In-silico Genomics Benchmarking for Neural Models

Heng Yang[1], Jack Cole[1], Krasimira Tsaneva-Atanasova[2], Stefano Pagliara[3], Yiliang Ding[4], and Ke Li[1]

[1] Department of Computer Science, University of Exeter, Exeter, EX4 4QF, UK
[2] Living Systems Institute and Biosciences, University of Exeter, Exeter, EX4 4QD, UK
[3] Department of Mathematics and Statistics, University of Exeter, Exeter, EX4 4QJ, UK
[4] Department of Cell and Developmental Biology, John Innes Centre, Norwich Research Park, 6 Norwich NR4 7UH, UK

1 Tasks and Data

In this challenge, we focus on predicting the structure and function of RNA sequences, exploring the learning capabilities of GFMs and other DL methods in this domain. We employ the RNA Genomic Benchmark (RGB), which emphasizes the model's sensitivity to single-nucleotide (instead of multi-nucleotide) resolution modeling, for example, allowing for the detection of individual nucleotides' impact on RNA molecules. The RGB suite encompasses six species and various genomic sequences, aiming to evaluate models' proficiency in RNA sequence modeling. Each task is meticulously designed to reflect the inherent complexity and diversity of genomic data, offering a robust framework for assessing state-of-the-art RNA models. Table 1 on page 2 presents detailed statistics for each task, including the number of examples, categories, evaluation metrics, and sequence lengths.

We elaborate on the task description as follows:

- **mRNA Degradation Rate Prediction.** This task aims to predict nucleotide-level decay rates in mRNA sequences to understand gene expression dynamics. The dataset is sourced from the Kaggle COVID-19 Vaccine Design Competition, focusing on sequence-only data without structural features. It is a token-level regression task where the objective is to predict degradation rates at each nucleotide position. Accurate predictions in this task can enhance the optimization of mRNA stability, which is crucial for RNA-based therapeutics and vaccine development.
- **Single-Nucleotide Mutation Detection (SNMD).** This task focuses on identifying single-nucleotide mutations in plant RNA sequences to aid in understanding genetic variations. The dataset comprises synthetic plant RNA sequences with up to 10 random single-nucleotide mutations per sequence. It is a binary token classification task designed to distinguish between mutated and non-mutated nucleotides. Successfully detecting genetic mutations in plants is essential for agricultural genomics and crop improvement.
- **Single-Nucleotide Mutation Repair (SNMR).** Building upon the SNMD task, this task aims to predict the correct nucleotides to repair single-nucleotide mutations, supporting gene therapy research. The dataset is similar to the SNMD dataset but requires models to suggest the correct nucleotide (A, U, C, or G) at mutated positions. It is a four-way token classification task to determine the appropriate nucleotide for mutation correction. This task evaluates a model's ability to detect and correct mutations, impacting precision medicine and synthetic biology.
- **RNA Secondary Structure Prediction.** This task aims to predict the secondary structures (base-pairing patterns) of RNA sequences, which is fundamental for understanding RNA function. The sub-datasets include bpRNA, a benchmark dataset with sequences up to 512 nucleotides using simplified structural symbols ('(', '.', ')'), and ArchiveII & RNAStrAlign, additional datasets processed to match the 512-nucleotide length limit and simplified structural annotations. Accurate secondary structure prediction is vital for understanding RNA functionality and interactions.
- **Predicting Antibiotic Resistance in Genes.** This task aims to predict the class of antibiotics to which each gene, represented by nucleotide sequences, is resistant. The dataset comprises a collection of Antibiotic Resistance Genes (ARGs), with each gene associated with resistance to a specific class of

antibiotics. This sequence-level classification task involves 14 distinct classes, enabling the prediction of antibiotics that are ineffective against each gene. Such predictions are crucial for antimicrobial resistance (AMR) research, facilitating the timely identification of antibiotics that remain effective against resistant strains of pathogens.

Table 2 on page 2 shows the virtual examples of different sub-datasets in RGB.

Table 1. The brief statistics of subtasks in the RGB. These benchmark datasets are held out or not included in the pretraining database. The numbers of examples in training, validation and testing sets are separated by "/". * indicates the datasets are used for zero-shot performance evaluation only.

Task	Task Type	# of examples	# of classes	Metric	Sequence length	Source
SNMD	Token classification	8,000/1,000/1,000	2	AUC	200	[6]
SNMR	Token classification	8,000/1,000/1,000	4	macro F1	200	[6]
mRNA	Token regression	1,735/193/192	—	RMSE	107	Kaggle
bpRNA	Token classification	10,814/1,300/1,305	3 å	macro F1	≤ 512	[7]
AchiveII	Token classification	2,278/285/285	3	macro F1	≤ 500	[8]
RNAStrAlign	Token classification	17,483/2,186/2,185	3	macro F1	≤ 500	[9]
ARG-AP	Seq classification	24,751/3,089/3,108	14	macro F1	$\leq 8,673$	[10][11][12][13][14]

Table 2. The virtual input and output examples in RGB. The "..." represents the sequences that are omitted for better presentation, and the red color indicates incorrect predictions in classification tasks. In the mRNA dataset, all single nucleotide bases have three values to predict. Note that "T" and "U" can be regarded as the same symbol in RNA sequences, depending on the dataset.

Dataset		Examples
SNMD	Input Sequence	G A G T A ... T T G A G
	True Label	0 0 1 0 0 ... 0 0 1 0 0
	Prediction	0 0 0 0 0 ... 0 0 1 0 0
SNMR	Input Sequence	T A C G A ... C T G A T
	True Label	T A C A A ... G T A A T
	Prediction	T A C A A ... C T G A T
mRNA	Input Sequence	G G ... A C
	True Label	[0.1,0.3,0.2] [0.8,0.4,0.1]... [0.9,0.4,0.3] [0.5,0.2,0.6]
	Prediction	[0.1,0.3,0.2] [0.8,0.4,0.1]... [0.9,0.4,0.3] [0.5,0.2,0.6]
bpRNA Archive2 RNAStralign	Input Sequence	G G C G A ... C U U U U
	True Label	(((· ·... · ·)))
	Prediction	((((·... ·))))
Antibiotic Prediction	Input Sequence	G T A C G ... C A T G C
	True Label	Beta-Lactam
	Prediction	Tetracycline