Supplementary Material: Appendices

A Information-theoretic Analysis of DNA/RNA Tokenization

In this section, we compare the information content of a nucleotide token and a BPE token inspired by key empirical observations in the training data. Information-theoretic analysis of biological sequences is a well-studied field of research where the key challenges include determining the prior distribution of nucleotides or k-mers, the fact that only a fraction of possible biological sequences occurs in nature and the difficulty in comparing results from biological sequences with those from linguistics due to significant differences morphology.

Information content of a Nucleotide token We consider each nucleotide in a sequence as an independent variable that carries some amount of information. We wish to quantify the maximum amount of information for each new nucleotide in a sufficiently long sequence. We can derive the per-token upper bound of the Shannon Entropy of a DNA/RNA sequence as follows.

$$\begin{aligned} H(X_{NUC}) &= -\sum_{i=1}^{4} P(x_i) \log_2 P(x_i) \\ &\leq -\left(4 \cdot \frac{1}{4} \log_2 \frac{1}{4}\right) \\ &= 2 \text{ bits} \end{aligned} \tag{1}$$

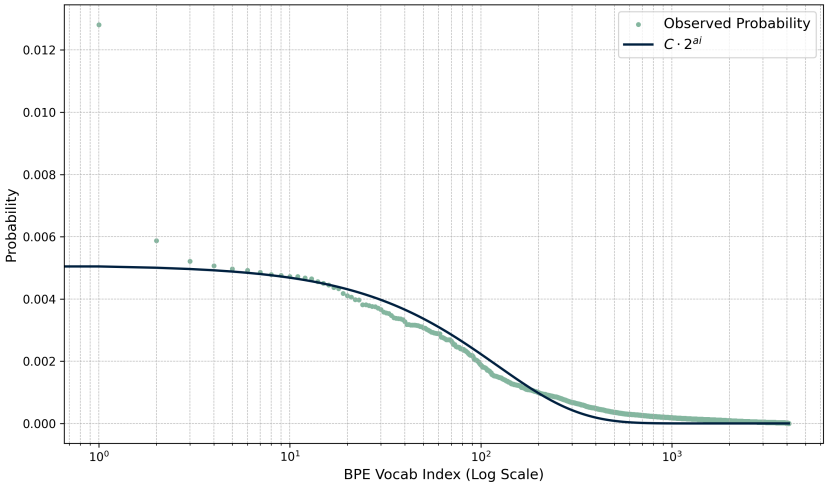

Figure 1: Fitting the exponential function $C \cdot 2^{-ai}$ to the empirically observed BPE token probabilities on pretraining datasets. Since the index assigned to a token is arbitrary, we reverse sort by probability/frequency and reindex tokens to represent the probability as a function of the index. We determined best-fit when $C \approx 0.005086$ and $a \approx 0.011909$.

Information content of a BPE token As before, we consider each BPE token in a sequence as an independent variable that carries some amount of information. Let the size of the vocabulary be N. On our pretraining datasets, we observe that the frequency of the BPE tokens is exponentially distributed, and as a result, the probability of a token can modeled by an exponential function.

$$P(x_i) = \frac{C}{2^{ai}} \tag{2}$$

Since the index assigned to a token is arbitrary, tokens can be sorted by descending probability and reindexed without issue. Under this formulation, the token with the index $i = 1$ is the most frequent

(the *GGG* token in our pretraining dataset) and $i = 4095$ is the least frequent (the *TTGTCGGGTAAG* token).

$$\begin{aligned}
&\sum_{i=1}^{N} \frac{C}{2^{ai}} = C \sum_{i=1}^{N} \frac{1}{2^{ai}} = 1 \\
&\Leftrightarrow \frac{C}{2^a} \cdot \frac{1 - \left(\frac{1}{2^a}\right)^N}{1 - \frac{1}{2^a}} = 1 \\
&\Leftrightarrow \frac{C\left(1 - \frac{1}{2^{aN}}\right)}{2^a - 1} = 1 \\
&\Leftrightarrow C = \frac{2^a - 1}{1 - \frac{1}{2^{aN}}}
\end{aligned} \tag{3}$$

