picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

This is a title

Corresponding Author [1,*], Co-Author [2] and Co-Author [2,*]

[1]Department, Institution, City, Post Code, Country and
[2]Department, Institution, City, Post Code, Country.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Text Text Text Text Text Text.
Results: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text
Availability: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text
Contact: name@bio.com
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Clustered regularly interspaced short palindromic repeats (CRISPR)/CRIPSR-associated protein 9 (Cas9) systems is preferred over other biological research and human medicine technologies now, beacuse of it's efficiency, robustness and programmability. Cas9 nucleases can be directed by short guide RNAs (sgRNAs) to introduce site-specific DNA double-stranded breaks in target, so to enable editing site-specific within the mammalian genome (Jinek *et al.*, 2012; Cong *et al.*, 2013; Mali *et al.*, 2013). CRISPR/Cas9, to a large extent, has developed genetic therapies at the cellular level, while there are still severe medical disadvantage even now which has greatly hindered the further clinical application of the CRISPR/Cas9 systems. One of these disadvantage is due to point mutations caused by off-target effects (Rubeis and Steger, 2018; Kang *et al.*, 2016; Ishii, 2017; Liang *et al.*, 2015). To overcome this disadvantage, a solution is to engineer CRISPR/Cas9 with higher specificities. That's why more and more higher specificities Cas9 variants, such as enhanced SpCas9 (eSpCas9(1.1)), Cas9-High Fidelity (SpCas9-HF1) (Ishii, 2017; Slaymaker *et al.*, 2016), hyper-accurate Cas9 (HypaCas9) (Kleinstiver *et al.*, 2016), been developed and bring a significant volume of experimental data, that is to say researchers have to face the difficulty of analyzing such huge and heterogeneous data.

The activity of chosen sgRNA sequence determines the success of genome editing, however distinctly fluctuant behaviors have been observed for the performance of different sgRNAs, even in the same Cas9 system. Some optimum sgRNAs can hit almost all targers alleles, while anothers don't even show activity (Wang *et al.*, 2019). This fact indicates that it's meaningful to explore an efficient approaches to guide sgRNA design.

In practice, there have been a number of application and toolkit applied in this task. In the earlier studies, methods in silico are categorized into three types: (1) alignment-based, (2) hypothesis-driven and (3) learning-based (Chuai *et al.*, 2018). Recently we noticed that the last type of method seems to be getting more attention because of huger and huger data set (Liu *et al.*, 2019a).

Learning-based method, which designed to predict the activity of sgRNAs, is essentially a computational model built by machine learning algorithm, not only conventional machine learning but also deep learning algorithm. Some studies on HT_ABE and HT_CBE have shown that deep learning based models often outperformed conventional machine learning, when the number of sgRNAs in the data set reached a certain level (Song *et al.*, 2020; Kim *et al.*, 2018, 2019). Per contra, conventional machine learning algorithms, such as linear regression, logistic regression and the decision tree, are often more interpretable due to the fewer parameters and clearer mathematical assumptions. In short, what was needed for developer is to trade-off accuracy and interpretability. Muhammad Rafidet al. consider deep-learning-based models as black boxs and believe they lack interpretability, motivated by the empirical assertion, they turn to build a model based conventional machine learning to compete with state of the art deep learning models (Muhammad Rafid *et al.*, 2020). On the other hand, input perturbation based feature importance analysis become a preferred components to reveal the importance of features in deep learning models. Liu et al. use a sliding window of length 2 to extract dimeric as input and rank the position of dimeric by contribution to final output (Liu *et al.*, 2019b). One regret is that subject to the processing of the input sgRNA sequence, their analysis can not exactly on the nucleotide class. Further, SHAP, one of the most prominent of model explain techniques, has been widely used to understand the decision made by the model.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

Wang et al. develope DeepHF, a deep learning based model, and use Deep SHAP to revealed nucleotide contributions (Wang *et al.*, 2019). Deep SHAP is a compositional approximation of SHAP values since it is challenged to computate SHAP values exactly, especially for a complex deep neural networks (Lundberg and Lee, 2017). In our understanding, the method based on input perturbation often requires better generalization ability of the model (even for artificial ridiculous noise data). Moreover, recent work indicates that model explain techniques, which based post hoc explanations techniques and input perturbations, could be fooled to generate meaningless explanations instead of reflecting the underlying biases (Slack *et al.*, 2019), in other word, they could be unreliable and misleading, even on model with excellent performance. In light of the above, we believe it is essential to develope a model which can not only match deep learning based model in performance, but also be comparable to conventional machine learning algorithms in interpretability.

