INTERPRETABLE PREDICTION OF DNA REPLICATION ORIGINS IN *S. cerevisiae* USING ATTENTION-BASED MOTIF DISCOVERY

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

[1]Forschungszentrum Jülich, JSC, Jülich, Germany
[2]Helmholtz AI, Munich, Germany
[3]Saarland University, ZBI, Saarbrücken, Germany
[4]University of Birmingham, Institute of Biomedical Research, Birmingham, United Kingdom

ABSTRACT

In a living cell, DNA replication begins at multiple genomic sites, called replication origins. Identifying these origins and their underlying base sequence composition is crucial for understanding replication process. Existing machine learning methods for origin prediction often require labor-intensive feature engineering or lack interpretability. Here, we employ DNABERT to predict yeast replication origins and uncover sequence motifs by combining attention maps with MEME, a classical bioinformatics tool. Our approach eliminates manual feature extraction and identifies biologically relevant motifs across datasets of varying complexity. This work advances interpretable machine learning in genomics, offering a potentially generalizable framework for origin prediction and motif discovery.

1 INTRODUCTION

DNA replication is a biological process of producing two identical copies of DNA from one original DNA molecule. It is essential for all living organisms, ensuring accurate transmission of the genetic material from one generation to the next. A thorough understanding of this process is crucial, since any impairment in the DNA replication may lead to organism dysfunction and disease (Schmit & Bielinsky, 2021). In eukaryotes, DNA replication begins at multiple genomic sites called replication origins (ORIs), distributed across the genome. Identifying their locations and sequence base composition behind them is key to understanding DNA replication mechanisms.

Although in most eukaryotes, the underlying origin motif sequence remains elusive, there are hundreds of well-characterized replication origins in *S. cerevisiae* (budding yeast) with known locations and consensus sequences (Ekundayo & Bleichert, 2019), referred to as autonomously replicating sequences (ARSs).

The ARS consensus sequence (ACS) was first identified by Broach et al. (1983) through the alignment of known essential elements and later refined by Marahrens & Stillman (1992). An extended representation as a 17-basepair motif, WWW-WTTTAYRTTTW-GTT[1] , was proposed by Theis & Newlon (1997). However, not all experimentally identified origins contain an ACS, and fewer than 5% of ACS matches in the budding yeast genome correspond to actual replication origins, as noted by Siow et al. (2012). Thus, the presence of ACS is neither sufficient nor necessary for determining origin locations, leaving this as an unresolved issue and presenting a compelling challenge for machine learning (ML) techniques.

Current ML approaches for identifying ORIs often depend on manually extracted features. For example, Do & Le (2020) uses XGBoost (Chen & Guestrin, 2016) on a hybrid feature set combining

*Correspondence to z.piroozeh,al.bazarova@fz-juelich.de
[1]Here, 'W' denotes A or T, 'Y' denotes C or T, and 'R' denotes G or A.

biological and sequence-based properties, improving accuracy over a number of methods (Chen et al., 2012; Li et al., 2015; Dao et al., 2018) but requiring intensive feature selection. Similarly, yORIpred (Manavalan et al., 2021) employs multiple classifiers and feature encoding techniques, achieving high accuracy but relying on manual intervention, high-dimensional feature engineering, and costly iterative steps. To address the limitations of the previous research, Wu et al. (2021) developed a convolutional neural network incorporating sequence segmentation and the Word2Vec representations for ORI identification. The proposed method achieved an impressive accuracy of 97% for *S. cerevisiae* but required training of the model from scratch, making it computationally expensive and negatively affecting the model's interpretability, the latter being of crucial importance for DNA replication origin motif identification.

In this study, we aim to tackle the problem of identifying budding yeast replication origins using large language models (LLMs). By leveraging pre-training, LLMs automatically capture complex genomic patterns, eliminating the need for extensive feature engineering and manual extraction. The attention mechanism further enables the recognition of important sequence dependencies and patterns. Recently, genome-based pre-trained LLMs have gained popularity, allowing fine-tuning for a wide range of downstream genomic tasks.

Here, we fine-tune the pre-trained DNABERT model (Ji et al., 2021) to predict replication origins in budding yeast and develop a comprehensive pipeline for identifying the underlying sequence motifs using attention maps and bioinformatics post-processing. To ensure robustness, we employ a data engineering strategy, evaluating the model's performance across datasets of varying complexity.

2 Materials and Methods

2.1 Datasets

OriDB (Siow et al., 2012) offers highly curated, experimentally validated data, ideal for benchmarking across studies. For our training dataset, we selected 325 confirmed ORIs no longer than 500 bp as positive instances, meeting DNABERT's input size requirements. To standardize the lengths, we randomly asymmetrically extended shorter sequences to 500 bp.

