DNABART: A Genomic LLM Foundational Model for Sequence Correction and Classification

Matthew Gaston
Department of Computer Science
Missouri University of Science and Technology
Rolla, Missouri
mpg8nm@mst.edu

Mukund Telukunta
Department of Computer Science
Missouri University of Science and Technology
Rolla, Missouri
mt3qb@mst.edu

***Abstract*—Advances in next-generation DNA sequencing have generated a vast number of genomic sequences for model organisms and beyond, yet effectively interpreting these data remains a challenge and critical biological problem. Existing computational models often fail to achieve high-accuracy classifications and require extensive training data, hindering broader implementation in diverse research settings.**

I. Introduction

DNA sequencing technologies have undergone a dramatic transformation since the inception of the Human Genome Project, evolving from a billion-dollar endeavor to one that can now be completed for a few hundred dollars [1]. As costs have plummeted, genomic sequencing has become accessible to a wide range of stakeholders, including independent research laboratories and even individual enthusiasts. This democratization of sequencing has spurred a multitude of large-scale studies aimed at capturing genetic diversity across populations, diseases, and phenotypes, resulting in an unprecedented accumulation of genomic data. However, the rate at which this data is generated now far outstrips our capacity to interpret it effectively. Improved methods for accurate and rapid genomic interpretation could revolutionize precision medicine, especially in areas directly influenced by genetic variants, including complex disorders such as Postural Orthostatic Tachycardia Syndrome (POTS), Ehlers-Danlos syndromes, and various cancers.

Despite these advances, the inherent complexity, massive scale, and nuanced variability of genomic sequences pose substantial analytic challenges. Historically, genomic data analysis relied on expert knowledge from clinicians and statisticians, but recent developments in machine learning—particularly deep learning—have opened avenues to more automated and scalable solutions. Convolutional neural networks and Transformer-based models have both been applied to tasks ranging from classification to sequence generation. Models such as DNABERT [2], HyenaDNA [3], and Caduceus [4] exemplify state-of-the-art architectures, each excelling in their specific domains. However, these models often struggle when applied to unfamiliar tasks or other models' benchmarks. Encoder-only architectures, for instance, excel at identifying patterns through bi-directional attention but cannot generate new sequences. Conversely, decoder-only models specialize in sequence generation but lose access to global contextual information due to their inherently uni-directional or auto-regressive structure.

To address these limitations, this paper introduces DNABART, a new foundational model that leverages an encoder-decoder Transformer architecture optimized for genomic data. By integrating the strengths of both encoders and decoders, DNABART aims to deliver a more flexible and powerful solution capable of both interpreting and generating complex genomic sequences.

II. Related Work

A. Sequence Generation Models

Over the past few years, researchers have proposed a range of methodologies focus on generating or refining genomic sequences to ensure higher accuracy and reliability. For short-read data, tools such as Quake [5] serve as error correction frameworks, employing statistical models of k-mer frequencies to identify and rectify sequencing errors that arise from high-throughput short-read technologies like Illumina [6]. These methods effectively improve read quality without necessitating reference genomes, thereby facilitating more accurate downstream analyses. On the other hand, emerging long-read sequencing platforms—such as PacBio [7] and Nanopore [8]—prioritize read length over initial accuracy, producing substantially longer sequences that simplify genome assembly and resolve complex structural variations. Although they typically introduce higher per-base error rates, their capacity to span challenging genomic regions often compensates for this limitation. Together, these varied approaches—error correction tools for short reads and cutting-edge long-read technologies—tackle complementary aspects of the sequence generation problem, ultimately converging on the shared goal of producing more accurate and complete genomic assemblies.

