Can genome-based Large Language Models predict gene expression?

Sofiane Sadat, Arnaud Ferré, Guillaume Kon Kam King, Sofia Lotfi

Can genome-based Large Language Models predict gene expression?

Sofiane SADAT[1,2], Arnaud FERRÉ[2], Guillaume KON KAM KING[2], Sofia LOTFI[1]

1.Objective

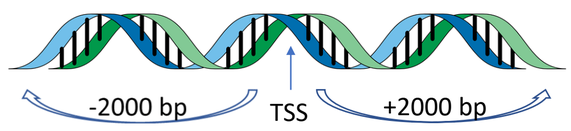
This study aims to determine whether genome-based large language models can estimate gene expression as accurately as models specifically developed for this purpose. Our LLM approach involves using a representative large language model, **DNABERT-2[1]**, to predict the median gene expression in the human pituitary. We will then compare our results with those obtained using **DExTER[2]**, a method specially designed for estimating gene expression.

2. Retrieval of Gene expression data

Gene Sequences from:

e!GRCh37

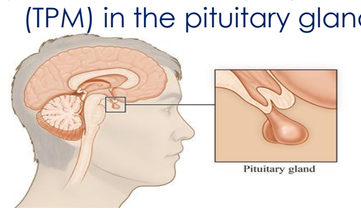
- More than 22 000 protein coding genes
- We take only 4000 base pair centred around the TSS (transcript start site)



Gene Expression from:

GTEx Portal

Log-transformed median expression in transcripts per million (TPM) in the pituitary gland



Gene Sequence	Gene expression
ATCGATCCATGGATAAATTTTATATCGATAAAAAAAACC...	1.5
ATAAATTTTAATAAATTTTAATAAATCCCCCCCCCCTTTA...	2
CGATCCATGGATAATAAATGGGGGGGGAAATAAATGAA...	1

3. DExTER Domain Exploration To Explain gene Regulation

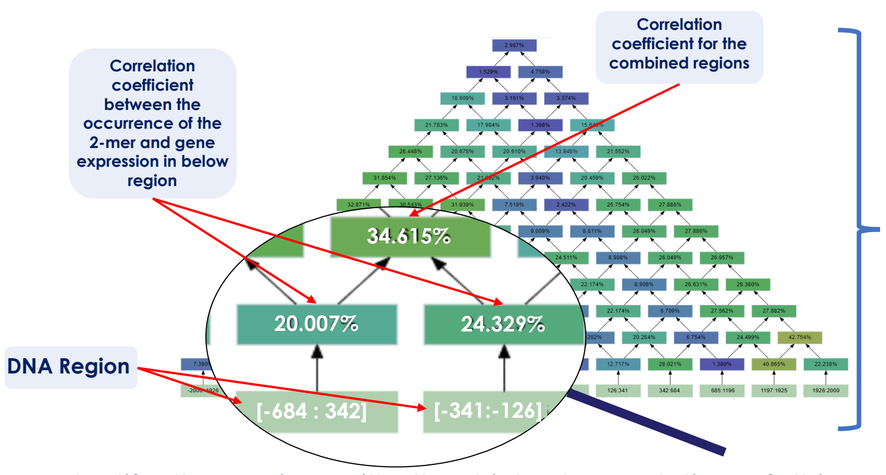
General idea: By calculating the occurrence of k-mers (AA, AT, AC,... AAA,...) in sub-segments of the 4000 base pairs, the aim is to predict the gene expression.

Methodology

1. Feature extraction

- Segmentation : 1st step is to segment the 4000 bp of the genes into 13 bins.

- For each k-mer calculate a lattice, Example :



- Identify the region with the highest correlation. If this correlation is significantly higher than the top node's correlation, select the region with the associated k-mer.
- Expansion: Perform the same process for (k+1)-mers to identify (k+1)-mers with higher correlations in the identified region than the original k-mers. Repeat this step until no further improvement is observed.
- Return the list of identified variables, where a variable is simply a pair of a k-mer and its associated region. Example :

The k-mer: AA — DNA Region: [-684:-342]

2. Feature selection and learning

- Train a LASSO regression (a robust predictive model) with the normalized occurrence of the identified variables as features and expression as the target variable for each gene in the training set.

	sequence	expression	AA.-125_684	AA.-2000_2000	AC.-2000_2000
0	ENSG00000243485.2	-1.342084	0.071605	0.066983	0.045489
1	ENSG00000233750.3	-0.772885	0.066667	0.070232	0.048738
2	ENSG00000237973.1	1.312389	0.081481	0.066233	0.024744
3	ENSG00000230368.2	-0.695509	0.034568	0.053487	0.053237
4	ENSG00000187961.9	1.330819	0.014815	0.025744	0.044739
...	...	...	...	...	...

