Poster No. 1000

An Optimized Convolutional Neural Network Framework for Accurate Prediction of CRISPR/Cas9 Repair Outcomes in Primary Human T Cells

Hien Quang Kha[1,2], Minh Huu Nhat Le[1,2], Phat Ky Nguyen[1,2], Uyen Khoi Minh Huynh[4], and Nguyen Quoc Khanh Le[2,3]

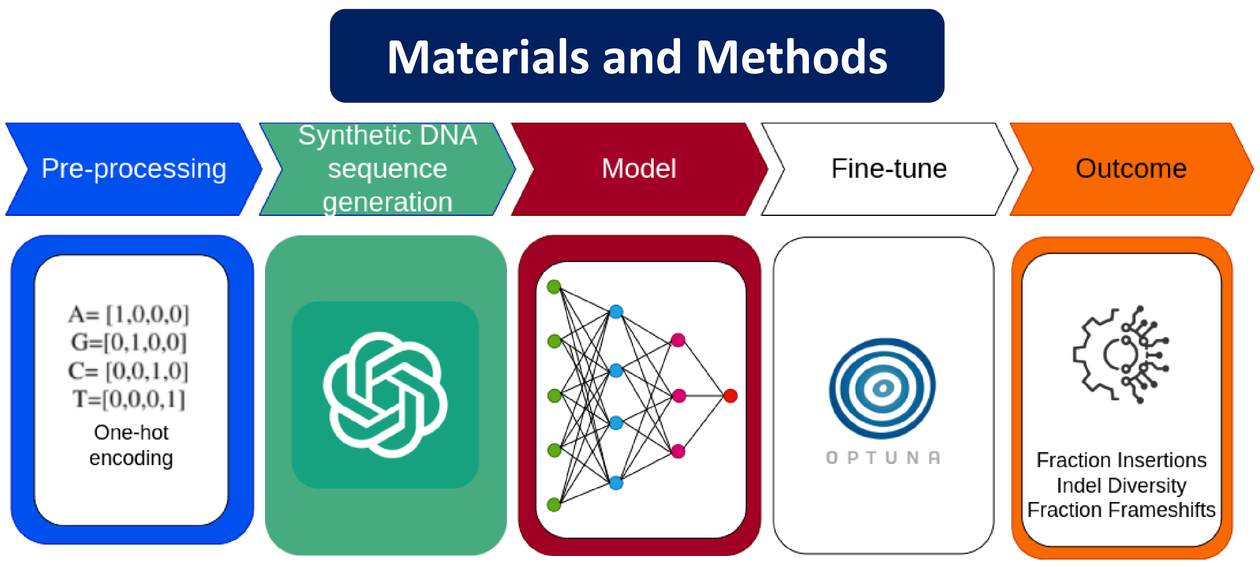
[1]International Master/Ph.D. Program in Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
[2]AIBioMed Research Group, Taipei Medical University, Taipei, Taiwan
[3]In-Service Master Program in Artificial Intelligence in Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan
[4]Faculty of Biological Sciences, Nong Lam University, Ho Chi Minh City, Vietnam

Introduction

Recent machine learning models, such as inDelphi [1], FORECasT [2], SPROUT [3], and CROTON [4], predicted CRISPR/Cas9 repair outcomes based on DNA repair data, but often underperform with primary human T cells, particularly in predicting indel diversity and fraction frameshift outcomes. This limitation is primarily attributed to insufficient training data and a lack of robust hyperparameter optimization.

In this work, we introduce **CRISPRON**, a novel framework that combines *GPT-Neo-1.3B* [5] for synthetic data generation and *Optuna* [6] for hyperparameter optimization, significantly improving CRISPR/Cas9 repair outcome prediction in primary human T cells.

Materials and Methods



The SPROUT dataset [3] contains CRISPR/Cas9 repair outcomes from 18 CD4+ T cell donors across 1,656 genomic sites. This study focuses on predicting fraction insertion, indel diversity, and frameshift outcomes.

DNA sequences were one-hot encoded into 3D arrays, and targets were min-max normalized. Data was randomly split into 80% training, 10% validation, and 10% testing.

