Splicing Up Your Predictions with RNA Contrastive Learning

Philip Fradkin [1 2] **Leo J. Lee** [3 2] **Bo Wang** [1 3 2] **Brendan Frey** [1 2 4]

Abstract

In the face of rapidly accumulating genomic data, our understanding of the genetic regulatory code remains limited. For example, the majority of mutations recorded in clinical databases are currently undefined in terms of their pathogenicity. Drawing inspiration from the success of self-supervised methodologies, we extend contrastive learning techniques to genomic data, focusing on RNA sequences generated through splicing, gene duplication, and speciation. Our novel dataset and contrastive objective enable the learning of generalized RNA representations. We demonstrate their utility on downstream tasks such as RNA half-life prediction, and gene ontology classification. Our pre-training strategy yields competitive results using linear probing on both tasks, along with up to a two-fold increase in Pearson correlation under low-data conditions. Importantly, our exploration of the learned latent space reveals that the self-supervised contrastive task yields semantically meaningful representations, solidifying its potential as a valuable pre-initialization method for RNA property prediction.

1. Introduction

Contrastive learning techniques can be used to learn effective representations but they require domain-specific augmentations that preserve key information. To generate RNA augmentations, we rely on naturally occurring cellular and evolutionary processes: splicing, gene duplication, and speciation. Byproducts of these processes generate RNAs with different sequences and similar functions. First is RNA splicing, a process for assembling mature RNA from precursor RNA (pre-RNA). At a high level, splicing involves the removal of non-coding regions, called introns, from the precursor RNA molecule and then joining together the remaining coding regions, called exons, to create the final RNA molecule (Baralle & Giudice, 2017). Importantly, alternative splicing is a prevalent phenomenon in which different combinations of exons can be joined together, leading to the production of multiple RNA isoforms from a single gene. This process greatly increases the diversity of proteins that can be generated from a limited number of genes. The second process is gene duplication and speciation events which generate homologous genes. At a high level, homologous genes are those found in different organisms that have descended from a common ancestral gene. These genes are related by virtue of their shared ancestry and often retain similar functions, structures, or sequences. Utilizing these function-preserving processes, we identify different RNA sequences and use them as augmentations for learning a general RNA embedding. By minimizing the distance between functionally similar sequences, the model is able to learn regulatory regions critical for RNA property and function prediction.

In this work, we propose IsoCLR, a domain-specific method for learning general isoform RNA representations. We pretrain a dilated convolutional residual model which has been demonstrated to be successful in applications for cellular property prediction and able to generalize to long variable length sequences (Kelley et al., 2018; Linder et al., 2022; Chen et al., 2016; He et al., 2015). Utilizing biologically inspired RNA augmentations allows us to generate robust multi-purpose RNA representations. We investigate the effectiveness of these representations by evaluating the models on RNA half-life predictions and gene ontology classification tasks (Agarwal & Kelley, 2022; authors listed, 2019). Our main contributions are:

- We create a novel dataset by proposing augmentations for genomic sequences produced through speciation, gene duplication, and splicing processes.
- We propose IsoCLR, a novel method that employs a contrastive learning objective to learn robust RNA isoform representations.

[1]Department of Computer Science, University of Toronto, Canada [2]Vector Institute, Toronto, Canada [3]Department of Electrical & Computer Engineering, University of Toronto, Canada [4]Peter Munk Cardiac Center UHN, Toronto, Canada. Correspondence to: Philip Fradkin <phil.fradkin@mail.utoronto.ca>.

The 2023 ICML Workshop on Computational Biology. Baltimore, Maryland, USA, 2023.

