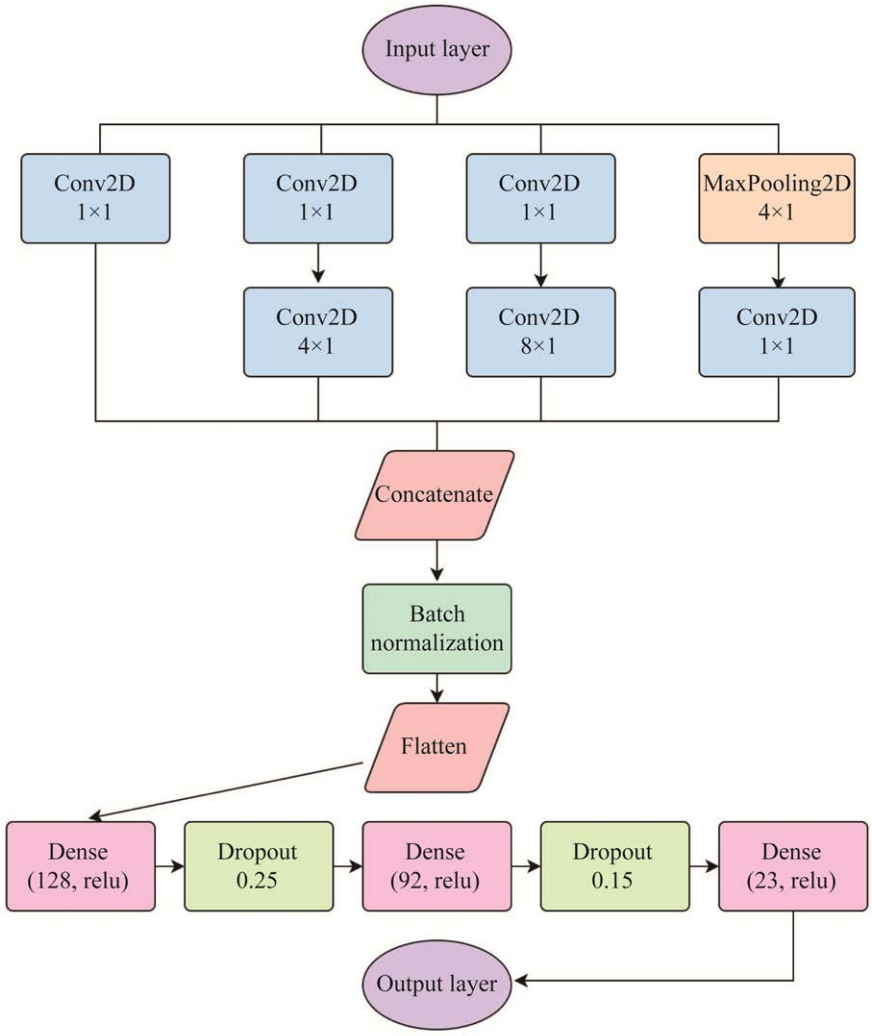

图 3　CnnCRISPR 架构

Figure 3　Architecture of CnnCRISPR.

(3) CnnCRISPR_nd：将 CnnCRISPR 去掉 dropout 层；(4) CnnCRISPR_nb：将 CnnCRISPR 删除 BN 层。图 4 展示所有模型的 ROC 曲线和 PRC 曲线的趋势，可以看到 CnnCRISPR 的 ROC 和 PRC 曲线均位于所有曲线的上方，并且 CnnCRISPR 的 ROC 曲线下面积和 PRC 曲线下面积均优于其他模型变体。另外，表 2 记录的各个模型对测试集预测后的评价结果可以看出，无论是精确率、召回率还是 F1 分数，CnnCRISPR 模型的所有评价指标均优于其他模型，表明了 CnnCRISPR 标准模型预测 CRISPR/Cas9 脱靶情况具有最好的预测性能。此外，没有变换卷积尺寸的 Origin_Inception 模型的预测效果最差，因此，优化卷积核尺寸能够提升 CnnCRISPR 脱靶预测能力。

为了进一步验证多尺度卷积神经网络的可行性，本文在基准数据集上评估了 CnnCRISPR 和其他卷积组合变体模型，包括：(1) CnnCRISPR_1×1：将所有卷积层的尺寸大小都设置为 1×1；(2) CnnCRISPR_4×1：将所有卷积层的尺寸大小都设置为 4×1；(3) CnnCRISPR_8×1：将所有的卷积层尺寸大小都设置为 8×1。如图 5 所示，CnnCRISPR

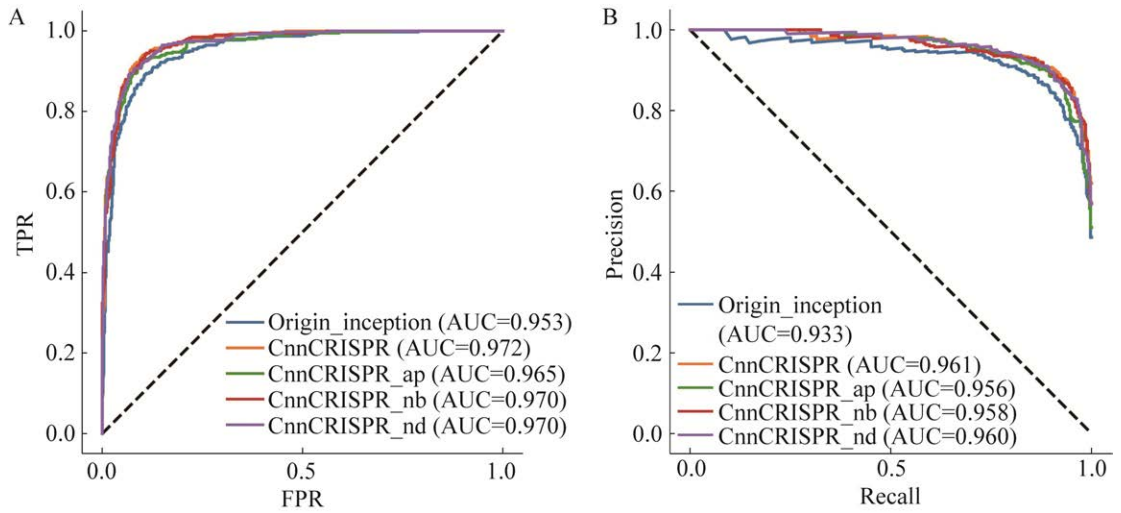

图 4　CnnCRISPR 与其模型变体的 ROC 曲线(A)和 PRC 曲线(B)

Figure 4　ROC curve (A) and PRC curve (B) of CnnCRISPR and its model variants.