B. Genomic LLM Models

DNABERT [2] introduced the concept of adapting BERT-style Transformer models to the genomic domain by addressing the need for specialized tokenization to handle the complexity and repetitiveness of DNA sequences, ultimately surpassing conventional sequence-based and domain-specific baselines on numerous tasks. Building upon its foundational

principles, DNABERT-2 [9] integrated multiple k-mer resolutions to better capture the hierarchical complexity and long-range dependencies of genomic sequences, while refining pretraining, architecture, and fine-tuning strategies to improve computational efficiency and predictive accuracy. More recently, the Nucleotide Transformer [10] advanced these efforts further by introducing a scalable, general-purpose foundation model for human genomics that employs unsupervised pretraining and masked nucleotide prediction to learn rich, context-aware representations of nucleic acid sequences without heavy reliance on domain-specific features. By coupling carefully optimized tokenization methods with multi-head self-attention, it excels at identifying both local and global genomic patterns. Collectively, these advancements mark a growing trend toward universally applicable, high-capacity Transformer architectures that redefine performance benchmarks and accelerate discovery in computational genomics.

III. Proposed Methodology

In this work, we adapt the BART language model [11]—originally developed by Facebook AI Research—to the task of genomic sequence correction and classification. The resulting model, referred to as DNABART, closely follows the Transformer architecture proposed by Vaswani et al. [12], but incorporates modifications introduced by BART. Specifically, DNABART utilizes GELU activations in the encoder and decoder layers instead of ReLU, improving convergence and empirical performance on sequence reconstruction tasks.

A. Model Architecture

DNABART maintains the standard BART encoder-decoder framework. Let $Y = (y_1, y_2, \ldots, y_n)$ represent an input DNA sequence, where each $y_i \in \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$. The encoder produces contextualized representations $H = (h_1, h_2, \ldots, h_n)$, and the decoder, conditioned on H, autoregressively generates the corrected sequence $\hat{Y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$. Throughout the model, multi-head self-attention layers capture both local and long-range dependencies in genomic data. The model produces a probability distribution over the vocabulary at each timestep. Let $f_\theta(y_i)$ represent the predicted logits for sequence y_i at the output layer of the decoder, where $[f_\theta(y_i)]_{jk}$ is the logit probability of the k-th token in the vocabulary for the j-th position of sequence i. To form predictions, we select the token with the highest logit probability at each position (i.e., apply $\arg\max_k$ over the logits).

We measure accuracy by comparing the predicted tokens to the ground-truth sequence. Given N samples, let each sample i have length L_i and ground-truth labels $y_i = (y_{i1}, y_{i2}, \ldots, y_{iL_i})$. The accuracy is defined as:

$$Acc = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L_i}\sum_{j=1}^{L_i}\mathbb{1}(\arg\max_k [f_\theta(y_i)]_{jk} = y_{ij}) \quad (1)$$

B. Tokenization Strategy

Tokenization is critical for large language models (LLMs) and can profoundly influence their representational power.

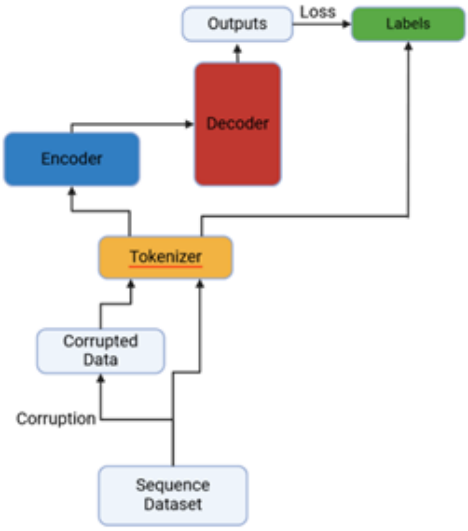


Fig. 1: Proposed DNABART Pretraining Workflow

Unlike natural language, genomic data do not present clear word boundaries, and the four basic nucleotides $\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}$ provide limited semantic richness if used as tokens directly. K-mer tokenization, although common in bioinformatics tasks such as error correction and assembly, leads to a proliferation of arbitrary tokens with limited semantic value. To address these challenges, we adopt Byte-Pair Encoding (BPE) as recommended by Zhou et al. [9], which has proven effective in learning meaningful token representations from genomic data. Formally, given a sequence Y, BPE iteratively merges frequent symbol pairs, producing a subword vocabulary V of size $|V| = 4096$, as suggested by prior research. This approach yields token embeddings that are both semantically dense and efficient, improving downstream performance and enabling longer context lengths.