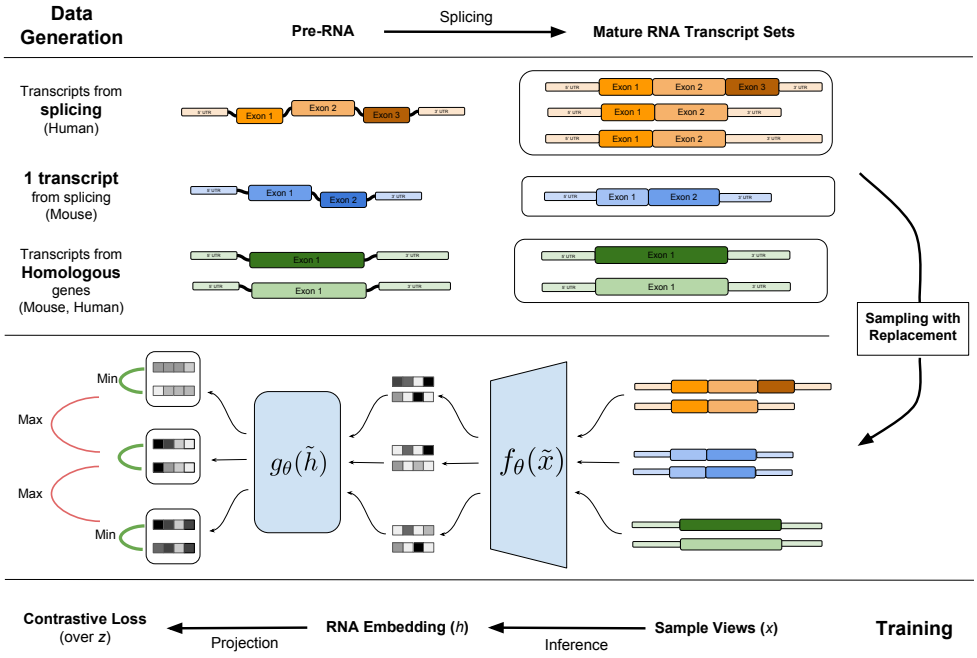


Figure 1. Description of the data generation and training processes for IsoCLR. The **upper** half of the figure demonstrates three hypothetical examples for creation of the mature RNA transcript sets. The **lower** half of the figure demonstrates the training process utilizing the generated mature RNA sets. First RNAs are sampled with replacement from the set and an RNA embedding is generated using a dilated convolutional residual encoder f. Then the representations are passed through a projector g, the normalized output of which is used to compute contrastive loss.

- We conduct extensive evaluations of IsoCLR on tasks such as RNA half-life prediction and gene ontology classification. Our results demonstrate improvements, particularly in scenarios with limited available data.

2. Methods

Our proposed dataset used for the contrastive learning objective is composed of annotated mRNA transcriptomes. Gencode and Refseq databases compile mRNA isoforms for different species, indicating relative positions of exonic coordinates and other important genomic features such as 5'UTR, CDS, and 3'UTR regions (Frankish et al., 2021; O'Leary et al., 2016). Using this information, we generated a six-track mature RNA representation, consisting of four one-hot encoded tracks representing genomic sequence, a track indicating the 5' location of splice sites, and a track indicating the first nucleotide of every codon. The addition of extra tracks indicating splice site and coding sequence locations has been shown to be beneficial for downstream genomic tasks (Agarwal & Kelley, 2022). Depending on the species analyzed and the transcriptome annotation resource used, between 25% and 50% of genes contain multiple isoforms which we then sample to use as augmentations. Additionally, we used homology as a source of RNA isoform invariances. To annotate these relationships, we used the Homologene database (Sayers et al., 2023).

2.1. Contrastive Learning Objective

During our contrasting training phase, we pool together sequences of splicing isoforms from homologous genes and treat them as views of the same object. Given a batch of N sequences (e.g. RNA isoforms) $x_1, ...x_N$ let x_i^1, x_i^2 be two splicing isoforms coming from a set of homologous genes. We pass these augmented views through a dilated convolutional encoder f resulting in the outputs h_i^1 and h_i^2. These representations are then fed into a multi-layer perceptron projection head, g the output of which is used to calculate normalized projections z_i as shown in figure 1: $z_i^1 = \frac{g(h_i^1)}{\|g(h_i^1)\|}$ and $z_i^2 = \frac{g(h_i^2)}{\|g(h_i^2)\|}$. Normalized projections z_i are used to compute the decoupled contrastive loss, utilizing samples from the rest of the batch as negatives (Yeh et al., 2021).

