Interpretable prediction of DNA replication origins in *S.cerevisiae* using attention-based motif discovery

Zohreh Piroozeh[1,2], Ildem Akerman[5], Stefan Kesselheim[1,3], Olga Kalinina[2,4], Alina Bazarova[1,3]

[1] Forschungszentrum Jülich, JSC, Jülich, Germany, [2] Saarland University, ZBI, Saarbrücken, Germany, [3] Helmholtz AI, Munich, Germany

[4] Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Saarbrücken, Germany,

[5] University of Birmingham, Institute of Biomedical Research, Birmingham, United Kingdom

INTRODUCTION

DNA replication: Biological process of producing two identical replicas of DNA from one original DNA molecule.
DNA replication origins: Genomic locations initiating replication.

S.Cerevisiae Genome:
- Many Origins
- Known Locations
- Identified Consensus Sequences

ARS
- Autonomously Replicating Sequence in the *S.Cerevisiae* genome
- 200-1000 bp

ACS
- ARS Consensus Sequences
- Recognized by the Origin Recognition Complex (ORC)
- 12-17 bp

5'-WWW-WTTTAYRTTTW-GTT-3'

(Theis and Newlon, 1997 [1])
'W' denotes A or T, 'Y' denotes C or T, 'R' denotes G or A.

Problem:
- How to discriminate origin sequences from non-origins the by Genome LLMs?
- How to interpret the results?
- Is there anything on top of ACS which determines the replication origin?

METHOD

DNABERT: A multi-layer Transformers encoder following the same training process as BERT [2].
- Pre-trained on Human Genome data.
- Overlapping k-mers tokenization.

- Four balanced datasets are curated based on OriDB [3]
- Confirmed Origins < 500 bp, as positive instances.
- Negative instances (non-origins) subsampled by multiple levels of complexity as follow:

Subsample randomly from the genome, 500 bp instances, non-overlapping positive ranges

Random-Neg (OriDB)

Select non-replicating ACS motifs, extend them to 500 bp, non-overlapping positive ranges

ACS-Neg (OriDB)

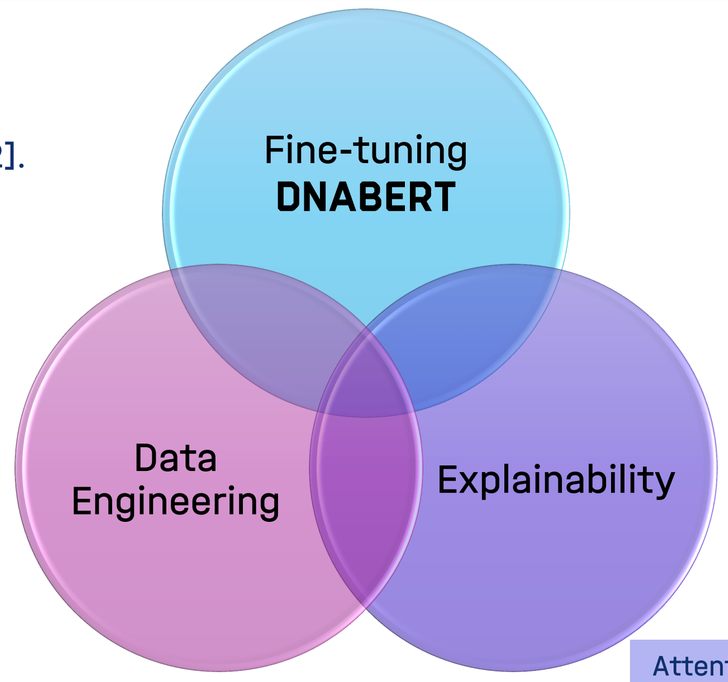
ACS-Neg dataset: *presence of ACS matches in both origin and non-origins, making discrimination more challenging*
The aim: *identifying discriminative factors beyond ACS motifs.*

*Shuffling randomly origin sequences to generate negative samples, breaking both **local** and **global** patterns*

Shuffled-Neg (OriDB)

*Shuffling randomly blocks of 5 bp of origin sequences to generate negative samples, disrupting only the **global** patterns*

Block-5-Shuffled-Neg (OriDB)



