DREAM Challenge 2022

Predicting gene expression using millions of random promoter sequences

by

BMDS Lab (Queensland University of Technology, Australia)

Abstract

The submission is based on a simple idea: apply the dna2vec sequence embedding model [1] to the promoter sequences (in a running manner on short k-mers), which are then subsequently used as features for a transformer-based deep neural network model. As the dna2vec embeddings are expected to be roughly additive (e.g., embed('CAT') + embed('GGA') has the same embedding as 'CATGGA' or 'GGACAT'), the motivation for such a feature extraction approach for the provided sequences was that such additive properties could yield the ability to learn difficult patterns.

1. Description of data usage

Split: The data were divided into a training and validation set. The first 5.72 million observations comprised the training set, and the rest the validation set.

Encoding: The promoter sequences were converted into sequences of length-100 real-valued vectors by encoding running 3-mers each as a 100-dimensional vector using the dna2vec algorithm [1]. All sequences were padded with zeros at the end where necessary so to meet the length of the maximum sequence length (110).

2. Description of the model

Following a positional encoding layer, three (3) sequential transformer encoder [2] (i.e., self multi-head attention followed by a feedforward layer) were used. Each multi-head attention had 10 heads, and the each feedforward layer had 1000 neurons. ReLU activation functions were used.

3. Training Procedure

Loss Function: Mean-Squared Error
Regularization: None
Optimizer: Adam, with default PyTorch parameters

4. Other Important Features

N/A

5. Contributions and Acknowledgement

5.1. Contributions

Name	Affiliation	Email
Jake Bradford	Queensland University of Technology	jake.bradford@qut.edu.au
Dimitri Perrin	Queensland University of Technology	dimitri.perrin@qut.edu.au

Robert Salomone	Queensland University of Technology	robert.salomone@qut.edu.au
Carl Schmitz	Queensland University of Technology	c.schmitz@qut.edu.au

5.2. Acknowledgement

None.

6. References

[1] Ng, Patrick. "dna2vec: Consistent vector representations of variable-length k-mers." *arXiv preprint arXiv:1701.06279* (2017).

[2] Vaswani, Ashish, et al. "Attention is all you need." *Advances in Neural Information Processing Systems* 30 (2017).