C. Corruption and Pre-training Objective

To train DNABART as a denoising autoencoder for genomic sequences, we introduce synthetic noise into the input sequences. Specifically, each nucleotide is corrupted with a 30% substitution probability, replaced with a random nucleotide from $\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}$. This corruption scheme, though simplistic, mimics realistic sequencing errors. Formally, for each position i in Y, we sample $\hat{y}_i$ from a uniform distribution over the nucleotide set with probability 0.3, and leave it unchanged otherwise. The resulting corrupted sequences, $\hat{Y}$, are paired with the original Y to form the input-target pairs for training.

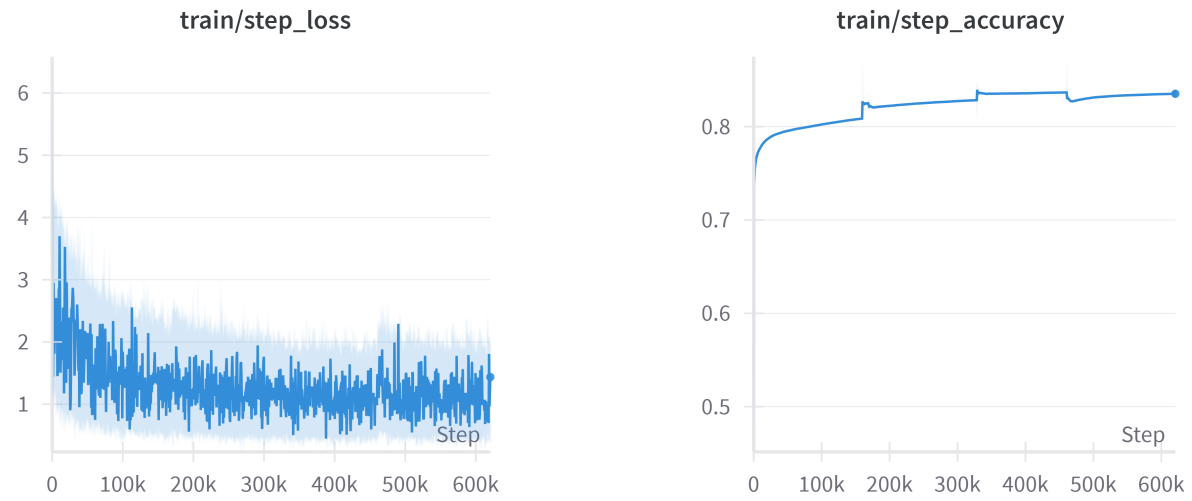
IV. Experiments

A. Datasets and Model Configurations

In this paper, we pretrain the BART language model using two different model configurations each trained on different datasets and model parameters, as shown in Figure 1. Note that the two models follow similar tokenization strategy and corruption methodology.

TABLE I: Pretraining Performance of Model $\mathcal{T}$ using Different Corruption Techniques

Model	**Saccharomyces 1M**		
	Acc	F1	PPL
BART Base			
w/ Substitution	98.2	90.6	1.18
w/ Deletion + Insertion	83.5	83.5	3.05
w/ Substitution + Deletion + Insertion	89.1	86.2	5.03



(a) Step Loss

(b) Step Accuracy

Fig. 2: Loss and Accuracy Performance of the Model $\mathcal{T}$ under Insertion + Deletion Corruption

Model $\mathcal{G}$: The objective of this model is to train sequences derived from the DNABERT2 dataset [9], chosen for its substantial genetic diversity and size. This dataset, originally on the order of tens of gigabytes, was randomly shuffled and subsequently halved, resulting in a final training corpus of approximately 15 GB. These preprocessing steps ensured both manageable computational demands and a representative sample of genomic variability. Then, the sequences are corrupted as discussed above. We initialize DNABART from a base BART configuration and train it to reconstruct the original sequence from the corrupted input. Cross-entropy loss is employed, and the model is optimized using the AdamW optimizer with a learning rate of 5×10^{-5} and a weight decay of 0.1. A linear scheduler gradually decreases the learning rate over time. Training is performed for 2 epochs with a batch size of 128 on the Missouri S&T supercomputing cluster ("The Mill"), utilizing an NVIDIA H100 100GB GPU. Although exact timing estimates vary due to job preemptions, total pretraining time is approximately 4 days.

