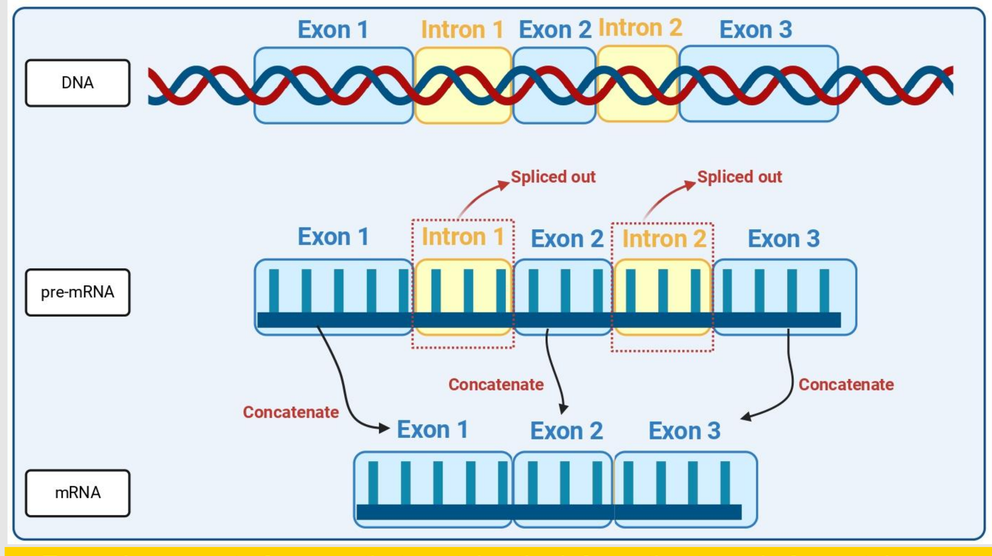
CENTER FOR BIOSYSTEMS AND BIOTECH DATA SCIENCE, GHENT UNIVERSITY GLOBAL CAMPUS
Espoir Kabanga, Jiwon Koh, Youngmin Koh, Kangmin Kim, Nina Geiregat, Arnout Van Messem, Wesley De Neve

GENERATING AND EVALUATING SYNTHETIC DNA SEQUENCES WITH SPLICE SITES USING GENERATIVE ADVERSARIAL NETWORKS

Abstract

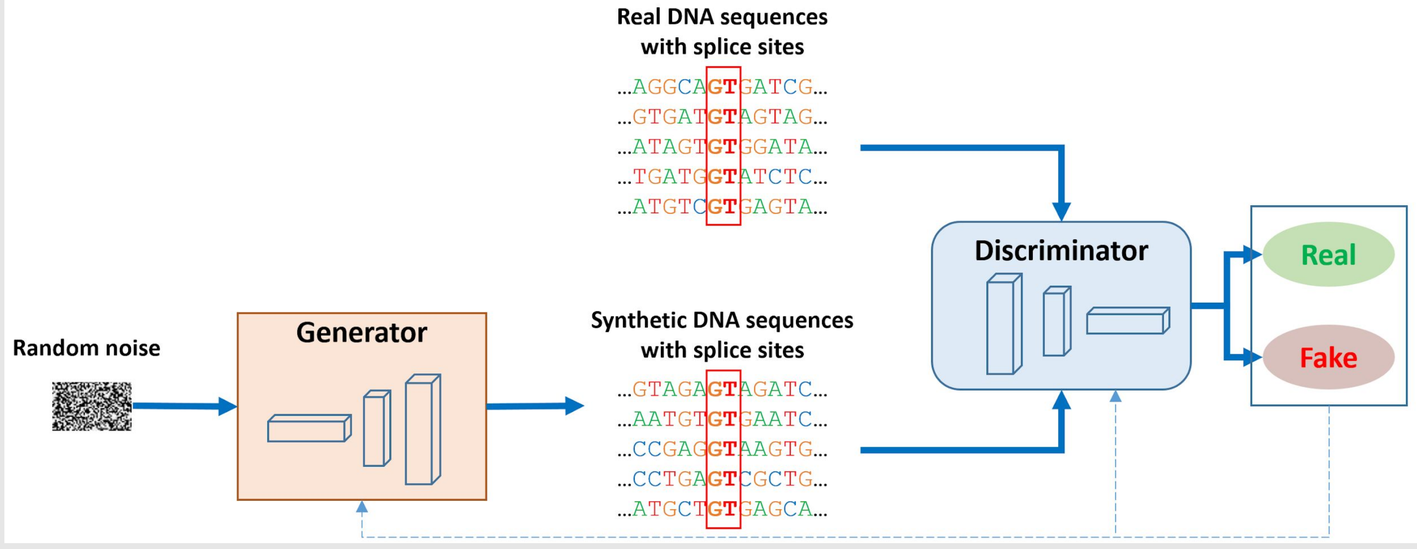
We present a GAN-based approach to generate synthetic DNA sequences with splice sites, evaluated using direct and indirect strategies. Our study shows that GANs can effectively generate synthetic DNA sequences that closely mimic real DNA sequences, with similar nucleotide distributions and patterns. However, models trained on synthetic sequences still exhibit lower effectiveness when tested on real sequences, highlighting the need for further refinement to improve the robustness of our GAN model.