$$\mathcal{L}_{DCL,i}(\theta) = \log \sum_{z_k \in \mathcal{Z}, l \in 1,2}^{N} \mathbb{1}_{k \neq i} \exp(\langle z_i^1 \cdot z_k^l \rangle / \tau) - w_i \langle z_i^1, z_i^2 \rangle / \tau.$$

τ is the temperature parameter set to 0.1, and $\mathbb{1}_{k \neq i}$ is an indicator function that evaluates to 1 when $k \neq i$. Due to the non-uniform number of views per set of homologous genes, we use the term w_i to up-weight the importance of difficult examples containing more transcripts and down-weight the importance of the positive loss when a gene has only a single transcript. The above loss is computed for all the samples in the batch for both the sampled views $l \in 1, 2$.

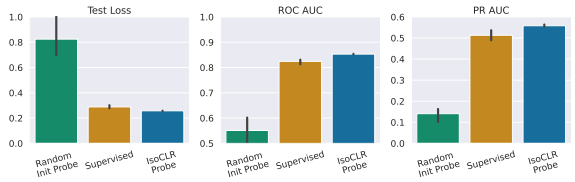
3. Experimental Results

We demonstrate that contrastive pre-training across homologous genes and splicing isoforms improves downstream prediction for both mRNA half-life evaluation and multi-label GO term classification tasks. We evaluate the effectiveness of the learned representation with three strategies: linear probing, full model fine-tuning, and latent space visualization. In addition, we highlight the effectiveness of pre-training in low-data settings for both linear probing and fine-tuning results. For latent space visualization, we generate the results from the output of the encoder f and visualize RNA representations utilizing the corresponding gene ontology categories. We demonstrate that our learned embedding has information regarding cellular component function annotation, without ever explicitly learning that information.

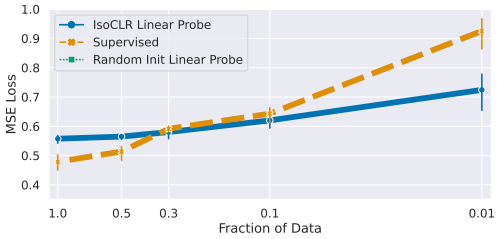
3.1. Linear probing approaches supervised model performance

To evaluate the effectiveness of our pre-trained representation, we followed the conventional evaluation strategy of linear probing. We freeze the weights of the dilated convolutional encoder f and train a linear layer to predict the corresponding task. We compare the performance of the linear probing strategy with a supervised model with matched architecture. We demonstrate competitive results using only a linear probe on the RNA half-life prediction task 2b, 2c, and show that only training the linear layer exceeds the performance of the supervised model on the gene ontology prediction task 2a.

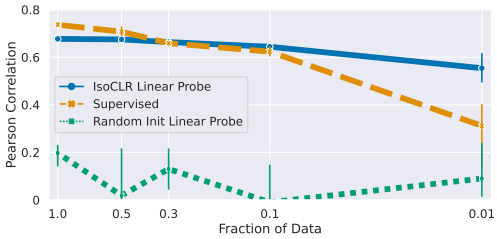
We identify that our model performs well in the low data regime by sub-sampling the amount of data to 10% and 1% which corresponds to 2000 and 200 data points respectively 2b, 2c. The linear probe performance matches that of the supervised model in the 30% training task and significantly exceeds it in the 1% task 2c. We also observe that the model is robust to the two orders of magnitude decrease in the amount of data available with respect to the correlation coefficient.



(a) Gene ontology multi-label classification comparison with the supervised model. This is a ten-class multi-label classification task for the molecular function gene ontology category.



(b) IsoCLR RNA HL Linear Probe and supervised model performance test set MSE. Randomly initialized linear probe MSE is 1.51 for human loss.