Model $\mathcal{T}$: On the other hand, the objective of this model is to train Saccharomyces genome sequences[1]. The Saccharomyces Genome Database (SGD) provides comprehensive integrated biological information for the budding yeast Saccharomyces cerevisiae. Its extensive characterization and stability make it an ideal benchmark for evaluating new computational genomics methodologies. We generated 1 million sequences from the reference sequence each of length 150. Three different models are trained based on a combination of corruption techniques as follows: only substitution, and deletion + insertion, substitution + deletion + insertion. Upon corrupting the sequences, base BART model is initialized and trained it to reconstruct the original sequence. This model adopts cross-entropy loss with AdamW optimizer and a learning rate of 10^{-4}. Training is performed for 3 epochs with a batch size of 5 using 3 V100 32GB GPUs (from Missouri S&T's Mill Cluster). Total pretraining time is approximately 18 hours.

[1]Dataset is available at: http://sgd-archive.yeastgenome.org/sequence/S288C_reference/genome_releases/

B. Pretraining

Both models $\mathcal{G}$ and $\mathcal{T}$ were evaluated on a corresponding test dataset to see if it could accurately regenerate sequences with 30% corruption rates. Table I demonstrates the pretraining performance (accuracy, f1 score, and perplexity) of the model $\mathcal{T}$ based on different corruption techniques. As observed, the BART-base model with substitution corruption alone results in higher accuracy and F1 score compared to other corruption techniques. Though the loss is decreasing significantly during pretraining (as shown in Figure 2), we observe that the model is not accurately recreating the original sequences while testing

TABLE II: Epigenetic Marks Prediction

	H3	H3K14ac	H3K36me3	H3K4me1	H3K4me2	H3K4me3	H3K79me3	H3K9ac	H4	H4ac
NT-2500M	78.77	56.2	61.99	55.3	36.49	40.34	64.7	56.01	81.67	49.13
DNABERT2	80.17	57.42	61.9	53	39.89	41.2	67.39	57.07	81.86	50.43
DNABART	78.29	**63.51**	**66.4**	**63.79**	**64.94**	**58.32**	**72.98**	**66.97**	**80.97**	**60.85**

TABLE III: Promoter Detection

	all	notata	tata
NT-2500M	91.01	94	79.43
DNABERT2	86.77	94.34	71.59
DNABART	**85.34**	89.03	62.97

TABLE IV: Core Promoter Detection

	all	notata	tata
NT-2500M	70.33	71.58	72.97
DNABERT2	69.37	69.53	76.18
DNABART	**77.75**	**79.8**	71.13

TABLE V: Transcription Factor Prediction (Human)

	0	1	2	3	4
NT-2500M	66.64	70.28	58.72	51.65	69.34
DNABERT2	71.99	76.06	66.52	58.54	77.43
DNABART	**79.6**	**81.9**	**72.9**	**65.8**	**77.5**

TABLE VI: Transcription Factor Prediction (Mouse)

	0	1	2	3	4
NT-2500M	66.64	70.28	58.72	51.65	47.07
DNABERT2	64.23	86.28	81.28	73.49	52.66
DNABART	52.84	79.73	78.05	56.9	**54.54**

TABLE VII: Splice and Virus Predictions

	Splice	Virus
NT-2500M	89.35	73.04
DNABERT2	85.93	71.02
DNABART	56.51	14.15

phase. Moreover, the model is unable to generate full length of sequence (i.e., 150).

Because the pretraining dataset was heavily corrupted and differed substantially from realistic raw sequencing reads, a more representative test dataset was constructed to evaluate the model's performance on real-world scenarios. To achieve this, a modified version of the wgsim simulator was employed to generate 10,000 short-read samples from the T2T reference genome, reflecting Illumina sequencing characteristics. Consistent with previous observations, results indicated that the model's performance remained far from producing usable error corrections on realistic raw reads. We hypothesize that this discrepancy arises from the significantly lower error rates in actual sequencing data compared to the artificially high corruption levels used during pretraining. Consequently, when the model attempts to correct realistic sequences, it overcorrects, inadvertently increasing the overall error rate instead of reducing it.