Splicing mechanism

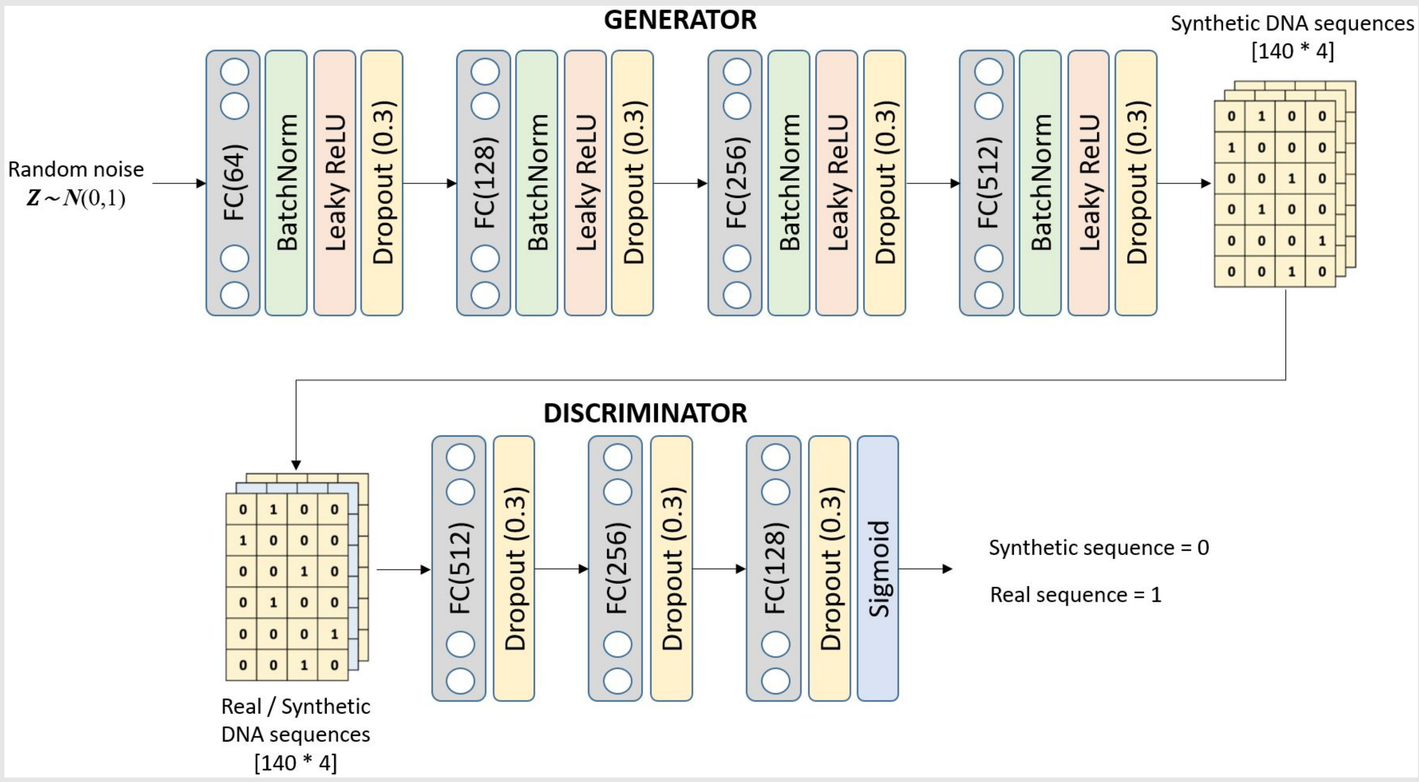


- **Exon**: A segment of a gene that contains the information to code for proteins; exons are spliced together during RNA processing to form mature mRNA
- **Intron**: A non-coding segment of a gene that is removed during RNA processing; introns are not involved in coding for proteins
- **Pre-mRNA**: The initial RNA transcript produced from a gene, which contains both exons and introns; it is processed to form mature mRNA
- **mRNA (messenger RNA)**: The final processed RNA molecule that carries the genetic information from the DNA to the ribosome, where proteins are synthesized