Deep neural network has shown its power in the study of CRISPR/Cas9 and its improved Systems (Liu *et al.*, 2019a). Most of the deep neural network existing are the combination of recurrent neural network (RNN), convolutional neural network (CNN), fully connected neural network(FNN), and their variants. We found that the deep learning models used in sgRNA on-target activity (even for off-target effect) prediction tasks in recent years can be divided into the following two categories according to the encoding approach of the sgRNA sequence (sgRNA-DNA sequence pair, for off-target effect prediction):

(1) Methods in spatial domain. Some previous studies have used model based CNN to predict gRNA on-target activity or off-target effects (Lin and Wong, 2018; Kim *et al.*, 2018; Chuai *et al.*, 2018). They process sgRNA base sequence inputs with the help of one-hot encoding idea. In other words, they regard it as two-dimensional image data, and use convolution layer to extracte potential features in spatial domain, It is worth noting that Zhang et al. adds bidirectional gated recurrent unit (BGRU, in short), a RNN variant, after pooling layer of classic CNN network (Zhang *et al.*, 2020). Our explanation is that BGRU assists CNN to extract spatial features in one dimension, under this belief it belong to this category.

(2) Methods in temporal domain. Although RNN-based network have been shown effective to improve the performance of the model with temporal sequential input, especially in natural language processing and sequential recommendation(Huang *et al.*, 2018), RNN be not used for gRNA activity prediction, until recently (Wang *et al.*, 2019; Liu *et al.*, 2020, 2019b). They consider the nucleotides(can also dimer or polymer) in the sgRNA sequence as word, and the sgRNA sequence itself as a sentence (from 5' to 3'), then a trainable matrix (could be either supervised or unsupervised) is used to project the word to the dense real-valued space. This technology is called embedding, which generates the base embedding. RNN further encoding the base embedding into a sequence of hidden state vector. Specially, Liu et al. use RNN and CNN in parallel to extract features in base embedding (Liu *et al.*, 2019b). However, base embedding is not spatially interpretable (different from one-hot encode), and they have no way to further explore the correlation between CNN and RNN output. Almost all RNN based models used in sgRNA on-target activity or off-target effect flatten the hidden state vector into a one-dimensional vector as the input of the fully connected layer. It is a pity that the temporal sequential dependency of hidden state vector are rarely noticed. To summarise, RNN has limited representation power in capturing spatial feature. Furthermore, hidden state vector representation is usually hard to understand and explain.

2 Approach

Equation (1) Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text .....

…… text follows.

$$\sum x + y = Z \tag{1}$$

3 Methods

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text

- for bulleted list, use itemize
- for bulleted list, use itemize
- for bulleted list, use itemize

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method

Table 1. This is table caption

head1	head2	head3	head4
row1	row1	row1	row1
row2	row2	row2	row2
row3	row3	row3	row3
row4	row4	row4	row4

This is a footnote

Fig. 1. Caption, caption.

Text Text Text Text Text Text Text Text Text Text Text Text.

3.1 This is subheading

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text

3.1.1 This is subsubheading

Text Text Text Text Text Text Text Text Text Text Text Text Text

Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text

3.2 Test1

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text

4 Discussion

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text.

Table 1 shows that Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text. Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text. Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text.

5 Conclusion

(Table 1) Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text

Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text

1. this is item, use enumerate
2. this is item, use enumerate
3. this is item, use enumerate

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above

method Text Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bauer *et al.*, 2007 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text

Acknowledgements

Text Text Text Text Text Text Text Text. Jinek *et al.*, 2012 might want to know about text text text text

Funding

This work has been supported by theLiu *et al.* (2019b)... Text Text Text Text.

References

Bauer, M., Klau, G. W., and Reinert, K. (2007). Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. *BMC Bioinformatics*, **8**, 271.

Chuai, G., Ma, H., Yan, J., Chen, M., Hong, N., Xue, D., Zhou, C., Zhu, C., Chen, K., Duan, B., *et al.* (2018). Deepcrispr : optimized crispr guide rna design by deep learning. *Genome Biology*, **19**(1), 1–18.

Cong, L., Ran, F. A., Cox, D. D., Lin, S., Barretto, R. P. J., Habib, N., Hsu, P., Wu, X., Jiang, W., Marraffini, L. A., *et al.* (2013). Multiplex genome engineering using crispr/cas systems. *Science*, **339**(6121), 819–823.

Huang, J., Zhao, W. X., Dou, H., Wen, J., and Chang, E. Y. (2018). Improving sequential recommendation with knowledge-enhanced memory networks. pages 505–514.