表 2　CnnCRISPR 标准模型及其模型变体在基准数据集的预测性能比较

Table 2　Comparison of prediction performance of CnnCRISPR standard model and its model variants on benchmark datasets

	Origin_Inception	CnnCRISPR	CnnCRISRP_ap	CnnCRISPR_nb	CnnCRISPR_nd
Accuracy	0.891	**0.918***	0.907	0.912	0.911
Precision	0.863	**0.905**	0.886	0.895	0.895
Recall	0.885	**0.902**	0.897	0.900	0.897
F1	0.874	**0.903**	0.892	0.898	0.896
AUROC	0.953	**0.972**	0.965	0.970	0.970
AUPRC	0.933	**0.961**	0.956	0.958	0.960

*: Values in bold indicate the best score for each indicator.

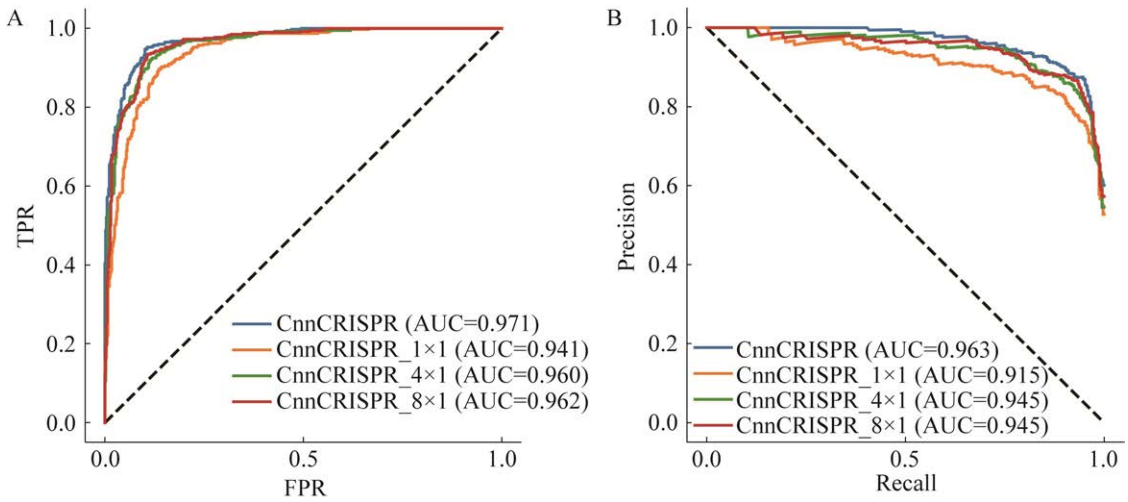

图 5　CnnCRISPR 与其模型变体的 ROC 曲线(A)和 PRC 曲线(B)

Figure 5　ROC curve (A) and PRC curve (B) of CnnCRISPR and its model variants.

模型的 ROC 曲线和 PRC 曲线均位于其他模型变体的曲线上方。表 3 总结了标准 CnnCRISPR 及其 3 种仅含单一尺度的卷积核的模型变体预测 CRISPR/Cas9 脱靶情况的实验结果，可以看出，CnnCRISPR 标准模型在所有的评价指标上都优于其他模型变体，另外，所有卷积层的尺寸大小都设置为 1×1 的模型变体，其预测性能最差，这说明了该模型变体不能很好学习到序列的特征。上述结果进一步验证了多尺度卷积神经网络的设计的优越性。

综上，多尺度卷积神经网络的设计更能从不同的角度捕捉到脱靶数据的特征，从而有助于提高模型的预测性能。

2 结果与分析

2.1 模型比较

本节在基准数据集上比较 CnnCRISPR 与现有的基于深度学习的 CRISPR/Cas9 脱靶预测方法，包括 Charlier 等[7]提出的 CNN、FNN 和 RNN 脱靶预测模型、Lin 等[13]提出的 CNN_std 模型以及 Lin 等[14]提出的 CRISPR-Net 模型，为了能很好地做对比，CRISPR-Net 模型的输入层做了调整，以满足本文所提出的编码方式。实验结果如图 6 所示，CnnCRISPR 的 ROC 曲线位于最左上方，PRC 曲线位于最右上方。表 4 描述上述模型的各个评价指标，可以看出，

表 3 CnnCRISPR 标准模型及 3 种仅含单一尺寸卷积核的模型变体在基准数据集的预测性能比较

Table 3 Performance comparisons for CnnCRISPR and three variants with single convolution kernel size on benchmark dataset

	CnnCRISPR	CnnCRISPR_1×1	CnnCRISRP_4×1	CnnCRISPR_8×1
Accuracy	**0.913***	0.875	0.896	0.905
Precision	**0.894**	0.825	0.863	0.878
Recall	**0.905**	0.897	0.900	0.902
F1	**0.899**	0.860	0.881	0.890
AUROC	**0.971**	0.941	0.960	0.962
AUPRC	**0.963**	0.915	0.945	0.945

*: Values in bold indicate the best score for each indicator.

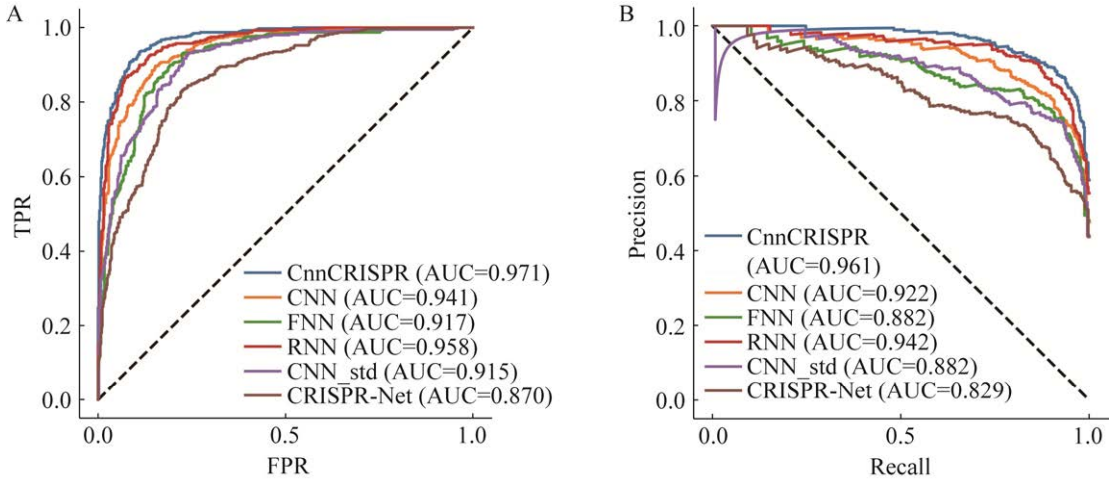

图 6 CnnCRISPR 与其他深度学习脱靶预测方法在 8×23 编码方式下的 ROC 曲线(A)和 PRC 曲线(B)

Figure 6 ROC curve (A) and PRC curve (B) of CnnCRISPR and other deep learning-based off-target prediction methods under 8×23 coding mode.

