Your Paper

You

December 1, 2023

Abstract

Your abstract.

1 Introduction

TODO

2 Use-case 1

Transformers have been used to encode proteins as vectors in a high-dimensional space, where sequences with similar properties are mapped closely together. This study [1] demonstrates that after training, the representation space of a transformer model clusters orthologous genes effectively. This is visualized through t-SNE and PCA (figure 1), which show that species and orthology become principal axes of variation, indicating that unsupervised learning captures biological variations. This learned structure allows for improved recovery of proteins based on biological properties using vector similarity queries, confirming that biological information is encoded within the representation space.

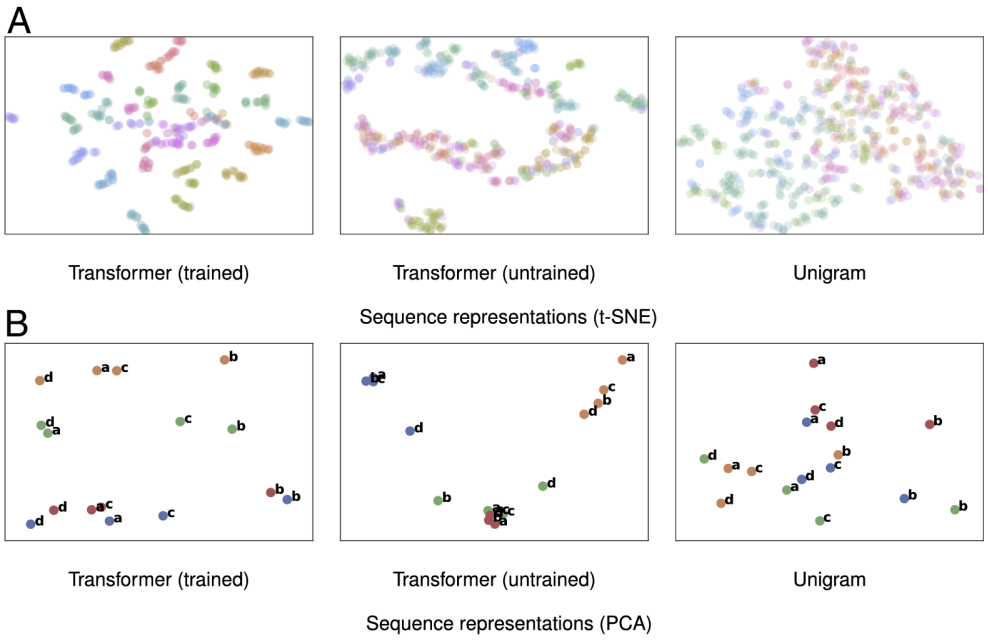

Figure 1: Enformer's analysis of a genetic variant (rs11644125 C/T) showing its impact on NLRC5 gene expression, with model predictions suggesting the T allele reduces expression, possibly through altered SP1 transcription factor binding, as validated by cap analysis gene expression (CAGE) data in peripheral blood mononuclear cells (PBMCs).

3 DNABERT

DNABERT is a bidirectional encoder pre-trained on genomic DNA sequences with up- and downstream nucleotide contexts. For short DNA sequences

4 Enformers

The Enformer model [2] utilizes transformer modules, known for their effectiveness in natural language processing, to analyze DNA sequences. Its key features include:

- Transformer Layers: These enable the model to consider each part of the DNA sequence in relation to the entire sequence, crucial for integrating distant genomic elements.
- Extended Receptive Field: Enformer can analyze elements up to 100 kb from the transcription start site, much further than previous models, allowing it to capture a broader range of regulatory elements like distant enhancers.
- Attention Mechanism: This allows the model to weigh different parts of the sequence differently, depending on their relevance to gene expression.

By utilizing existing gene expression data, Enformer can be adapted to Dicty's genome, enabling predictions of gene expression levels based on genomic sequences.

4.1 Effects of genetic variants

Enformer is a computational model that can predict the effects of genetic variants on gene expression in a cell-type-specific manner, which is valuable for fine-mapping noncoding associations from genome-wide association studies (GWAS). Figure 2 depicts Enformer's predictive analysis of a genetic variant (rs11644125 C/T) showing its impact on NLRC5 gene expression, with model predictions suggesting the T allele reduces expression, possibly through altered SP1 transcription factor binding, as validated by cap analysis gene expression (CAGE) data in peripheral blood mononuclear cells (PBMCs).

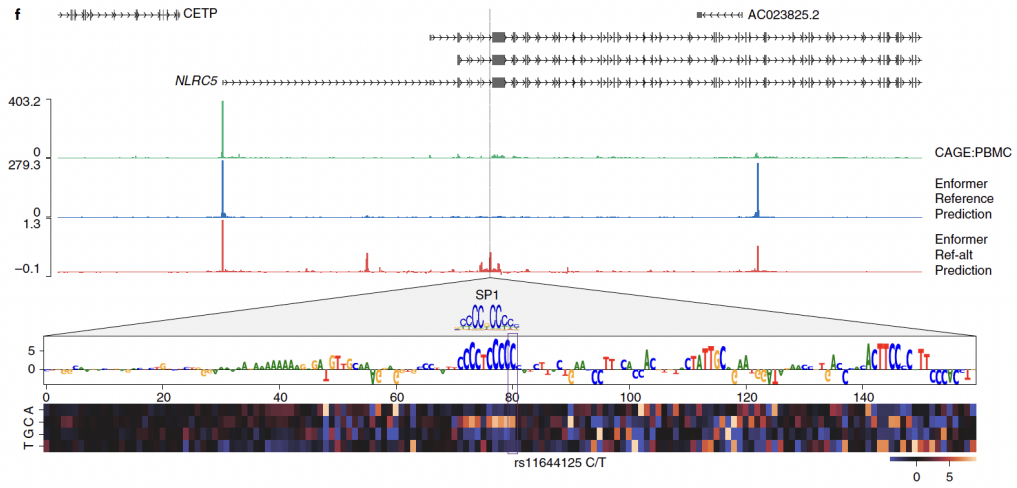

Figure 2: Enformer's analysis of a genetic variant (rs11644125 C/T) showing its impact on NLRC5 gene expression, with model predictions suggesting the T allele reduces expression, possibly through altered SP1 transcription factor binding, as validated by cap analysis gene expression (CAGE) data in peripheral blood mononuclear cells (PBMCs).

References

[1] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, April 2021.

[2] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203, October 2021.