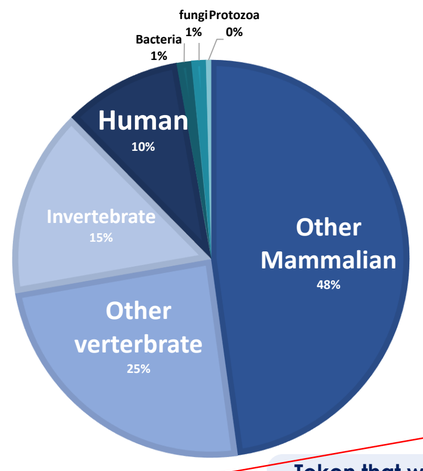
Target variable

Features: occurrences of the k-mers in corresponding regions

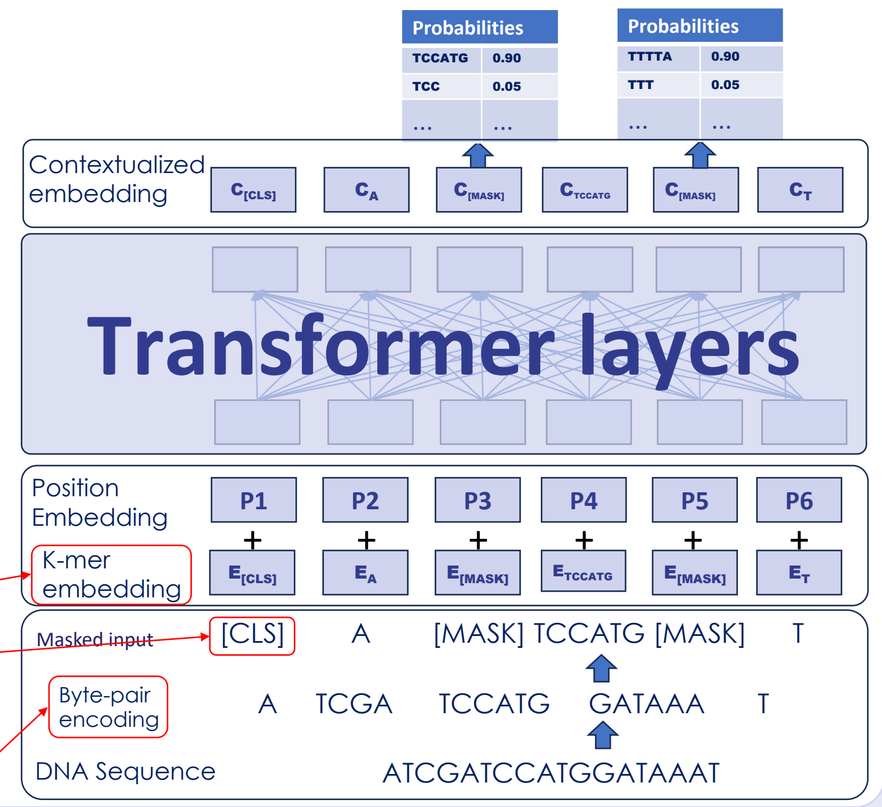
4. DNABERT-2 is an LLM trained extensively on genetic sequences. DNABERT2 learns to understand those sequences by predicting masked parts of the sequence based on the context of the surrounding information, a process known as Masked Language Modeling (**MLM**). We adjust the model further to better suit the specific task, a process typically referred to as **fine-tuning**. We hypothesize that DNABERT-2 encodes meaningful representations of DNA sequences that are also relevant to predict genetic expression.

MLM Training corpus

Proportion of DNA Used in DNABERT-2 Training Corpus by Organism Type



How large language models are trained



convert one-hot encoded tokens to embeddings by multiplying the one-hot encoded vectors with an embedding matrix.

Token that will represent the entire sequence

Tokenizes the most frequent sub-sequences.

5. Experimentation

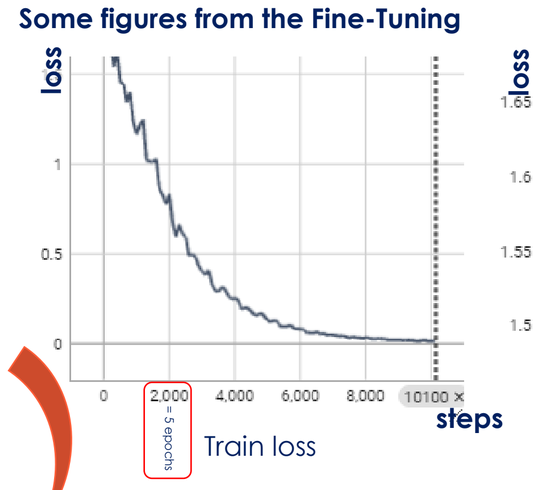
DExTER

We used the same dataset for human pituitary described in the paper and obtained similar results.