表 4　在基准数据集上比较 CnnCRISPR 与其他方法的性能

Table 4　Performance comparison of CnnCRISPR and other methods on benchmark dataset

	CnnCRISPR	CNN	FNN	RNN_modified	CNN_std	CRISPR-Net
Accuracy	**0.907***	0.866	0.845	0.892	0.824	0.800
Precision	**0.873**	0.817	0.782	0.858	0.791	0.759
Recall	**0.915**	0.885	0.882	0.895	0.797	0.779
F1	**0.894**	0.850	0.829	0.876	0.794	0.769
AUROC	**0.971**	0.941	0.917	0.958	0.915	0.870
AUPRC	**0.961**	0.922	0.882	0.942	0.882	0.829

*: Values in bold indicate the best score for each indicator.

CnnCRISPR 的综合表现是最好的，精确率、召回率和 F1 值均最大，分别为 0.873、0.915 和 0.894，说明 CnnCRISPR 对脱靶数据分类正确的概率最高。

另外还使用了独立测试集 II1、II2、II3 以及 II4 测试 CnnCRISPR 与其他现有模型的泛化能力。图 7 展示了所有模型在 4 个测试集上的 ROC 曲线和 PRC 曲线的趋势，表 5 记录了各个模型在所有测试集上的评价结果。具体来说，对比所有模型在 II1 测试集上的预测效果，CnnCRISPR 的准确率、F1 分数、AUROC 值和 AUPRC 值是最优的，因此，CnnCRISPR 在 II1 上的预测性能优于其他模型；对比所有模型在 II2 测试集上的预测效果，CnnCRISPR 的精确率、召回率、F1 分数和 AUROC 值是最优的，而 RNN_modified 的召回率和 AUPRC 值是最高的，但整体泛化能力不如 CnnCRISPR；而对 II3 和 II4 的预测，RNN_modified 的泛化能力是最优的，CnnCRISPR 的泛化能力仅次于 RNN_modified。综上分析，CnnCRISPR 在 II1 和 II2 上的泛化能力优于其他深度学习方法，而在 II3 和 II4 上的泛化能力仅次于 RNN_modified，RNN_modified 在这 2 个测试集上的泛化能力是最优的。

2.2　重采样策略对结果的影响

为了验证本文采用的重采样策略对 CnnCRISPR 模型脱靶预测的显著性，首先将 5 个开源数据集整合为一个大的数据集，名为 Rawdata，样本总量达到了 1 077 446 条，其中有 2 616 个正样本，有 1 074 830 个负样本，正负样本比例约为 1:410。接着将 Rawdata 数据集和基准数据集分别输入到 CnnCRISPR 模型中进行训练。除数据集不一样外，其余无论是数据划分方式还是模型训练超参数等都保持一致，实验结果如表 6 所示，Rawdata 数据集训练的 CnnCRISPR 模型，其预测的准确率和 AUROC 值要高于基准数据集训练的 CnnCRISPR 模型的预测，但这不能说明 Rawdata 数据集训练的 CnnCRISPR 模型要比基准数据集训练的 CnnCRISPR 模型的预测效果好，因为在数据极端不平衡的情况下，准确率和 AUROC 值不能反映模型的好坏，在这种情况下，应该观察精确率、召回率和 F1 分数这 3 个评价指标，而基准数据集训练的 CnnCRISPR 模型，其预测结果的精确率、召回率和 F1 分数均明显优于 Rawdata 数据集训练的 CnnCRISPR 模型的预测结果。实验结果说明了本文所采用的重采样策略能明显提高 CnnCRISPR 模型的脱靶预测性能。