Our earlier efforts to identify ORIs using negative instances from the whole genome with DeepGRN (Chen et al., 2021) yielded a low true positive (TP) rate, primarily due to a significant class imbalance (data not shown). To address this issue, we curated four balanced datasets of varying complexity, designed to answer specific research questions regarding the ORI base composition, described as follows:

- **Random-Neg dataset**: 325 negative instances were randomly selected from regions of the genome that do not intersect with any of the 325 positive instances.
- **ACS-Neg dataset**: Only the positive instances containing ACS were considered (298 overall). The same number of negative instances were subsampled from 11500 ACS matches found by Homer (Heinz et al., 2010) which were not in proximity of any ORI. We thereby aimed to identify discriminative factors beyond ACS motifs.
- **Shuffled-Neg Dataset**: 325 negative instances were created by randomly shuffling the positive instances, thereby breaking both local and global patterns within the positive sequences. Here, we assess the importance of the base order for ORI identification, while keeping the base frequency the same across positive and negative instances.
- **Block-5-Shuffled-Neg Dataset**: 325 negative instances were created by splitting the positive instances into the blocks of five bp and subsequently shuffling these blocks, thereby disrupting only the global patterns within the positive instances, while the local ones stayed preserved. This way we assess the relative importance of short- versus long-range sequence features in origin recognition.

2.2 Methods

For ORI classification, we selected DNABERT, an early genome-focused LLM adapted from the original BERT architecture for genomic tasks and pre-trained on human genome using overlapping k-mer tokenization.

While larger DNA LLMs are available (Dalla-Torre et al., 2025; Nguyen et al., 2024), we opted for DNABERT due to the straightforward process for extracting attention maps, often regarded as a tool to focus on the most relevant input components, potentially capturing meaningful correlations that explain model predictions (Vaswani et al., 2017). This is of crucial importance, as the main focus of our study is the explainability of the results, rather than the performance analysis. We used the DNABERT model pre-trained on the overlapping 4-mers. Notably, varying the k-mer length did not result in any significant difference in model performance.

DNABERT's built-in motif analysis tool, effective for transcription factor binding sites (Ji et al., 2021), proved unsuitable for identifying ORI motifs due to their longer length. While it detected motifs resembling ACS or A/T-rich flanking regions, the results lacked robustness in p-value and motif length. This limitation likely stems from its reliance on continuous high-attention regions as the initial search space. Since the model does not consistently assign high attention across entire continuous regions, significant portions of motifs may be missed by the algorithm. Furthermore, the tool's lack of advanced clustering strategies, such as hierarchical or multiple sequence alignment, resulted in fragmented and less comprehensive motif representations. To overcome these issues, we developed an alternative pipeline using high-attention sequence fragments and the bioinformatics tool MEME (Multiple Expectation Maximization for Motif Elicitation by Bailey & Elkan (1994)).

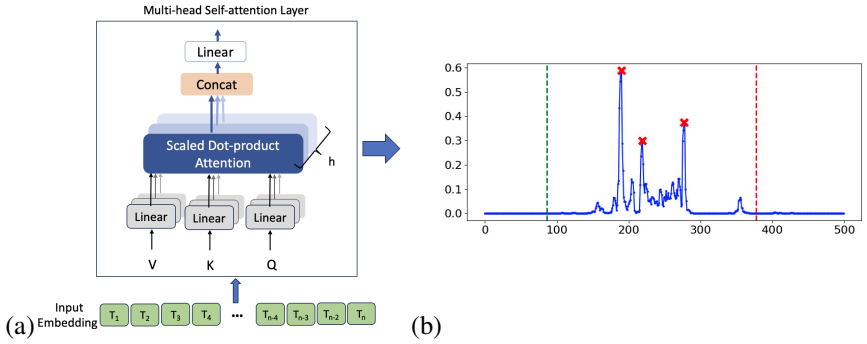

Figure 1: (a) **The schematic of the attention mechanism of DNABERT.** Attention scores being extracted from the multi-head attention layer, (b) **Attention scores across a positive origin instance**. Motif discovery targets are fragments of 20 bps around peaks (red crosses) of the attention scores (values along the Y axis). Vertical lines correspond to origin start (green) and end (red) with sequence position along the X axis.

We extracted the attention scores for the [CLS] token—which serves as a global representation of the input sequence—from the final hidden layer of our fine-tuned model (Figure 1). To obtain a single vector of attention scores per token and translate the attention scores from the token level to the resolution of individual nucleotides, we follow the approach proposed by Ji et al. (2021). Visualization of attention scores for origin and non-origin sequences revealed a recurring pattern, with sharp attention peaks specifically within the actual range of origin sequences. These peaks were identified as target regions for motif discovery, Figure 1b. Of note, is that the presence of more than one ACS within a positive instance has a biological meaning, corresponding to A and B2 elements of the ARS (Lee et al., 2023).