(c) IsoCLR RNA HL Linear Probe, supervised model, randomly initialized linear probe performance test set Pearson.

Figure 2. Comparison of IsoCLR linear probe with a supervised model matched by architecture on gene ontology multi-label classification and RNA HL regression tasks. Trained and evaluated on matched subsets of the data with the standard deviation computed across three folds. HL, half-life. MSE, mean squared error.

3.2. RNA half-life fine-tuning demonstrates effective low data regime performance

To assess whether the IsoCLR pre-training provides utility beyond an effective representation, we evaluate its performance by fully fine-tuning it and compare to a fully supervised model with matched architecture as well as the published method for the RNA half-life task, Saluki (Agarwal & Kelley, 2022). We evaluate IsoCLR in different data regimes and find that it demonstrates significant improvement in the lower data regime. At 10% and 1% of data used, it outperforms Saluki by 9.2% and 58.5% of test loss. For the Pearson correlation coefficient, the gains are even more stark where we observe IsoCLR competitive performance at only 0.5% of supervised data. When comparing the performance of the models using the entire dataset, we find

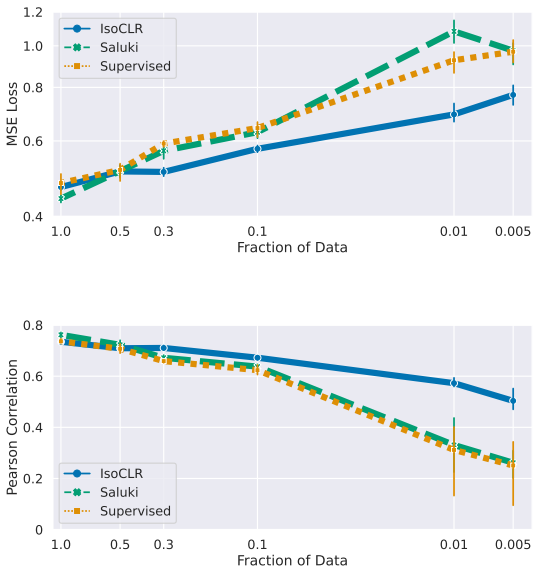


Figure 3. Fine-tuning evaluation of IsoCLR on RNA half-life prediction task.

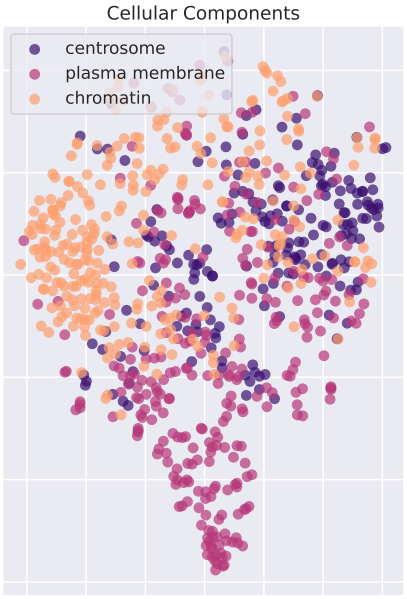


Figure 4. Visualization of the learned latent representations with stochastic neighbor embedding. Each dot is an RNA transcript from a unique gene colored by the correspondingly annotated gene ontology. The top three GO terms are sub-sampled to have an equal number of points.

that IsoCLR pre-training improves performance relative to that of the supervised model. However, when comparing IsoCLR with Saluki, we observe a decrease in performance. We attribute the difference in performance to specialized architectural choices made by Saluki authors, and since our goal was to create a general RNA representation, we instead opt to use a ResNet-like dilated convolutional architecture (He et al., 2015).