When the vocabulary size N is large, we can approximate $C \approx 2^a - 1$ and $a \approx log_2(C + 1)$. Now we can derive a general expression for the entropy of BPE tokens as,

$$\begin{aligned}
H(X_{BPE}) &= -\sum_{i}^{N} P(x_i) \log_2 P(x_i) \\
&= -\sum_{i=1}^{N} \frac{C}{2^{ai}} \log_2 \left(\frac{C}{2^{ai}}\right) \\
&= -\sum_{i=1}^{N} \frac{C}{2^{ai}} \left(\log_2 C - ai\right) \\
&= -\log_2 C \sum_{i=1}^{N} \frac{C}{2^{ai}} + aC \sum_{i=1}^{N} \frac{i}{2^{ai}}, \\
&= -\log_2 C + aC \sum_{i=1}^{N} \frac{i}{2^{ai}}, \\
&\approx -\log_2 C + aC \frac{2^a}{(2^a - 1)^2}\ , \qquad \text{When } N \text{ is large} \\
&= -\log_2 C + \log_2(C+1) C \frac{C+1}{(C)^2} \\
&= -\log_2 C + \log_2(C+1) \left(\frac{C+1}{C}\right) \\
&= \log_2 \left(\frac{(C+1)^{(C+1)/C}}{C}\right).
\end{aligned} \tag{4}$$

If the weighted average length of a BPE token is $\bar{L} = \sum_{i=1}^{N} P(x_i)\text{len}(x_i)$, the average character-level entropy of BPE representation of a sequence will be $\hat{H}(X_{BPE}) = \frac{H(X_{BPE})}{\bar{L}}$. Since nucleotides are one character each, the per-character entropy is $\hat{H}(X_{NUC}) = H(X_{NUC})$. The BPE tokenization will lead to less entropy if,

$$\begin{aligned}
&\frac{\hat{H}(X_{BPE})}{\hat{H}(X_{NUC})} < 1 \\
&\Rightarrow \frac{H(X_{BPE})}{\bar{L} \times H(X_{NUC})} < 1 \\
&\Rightarrow \log_2 \left(\frac{(C+1)^{(C+1)/C}}{C}\right) < 2 \times \bar{L} \\
&\Rightarrow \frac{(C+1)^{(C+1)/C}}{C} < 4^{\bar{L}}.
\end{aligned} \tag{5}$$

When $C << 1$, we can approximate $(C + 1) \approx 1$. Then the inequality in Equation 5 can be further simplified as

$$\frac{1}{C} < 4^{\bar{L}} \Rightarrow C > 4^{-\bar{L}}.$$

Empirical Entropy Ratio On our pretraining data mixture, we determine that $P(A) \approx 0.2726$, $P(A) \approx 0.2144$, $P(A) \approx 0.26642$, $P(A) \approx 0.2465$, and averge BPE token length $\bar{L} \approx 6.0768$. This yields the empirical entropy of nucleotide tokens $H_e(X_{NUC}) \approx 1.9939$ bits. As shown in Figure 1, the **empirical value of C is 0.005086** when determined on 33 million sequences of our pretraining dataset.

Plugging in this value in Eqn. 4, yields $H_e(X_{BPE}) \approx 9.1044$ bits. Therefore, the empirical per-character entropy ratio is

$$\frac{\hat{H}_e(X_{BPE})}{\hat{H}_e(X_{NUC})} = \frac{H_e(X_{BPE})}{\bar{L} \times H_e(X_{NUC})} \approx \frac{9.1044}{6.0768 \times 1.9939} \approx 0.7514 < 1.$$

The empirical per-character entropy ratio of 0.7514 indicates that the Byte-Pair Encoding (BPE) tokenization technique effectively compresses the input sequence. Although compressed information is likely more difficult for Language Models to process, it is well-compensated by the ability to process sequences up to 6 times longer than the original input with the same GPU memory constraints. This also partially explains why we observed BPE underperforming their NUC counterparts on short-sequence downstream tasks from an Information-theoretic perspective.

Therefore, BPE tokenization is essentially a trade-off between information compression and computational efficiency, which BiRNA-BERT can dynamically adjust depending on the hardware constraints and sequence length.

