An mRNA and protein level transformer-based predictor

Emanuele Frascaroli
University of Modena and Reggio Emilia
emanuele.frascaroli@unimore.it

Mathurin Videau
Inria Paris Saclay
mvideau@meta.com

Vittorio Pippi
University of Modena and Reggio Emilia
vittorio.pippi@unimore.it

Vincenzo Mariano Scarrica
University of Naples, Parthenope
vincenzomariano.scarrica001@studenti.uniparthenope.it

Brandon Wily Viglianisi
University of Modena and Reggio Emilia
brandon.viglianisi@unimore.it

Teemu Sarapisto
University of Helsinki
teemu.sarapisto@helsinki.fi

Abstract

This research investigates the application of Transformer-based architectures in predicting gene and protein expression levels directly from genomic and proteomic sequences. By combining deep learning (DL) and natural language processing (NLP) techniques, we exploit Transformers' capacity to capture intricate sequence patterns. We train our models from scratch on the Xpresso dataset. Our investigation reveals that randomly masking 5% of the DNA sequence within the Xpresso dataset does not significantly affect the overall predictive accuracy (r^2), highlighting the robustness of the employed architecture in handling sequence variations.

1. Introduction

Accurate prediction of gene and protein expression levels from genomic and proteomic sequences has been a transformative effort in genomics research. The development of algorithms for predicting DNA characteristics solely based on primary sequence data marked a significant milestone [1]. Building upon this progress, the question arose: can we extend this predictive capability to elucidate gene expression levels directly from genome sequences? This inquiry led to the emergence of approaches, including the renowned Xpresso model, which used deep convolutional neural networks to predict steady-state mRNA levels with remarkable accuracy, even surpassing models relying on chromatic immunoprecipitation sequencing (ChIP-seq) data [1].

While Xpresso represented a groundbreaking advance, the pursuit of enhancing our ability to predict gene and protein expression levels has continued. In a contemporary context, the field has witnessed the emergence of novel approaches, such as the Perceiver architecture [2]. This framework addresses the shortcomings of previous models by leveraging a Transformer-based [5] architecture with an attention mechanism that enables the capture of long-range interactions within sequences, without the quadratic complexity associated with standard Transformer architectures [3]. Recent work of Stefanini et al. [3] introduces not only the DNAPerceiver model for mRNA level prediction but also the ProteinPerceiver model for protein level prediction, and a combined Protein&DNAPerceiver model to predict protein levels based on both Transcription Start Site (TSS) and protein sequences [3].

In light of these advancements, our research extends the utility of the Perceiver [2] architecture in the context of gene and protein expression prediction. Specifically, we explore the efficacy of this architecture in conjunction with various data augmentation techniques to mask specific portions of DNA sequences. This study aims to shed light on the impact of sequence masking on the predictive performance, elucidating whether the robustness demonstrated by the Perceiver architecture remains intact when sequence variations are introduced. In doing so, we contribute to the ongoing dialogue surrounding the prediction of gene and protein expression levels, ultimately striving to enhance our under-

standing of the intricate regulatory mechanisms governing gene expression.

2. Methodology & materials

In this study, we employed a variation of the Transformer architecture known as the Perceiver to predict gene and protein expression levels from mRNA and protein sequences. The Perceiver architecture has demonstrated its effectiveness in various machine learning tasks, including sequence-to-sequence tasks in the field of genomics. For our experiments, we utilized mRNA sequences with a fixed length of $20,000$ nitrogenous bases as input data. To enhance the model's ability to capture intricate patterns within these sequences, we applied a modification step to the mRNA sequences before feeding them into the Perceiver architecture. In particular, we randomly mask consecutive nitrogenous bases within each mRNA sequence as a data augmentation technique. This masking process was carried out to simulate the inherent noise and complexity present in biological data, as well as to challenge the model to extract relevant features from partially obscured sequences. It can also be considered a form of regularization, which is beneficial due to the scarcity of biological data. We drew inspiration from the work on the Perceiver architecture in our study, and as a part of our model evaluation, we utilized the "explained variance" r^2 as a key metric. Much like the Perceiver architecture's ability to comprehend complex relationships in diverse data domains, our use of explained variance as an evaluation metric helped us gauge the models' proficiency in capturing and explaining the variance present in our genomic and proteomic datasets. For all of the experiments conducted in this study, we employed a training protocol consisting of 400 epochs, with the inclusion of early stopping criteria following hyper-parameters from Stefanini et al. [3].

3. Ablation study

To further evaluate the impact of sequence modification on the predictive performance of the Perceiver architecture in our study, we conducted an ablation study by exploring different masking strategies. In addition to the initial experiment where we randomly masked $1,000$ consecutive nitrogenous bases within the mRNA sequences, we systematically investigated three additional masking strategies.

