Recognition of CRISPR/Cas9 off-target sites through ensemble learning of uneven mismatch distributions

Hui Peng[1] , Yi Zheng[1] , Zhixun Zhao[1] , Tao Liu[2,3] and Jinyan Li[1]

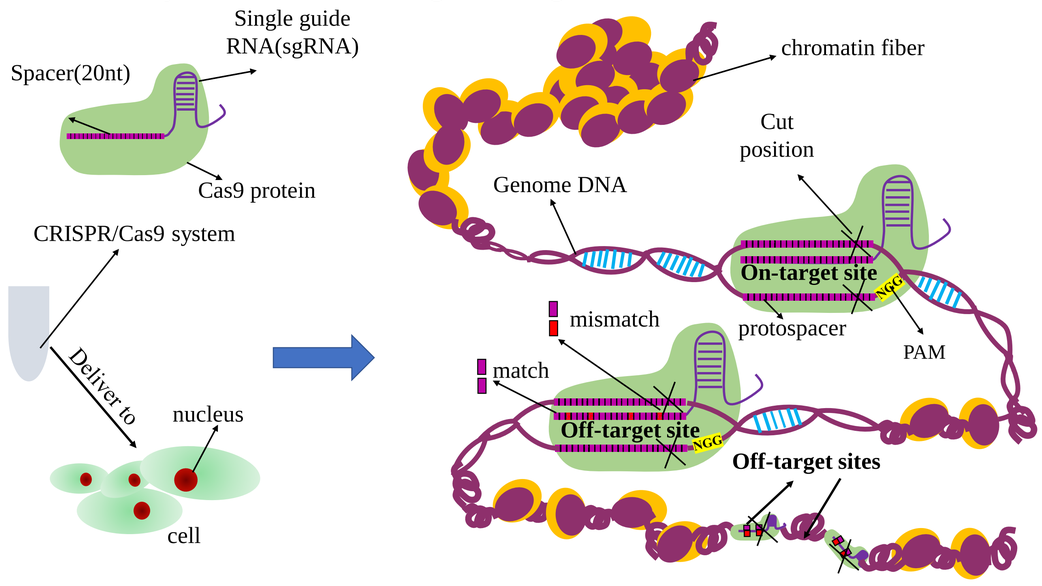
1.The Advanced Analytics Institute, Faculty of Engineering and Information Technology ,University of Technology Sydney, Australia
2.Centre for Childhood Cancer Research, University of New South Wales, Australia
3.Children's Cancer Institute Australia

ABSTRACT

CRISPR/Cas9 is driving a broad range of innovative applications from basic biology to biotechnology and medicine. One of its current issues is the effect of off-target editing that should be critically resolved and should be completely avoided in the ideal use of this system. Here, we developed an ensemble learning method to detect the off-target sites of an sgRNA from its thousands of genome-wide candidates. Nucleotide mismatches between on-target and off-target sites have been studied recently. We confirm that there exist strong mismatch enrichment and preferences at the 5'-end close regions of the off-target sequences. Comparing with the on-target sites, sequences of no editing sites can be also characterized by GC composition changes and position-specific mismatch binary features. Under this novel space of features, an ensemble strategy was applied to train a prediction model. The model achieved a mean score 0.99 of Area Under Receiver Operating Characteristic curve (AUROC) and a mean score 0.45 of Area Under Precision-Recall curve (AUPRC) in cross-validations on big data sets, outperforming state-of-the-art methods in various test scenarios. Our predicted off-target sites also correspond well to those detected by high-throughput sequencing techniques. Especially, two case studies for selecting sgRNAs to cure hearing loss and retinal degeneration partly prove the effectiveness of our method.

BACKGROUND

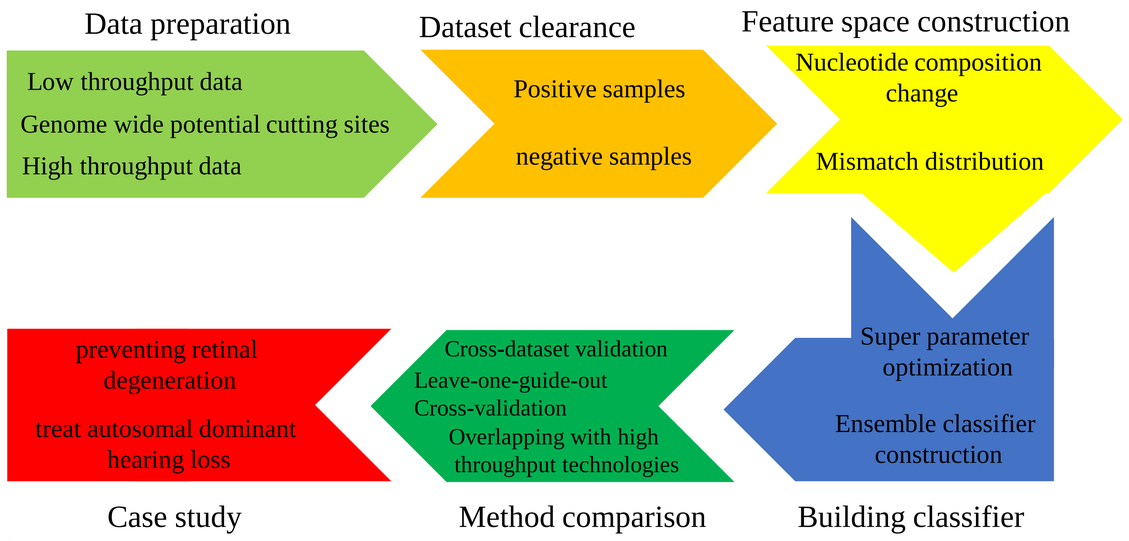
❑ CRISPR/Cas9 System and On/Off-target Editing