2.3　编码方式对结果的影响

本文选择的编码方式是 8×23，这种编码方式的优势在于可以把 sgRNA 与 DNA 的数据特征完整地保留下来，但是却无法很好地表示

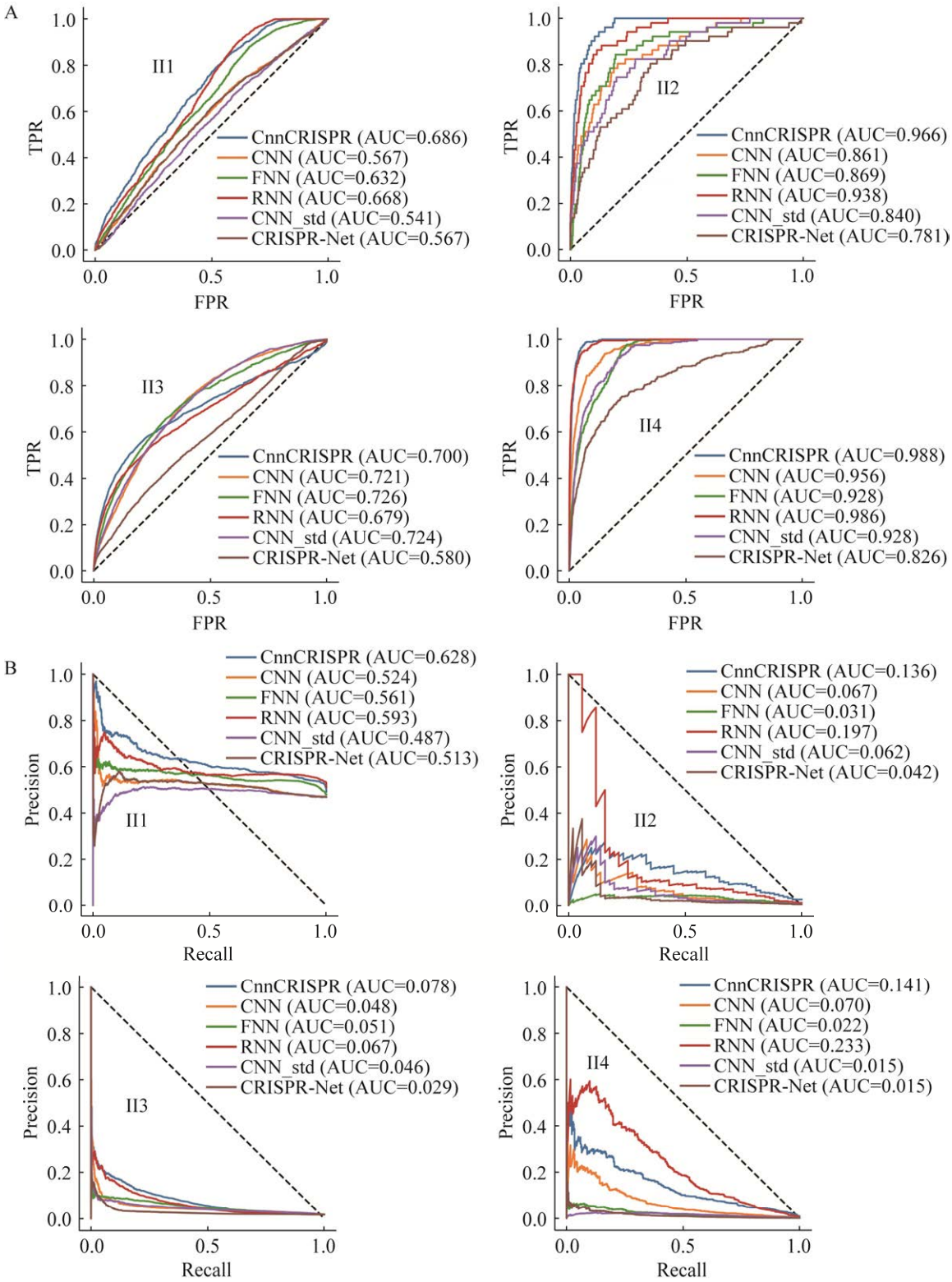

图 7　8×23 编码方式下 CnnCRISPR 和现有模型在独立测试集上的 ROC 曲线(A)和 PRC 曲线(B)

Figure 7　ROC (A) and PRC (B) of CnnCRISPR and existing models on independent test set under 8×23 encoding mode.

表 5　比较 CnnCRISPR 与其他方法在独立测试集上的泛化能力

Table 5　Generalizability comparison of CnnCRISPR with other methods on independent test sets

	CnnCRISPR	CNN	FNN	RNN_modified	CNN_std	CRISPR-Net
II1						
Accuracy	**0.595***	0.541	0.555	0.536	0.527	0.546
Precision	0.537	0.536	**0.607**	0.502	0.475	0.546
Recall	0.965	0.138	0.141	**1.000**	0.105	0.177
F1	**0.690**	0.220	0.229	0.669	0.172	0.267
AUROC	**0.686**	0.567	0.632	0.668	0.541	0.567
AUPRC	**0.628**	0.524	0.561	0.593	0.487	0.513
II2						
Accuracy	0.747	0.788	0.721	0.711	0.833	**0.839**
Precision	**0.020**	0.019	0.016	0.017	**0.020**	0.017
Recall	**1.000**	0.804	0.902	**1.000**	0.647	0.549
F1	**0.039**	0.037	0.032	0.034	0.038	0.034
AUROC	**0.966**	0.861	0.869	0.938	0.840	0.781
AUPRC	0.136	0.067	0.031	**0.197**	0.062	0.042
II3						
Accuracy	**0.956**	0.926	0.930	**0.956**	0.924	0.886
Precision	0.116	0.053	0.083	**0.146**	0.056	0.031
Recall	0.235	0.194	0.305	**0.314**	0.215	0.185
F1	0.155	0.083	0.131	**0.199**	0.089	0.053
AUROC	0.700	0.721	**0.726**	0.679	0.724	0.580
AUPRC	**0.078**	0.048	0.051	**0.067**	0.046	0.029
II4						
Accuracy	0.849	0.739	0.777	**0.895**	0.735	0.637
Precision	0.008	0.004	0.005	**0.011**	0.004	0.003
Recall	**1.000**	0.972	0.991	0.997	0.966	0.818
F1	0.016	0.009	0.011	**0.022**	0.009	0.005
AUROC	**0.988**	0.956	0.928	0.986	0.928	0.826
AUPRC	0.141	0.070	0.022	**0.233**	0.015	0.015

*: Values in bold indicate the best score for each indicator.

表 6　比较重采样前后 CnnCRISPR 的性能

Table 6　Performance comparison of CnnCRISPR before and after resampling

	Accuracy	Precision	Recall	F1	AUROC	AUPRC
Rawdata	**0.998***	0.622	0.265	0.372	**0.985**	0.406
Benchmark	0.917	**0.892**	**0.915**	**0.903**	0.973	**0.963**

*: Values in bold indicate the best score for each indicator.

sgRNA 和 DNA 序列之间的差异关系。因此，本节将采用一种新的编码方式对基因数据进行编码处理，以此来研究不同的编码方式对脱靶预测效果的影响。

正常的碱基都是以 A-T、C-G 的方式进行配对的，考虑到碱基错配的现象可能会对脱靶效应有很大的影响，因此引入一种新的编码方式来展示错配特征。首先依旧沿用独热编码的

方式对 sgRNA、DNA 序列进行编码，不同之处在于后续的处理。接着按照位或(bitwise or)的方法将两个矩阵对应位置的值相或，组合成为一个 4×23 的矩阵，结果如图 8 所示。

此外，由于输入数据的维度发生了变化，因此对现有的各个模型的输入端口做了适当的调整。为了适应数据尺寸的变化，CnnCRISPR 的卷积核尺寸也做了调整，将原来 8×1 的卷积核和 4×1 的最大池化层都改为了 2×1 的尺寸。

对采用位或方式编码的数据进行训练后，得出 CnnCRISPR 模型与其他深度学习模型的各项评价指标，通过观察图 9 和表 7 的各个评价结果可知，数据预处理的方式对于模型的训练也有很大的影响，4×23 的编码方式更能体现