Ishii, T. (2017). Reproductive medicine involving genome editing: clinical uncertainties and embryological needs. *Reproductive Biomedicine Online*, **34**(1), 27–31.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-rna-guided dna endonuclease in adaptive bacterial immunity. *Science*, **337**(6096), 816–821.

Kang, X., He, W., Huang, Y., Yu, Q., Chen, Y., Gao, X., Sun, X., and Fan, Y. (2016). Introducing precise genetic modifications into human 3pn embryos by crispr/cas-mediated genome editing. *Journal of Assisted Reproduction and Genetics*, **33**(5), 581–588.

Kim, H. K., Min, S., Song, M., Jung, S., Choi, J. W., Kim, Y., Lee, S., Yoon, S., and Kim, H. (2018). Deep learning improves prediction of crispr-cpf1 guide rna activity. *Nature Biotechnology*, **36**(3), 239–241.

Kim, H. K., Kim, Y., Lee, S., Min, S., Bae, J. Y., Choi, J. W., Park, J., Jung, D., Yoon, S., and Kim, H. (2019). Spcas9 activity prediction by deepspcas9, a deep learning–based model with high generalization performance. *Science Advances*, **5**(11).

Kleinstiver, B. P., Pattanayak, V., Prew, M. S., Tsai, S. Q., Nguyen, N. T., Zheng, Z., and Joung, J. K. (2016). High-fidelity crispr–cas9 nucleases with no detectable genome-wide off-target effects. *Nature*, **529**(7587), 490–495.

Liang, P., Xu, Y., Zhang, X., Ding, C., Huang, R., Zhang, Z., Lv, J., Xie, X., Chen, Y., Li, Y., *et al.* (2015). Crispr/cas9-mediated gene editing in human tripronuclear zygotes. *Protein & Cell*, **6**(5), 363–372.

Lin, J. and Wong, K. (2018). Off-target predictions in crispr-cas9 gene editing using deep learning. *Bioinformatics*, **34**(17).

Liu, G., Zhang, Y., and Zhang, T. (2019a). Computational approaches for effective crispr guide rna design and evaluation. *Computational and structural biotechnology journal*, **18**, 35–44.

Liu, Q., He, D., and Xie, L. (2019b). Prediction of off-target specificity and cell-specific fitness of crispr-cas system using attention boosted deep learning and network-based gene feature. *PLoS computational biology*, **15**(10), e1007480–e1007480. 31658261[pmid].

Liu, Q., Cheng, X., Liu, G., Li, B., and Liu, X. (2020). Deep learning improves the ability of sgrna off-target propensity prediction. *BMC Bioinformatics*, **21**(1), 51.

Lundberg, S. and Lee, S. (2017). A unified approach to interpreting model predictions. pages 4768–4777.

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., Dicarlo, J. E., Norville, J. E., and Church, G. M. (2013). Rna-guided human genome engineering via cas9. *Science*, **339**(6121), 823–826.

Muhammad Rafid, A. H., Toufikuzzaman, M., Rahman, M. S., and Rahman, M. S. (2020). Crisprpred(seq): a sequence-based method for sgrna on target activity prediction using traditional machine learning. *BMC Bioinformatics*, **21**(1), 223.

Rubeis, G. and Steger, F. (2018). Risks and benefits of human germline genome editing: An ethical analysis. *Asian Bioethics Review*, **10**(2), 133–141.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2019). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. *arXiv: Learning*.

Slaymaker, I. M., Gao, L., Zetsche, B., Scott, D. A., Yan, W. X., and Zhang, F. (2016). Rationally engineered cas9 nucleases with improved specificity. *Science*, **351**(6268), 84–88.

Song, M., Kim, H. K., Lee, S., Kim, Y., Seo, S.-Y., Park, J., Choi, J. W., Jang, H., Shin, J. H., Min, S., Quan, Z., Kim, J. H., Kang, H. C., Yoon, S., and Kim, H. H. (2020). Sequence-specific prediction of the efficiencies of adenine and cytosine base editors. *Nature Biotechnology*.

Wang, D., Zhang, C., Wang, B., Li, B., Wang, Q., Liu, D., Wang, H., Zhou, Y., Shi, L., Lan, F., *et al.* (2019). Optimized crispr guide rna design for two high-fidelity cas9 variants by deep learning. *Nature Communications*, **10**(1), 4284–4284.

Zhang, G., Dai, Z., and Dai, X. (2020). C-rnncrispr: Prediction of crispr/cas9 sgrna activity using convolutional and recurrent neural networks. *Computational and structural biotechnology journal*, **18**, 344–354.