Exploring hierarchical models for gene expression prediction from sequence

Kanav Mittal[1], Ruchir Rastogi[1], Nilah M. Ioannidis[1]

[1]University of California, Berkeley, CA, USA.

Introduction

Recently, several deep learning models have been developed to predict gene expression from the surrounding regulatory DNA sequence. To better capture regulatory grammar, these models are often trained not only to predict gene expression, but also chromatin regulation data, such as transcription factor binding, DNA accessibility, and histone modifications. Two main architectural approaches have emerged to jointly incorporate gene expression and regulatory data. The first approach, epitomized by Basenji [1, 2], Enformer [3], and Borzoi [4], predicts both gene expression and regulatory data in a multitask fashion. Conversely, ExPecto [5] proposes a hierarchical strategy. In ExPecto, a convolutional neural network (CNN) called Beluga (a successor to DeepSEA [6]) first predicts regulatory features from DNA sequence in small 200 bp bins. Then, these predicted regulatory features from a broader genomic context (40 kb surrounding the TSS) are integrated using predefined spatial basis functions and input into tissue-specific regularized linear models to predict gene expression.

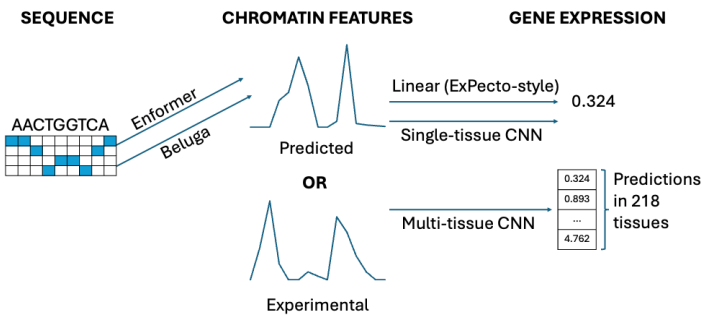


Fig. 1: Overview of hierarchical architectures explored in this paper. Enformer and Beluga models are used to predict epigenomic regulatory features from DNA sequence. These predictions (or experimental regulatory features) are input to ExPecto-style tissue-specific linear models or single-tissue CNNs predicting tissue-specific gene expression, or to multi-tissue CNNs predicting gene expression over all 218 tissues.

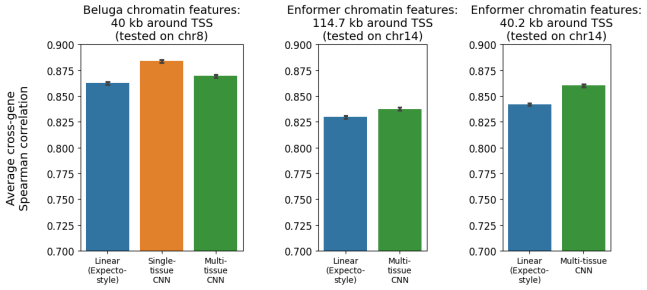
This hierarchical approach, which decomposes the sequence-to-expression prediction problem into a sequence-to-regulation module and a regulation-to-expression module, has several potential advantages. First, it may better approximate the biology, as most expression changes are thought to be mediated by corresponding regulatory changes. Second, we hypothesize that predicting regulatory marks requires only local DNA sequence context, whereas predicting gene expression requires a broader context. Attempting to predict both simultaneously may lead to negative transfer, potentially explaining previous findings that multitask models struggle to capture distal regulatory effects [7]. Third, having two distinct modules allows users to train custom regulation-to-expression models on new gene expression datasets without retraining the sequence-to-regulation model.

Despite these potential advantages, there is limited follow-up work on hierarchical modeling of gene expression. In this paper, we introduce two new regulation-to-expression architectures: single and multi-tissue CNNs. We compare their performance to the linear model proposed in ExPecto across many settings, including using experimental versus predicted regulatory marks as inputs, using predicted regulatory marks from Beluga or Enformer, and using regulatory marks from different genomic context sizes.

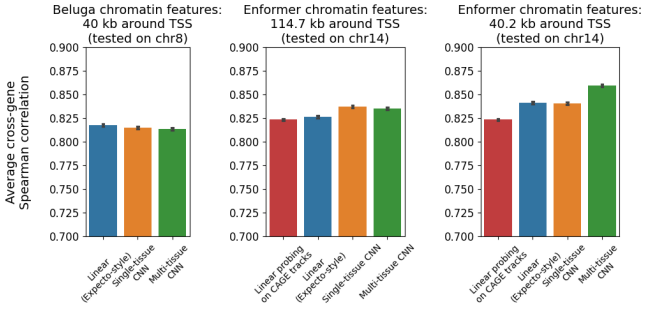
Methods

We used Beluga and Enformer to compute predictions for regulatory marks for a dataset of genes centered around the TSS (same dataset used to train the original ExPecto). We then trained ExPecto-style linear models (tissue-specific ridge regression models) and single and multi-tissue convolutional neural networks, using either Beluga/Enformer-predicted or experimentally-measured regulatory tracks as inputs and 218 tissue expression profiles as output targets. ExPecto-style linear models used the same spatial basis functions as ExPecto. Linear models and single-tissue CNNs were trained to predict a scalar gene expression value for an input gene while multi-tissue CNNs predicted a vector of 218 gene expressions. Figure 1 shows the setup for the hierarchical architecture. We employed a hyperparameter search over ridge regression penalty for the linear models and learning rate and dropout for the CNNs. Additionally, for Enformer models, we trained models that took in as input the full Enformer output context length (114,688 bp) and models that took in as input a shorter Enformer output context length (40,192 bp) to be comparable to Beluga.