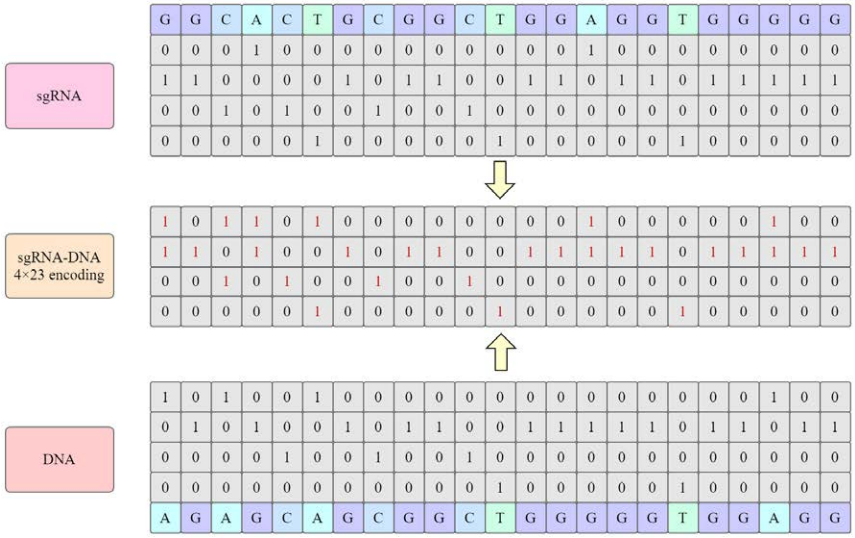

图 8　sgRNA-DNA 的 4×23 位或编码

Figure 8　sgRNA-DNA 4×23 bitwise or encoding.

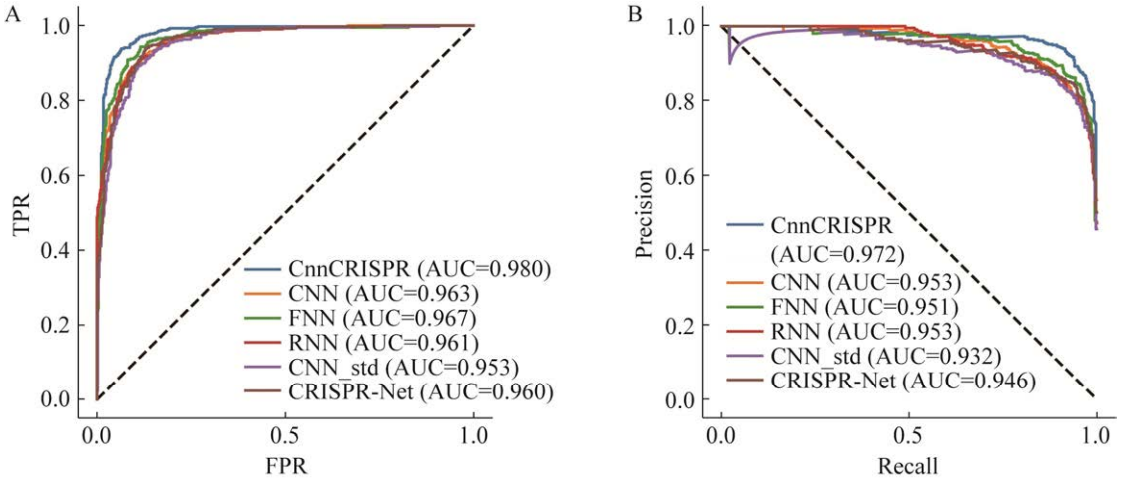

图 9　CnnCRISPR 与其他深度学习脱靶预测方法在 4×23 编码方式下的 ROC 曲线(A)和 PRC 曲线(B)

Figure 9　ROC curve (A) and PRC curve (B) of CnnCRISPR and other deep learning-based off-target prediction methods under 4×23 coding mode.

表 7　4×23 编码方式下模型间的性能比较

Table 7　Performance comparison among models under 4×23 encoding method

	CnnCRISPR	CNN	FNN	RNN_modified	CNN_std	CRISPR-Net
Accuracy	**0.936***	0.898	0.908	0.892	0.882	0.890
Precision	**0.926**	0.888	0.902	0.909	0.879	0.889
Recall	**0.926**	0.877	0.884	0.835	0.845	0.520
F1	**0.926**	0.882	0.893	0.870	0.862	0.870
AUROC	**0.980**	0.963	0.967	0.961	0.953	0.960
AUPRC	**0.972**	0.953	0.951	0.953	0.932	0.946

*: Values in bold indicate the best score for each indicator.

出数据样本的特征，无论是哪种模型，在评估中都取得了比 8×23 拼接编码方案更好的效果，并且采用新编码方式的数据训练出来的 CnnCRISPR 模型的性能表现仍然是最好的。

同样用上述4个独立测试集测试CnnCRISPR与其他现有模型的泛化能力。图 10 展示了所有模型在各个测试集上的 ROC 曲线和 PRC 曲线的趋势，表 8 记录了各个模型在所有测试集上的评价指标。具体来说，比较所有模型在 II1 测试集上的预测效果，FNN 的准确率、精确率和 F1 分数是最优的，RNN_modified 的 AUROC 值和 AUPRC 值是最高的，CnnCRISPR 的泛化能力仅次于 FNN 和 RNN_modified；比较所有模型在 II2 测试集上的预测效果，CnnCRISPR 的召回率和 AUROC 值是最高的，RNN_modified 的召回率和 AUPRC 值是最高的，CNN_std 的准确率、精确率和 F1 分数是最高的，这 3 个模型在重要指标上各有突出，泛化能力优于其他模型；而对 II3 和 II4 的预测，CnnCRISPR 的预测性能要优于其他模型。综上分析，CnnCRISPR 在 II1 上的泛化能力仅次于泛化能力最优的 FNN 和 RNN_modified，在 II2 上的泛化能力与 RNN_modified 和 CNN_std 相近，优于其他模型，而在 II3 和 II4 上的泛化能力优于其他深度学习方法。

表 9 和表 10 单独比较了 CnnCRISPR 模型在 2 种编码方式上的性能表现。具体来说，一方面，对比在基准数据测试集上的预测效果，4×23 编码方式的 CnnCRISPR 的预测性能要明显优于 8×23 编码方式的 CnnCRISPR。另一方面，在 4 个独立测试集上的预测效果，从重要指标 F1 分数、AUROC 值和 AUPRC 值的表现来看，4×23 编码方式的 CnnCRISPR 在这些重要指标上最优的有 7 个，分别在 II1、II2 和 II4，而 8×23 编码方式的 CnnCRISPR 在重要指标上最优的则占了 5 个，在对 II3 的预测评价指标就占了 3 个。换言之，8×23 编码方式的 CnnCRISPR 在 II3 上的泛化能力要优于 4×23 编码方式的 CnnCRISPR，而在 II1、II2 以及 II4 上的泛化能力不及 4×23 编码方式的 CnnCRISPR。综上分析说明了 bitwiseOR 编码对模型的性能提升有帮助，并且更适合用于脱靶预测。