In the first variant, we masked the first half ($10,000$ bases) of the sequences, effectively removed the early portions of the genetic information. In the second variant, we focused on the intermediate section of the sequences and masked the central $10,000$ bases. Finally, in the third variant, we masked the last $10,000$ bases of the mRNA sequences, emphasizing the latter portion of the genetic code. Here the data are suppressed and the model has been modified in order to accept the new dimension. The rest of the experiments have been made in such a way that for different masking size the length of the sequence has remained unchanged, fixed to $20,000$. For every experiment, 10 runs have been launched and the values have been average all over the number of runs.

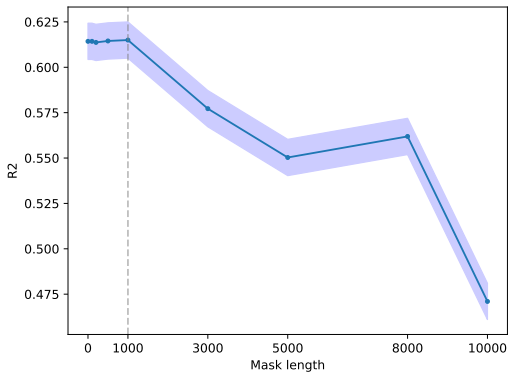

Figure 1. This graph represents on the y-axis the mask size applied to the sequences, while on the x-axis, it shows the explained variance r^2 as a function of the mask size.

4. Hardware & implementation

For our experiments, we harnessed the computational power of the Leonardo supercomputer from CINECA, Italy's premier supercomputing center. Leonardo, boasting state-of-the-art hardware and substantial computational capabilities, provided the ideal infrastructure for our research.

To efficiently utilize this high-performance computing resource and facilitate our deep learning experiments, we employed the PyTorch library, a powerful framework for machine learning and deep neural network development. Leveraging PyTorch's flexibility and GPU support, we orchestrated our experiments to run in parallel across a formidable array of 160 GPUs. This parallelized approach not only accelerated the computational processes but also ensured the scalability and efficiency required for our gene and protein expression prediction tasks. The combined utilization of PyTorch and the Leonardo supercomputer underscored our commitment to employing cutting-edge tools and infrastructure to advance our research in genomics and bioinformatics.

5. Results & discussion

As shown in Table 1 the results of our study provide some evidence in support of our thesis that the strategic use of a shorter mask, covering only $1,000$ bases within the genomic sequences, yields improved predictive performance compared to larger masks or the absence of masking the se-

Masking strategies	$\mathbf{r^2}[\mu \pm \sigma]$
Original (no mask)	0.81 ± 0.003
Masking (first $10,000$ bases)	0.81 ± 0.003
Masking (intermediate $10,000$ bases)	0.81 ± 0.003
Masking (last $10,000$ bases)	0.81 ± 0.003
Masking (random consecutive 1000 bases)	\$0.81 $\pm$0.003

Table 1. Expressed variance and epochs of the experiments, listed for each dataset employed in the manuscript

quence altogether. This finding aligns with a fundamental concept in deep learning, inspired by models like BERT [4], which demonstrates the utility of injecting controlled noise into input data to enhance a model's ability to generalize and capture important features.

By applying a $1,000$-base mask to the mRNA sequences, we effectively introduced a controlled level of noise into the data. This noise mimics the inherent complexity, uncertainty, and possible ambiguity present in biological sequences, challenging the model to focus on critical sequence segments and learn meaningful representations. Larger masks covering substantial portions of the sequence or the absence of masking may have overwhelmed the model with excessive noise, hindering its capacity to discern essential patterns and relationships within the data and this way hurting the model.

The strategic use of a shorter mask, covering only $1,000$ bases within the genomic sequences, not only improved predictive performance but also served as a form of regularization in our study. Regularization techniques are widely employed in machine learning to prevent overfitting and enhance the generalization capabilities of models. In our case, the masking of sequences can be viewed as a unique form of regularization tailored to genomics.

In the context of the work, as depicted in Figure 1, we see that the best-performing mask size corresponds to $1,000$ bases. This demonstrate the advantages of masking as a regularization technique.

6. Conclusion

This study employed Transformer-based architectures, inspired by the Perceiver model, to predict gene and protein expression from genetic sequences. The introduction of a $1,000$-base mask into mRNA sequences, following noise injection principles akin to BERT, improved predictive accuracy and served as an effective regularization technique. A thorough ablation study indicated that alternative masking strategies did not significantly enhance performance, highlighting the importance of appropriate noise levels. This research showcase Transformer models' potential and the utility of noise injection for more accurate gene and protein expression prediction, with implications for understanding complex biological processes and diseases.

References

[1] Vikram Agarwal and Jay Shendure. Predicting mrna abundance directly from genomic sequence using deep convolutional neural networks. *Cell reports*, 31(7), 2020. 1

[2] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 1

[3] Matteo Stefanini, Marta Lovino, Rita Cucchiara, and Elisa Ficarra. Predicting gene and protein expression levels from dna and protein sequences with perceiver. *Computer Methods and Programs in Biomedicine*, 234:107504, 2023. 1, 2

[4] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019. 3

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1