- **CRISPR/Cas9 complex**:
  part 1: a single-guide RNA(sgRNA) -- find target region (rule: 3nt PAM+20nt spacer, mismatches and indels can be tolerated); ***part 2***: a Cas9 protein -- cut the DNA.
- **On-target site**: the expected cutting region. **Off-target site**: the unintended cutting region.

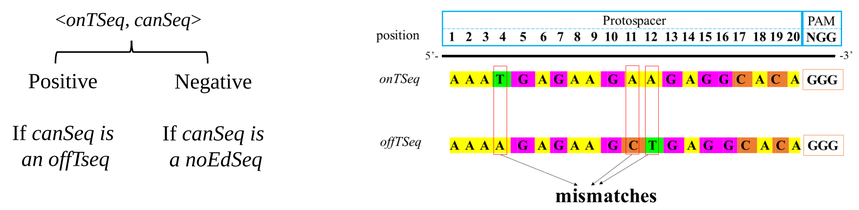
METHODOLOGY

❑ Flowchart



❑ Dataset

❖ Sequence pair (sample)



Positive Sets:

- D_{+}^{low}: Low throughput technologies (PCR) (215 positive samples involving 29 sgRNAs)
- D_{+}^{high}: High throughput technologies (NGS)(527 positive samples involving 11 sgRNAs)

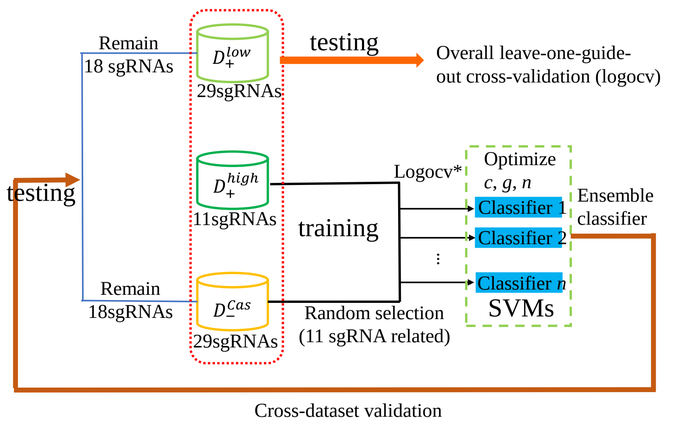
Negative Set:

- D_{-}^{Cas}: Cas-OFFinder tool (408260 negative samples involving 29 sgRNAs)

❑ Feature

- Nucleotide composition change features: <GC count change>, <GC percent change>, <GC skew change>, <AT skew change>, <Change of ratio of GC skew and AT skew>
- Position-specific binary mismatch features: <mismatch binary vector>

❑ Ensemble Learning Scheme



Optimization:

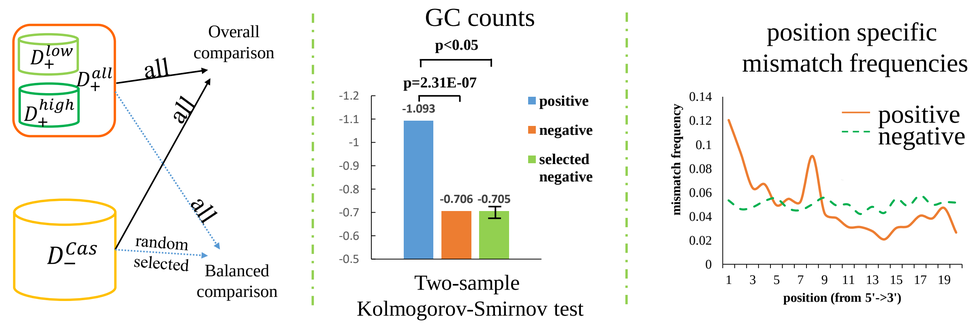
- ***Objective***: optimize super parameters
- ***Method***: leave-one-guide-out cross-validation (logocv), best AUROC
- ***Dataset***: D_{+}^{high} and D_{-}^{Cas} (11 sgRNAs, 11 folds)

Evaluation and comparison:

- ***Objective***: evaluate the classifier and compare with state-of-the-art methods
- ***Method***: Cross-dataset validation (cdv) and overall logocv, AUROC and AUPRC
- ***Dataset***: cdv—all samples exclude above Dataset (18sgRNAs); overall logocv—all samples (29 sgRNAs)

RESULTS

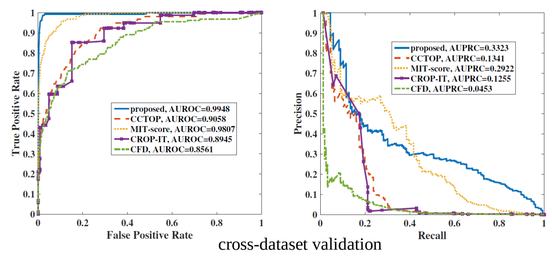
❑ Important sample difference



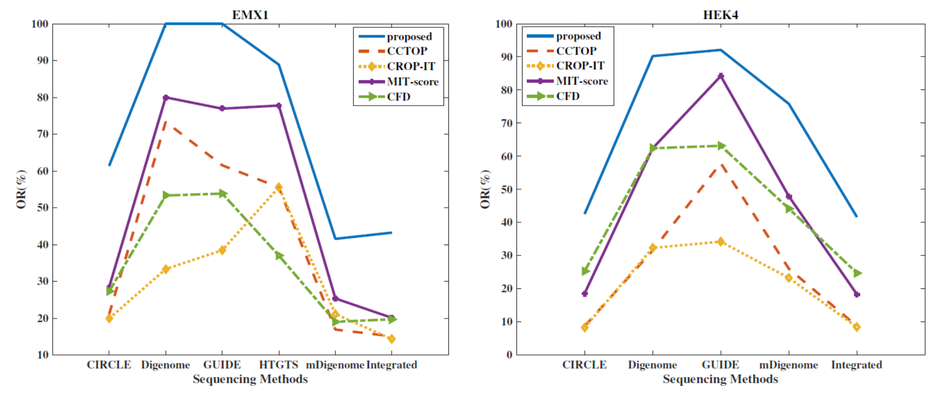
❑ Performance Comparison

- Performance Comparison in cross-dataset validation and leave-one-guide-out cross validation

Methods	cross-dataset validation		logocv[a]	
	AUROC	AUPRC	AUROC	AUPRC
proposed	0.9948	0.3323	0.9926	0.4571
CCTOP	0.9058	0.1341	0.9021	0.1407
MIT-score	0.9807	0.2922	0.9783	0.2960
CROP-IT	0.8945	0.1255	0.9160	0.1086
CFD	0.8561	0.0453	0.8835	0.0844



- Comparison of the off-target sites detected by the computational methods and those by the high-throughput sequencing methods



Overlap rate (OR): $\mathrm{OR}\% = \frac{overlapped\ number\ (com \cap wet\text{-}lab)}{number\ of\ wet\text{-}lab\ technique\ detections\ (wet\text{-}lab)} \times 100\%$

❑ Case Study: Choose optimal sgRNA for curing diseases

- Case One: editing Nrl gene to treat retinal degeneration
- Case Two: editing TMC1 gene to treat human autosomal dominant hearing loss gene

Table 3. The ranks of the sgRNAs by considering both of their cutting efficiencies and off-target potentials.

	sgRNA	literature			proposed			CRISPR Design			sgRNA Designer		
		Indel(%)	efficiency rank	final rank	ot number	ot rank	final rank	ot number	ot rank	final rank	ot number	ot rank	finak rank
Case one	NT1	21.9	4	-	264	5	5	101	2	2	-	1	3
	NT2	22.7	2	**1**	83	1	**1**	69	1	1	-	2	**1**
	NT3	22.5	3	-	139	4	3	159	4	4	-	5	4
	NT4	23.2	1	-	119	3	2	146	5	5	-	3	**1**
	NT5	18.3	5	-	95	2	3	115	3	3	-	4	5
Case two	Tmc1-mut1	4.1	2	-	613	3	3	337	3	3	-	1	**1**
	Tmc1-mut2	0.74	3	-	183	1	2	318	2	2	-	3	2
	Tmc1-mut3	10	1	**1**	247	2	**1**	197	1	**1**	-	2	3

- **ot number**: predicted off-target sites numbers
- **efficiency rank**: rank sgRNA by on target cutting efficiency; **ot rank**: rank sgRNA by off-target site number; **final rank**: (efficiency rank + ot rank)/2, final selection guidance

❑ Conclusion

Contribution:

- Turn the off-target site detection problem into a binary classification issue
- Define the sample as a sequence pair and take nucleotide composition changes and mismatch distribution properties as novel features
- Propose an ensemble learning scheme and improve the prediction performance

Future work:

- Collect more reliable negative samples
- Consider about bulges in the target sites
- Develop a complete sgRNA designing tool with the on-target cutting efficiency prediction with off-target site detection

CONTACT

Email: Jinyan.Li@uts.edu.au

UTS:AAI THE ADVANCED ANALYTICS INSTITUTE

ECCB 2018