DREAM Challenge 2022
Predicting gene expression using millions of random promoter sequences
by
Davuluri Lab

Abstract

In this challenge, we have utilized a transformer-based representation model named DNABERT [1] for predicting gene expression using million of the promoter sequences. We choose DNABERT, as it is pretrained on human genome so there is no chance of the data-leakage. The training data is preprocessed and converted into 6-mer sequences and then feed to the DNABERT model for finetuning the yeast promoter sequences. The detailed preprocessing step of the dataset and the model description in described in the subsequent sections. During validation, we got Spearman's rho 0.76 and Pearson's r as 0.74.

1. Description of data usage

We have preprocessed the data (both training and testing data) before feeding to our developed model. Instead of considering each base as a single token, we tokenized a DNA sequence with the k-mer representation, an approach that has been widely used in analyzing DNA sequences. The k-mer representation incorporates richer contextual information for each deoxynucleotide base by concatenating it with its following ones. The concatenation of them is called a k-mer. For example, a DNA sequence "ATGGCT" can be tokenized to a sequence of four 3-mers: {ATG, TGG, GGC, GCT} or to a sequence of two 5-mers: {ATGGC, TGGCT}. In our experiments, we set k as 6 and train the entire model. In our architecture, the vocabulary of it consists of all the permutations of the k-mer as well as 5 special tokens: [CLS] stands for classification token: [PAD] stands for padding token, [UNK] stands for unknown token, [SEP] stands for separation token and [MASK] stands for masked token. Thus, there are (4^6) + 5 tokens in the vocabulary of our model. After converting the sequence to k-mer format, we split the training and validation set with ratio of 90:10. This data is finally feed to the developed architecture for the finetuning stage.

2. Description of the model

In this challenge, we have utilized DNABERT [1] which we have finetuned with the training data. ***As the DNABERT is pretrained on the human genome, there is no point of data leakage at all.*** DNABERT applies Transformer, an attention-based architecture that has achieved state-of-the-art performance in most natural language processing tasks. DNABERT takes a set of sequences represented as k-mer tokens as input. Each sequence is represented as a matrix by embedding each token into a numerical vector. Formally, DNABERT captures contextual information by performing the multi-head self-attention mechanism on *M*

$$MultiHead(M) = Concat(head_1, \dots, head_h)W^0$$

where $$head_i = softmax\left(\frac{MW_i^Q MW_i^{K^T}}{\sqrt{d_k}}\right).MW_i^V$$

W^0 and $\{W_i^Q, W_i^K, W_i^V\}$ are the learned parameters for linear projection. *head* calculates the next hidden states of M by first computing the attention scores between every two tokens and then utilizing them as weights to sum up lines in MW_i^V. $Multihead()$ concatenates results of independent with different set of $\{W_i^Q, W_i^K, W_i^V\}$. We have used same architecture of as the BERT base model, which consists of 12 Transformer layers with 768 hidden units and 12 attention heads in each layer.

3. Training procedure

In the training stage, we have fine-tuned the model with the provided training dataset. The dataset is first preprocessed to make 6-mer sequences (refer section-1). The training is done by 4 GPUs, where the batch size in each GPU is 32. We trained for 10 epochs with learning rate 1e-4.

4. Contributions and Acknowledgements

4.1. Contributions

Name	Affiliation	Email
Zhihan Zhou	Department of Computer Science, Northwestern University	zhihanzhou2020@u.northwestern.edu
Pratik Dutta	Department of Biomedical Informatics, Stony Brook University	pratik.dutta@stonybrook.edu
Rekha Sathian	Department of Biomedical Informatics, Stony Brook University	rekha.sathian@stonybrook.edu
Pallavi Surana	Department of Biomedical Informatics, Stony Brook University	pallavi.surana@stonybrook.edu
Yanrong Ji	Division of Health and Biomedical Informatics, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine	yanrongji2021@u.northwestern.edu
Han Liu	Associate Professor of Statistics and Electrical Engineering and Computer Science, Northwestern University	hanliu@northwestern.edu
Ramana V Davuluri	Professor, Department of Biomedical Informatics, Stony Brook University	Ramana.Davuluri@stonybrookmedicine.edu

4.2. Acknowledgement

We profoundly thank Stony Brook University, Stony Brook cancer center, and Northwestern University for providing all computational and technical support throughout the challenge completion.

5. References

[1] Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics, 37(15), 2112-2120.