2.4　CnnCRISPR 模型泛化性能测试

为了进一步验证 CnnCRISPR 具有较好的泛化能力，本节在数据集 k562 (样本总数=20 319) 和 hek (样本总数=132 914)上将其与 RNN 和 CNN_std 以及 CRISPR-Net 进行比较。对数据的处理方式沿用了 2.3 节的 4×23 编码方式。为了缓解源数据集样本不均衡的问题，本节采用过采样[22]法进行模型训练。具体思路如下：首先，将 k562 数据集分为正类子集以及负类

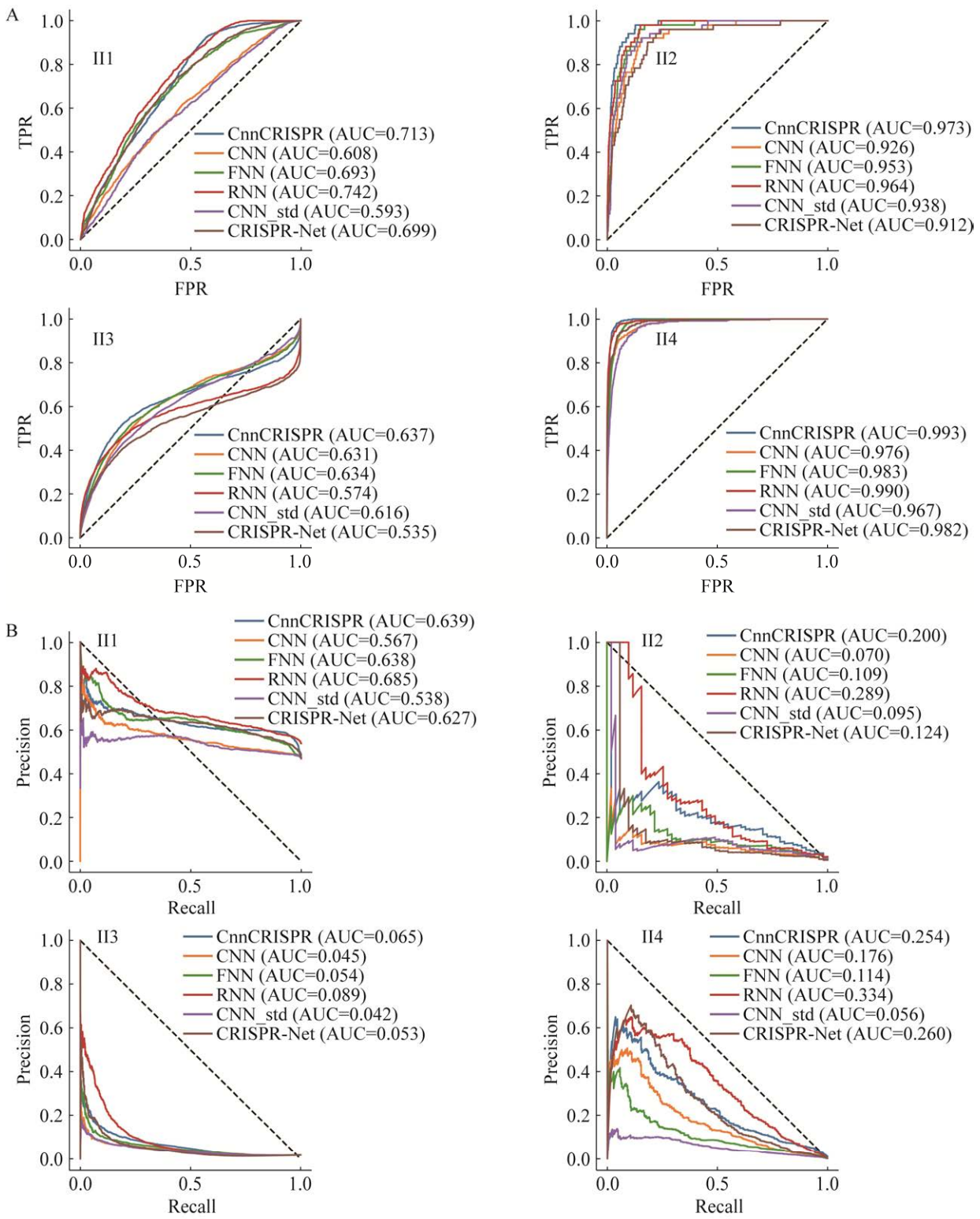

图 10　4×23 编码方式下 CnnCRISPR 和现有模型在独立测试集上的 ROC 曲线(A)和 PRC 曲线(B)

Figure 10　ROC (A) and PRC (B) of CnnCRISPR and existing models on independent test set under 4×23 encoding mode.

表 8　4×23 编码方式下比较所有模型在所有测试集上的泛化能力

Table 8　Generalizability comparison of all models on all test sets under 4×23 encoding method

	CnnCRISPR	CNN	FNN	RNN_modified	CNN_std	CRISPR-Net
II1						
Accuracy	0.542	0.521	**0.617***	0.483	0.545	0.494
Precision	0.506	0.494	**0.560**	0.475	0.511	0.480
Recall	0.991	0.924	0.848	1.000	0.681	**0.996**
F1	0.669	0.644	**0.675**	0.644	0.583	0.648
AUROC	0.713	0.608	0.693	**0.742**	0.593	0.699
AUPRC	0.639	0.567	0.638	**0.685**	0.538	0.627
II2						
Accuracy	0.656	0.815	0.688	0.569	**0.844**	0.767
Precision	0.015	0.025	0.016	0.012	**0.029**	0.020
Recall	**1.000**	0.922	0.980	**1.000**	0.902	0.941
F1	0.029	0.048	0.031	0.023	**0.056**	0.039
AUROC	**0.973**	0.926	0.953	0.964	0.938	0.912
AUPRC	0.200	0.070	0.109	**0.289**	0.095	0.124
II3						
Accuracy	0.944	0.926	0.936	0.918	0.916	**0.949**
Precision	**0.088**	0.060	0.071	0.072	0.055	0.075
Recall	0.241	0.223	0.224	**0.316**	0.236	0.173
F1	**0.129**	0.095	0.107	0.117	0.089	0.104
AUROC	**0.637**	0.631	0.634	0.574	0.616	0.535
AUPRC	0.065	0.045	0.054	**0.089**	0.042	0.053
II4						
Accuracy	**0.902**	0.852	0.880	0.870	0.818	0.819
Precision	**0.012**	0.008	0.010	0.009	0.006	0.007
Recall	**0.997**	0.963	0.989	0.991	0.977	0.989
F1	**0.024**	0.015	0.019	0.018	0.013	0.013
AUROC	**0.993**	0.976	0.983	0.990	0.967	0.982
AUPRC	0.254	0.176	0.114	**0.334**	0.056	0.260

*: Values in bold indicate the best score for each indicator.