3.3. IsoCLR learns semantically meaningful latent representations

Finally, we evaluate whether the pre-training results in the model learning general biological functions. We examined high-level biological labels associated with specific genes. More specifically we examined how well do IsoCLR representation capture differences between GO terms from hierarchies associated with cellular components. We generate the representation with encoder f and reduce the dimensionality of the embedding with t-sne (van der Maaten & Hinton, 2008). We find that the model indirectly learns high-level associations between RNAs and the cellular compartment with which they are associated. Analysis of IsoCLR's latent space and linear probing results demonstrates that the model infers information on gene ontology classes although it never gets directly trained on this information. This demonstrates that contrastive pre-training based on homology and splicing isoforms has the potential to learn fundamental biological properties creating an efficient downstream training initialization.

4. Conclusions

An important question to assess is why do we expect that minimizing distances between RNA isoforms would be helpful at all for seemingly unrelated phenotypes like RNA half-life prediction and gene ontology classification. One hypothesis is that diversity-generating but function-preserving processes select for RNA regions that are essential for biological processes. Through the contrastive pre-training procedure, the hypothesis space for a causal signal is reduced to the regions that are conserved over evolutionary and splicing processes.

In this work, we propose a novel, self-supervised contrastive objective for learning mature RNA isoform representations. We show that this approach is an effective strategy to address two major challenges for cellular property prediction: data efficiency, and model generalizability. We demonstrate that IsoCLR representations are effective in the low data setting, paving the path to true few-shot learning for RNA property prediction.

References

Agarwal, V. and Kelley, D. R. The genetic and biochemical determinants of mRNA degradation rates in mammals. *Genome Biol*, 23(1):245, Nov 2022.

authors listed, N. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res*, 47(D1):D330–D338, Jan 2019.

Baralle, F. E. and Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nat Rev Mol Cell Biol*, 18(7):437–451, Jul 2017.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *arXiv e-prints*, art. arXiv:1606.00915, June 2016. doi: 10.48550/arXiv.1606.00915.

Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., Sisu, C., Wright, J. C., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Boix, C., Carbonell Sala, S., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., n, C., Gonzalez, J. M., Grego, T., Hardy, M., Hourlier, T., Howe, K. L., Hunt, T., Izuogu, O. G., Johnson, R., Martin, F. J., nez, L., Mohanan, S., Muir, P., Navarro, F. C. P., Parker, A., Pei, B., Pozo, F., Riera, F. C., Ruffier, M., Schmitt, B. M., Stapleton, E., Suner, M. M., Sycheva, I., Uszczynska-Ratajczak, B., Wolf, M. Y., Xu, J., Yang, Y. T., Yates, A., Zerbino, D., Zhang, Y., Choudhary, J. S., Gerstein, M., ó, R., Hubbard, T. J. P., Kellis, M., Paten, B., Tress, M. L., and Flicek, P. GENCODE 2021. *Nucleic Acids Res*, 49(D1):D916–D923, Jan 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. *arXiv e-prints*, art. arXiv:1512.03385, December 2015. doi: 10.48550/arXiv.1512.03385.

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

Linder, J., Koplik, S. E., Kundaje, A., and Seelig, G. Deciphering the impact of genetic variation on human polyadenylation using APARENT2. *Genome Biol*, 23(1):232, Nov 2022.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, 44(D1):D733–745, Jan 2016.

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Farrell, C. M., Feldgarden, M., Fine, A. M., Funk, K., Hatcher, E., Kannan, S., Kelly, C., Kim, S., Klimke, W., Landrum, M. J., Lathrop, S., Lu, Z., Madden, T. L., Malheiro, A., Marchler-Bauer, A., Murphy, T. D., Phan, L., Pujar, S., Rangwala, S. H., Schneider, V. A., Tse, T., Wang, J., Ye, J., Trawick, B. W., Pruitt, K. D., and Sherry, S. T. Database resources of the National Center for Biotechnology Information in 2023. *Nucleic Acids Res*, 51(D1):D29–D38, Jan 2023.

van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Yeh, C.-H., Hong, C.-Y., Hsu, Y.-C., Liu, T.-L., Chen, Y., and LeCun, Y. Decoupled Contrastive Learning. *arXiv e-prints*, art. arXiv:2110.06848, October 2021. doi: 10.48550/arXiv.2110.06848.