C. Finetuning

The pretrained DNABART model was further evaluated on classification tasks from the Genome Understanding and Evaluation (GUE) benchmark. For each dataset, DNABART was fine-tuned by adding a classification head consisting of a linear layer, ReLU activation, a 30% dropout layer, and a final linear layer with output dimensionality matching the task at hand. Each model was trained for 8 epochs using a learning rate of 2e-5, with the same optimizer and linear scheduling parameters as in pre-training. The fine-tuning experiments were conducted on an NVIDIA RTX 4070 12GB GPU, with total training times under 30 minutes per dataset.

The GUE benchmark includes 7 tasks spanning 28 distinct datasets. After fine-tuning on 26 of these datasets, DNABART achieves state-of-the-art accuracy on 16 of them, with improvements of up to 24%. In tasks where DNABART does not surpass current state-of-the-art results, it generally achieves comparable performance. One exception is the COVID dataset in the virus classification task (20-class classification), where performance is notably lower. This shortfall may be due to the high corruption levels used during pre-training, limiting the model's ability to capture subtle genomic variations that distinguish closely related viral strains, which sometimes differ by a single nucleotide.

Observations suggest that DNABART excels in tasks characterized by longer-range dependencies and data that span multiple tokens. As shown in the accompanying table, DNABART's results are compared with the best-performing variants of the Nucleotide Transformer ($\sim$2.5B parameters) and DNABERT2 ($\sim$117M parameters), while DNABART itself has approximately 103M parameters. This comparison highlights the efficiency and competitive performance of DNABART in the computational genomics domain.

REFERENCES

[1] K. Wetterstrand, "Dna sequencing costs: Data from the nhgri genome sequencing program (gsp)." https://www.genome.gov/sequencingcosts/. Accessed: 2024-12-12.

[2] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri, "DNABERT: Pre-Trained Bidirectional Encoder Representations from Transformers Model for DNA-language in Genome," *Bioinformatics*, vol. 37, no. 15, pp. 2112–2120, 2021.

[3] E. Nijkamp, P. Yin, X. Liu, Y. Yue, G. Kreiman, and Y. N. Wu, "Hyenadna: A biologically interpretable language model for long-range genomics," *bioRxiv*, 2023.

[4] P. K. Zhang, S. Kelly, L. Stankovic, M. Matei, D. Wingate, and A. Faulconbridge, "Caduceus: A language model for genomes," *bioRxiv*, 2023.

[5] D. R. Kelley, M. C. Schatz, and S. L. Salzberg, "Quake: Quality-Aware Detection and Correction of Sequencing Errors," *Genome biology*, vol. 11, pp. 1–13, 2010.

[6] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, *et al.*, "Accurate whole human genome sequencing using reversible terminator chemistry," *nature*, vol. 456, no. 7218, pp. 53–59, 2008.

[7] C.-S. Chin, D. H. Alexander, P. Marks, A. A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E. E. Eichler, *et al.*, "Nonhybrid, finished microbial genome assemblies from long-read smrt sequencing data," *Nature methods*, vol. 10, no. 6, pp. 563–569, 2013.

[8] M. Jain, H. E. Olsen, B. Paten, and M. Akeson, "The oxford nanopore minion: delivery of nanopore sequencing to the genomics community," *Genome biology*, vol. 17, pp. 1–11, 2016.

[9] Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri, and H. Liu, "Dnabert-2: Efficient Foundation Model and Benchmark for Multi-Species Genome," *arXiv preprint arXiv:2306.15006*, 2023.

[10] H. Dalla-Torre, L. Gonzalez, J. Mendoza-Revilla, N. Lopez Carranza, A. H. Grzywaczewski, F. Oteri, C. Dallago, E. Trop, B. P. de Almeida, H. Sirelkhatim, *et al.*, "Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics," *Nature Methods*, pp. 1–11, 2024.

[11] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.