Here, we assume tokens are independent and identically distributed random variables (i.i.d) to approximate the information content of NUC and BPE sequences. In reality, the information content of non-i.i.d sequences is much lower than Shannon Entropy due to the correlation between nearby symbols. Language entropy [Shannon, 1948] and Kolmogorov Complexity [Kolmogorov, 1998] take symbol correlation and order into account but are generally intractable.

B Downstream Task Finetuning Methodology

In this part, we describe the finetuning details for the RNA downstream tasks for the models RNA-FM, RiNALMo, and BiRNA.

miRNA-lncRNA Interaction

- **Embedding Strategy**:
  - Following Wang et al. [2023], we use frozen embeddings with a CNN head.
- **Parameter Grid Search**:
  - **Learning Rate (LR)**: [1e-3, 5e-4, 1e-4, 5e-5, 1e-5, 5e-6, 1e-6]
  - **Warmup Proportion**: [0.05, 0.1, 0.3]
  - **Number of Epochs**: [2, 3, 5, 10, 20, 30]
- **Default Selected Configuration**:
  - 3 convolutional layers, flatten, 2 dense layers
  - Batch size: 32
  - GPU: A6000 48GB
  - 3 epochs, 5e-4 learning rate, 0.1 warmup proportion
- **Specific Selected Configurations**
  - **30 epochs; 1e-3 LR; 0.05 WarmUp:** GMA-MTR: RNA-FM, RiNALMo, BiRNA
  - **20 epochs:** MTR-ATH: RNA-FM, RiNALMo, BiRNA
  - **5 epochs:** ATH-MTR: RiNALMo, MTR-ATH: RNA-FM
  - **1e-3 LR; 0.05 WarmUp:** GMA-MTH: RiNALMo

Torsion Angle Regression

- RNA-FM and BiRNA-BERT
  - Batch size: 32
  - Learning rate: 1e-5
  - Epochs: 20
  - Warmup ratio: 0.1
  - Gradient accumulation steps: 1
- RiNALMo
  - Batch size: 8
  - Learning rate: 1e-5
  - Epochs: 20
  - Warmup ratio: 0.1
  - Gradient accumulation steps: 2

RNA-Protein Interaction

- Learning rate: 1e-6
- Batch size: 64
- Epochs: 10
- Warmup ratio: 0.1
- Single prediction head
- Early stopping on best validation F1 score

N6-methyladenosine Site Prediction

- **Learning Rate (LR)**: 0.005
- **Warmup Proportion**: 0.1
- **Number of Epochs**: 3

C BiDNA

We test dual tokenization on DNA sequences by training three more BERT models on the Human Genome DNA Dataset similar to DNABERT. Similar to BiRNA, sequences are tokenized both with BPE and NUC when pretraining BiDNA. Pretraining details of DNA models are shown in Table 2.

we evaluate all three variants on human-genome-related downstream tasks from the GUE benchmark [Zhou et al., 2023]. The tasks are:

1. Promoter Site Detection: 3 datasets
2. Core Promoter Site Detection: 3 datasets
3. Transcription Factor Binding Prediction: 5 datasets

We evaluate the performances on all versions of the pretrained BiDNA and provide performance metrics from DNABERT-2 for reference. DNABERT-2 is pretrained with 66X more compute than BiDNA.

We see from Table 1 that, in promoter site detection task, BiRNA with NUC tokenization is comparable to DNABERT-2, within only 0.8% performance margin. In the core promoter site detection task, BiDNA-NUC **outperforms** DNABERT-2 with 4.4% MCC. In the transcription factor binding site task, BiDNA achieves a competitive performance within 1.2% margin.

Table 1: Comparison of BiDNA with DNABERT-2 on Three Downstream Tasks (MCC Metric)

Task	**Promoter Site Detection**			**Core Promoter Site Detection**		
Dataset	**All**	**No tata**	**Tata**	**All**	**No tata**	**Tata**
BPE-Only	0.916	0.954	0.799	0.806	0.824	0.765
BiDNA-BPE	0.918	0.954	0.789	0.812	0.828	0.758
NUC-Only	0.927	0.963	0.807	**0.832**	0.836	**0.876**
BiDNA-NUC	0.936	0.966	0.817	**0.832**	0.839	0.844
DNABERT-2	**0.941**	**0.971**	**0.830**	0.831	**0.849**	0.814