表 9　CnnCRISPR 在 8×23 和 4×23 编码方式下的性能比较

Table 9　Performance comparison of CnnCRISPR under 8×23 and 4×23 encoding methods

	Accuracy	Precision	Recall	F1	AUROC	AUPRC
8×23	0.907	0.873	0.915	0.894	0.971	0.961
4×23	**0.936***	**0.926**	**0.926**	**0.926**	**0.980**	**0.972**

*: Values in bold indicate the best score for each indicator.

子集，分别进行编码。然后，设置一定大小的批次容量，从正子集和负子集中随机抽取等量的样本，各占批次容量的一半，作为模型训练的生成器。最后，在训练时使用 keras 的训练函数 fit_generator 代替 fit，利用生成器分批次向模型送入数据。在训练完毕后，将已经抽取过的负样本从负类子集中剔除，不改变正类子集，

表 10 比较 CnnCRISPR 在 8×23 和 4×23 编码方式下在所有测试集上的泛化能力

Table 10 Generalizability comparison of CnnCRISPR under 8×23 and 4×23 encoding methods on all test sets

	Accuracy	Precision	Recall	F1	AUROC	AUPRC
II1						
8×23	**0.595***	**0.537**	0.965	**0.690**	0.686	0.628
4×23	0.542	0.506	**0.991**	0.669	**0.713**	**0.639**
II2						
8×23	**0.747**	**0.020**	**1.000**	**0.039**	0.966	0.136
4×23	0.656	0.015	**1.000**	0.029	**0.973**	**0.200**
II3						
8×23	**0.956**	**0.116**	0.235	**0.155**	**0.700**	**0.078**
4×23	0.944	0.088	**0.241**	0.129	0.637	0.065
II4						
8×23	0.849	0.008	1.000	0.016	0.988	0.141
4×23	**0.902**	**0.012**	0.997	**0.024**	**0.993**	**0.254**

*: Values in bold indicate the best score for each indicator.

并按批次重复训练，直到负类子集中的数据被抽取，从而保证每批次的数据中正负样本是平衡的。随后设定训练批次容量为 128，测试批次容量为 100，分别对 4 个模型进行训练。表 11 和表 12 分别总结了各模型在数据集 k562 和 hek 上的实验结果。可以看出，无论是在 k562 数据集训练的模型还是在 hek 数据集训练的模型，RNN 模型在召回率上最好，但其 F1 分数的表现却不理想，而 CnnCRISPR 在 F1 分数以及 AUC 值等指标上均优于 RNN、CNN_std 和 CRISPR-Net，说明其预测效果最好。上述实验结果表明，CnnCRISPR 在数据集 k562 和 hek 上的综合脱靶预测结果优于 RNN、CNN_std 和 CRISPR-Net，从而验证了 CnnCRISPR 具有较好的泛化性能。另外，还讨论了分别在基准数据集、k562 数据集和 hek 数据集上的 CnnCRISPR 模型的各项评价指标。从表 13 中看到，基准数据集上训练的CnnCRISPR的准确率和AUROC值不如另外 2 个模型，但这不能否定在基准数据集上训练的 CnnCRISPR 的性能。前面可知 k562 数据集和 hek 数据集的正负比例分别为 1:168.33 和

表 11 比较 CnnCRISPR 与其他方法在 k562 数据集上的性能

Table 11 Performance comparison of CnnCRISPR and other methods on k562 dataset

	CnnCRISPR	RNN	CNN_std	CRISPR-Net
Accuracy	**0.988***	0.976	0.797	0.981
Precision	**0.300**	0.196	0.018	0.209
Recall	0.833	**1.000**	0.611	0.778
F1	**0.441**	0.327	0.034	0.329
AUROC	**0.996**	0.991	0.845	0.981
AUPRC	**0.651**	0.618	0.060	0.562

*: Values in bold indicate the best score for each indicator.

表 12 比较 CnnCRISPR 与其他方法在 hek 数据集上的性能

Table 12 Performance comparison of CnnCRISPR and other methods on hek dataset

	CnnCRISPR	RNN	CNN_std	CRISPR-Net
Accuracy	**0.995***	0.965	0.949	0.978
Precision	**0.421**	0.102	0.063	0.144
Recall	0.790	**0.963**	0.827	0.914
F1	**0.549**	0.184	0.117	0.248
AUROC	**0.996**	0.990	0.970	0.992
AUPRC	**0.698**	0.560	0.201	0.587

*: Values in bold indicate the best score for each indicator.

表 13 比较 CnnCRISPR 在不同数据集上的性能

Table 13 Performance comparison of CnnCRISPR on different datasets

	Accuracy	Precision	Recall	F1	AUROC	AUPRC
Benchmark	0.936	**0.926***	**0.926**	**0.926**	0.980	**0.972**
k562	0.988	0.300	0.833	0.441	**0.996**	0.651
hek	**0.995**	0.421	0.790	0.549	**0.996**	0.698

*: Values in bold indicate the best score for each indicator.

