DREAM Challenge 2022

Predicting gene expression using millions of random promoter sequences by

SYSU-SAIL-2022

Abstract

Givens that the there are nearly 6.7 million DNA sequences in train dataset, we consider to take the most advantage of these large amounts of sequences. Inspired by Language Model, we first train a pretrain model with 3-layers Bert using part of the train data. Then, we directly build a expression predictor from the embedding that obtain from the pretrain model.

1. Description of data usage

In both pretrain stage and finetune stage, we first, extend the given sequences 150bp(upstream) and 150bp(downstream), and then padding the sequences in each batch to the same length.

(1) Pretrain stage: We sort the sequences in training dataset by their expression. And we collect the top-20% sequences to pretrain a Bert Model. As for the top-20% sequences. We randomly divide into train(n=1675053) and validation(n=88160)

(2) Finetune stage: Randomly divide into train(n=6510079), validation(n=152786) and test(n=76393).

(3) We encode the sequences using k-mers. Apart from these, we add 5 special tokens (‘[PAD]’, ‘[CLS]’, ‘[UNK]’, ‘[SEP]’, ‘[MASK]’). Totally we get 130 tokens.

2. Description of the model

BertEmbeddings
BertLayer
BertLayer
BertLayer
BertPooler
Dropout
Linear

3. Training procedure

(1) Pretrain Stage: We mask part of each sentence and compute CossEntropyLoss for each sentence. Learning rate during pretrain stage first set to 0.0001 and use *torch.optim.lr_scheduler.ReduceLROnPlateau* to Reduce learning rate when a metric has stopped improving. For pretrain stage, we get mask part AUC=0.196

(2) Finetune Stage: During finetune stage, we input the output embedding of the pretrain model, compute the mseloss for each sequences. The optimizer we used is Adam. The learning rate in this stage is the same as pretrain stage. In final finetune, we get pearson score=0.7433, spearman score=0.7544

4. Other important features

None

5. Contributions and Acknowledgement

Name	Affiliation	Email
Ding Maolin	---	dingmlin3@mail2.sysu.edu.cn
Chen Ken	---	chenk87@mail2.sysu.edu.cn

6. References