Table 1: Comparison of BiDNA with DNABERT-2 on Three Downstream Tasks (MCC Metric) (Continued)

Model	**Transcription Factor Binding Dataset**					
	tf0	**tf1**	**tf2**	**tf3**	**tf4**	**Avg tf**
BPE-Only	0.847	0.852	0.807	0.738	0.847	0.818
BiDNA-BPE	0.847	0.848	0.803	0.753	0.852	0.821
NUC-Only	0.844	0.871	0.838	0.759	0.877	0.838
BiDNA-NUC	0.850	0.874	**0.843**	0.770	0.874	0.842
DNABERT-2	**0.856**	**0.886**	0.841	**0.790**	**0.888**	**0.852**

D Compute Analysis

RiNALMo

- **GPU:** 7XA100 (80GB)
- **Training Time:** 14 days
- **Peak FP16 Performance:** 624 TFLOPS (Nvidia A100 Datasheet)

BiRNA-BERT

- **GPU:** 8×3090 (24GB)
- **Training Time:** 48.42 hours
- **Peak FP16 Performance:** 142 TFLOPS (Nvidia Ampere Datasheet)

Ratio Calculation:

$$\text{Ratio} = \frac{7 \times 624 \times 14 \times 24}{8 \times 142 \times 48.42} = 26.682$$

DNABERT2

- **GPU:** 8×2080Ti
- **Training Time:** 14 days
- **Peak FP16 Performance:** 113.8 TFLOPS (Nvidia Ada Datasheet)

BiDNA

- **GPU:** 1× 4090
- **Training Time:** 14 hours
- **Peak FP16 Performance:** 330.3 TFLOPS (Nvidia Ada Datasheet)

Ratio Calculation:

$$\text{Ratio} = \frac{8 \times 113.8 \times 14 \times 24}{330.3 \times 14} = 66.150$$

Model	Train Tokens	Train Time	Hardware
RNA-BPE only	4.384B	3 hours	8×3090
RNA-NUC only	27.87B	45.1 hours	8×3090
RNA-BiRNA	32.254B	48.42 hours	8×3090
DNA-BPE only	586.854M	2.42 hours	1×4090
DNA-NUC only	3.027B	12 hours	1×4090
DNA-BiDNA	3.614B	14.33 hours	1×4090

Table 2: Pretraining details for various models

E Downstream Tasks Dataset Description

miRNA-lncRNA Interaction Prediction For evaluation, one benchmarking dataset is used as the training set, and another dataset is used for validation following the strategy used in Wang et al. [2023]. Thus, we have 6 train-test combinations and we report performance in all these combinations. The lengths of the sequences in the miRNA dataset are 10 to 50, whereas the lncRNA dataset ranges from 200 to 4000. The length distributions of sequences for the miRNA-lncRNA Interaction Prediction task are shown in Figure 2.