GPT-Neo-1.3B [5] was used to generate 200 synthetic 23-nucleotide sequences using a defined prompt "Generate a synthetic DNA sequence with exactly 23 nucleotides (A, C, G, T)." These were preprocessed like real data, with targets randomly sampled from real outcome ranges, and added only to the training set.

The CNN consists of a 1D conv layer (64 filters, kernel size 3), max pooling, and a dense layer with ReLU, L2 regularization, and 0.2 dropout. A linear output layer handles regression. Training used Adam with MSE loss, early stopping, and learning rate reduction for better generalization.

We used *Optuna* [6] to tune the CNN hyperparameters over 70 trials with 10-fold cross-validation (CV). Parameters included filter size, kernel size, dropout rate, dense units, learning rate, and batch size.

References

[1] Shen et al. (2018). *Nature*, 563(7733):646–651.

[2] Allen et al. (2019). *Nature Biotechnology*, 37(1):64–72.

[3] Leenay et al. (2019). *Nature Biotechnology,* 37(9):1034–1037.

[4] Li et al. (2021). *Bioinformatics*, 37(Supplement_1):i342–i348.

[5] S. Black, L. Gao, et al. (2021)

[6] Bergstra et al. (2021). *Advances inneural information processing systems*, 24.

Results

Ablation studies: The best results are achieved by integrating CNN, GPT-Neo-1.3B, and Optuna into one framework, with fraction insertions, indel diversity, and fraction frameshifts reaching the R^2 scores of 0.7235, 0.3009, and 0.1583, respectively (Table 1).

Comparative results: Leveraging synthetic DNA sequences generated by GPT-Neo-1.3B and fine-tuned using Optuna, **CRISPRON** outperforms existing models in predicting fraction insertions, indel diversity, and fraction frameshifts (Table 2).

Table 1: Performance of **CRISPRON** across different settings

Outcomes	CNN	GPT-Neo-1.3B	Optuna	MSE	R^2	$KTau$	ACC
Fraction insertions	✓			0.0086	0.6680	0.6281	0.8170
	✓	✓		0.0083	0.6808	0.6352	0.8105
	✓	✓	✓	**0.0071**	**0.7235**	**0.6789**	**0.8431**
Indel diversity	✓			0.4350	0.2672	0.3915	0.6667
	✓	✓		0.4293	0.2768	0.3932	0.6536
	✓	✓	✓	**0.4150**	**0.3009**	**0.4207**	**0.7059**
Fraction frameshifts	✓			0.0185	0.1137	0.2570	0.5817
	✓	✓		0.0183	0.1238	0.2626	0.5948
	✓	✓	✓	**0.0175**	**0.1583**	**0.3099**	**0.6667**

Table 2: Performance comparison of **CRISPRON** and other CRISPR/Cas9 repair outcome prediction models.

	Outcomes	Fraction insertions			Indel diversity	Fraction frameshifts
Model	**Methods**	$KTau$	$Pearson's R$	R^2	ACC	ACC
inDelphi (2018) [8]	Deep Neural Network k-Nearest Neighbor	0.3400	_	-4.3	_	0.6000
FORECasT (2019) [9]	Multi-Class Logistic Regression	0.3000	_	0.34	_	0.5200
SPROUT (2019) [10]	Gradient Boosting Decision Tree	0.6200	0.7700	0.59	0.6800	0.6600
CROTON (2021) [11]	CNN NAS	0.6522	0.8112	_	_	_
Apindel (2022) [18]	GloVe Positional Encoding BiLSTM Attention	0.6300	0.8000	_	_	_
CRISPRON (Ours)	CNN GPT-Neo-1.3B Optuna	**0.6550**	**0.8561**	**0.7231**	**0.7059**	**0.6667**

Conclusion

We propose **CRISPRON**, a CNN-based framework enhanced by *GPT-Neo-1.3B*–generated synthetic DNA and *Optuna* tuning for predicting CRISPR/Cas9 repair outcomes. CRISPRON outperformed existing models, especially in the fraction insertion predicting task. Future work will explore larger datasets, better augmentation, and improved architectures.

Contact us: http://aibiomed.tmu.edu.tw/