To enhance the robustness of attention-based motif discovery, we refined the search space by concentrating on 20 bp fragments around high-attention peaks. For each input sequence, we selected up to four top peaks based on an attention score threshold, ensuring that the total length of selected fragments stayed below 16% of the sequence. Only peaks separated by at least 10 bp were included, limiting the fragment overlap to a maximum of 10 bp. High-attention fragments were extracted from both train and test sets and independently processed using MEME. This approach was evaluated on the Rand-Neg and ACS-Neg datasets, using models fine-tuned for each, since they better represent real-world origin prediction with diverse non-origin samples. In contrast, non-origin samples in Shuffled-Neg and Block-5-Shuffled-Neg were artificially generated through shuffling. While performance analysis of DNABERT on these datasets is important for understanding model learning, the results of the motif analysis may be misleading, and therefore we omit them.

3 PERFORMANCE ANALYSIS

We evaluated DNABERT fine-tuned models on four datasets as described in section 2.1, employing seven distinct data splits (70% / 10% / 20% for train/val/test respectively) per dataset to ensure robustness and reliability. For each split, the model was fine-tuned independently, yielding seven fine-tunned models per dataset (training parameters were 400 epochs, using a batch size of 32 and a learning rate of 2e-4). Performance was assessed on the corresponding test sets for each split, and the average results across all splits are reported in Table 1.

Table 1: DNABERT Performance on datasets

Dataset	Accuracy	AUC	Precision	Recall
Random-Neg	0.83	0.90	0.83	0.83
ACS-Neg	0.74	0.82	0.74	0.74
Shuffled-Neg	0.90	0.96	0.90	0.90
Block-5-Shuffled-Neg	0.77	0.86	0.77	0.77

As anticipated, performance on the ACS-Neg dataset declined compared to the Random-Neg one. This is because origin and non-origin sequences in ACS-Neg share greater similarity due to the presence of ACS matches in both, making discrimination more challenging. This result confirms that DNABERT learns the ACS pattern as an important feature for classification. Nonetheless, the model achieved an accuracy of 0.74, indicating its ability to utilize additional discriminative features beyond the ACS matches.

DNABERT achieved its highest performance on the Shuffled-Neg dataset, with an accuracy of 0.90 and an AUC of 0.96. This indicates that ordered nucleotide patterns are critical for distinguishing origin sequences. Disrupting both local and global patterns in non-origin sequences creates a clearer boundary between classes, enhancing the model's discriminative ability. These results highlight the strength of BERT-style models with overlapping k-mer tokenization in capturing both local and global sequence dependencies, particularly for identifying ordered patterns in replication origins.

On the Block-5-Shuffled-Neg dataset, where local patterns are preserved but global ones are disrupted, performance declined compared to the Shuffled-Neg dataset. This suggests that long-range sequence patterns also play a significant role in DNABERT's ability to classify origin sequences effectively.

4 EXPLAINABILITY OF RESULTS

We applied our motif discovery pipeline to the train and test sets of the Rand-Neg and ACS-Neg datasets, and present the most significant motifs. High-attention fragments (20 bps) from TP cases were analyzed using MEME in the classic mode. For the Rand-Neg dataset, motifs from the test and train sets are shown in Figure 2 a, b, as 'TTTTTWTTTATRTTT' (E-value=2.6e-006) and 'TATATT-TATRTWTWT' (E-value=2.3e-032), respectively. These motifs closely resemble the ACS pattern, specifically the motif from test set shows a strong agreement with the experimentally confirmed ACS pattern 5'-WTTTATRTTTW-3' (Broach et al., 1983), underscoring its biological relevance. The difference stems from greater variability in the training set, resulting in broader motif representations than in the test set.

Figure 2: Motifs discovered from high-attention fragments, by MEME

For the ACS-Neg dataset, motif discovery on the test set yielded no significant results, likely due to the small sample size (<45 instances) and worse model performance. However, analysis of the train set revealed the motif 'ATATATATRTR' (E-value=5.7e-059, Figure 2c), characterized by alternating A/T bases, a pattern associated with the intergenic origins (Wang & Gao, 2019). These findings further validate DNABERT's ability to capture biologically relevant sequence patterns.

5 CONCLUSION

In this study, we present an interpretable pipeline combining DNABERT with the MEME motif discovery tool for identifying DNA replication origins, focusing on budding yeast due to the well-defined ACS motif, which serves as a gold standard for motif analysis. While our work centers on yeast, the principles of DNA replication initiation are highly conserved across eukaryotes (Wang & Gao, 2019), suggesting that the insights gained from *S. cerevisiae* and the developed pipeline could be potentially applied to higher eukaryotes as the logical next step.