Basic GAN structure



Generator and discriminator architectures



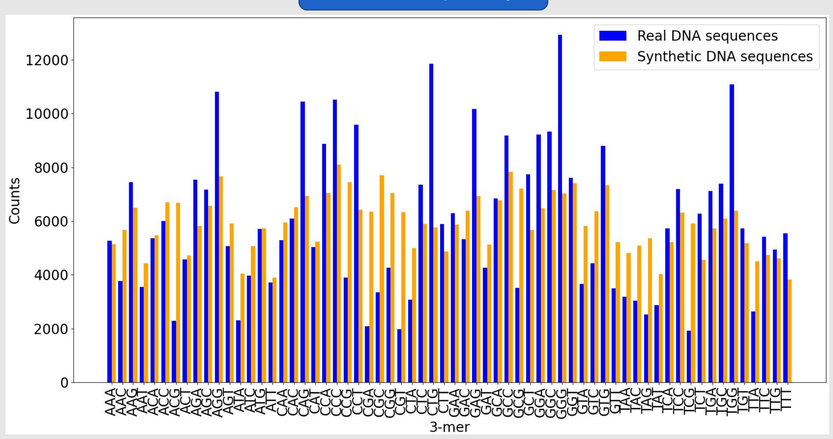
Data description

- The dataset is obtained from the Homo Sapiens Splice Sites Dataset (HS^3D)
- The dataset includes a positive set (DNA sequences with ture splice sites) and a negative set (DNA sequences without true splice sites)
- The dataset is balanced: 2,750 sequences for both the positive set and the negative set
- Each sequence has a length of 140 bp
- The splice sites are located at position 70 in each sequence

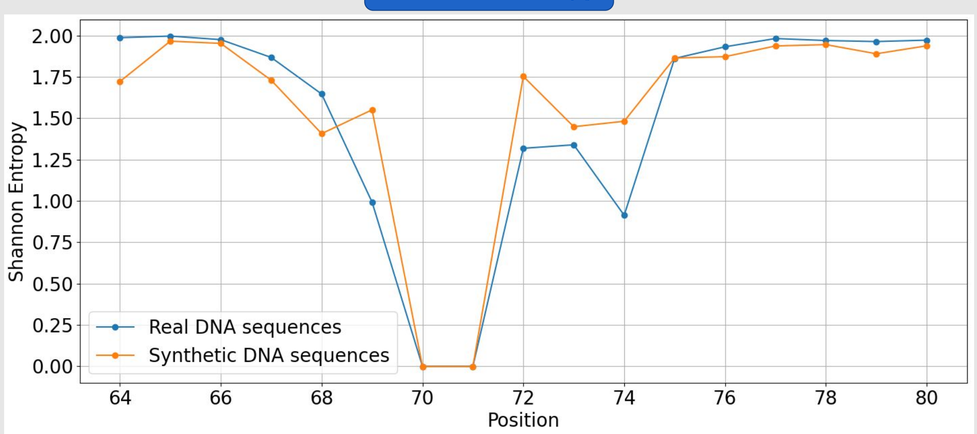
Evaluation techniques and results

Direct evaluation: Measurement of the fidelity (how accurate the GAN model captures important features of real data) and diversity (how effectively the GAN model represents the variety of real data).

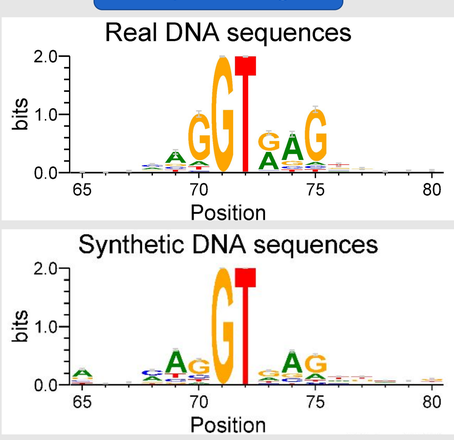
3-mer frequency



Shannon's entropy



Sequence logo



Indirect evaluation: Use of a splice site prediction model to indirectly evaluate the quality of the synthetic DNA sequences. We used the DeepSplicer[1] model.

	TRTR	TRTS	TSTS	TSTR
Recall	0.9384	0.6982	0.9855	0.3514
Precision	0.9453	0.9897	0.9963	0.9898
F1-score	0.9418	0.8188	0.9909	0.5187
MCC	0.8966	0.7233	0.9819	0.4551

- **TRTR** : Training and testing on real DNA sequences
- **TRTS** : Training on real DNA sequences and testing on synthetic DNA sequences
- **TSTS** : Training and testing on synthetic DNA sequences
- **TSTR** : Training on synthetic DNA sequences and testing on real DNA sequences

Conclusions and future research

- GANs effectively generate synthetic DNA sequences with splice sites, as shown by both direct and indirect evaluations using the DeepSplicer model
- Synthetic DNA sequences closely mimic real DNA sequences, demonstrating similar nucleotide distributions and patterns
- Despite the similarities, models trained on synthetic DNA sequences perform less effectively when tested on real DNA sequences
- The results indicate that while GANs are able to capture key characteristics of real DNA sequences, they may not yet fully replicate their complexity
- Further refinement of the GAN model is needed to improve its robustness and effectiveness

References

[1] V. Akpokiro, O. Oluwadare, J. Kalita, DeepSplicer: An Improved Method of Splice Sites Prediction using Deep Learning, 20th IEEE International Conference on Machine Learning and Applications, 2021

Contact: espoir.kabanga@ghent.ac.kr

GHENT UNIVERSITY GLOBAL CAMPUS
Center for Biosystems & Biotech Data Science
FACULTY OF ENGINEERING AND ARCHITECTURE
LIÈGE université Mathématique