Notes to paperes

Some subtitle for my thesis

Johan Ulstrup

July 9, 2025

Table of contents

1 index 1

2 DNA language models are powerful predictors of genome-wide variant effects 3

3 Genome-wide coancestry reveals details of ancient and recent male-driven reticulation in baboons 5

4 References 7

1 index

2 DNA language models are powerful predictors of genome-wide variant effects

The idea for this project was to evaluate the Genome-Pretrained Network (GPN) introduced in (Kantorovitz et al. 2024) and determine whether it could achieve greater accuracy than traditional methods in predicting genome-wide variant effects.

The model is designed as a convolutional neural network (CNN) and takes input sequences with a window size of 512. During training, 15% of the positions within each window are masked to enable the model to learn meaningful representations.

The architecture consists of 25 layers, each structured as follows: a dilated convolution layer, followed by an add-and-norm layer with a skip connection from before the dilated convolution. This is followed by a feedforward layer, another add-and-norm layer, and additional skip connections.

A feedforward layer is a fundamental component of neural networks, where inputs pass through one or more fully connected layers with activation functions, transforming data without looping back. This structure helps the model learn complex representations by applying weighted transformations and non-linearities.

A dilated convolution expands the receptive field of a convolutional layer without increasing the number of parameters or reducing resolution. By spacing kernel elements apart, it captures long-range dependencies in sequences, making it particularly useful in genomic data analysis. When combined, dilated convolutions and feedforward layers enhance a model's ability to recognize patterns across different scales efficiently.

After passing through the 25 layers, the model produces a contextual embedding with a dimension of 512 (D=512), followed by classification layers. The final layer outputs the probabilities of the four nucleotides at each masked position.

The GPN variant effect prediction score is calculated as the log-likelihood ratio between the alternate (ALT) and reference (REF) alleles. Here, L represents the window length in base pairs, and D denotes the embedding dimension.

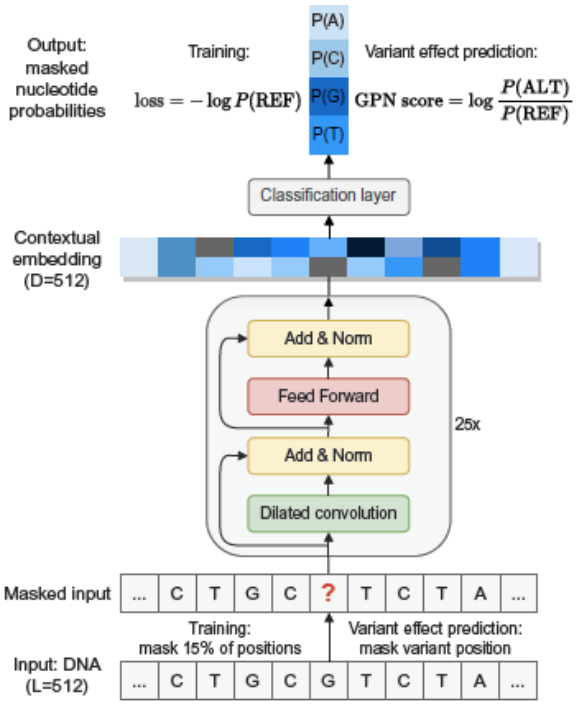
Output: masked nucleotide probabilities
Training:
$\text{loss} = -\log P(\text{REF})$
P(A)
P(C)
P(G)
P(T)
Variant effect prediction:
$\text{GPN score} = \log \frac{P(\text{ALT})}{P(\text{REF})}$
Classification layer
Contextual embedding (D=512)
Add & Norm
Feed Forward
Add & Norm
Dilated convolution
25x
Masked input
... C T G C ? T C T A ...
Training: mask 15% of positions
Variant effect prediction: mask variant position
Input: DNA (L=512)
... C T G C G T C T A ...

Figure 2.1

3 Genome-wide coancestry reveals details of ancient and recent male-driven reticulation in baboons

(Knuth 1984)

```
# Here is example python code
print("Hello world")
```

Here is a reference (Nielsen and Slatkin 2016)

4 References

Kantorovitz, Mitchell R. et al. 2024. "DNA Language Models Are Powerful Predictors of Genome-Wide Variant Effects." *Proceedings of the National Academy of Sciences* 121 (3): e2313724121. https://doi.org/10.1073/pnas.2313724121.

Knuth, Donald E. 1984. "Literate Programming." *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.

Nielsen, Rasmgb, and Montgomery Slatkin. 2016. *An Introduction to Population Genetics: Theory and Applications*.