Predicting gene expression using millions of random promoter sequences
DREAM Challenge 2022
SBME UBC
IBM Research
Google Research TPU Research Cloud
SageBionetworks

DREAM Challenge 2022

Predicting gene expression using millions of random promoter sequences by

auth

Abstract

To approach the problem, we explored many computational tools that have been developed by successfully applying deep learning techniques on genomic sequence data to study cis-regulatory logic[1,2]. To represent the input sequences, we experimented with a k-mer embedding computed using word2vec[3,4] and the standard one-hot encoding. We tried to adapt methods that are based on convolutional neural networks (CNNs), others that use recurrent neural networks (RNNs) and others that rely on hybrid architectures combining CNNs and RNNs. Our best performing model and the one submitted here takes one-hot encoded sequences as input and adopts a simple hybrid architecture combining a convolutional layer, a bidirectional LSTM layer and two fully connected layers. Furthermore, we were interested in exploring the possibility of applying powerful language representation models like transformers to the field of genomics. Inspired by previous works, we tried to adapt the idea of Bidirectional Encoder Representations from Transformers (BERT) model to DNA setting[5,6]. The limitations to fully implementing this idea were that both the pre-training and the finetuning of a BERT-like model are resource intensive (lasting for days even on TPUs) and we came up with this approach only a few days before the submission deadline. As we did not manage to train and test a BERT-like model on the challenge's data, we are curious about the potential of such an approach for gene expression prediction.

1. Description of data usage

We chose to use the 110-length sequences that did not contain any 'N' characters, limiting the original dataset to 5511002 sequences. We removed the flanking regions from each sequence and kept the random 80 bp region. Then, we performed one hot encoding. We performed train/test split by using *sklearn*'s '*train_test_split'* function and choosing 0.015 as '*test_size*'. The input shape for train and test/validation datasets are presented here: *X_train*: (5428336, 80, 4), *X_test*: (82666, 80, 4), *y_train*: (5428336, 1), *y_test*: (82666, 1).

2. Description of the model

- input_layer: Input (shape=(80,4))
- Convolution (filters=1000, kernel size=30, activation='relu')
- MaxPool (pool_size = 3, strides = 3)
- Dropout (0.2 probability)
- Bidirectional LSTM (320 units)

- Dropout (0.2 probability)
- Dense (64 units, activation='relu')
- Dense (64 units, activation='relu')
- output_layer: Dense (1 unit, activation='relu')

3. Training procedure

- Loss function: mean squared error
- Regularization: Dropout
- Optimizer: Adam
- Learning rate: -
- Batch size: 1024
- Epochs: 10
- The Pearson correlation coefficient (calculated with the function '*scipy.stats.pearsonr*') on our own test/validation set was ~ 0.73.

4. Other important features

-

5. Contributions and Acknowledgement

5.1 Contributions

Name	Affiliation	Email
Konstantinos Kardamiliotis	Laboratory of Biological Chemistry, Medical School, Aristotle University of Thessaloniki, GR-54124 Thessaloniki, Greece	k.kardamiliotis@gmail.com
Konstantinos Kyriakidis	Laboratory of Biological Chemistry, Medical School, Aristotle University of Thessaloniki, GR-54124 Thessaloniki, Greece	kokyriakidis@gmail.com
Andigoni Malousi	Laboratory of Biological Chemistry, Medical School, Aristotle University of Thessaloniki, GR-54124 Thessaloniki, Greece	andigoni@gmail.com

5.2 Acknowledgement

-

6. References

1. Trabelsi, A., Chaabane, M. & Ben-Hur, A. Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities. Bioinformatics 35, i269–i277 (2019).
2. Novakovsky, G., Fornes, O., Saraswat, M., Mostafavi, S. & Wasserman, W. W. ExplaiNN: interpretable and transparent neural networks for genomics. bioRxiv 2022.05.20.492818 (2022) doi:10.1101/2022.05.20.492818.
3. Asgari, E. & Mofrad, M. R. K. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. PLoS One 10, e0141287 (2015).
4. Distributed Representations of Words and Phrases and their Compositionality. https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

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

6. Bringing BERT to the field: Transformer models for gene expression prediction in maize | by Zihao Xu | Towards Data Science. https://towardsdatascience.com/bringing-bert-to-the-field-how-to-predict-gene-expression-from-corn-dna-9287af91fcf8.

7. Feedback (optional)

This was our first time experimenting with neural networks architectures and deep learning and we would like to thank you for motivating us to enter this fascinating world!