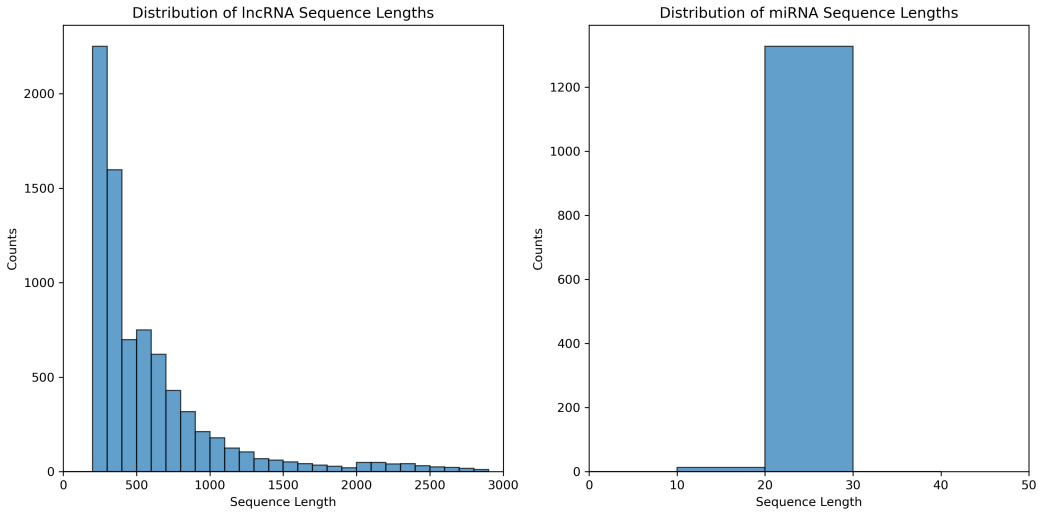

Figure 2: Sequence length distribution for lncRNA and miRNA datasets

Table 3: Benchmark of miRNA-lncRNA Interaction Prediction

Benchmarks of RNA–RNA interactions		No. of miRNAs	No. of lncRNAs	No. of molecule pairs
Arabidopsis thaliana (Ath)	Interacting Pairs	331	2014	2500
	Non-interacting Pairs	266	1964	2500
Glycine max (Gma)	Interacting Pairs	401	1770	2500
	Non-interacting Pairs	542	171	2500
Medicago truncatula (Mtr)	Interacting Pairs	335	1986	2500
	Non-interacting Pairs	424	2442	2500

RNA-Protein Interaction Prediction (short-sequence task) This focuses on finding the binding sites and interactions between RNA molecules and proteins to understand post-transcriptional regulation. Benchmark dataset for RNA-protein interaction prediction is collected from RBPsuit available at `http://www.csbio.sjtu.edu.cn/bioinf/RBPsuite/`. This database contains datasets for 154 different proteins and a collection of interacting and non-interacting RNA sequences for each protein. We consider a subset of 5 datasets for our evaluation (AARS, AATF, AKAP1, AGGF1, ABCF1). The length of RNA sequences used for this task is 101 across all the datasets. The number of sequences used for training, validation, and testing is shown in Table 4.

RNA N6-methyladenosine Prediction (short-sequence task) N6-methyladenosine (m6A) is a common and critical modification in eukaryotic mRNA, affecting various aspects of RNA metabolism. This includes stability, splicing, and translation. The prediction and detection of m6A sites are

Table 4: Dataset Specification for RNA-Protein Interaction Prediction

Dataset	Train	Valid	Test
AATF	26283	6571	8214
ABCF1	28768	7193	8991
AGGF1	76800	19200	24000
AKAP1	76800	19200	24000
AARS	76800	19200	24000

essential for understanding how this modification influences gene expression and cellular processes. In our work, we utilized datasets from human, rat, and mouse tissues, specifically focusing on brain, kidney, and liver samples for each species. Data sets were derived from the iRNA-m6A study available at `http://www.biolscience.cn/Deepm6A-MT/data/`, employing an antibody-independent m6A-REF-seq protocol, which is both high-throughput and accurate for m6A detection. Positive samples were selected based on the presence of m6A at the center of 41 continuous nucleotide residues, while negative samples were randomly selected from the same tissues but without m6A sites. The length of the sequences across all the datasets is 41. Dataset specifications are shown in Table 5.

Table 5: Dataset Specification for RNA N6-methyladenosine Prediction

Species	Tissue	Training Pos	Training Neg	Test Pos	Test Neg
Human	Liver	2634	2634	2634	2634
	Brain	2302	2303	1150	1150
	Kidney	2287	2287	1144	1143
Mouse	Brain	4013	4012	4013	4012
	Kidney	1977	1976	1976	1977
	Liver	2066	2067	2066	2067
Rat	Brain	1176	1176	1176	1176
	Kidney	1716	1716	1716	1716
	Liver	881	881	881	881

Multi Species RNA Splicing Site Prediction (short-sequence task) RNA splicing is a crucial process in eukaryotic gene expression, where introns are removed from precursor messenger RNAs (pre-mRNAs), and exons are joined together to form mature mRNAs. This process is essential for generating functional mRNAs that can be translated into proteins. Identifying splice sites—the donor sites at the 5' end of introns and the acceptor sites at the 3' end—is vital for accurately predicting gene structure and location. For this task, we consider the dataset proposed by Scalzitti et al. [2021]. Particularly we use the gold standard dataset GS_1 which contains an equal number of positive and negative samples. The dataset consists of "confirmed" error-free splice-site sequences from a diverse set of 148 eukaryotic organisms, including humans. We have tested the performance of the trained model on three independent test datasets containing the samples from 3 different species of fish (*Danio rerio*), fruit fly (*Drosophila melanogaster*), and plant (*Arabidopsis thaliana*). Here the sequence length is 400 and the train and independent test tests have 20000 sequences each for training and testing respectively.