1:246.97，在数据极端不平衡的情况下，精确率、召回率和 F1 分数更能反映模型的整体性能，而基准数据集上训练的 CnnCRISPR 在这几个评价指标上要优于其他 2 个模型的指标。另外，在基准数据集上训练的 CnnCRISPR，其 AUPRC 值也是最高的，体现了其模型较另外 2 个模型要更稳定。同时也体现了本文采用的重采样策略构建的基准数据集能有效提升模型的预测性能。

3 结论

CRISPR/Cas9 核酸酶自身的不稳定性易发生突变产生脱靶效应。本文基于 Inception 模块，通过优化卷积层的尺寸大小，构建多尺度卷积神经网络的 CnnCRISPR 脱靶预测模型。开源数据集的实验表明，与现有基于深度学习的 CRISPR/Cas9 脱靶预测算法相比，CnnCRISPR 模型具有较好的准确性和泛化能力。在后续的研究中，尝试从生物特征入手，了解更多表现脱靶的特征，为脱靶效应问题提出更好的解决方法。

REFERENCES

[1] ZHENG W, GU F. Progress of application and off-target effects of CRISPR/Cas9[J]. Yi Chuan, 2015, 37(10): 1003-1010.

[2] ISHINO Y, SHINAGAWA H, MAKINO K, AMEMURA M, NAKATA A. Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product[J]. Journal of Bacteriology, 1987, 169(12): 5429-5433.

[3] JANSEN R, EMBDEN JD, GAASTRA W, SCHOULS LM. Identification of genes that are associated with DNA repeats in prokaryotes[J]. Molecular Microbiology, 2002, 43(6): 1565-1575.

[4] JIANG WY, BIKARD D, COX D, ZHANG F, MARRAFFINI LA. RNA-guided editing of bacterial genomes using CRISPR-Cas systems[J]. Nature Biotechnology, 2013, 31(3): 233-239.

[5] 尹珅, 贺桂芳, 赖方秾, 谢凤云, 马俊宇. CRISPR/Cas9 系统的脱靶效应[J]. 生物技术通报, 2016, 32(3): 31-37.
YIN K, HE GF, LAI FN, XIE FY, MA JY. Off-target of CRISPR/Cas9 system[J]. Biotechnology Bulletin, 2016, 32(3): 31-37 (in Chinese).

[6] 徐海波. 基于机器学习的 CRISPR/Cas9 系统脱靶效应及靶向效率预测[D]. 桂林: 桂林电子科技大学, 2020.
XU HB. Off-target effect and target efficiency prediction of CRISPR/Cas9 system based on machine learning[D]. Guilin: Guilin University of Electronic Technology, 2020 (in Chinese).

[7] CHARLIER J, NADON R, MAKARENKOV V. Accurate deep learning off-target prediction with novel sgRNA-DNA sequence encoding in CRISPR-Cas9 gene editing[J]. Bioinformatics, 2021, 37(16): 2299-2307.

[8] 张晨, 雷展, 李凯, 商颖, 许文涛. CRISPR/Cas9 系统中的脱靶效应及检测技术研究进展[J]. 生物技术通报, 2020, 36(3): 78-87.
ZHANG C, LEI Z, LI K, SHANG Y, XU WT. Research progress on off-target effects and detection techniques in CRISPR/Cas9 systems[J]. Biotechnology Bulletin, 2020, 36(3): 78-87 (in Chinese).

[9] DOENCH JG, FUSI N, SULLENDER M, HEGDE M, VAIMBERG EW, DONOVAN KF, SMITH I, TOTHOVA Z, WILEN C, ORCHARD R, VIRGIN HW, LISTGARTEN J, ROOT DE. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9[J]. Nature Biotechnology, 2016, 34(2): 184-191.

[10] SINGH R, KUSCU C, QUINLAN A, QI YJ, ADLI M. Cas9-chromatin binding information enables more accurate CRISPR off-target prediction[J]. Nucleic Acids Research, 2015, 43(18): e118.

[11] HSU PD, SCOTT DA, WEINSTEIN JA, RAN FA, KONERMANN S, AGARWALA V, LI YQ, FINE EJ, WU XB, SHALEM O, CRADICK TJ, MARRAFFINI LA, BAO G, ZHANG F. DNA targeting specificity of RNA-guided Cas9 nucleases[J]. Nature Biotechnology, 2013, 31(9): 827-832.

[12] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[13] LIN JC, WONG KC. Off-target predictions in CRISPR-Cas9 gene editing using deep learning[J]. Bioinformatics, 2018, 34(17): i656-i663.

[14] LIN JC, ZHANG ZL, ZHANG SX, CHEN JY, WONG

KC. CRISPR-net: a recurrent convolutional network quantifies CRISPR off-target activities with mismatches and indels[J]. Advanced Science, 2020, 7(13): 1903562.

[15] CHUAI GH, MA HH, YAN JF, CHEN M, HONG NF, XUE DY, ZHOU C, ZHU CY, CHEN K, DUAN B, GU F, QU S, HUANG DS, WEI J, LIU Q. DeepCRISPR: optimized CRISPR guide RNA design by deep learning[J]. Genome Biology, 2018, 19(1): 1-18.

[16] CHEN L, WANG SP, ZHANG YH, LI JR, XING ZH, YANG JL, HUANG T, CAI YD. Identify key sequence features to improve CRISPR sgRNA efficacy[J]. IEEE Access, 2017, 5: 26582-26590.

[17] 张桂珊, 杨勇, 张灵敏, 戴宪华. 机器学习方法在CRISPR/Cas9 系统中的应用[J]. 遗传, 2018, 40(9): 704-723.
ZHANG GS, YANG Y, ZHANG LM, DAI XH. Application of machine learning in the CRISPR/Cas9 system[J]. Yi Chuan, 2018, 40(9): 704-723 (in Chinese).

[18] YU HL, MU CX, SUN CY, YANG WK, YANG XB, ZUO X. Support vector machine-based optimized decision threshold adjustment strategy for classifying imbalanced data[J]. Knowledge-Based Systems, 2015, 76: 67-78.

[19] KRIZHEVSKY A, SUTSKEVER I, HINTON GE. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[20] SZEGEDY C, LIU W, JIA YQ, SERMANET P, REED S, ANGUELOV D, ERHAN D, VANHOUCKE V, RABINOVICH A. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 7-12, 2015, Boston, MA, USA. IEEE, 2015: 1-9.

[21] PEPE MS. Receiver operating characteristic methodology[J]. Journal of the American Statistical Association, 2000, 95(449): 308-311.

[22] 陶朝杰, 杨进. 基于 BalanceCascade-GBDT 算法的类别不平衡虚假评论识别方法[J]. 经济数学, 2020, 37(3): 214-220.
TAO CJ, YANG J. Detection of class-imbalance spam reviews based on BalanceCascade-GBDT algorithm[J]. Journal of Quantitative Economics, 2020, 37(3): 214-220 (in Chinese).

(本文责编 郝丽芳)