The aim of subsampling non-origin instances by shuffling origins :
✓ *Assessing importance of the nucleotide order for model to identification origins.*

RESULTS

Performance Analysis

DATASET	ACC	AUC
Random-Neg	0.83	0.90
ACS-Neg	0.74	0.82
Shuffled-Neg	0.90	0.96
Block-5-Shuffled-Neg	0.77	0.86

ACS pattern is an important feature for model

Ability of model in capturing both local and global sequence dependencies

Motifs Discovered from High-attention Fragments, by MEME

Rand-Neg dataset

Test set
True Positives:

'TTTTTWTTTATRTTT' (E-value=2.6e-006)

Train set
True Positives:

'TATATTTATRTWTWT' (E-value=2.3e-032)

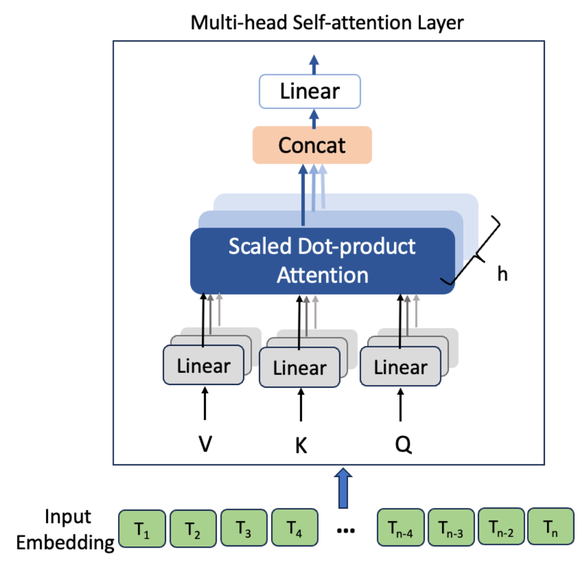
Closely resemble the ACS pattern

ACS-Neg dataset:
Train set
True Positives

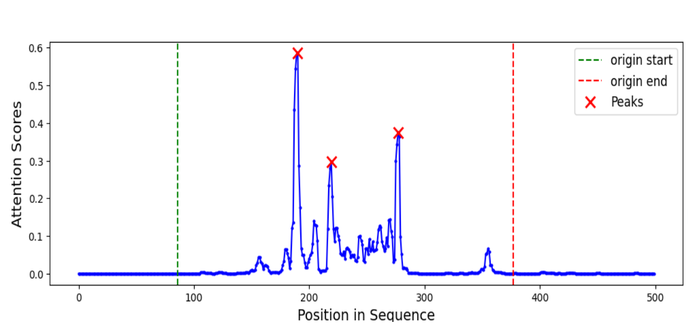
'ATATATATRTR' (E-value=5.7e-059)

A/T bases, the pattern associated with the intergenic origins [5]

Attention-based Motif discovery



A



B

MEME motif Discovery [4]

***A**: Attention scores being extracted from the multi-head attention layer for each sequence,*

***B**: Motif discovery targets are fragments of 20 bps around peaks (red crosses) of the attention scores.*

CONCLUSION:

- We developed an interpretable pipeline that integrates DNABERT attention maps with the MEME motif discovery tool.
- Motifs identified from high-attention regions validate DNABERT's ability in capturing biologically relevant sequence features through its attention mechanism.

REFERENCES

[1] J. F. Theis and C. S. Newlon. The ars309 chromosomal replicator of Saccharomyces cerevisiae depends on an exceptional ARS consensus sequence. Proceedings of the National Academy of Sciences of the United States of America, 94(20):10786–10791, 1997.
[2] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. Bioinformatics,37(15):2112–2120, aug 2021.
[3] Cheuk C. Siow, Sian R. Nieduszynska, Carolin A. MÅNuller, and Conrad A. Nieduszynski. Oridb, the dna replication origin database updated and extended. Nucleic Acids Research, 40(D1):D682–D686, 2012.
[4] Timothy L. Bailey and Charles Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28–36, Menlo Park, California, 1994. AAAI Press.
[5] D. Wang and F. Gao. Comprehensive analysis of replication origins in Saccharomyces cerevisiae genomes. Frontiers in Microbiology, 10:2122, 2019.