RNA 3D Torsion Angle Prediction (Nucleotide Level Task) There are seven torsion angles commonly referred to as α, β, γ, δ, ϵ, ζ, and χ. These angles describe the rotations around the bonds that connect the nucleotides within an RNA strand, influencing its overall structure and stability. These angles are mathematically represented as the dihedral angles between four consecutive atoms in the RNA backbone. For example, the α angle is measured as the dihedral angle between O5'-P-O3'-C3'. The dataset for RNA torsion angle prediction is collected from `https://sparks-lab.org/server/spot-rna-1d/`. The training (TR), validation (VL), and three test sets (TS1, TS2, and TS3) have 286, 30, 63, 30, and 54 RNA chains, with average sequence lengths of 122, 15, 30, 14, and 24 respectively.

F Detail Result Tables for Downstream Tasks

Table 6: Mean Absolute Error of RNA 3D torsion angle prediction. Here we report the mean squared error (MSE) between the predicted and actual RNA torsion angles. The "NUC-only" method refers to tokenization at the nucleotide level without any additional processing. We show the average error across different torsion angles for various methods. The best and second best results are shown in bold and italic, respectively.

Data	Method	Avg Error	Alpha	Beta	Gamma	Delta	Epsilon	Zeta	Chi
VL	BPE-NUC	*28.085*	29.132	24.018	25.692	21.345	25.718	35.918	34.771
	NUC-only	28.398	29.583	23.910	26.195	21.779	26.193	35.948	35.177
	RNA-FM	28.333	29.357	24.412	26.109	21.983	25.451	35.027	35.994
	RINALMo	**27.888**	28.861	23.188	25.866	22.486	25.078	35.132	34.603
TS1	BPE-NUC	**28.181**	30.223	21.856	19.553	29.193	34.232	26.764	35.449
	NUC-only	28.760	31.002	22.887	19.793	29.415	34.607	27.980	35.637
	RNA-FM	29.916	31.904	23.859	19.839	29.620	35.884	30.372	37.937
	RINALMo	*28.622*	31.127	22.658	20.199	29.686	32.880	27.607	36.195
TS2	BPE-NUC	*26.704*	24.728	17.363	23.836	19.104	31.875	36.048	33.973
	NUC-only	27.252	25.602	18.408	24.406	19.748	32.139	36.474	33.990
	RNA-FM	27.710	25.464	17.842	23.829	19.823	35.864	37.754	33.391
	RINALMo	**25.915**	22.677	16.964	20.958	18.356	31.906	38.238	32.304
TS3	BPE-NUC	*31.979*	34.728	17.363	23.836	19.104	31.875	36.048	33.973
	NUC-only	32.174	34.856	19.606	30.324	37.531	36.589	35.374	30.938
	RNA-FM	32.000	33.415	20.341	30.755	38.091	36.360	34.414	30.681
	RINALMo	**31.513**	34.016	19.292	30.343	37.841	35.763	33.960	29.575

References

A. Kolmogorov. On tables of random numbers. *Theoretical Computer Science*, 207(2):387–395, 1998. ISSN 0304-3975. doi: https://doi.org/10.1016/S0304-3975(98)00075-9. URL `https://www.sciencedirect.com/science/article/pii/S0304397598000759`.

N. Scalzitti, A. Kress, R. Orhand, T. Weber, L. Moulinier, A. Jeannin-Girardon, P. Collet, O. Poch, and J. D. Thompson. Spliceator: multi-species splice site prediction using convolutional neural networks. *BMC bioinformatics*, 22:1–26, 2021.

C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423, 1948.

Y. Wang, Z. Pan, M. Mou, W. Xia, H. Zhang, H. Zhang, J. Liu, L. Zheng, Y. Luo, H. Zheng, et al. A task-specific encoding algorithm for rnas and rna-associated interactions based on convolutional autoencoder. *Nucleic Acids Research*, 51(21):e110–e110, 2023.

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.