Evaluation datasets:

1. Genome Understanding Evaluation Benchmark:
    a) First, clone the GitHub repository of DNABert-2 paper https://github.com/MAGICS-LAB/DNABERT_2
    b) Follow the instructions mentioned under Section 6 of their README.md file, with subsection 6.1 titled "Evaluate models on GUE"
    c) We followed the instructions as given for DNABert model for 6-mer : *sh scripts/run_dnabert1.sh DATA_PATH 6*
    d) The scripts that we customized for our use is provided under the `scripts' directory. We provide the generated results folder for all the different variants.
2. Few-shot Evaluation Dataset
    a) Except for the silencer datasets, we use the datasets as made available in the GitHub repository of `GeneMask' paper, which are available at https://github.com/roysoumya/GeneMask/tree/main/Data-B_fewshot-task-datasets
    b) Please follow these steps to construct the silencer dataset used for evaluation.
        i. Please download the following FASTA format file from http://health.tsinghua.edu.cn/SilencerDB/download/Method/High_throughput_Homo_sapiens.fa
        ii. Run the Jupyter notebook present under the "scripts" directory named "create_silencer_data_from_fasta_file.ipynb" to generate the training and test splits.
        iii. We randomly create 10 sets of few-shot training sets, for ten different runs.
        iv. The evaluation data is provided under "/data/silencer/"

Proposed Models

1. Please clone the codebase of DNABert or GeneMask and follow their pretraining instructions.
2. Use the pretraining codes provided under "/src/pretraining-adaptive" directory

The pretrained model weights will be released upon acceptance.