DNABERT-2 fine-tuning with a new training objective: predict gene expression

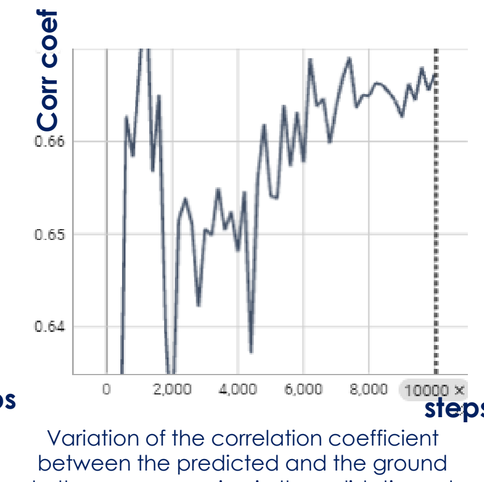
The fine-tuning process involved optimizing the model weights of the dense neural network connected to the contextualized embedding of the [CLS] token using a mean squared error loss function. We fine-tuned the model over 25 epochs

Some figures from the Fine-Tuning



Train loss

Validation loss



Variation of the correlation coefficient between the predicted and the ground truth gene expression in the validation set

6. Results

Train test datasets proportion

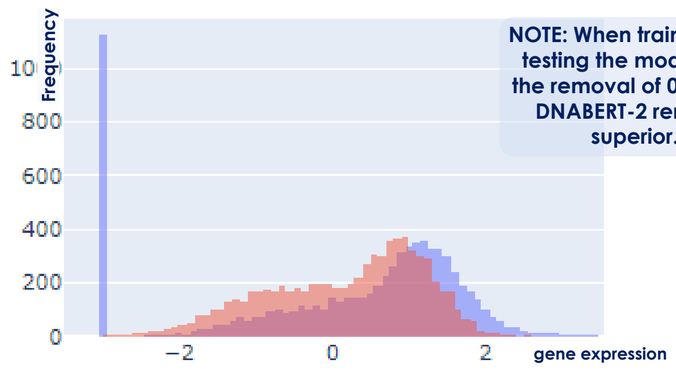
Train/Validation set	Test set
14 940 (2/3)	7 470 (1/3)

Dexter vs DNBERT-2 evaluation metrics

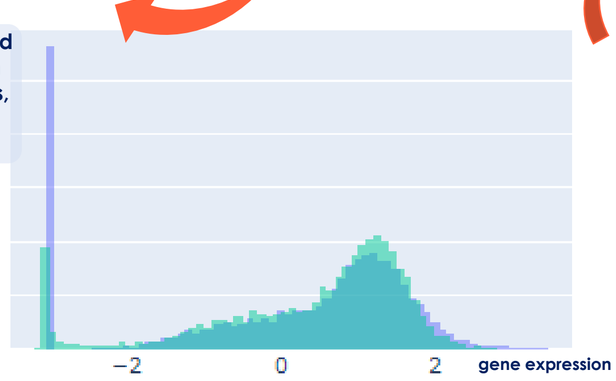
	DExTER	DNABERT-2
Pearson correlation coefficient	0.64	0.69

Prediction distribution in comparison to ground truth

NOTE: When training and testing the model with the removal of 0 values, DNABERT-2 remains superior.



DExTER prediction distribution vs ground truth distribution



Dnabert-2 prediction distribution vs ground truth distribution

7. Conclusion

The results from our study confirm that DNABERT-2 provides a way of representing sequences which is efficient for predicting gene expression. Impressively, its performance is on a par with methods that are specifically designed for gene expression prediction. This highlights the importance of genome-based large language models like DNABERT-2, creating new paths for future research.

References

[1] Zhou Z, Ji Y, Li W, Dutta P, Davuluri R, Liu H. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome [Internet]. arXiv; 2024 [cité 30 avr 2024]. Disponible sur: http://arxiv.org/abs/2306.15006
[2] Menichelli C, Guitard V, Martins RM, Lèbre S, Lopez-Rubio JJ, Lecellier CH, et al. Identification of long regulatory elements in the genome of Plasmodium falciparum and other eukaryotes. PLoS Comput Biol. avr 2021;17(4):e1008909.

SERVIER[1] Institut de R&D Servier
X
INRAE[2] Université Paris-Saclay, INRAE