We also trained a set of tissue-specific linear models on the CAGE tracks from Enformer predictions. Specifically, we extracted the middle four bins around the TSS and averaged their CAGE expression values for each of the CAGE tracks.



(a) Experimental regulatory features



(b) Predicted regulatory features

Fig. 2: **Performance of expression models using experimentally-measured regulatory features 2a and regulatory features predicted by Beluga and Enformer 2b**. For each of the 218 tissues, we calculate the Spearman correlations between expression predictions and experimentally-measured expression across all test set genes, and plot the average across all tissues. Error bars reflect ±1 standard error.

Results

Error Decomposition. When predicting gene expression from predicted Beluga regulation features, we can decompose the error into error from Beluga and error from the hierarchical linear/CNN models. The error from the hierarchical models specifically can be approximated by looking at the performance of the model when using experimental regulation features.

As seen in Figures 2a and 2b, CNNs outperform linear models with experimental Beluga regulation features - single-tissue CNN achieves an average Spearman coefficient of 0.884 versus linear's 0.862. However, CNNs achieve slightly lower performance than linear models when using predicted Beluga regulation features - single-tissue CNN achieves average Spearman coefficient of 0.815 versus linear's 0.818. The drop in performance of the CNNs can therefore be primarily attributed to the error from Beluga's predictions of regulation features. In particular, the drop in performance of CNNs relative to linear models reflects how the performance of more complex models like convolutional networks is more sensitive to errors in their inputs than simpler models like linear models. On the other hand, with Enformer, we see very little change in performance when going from using experimental regulation features to Enformer-predicted regulation features, shown in Figures 2a and 2b. For example, multi-tissue CNNs using a shorter context window go from an average Spearman coefficient of 0.8601 to 0.8596, and linear models using a shorter context window go from 0.8418 to 0.8417. Thus, most of the error with Enformer hierarchical models is coming from the hierarchical linear and CNN models rather than the Enformer model itself.

Improving Expression Prediction. With Enformer-predicted regulation features, moreover, we see CNNs outperform linear models (Figure 2b). With fewer errors in the epigenomic outputs of Enformer, we see that using deeper, more complex networks like CNNs can work well. Less error in the inputs (epigenomic regulation features) to the hierarchical models means that less error is being propagated/ amplified by complex architectures. More complex hierarchical models like CNNs may do a better job at extracting signal from these intermediate regulation features with less error. Indeed, we find that hierarchical epigenomic-to-expression models trained on Enformer's predictions, particularly CNNs, outperform simple baselines that manipulate the original Enformer model's expression outputs i.e., the CAGE tracks (Figure 2b). While linear probing of CAGE tracks achieved an average Spearman coefficient of 0.824, we see that when using a shorter context window, linear models achieve 0.842 and multitask CNNs 0.860. Unlike the Enformer model where epigenomic tracks and gene expression are predicted together, hierarchical models can extract additional signal from the predicted epigenomic tracks to help make predictions for gene expression.

We also see that using a shorter input context for the hierarchical models built on top of both experimental and predicted Enformer regulation features does not harm performance. Rather, in many cases, the Spearman coefficient increases when using a shorter context (Figures 2a and 2b). This phenomenon where increasing the context length does not help performance may be because the added context is extraneous and not biologically important for controlling the expression of the gene and/or because the added context of the predictions introduces additional error.

Conclusion

With sequence-to-expression models growing larger, applying a hierarchical framework can help break up these larger architectures into smaller, discrete components. Hierarchical architectures provide more structure grounded in biology, which can guide the model when learning the regulatory transcriptional code. We can also analyze the components of these hierarchical models, which can help create more interpretable models. Using hierarchical architectures, we can adopt an existing model and build additional models on top of these predictions for different tasks, such as expression prediction. As sequence-to-epigenomic predictors continue to improve, perhaps across a wider variety of epigenomic tracks, hierarchical model architectures can help improve (tissue-specific) expression predictions.

References

[1] Kelley, D.R., Reshef, Y.A., Bileschi, M., Belanger, D., McLean, C.Y., Snoek, J.: Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome research **28**(5), 739–750 (2018)

[2] Kelley, D.R.: Cross-species regulatory sequence activity prediction. PLoS computational biology **16**(7), 1008050 (2020)

[3] Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., Kelley, D.R.: Effective gene expression prediction from sequence by integrating long-range interactions. Nature methods **18**(10), 1196–1203 (2021)

[4] Linder, J., Srivastava, D., Yuan, H., Agarwal, V., Kelley, D.R.: Predicting rna-seq coverage from dna sequence as a unifying model of gene regulation. bioRxiv, 2023–08 (2023)

[5] Zhou, J., Theesfeld, C.L., Yao, K., Chen, K.M., Wong, A.K., Troyanskaya, O.G.: Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nature genetics **50**(8), 1171–1179 (2018)

[6] Zhou, J., Troyanskaya, O.G.: Predicting effects of noncoding variants with deep learning–based sequence model. Nature methods **12**(10), 931–934 (2015)

[7] Karollus, A., Mauermeier, T., Gagneur, J.: Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. Genome Biology **24**(1), 56 (2023)