ACKNOWLEDGMENTS

This work is supported by the Helmholtz Association Initiative and Networking Fund in the frame of Helmholtz AI.

REFERENCES

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.

J. R. Broach, Y. Y. Li, J. Feldman, M. Jayaram, J. Abraham, K. A. Nasmyth, and J. B. Hicks. Localization and sequence analysis of yeast origins of dna replication. *Cold Spring Harbor Symposia on Quantitative Biology*, 47(Pt 2):1165–1173, 1983. doi: 10.1101/sqb.1983.047.01.132. URL https://doi.org/10.1101/sqb.1983.047.01.132.

Chen Chen, Jie Hou, Xiaowen Shi, Hua Yang, James A. Birchler, and Jianlin Cheng. Deepgrn: prediction of transcription factor binding site across cell-types using attention-based deep neural networks. *BMC Bioinformatics*, 22(1):38, Feb 2021. ISSN 1471-2105. doi: 10.1186/s12859-020-03952-1. URL https://doi.org/10.1186/s12859-020-03952-1.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

Wei Chen, Pengmian Feng, and Hao Lin. Prediction of replication origins by calculating dna structural properties. *FEBS Letters*, 586(6):934–938, 2012. ISSN 0014-5793. doi: https://doi.org/10.1016/j.febslet.2012.02.034. URL https://www.sciencedirect.com/science/article/pii/S0014579312001573.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, Feb 2025. ISSN 1548-7105. doi: 10.1038/s41592-024-02523-z. URL https://doi.org/10.1038/s41592-024-02523-z.

Fu-Ying Dao, Hao Lv, Fang Wang, Chao-Qin Feng, Hui Ding, Wei Chen, and Hao Lin. Identify origin of replication in saccharomyces cerevisiae using two-step feature selection technique. *Bioinformatics*, 35(12):2075–2083, 11 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty943. URL https://doi.org/10.1093/bioinformatics/bty943.

Duyen Thi Do and Nguyen Quoc Khanh Le. Using extreme gradient boosting to identify origin of replication in *Saccharomyces cerevisiae* via hybrid features. *Genomics*, 112(3):2445–2451, 2020. ISSN 0888-7543.

B. Ekundayo and F. Bleichert. Origins of dna replication. *PLoS Genetics*, 15(9):e1008320, 2019.

Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and b cell identities. *Molecular Cell*, 38(4):576–589, 2010. doi: 10.1016/j.molcel.2010.05.004. URL https://doi.org/10.1016/j.molcel.2010.05.004.

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.

C. S. K. Lee, M. Weiß, and S. Hamperl. Where and when to start: Regulating dna replication origin activity in eukaryotic genomes. *Nucleus*, 14(1), 2023.

Wen-Chao Li, En-Ze Deng, Hui Ding, Wei Chen, and Hao Lin. iori-pseknc: A predictor for identifying origin of replication with pseudo k-tuple nucleotide composition. *Chemometrics and Intelligent Laboratory Systems*, 141:100–106, 2015. ISSN 0169-7439. doi: https://doi.org/10.1016/j.chemolab.2014.12.011. URL https://www.sciencedirect.com/science/article/pii/S0169743914002640.

Balachandran Manavalan, Shaherin Basith, Tae Hwan Shin, and Gwang Lee. Computational prediction of species-specific yeast dna replication origin via iterative feature representation. *Briefings in Bioinformatics*, 22(4), 2021.

Y. Marahrens and B. Stillman. A yeast chromosomal origin of dna replication defined by multiple functional elements. *Science (New York, N.Y.)*, 255(5046):817–823, 1992.

Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, Armin W. Thomas, Samuel H. King, Garyk Brixi, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D. Hsu, and Brian L. Hie. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336, 2024. doi: 10.1126/science.ado9336. URL https://www.science.org/doi/abs/10.1126/science.ado9336.

Megan Schmit and Anja-Katrin Bielinsky. Congenital diseases of dna replication: Clinical phenotypes and molecular mechanisms. *International Journal of Molecular Sciences*, 22(2), 2021. ISSN 1422-0067. doi: 10.3390/ijms22020911. URL https://www.mdpi.com/1422-0067/22/2/911.

Cheuk C. Siow, Sian R. Nieduszynska, Carolin A. Müller, and Conrad A. Nieduszynski. Oridb, the dna replication origin database updated and extended. *Nucleic Acids Research*, 40(D1):D682–D686, 2012.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Ilya Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010, 2017.

D. Wang and F. Gao. Comprehensive analysis of replication origins in *Saccharomyces cerevisiae* genomes. *Frontiers in Microbiology*, 10:2122, 2019.

F. Wu, R. Yang, C. Zhang, et al. A deep learning framework combined with word embedding to identify dna replication origins. *Scientific Reports*, 11:844, 2021.