CS 273B Literature Review: Human Splicing Code and Genetic Determinants of Disease

Joe Paggi
jpaggi@stanford.edu

Andrew Lamb
andrewl3@stanford.edu

Kevin Tian
kjtian@stanford.edu

Irving Hsu
irvhsu@stanford.edu

Pierre-Louis Cedoz
plcedoz@stanford.edu

Prasad Kawthekar
pkawthek@stanford.edu

The authors use deep learning to build a predictive model of the percentage spliced in (ψ) of cassette exons across 16 human tissues. Their method extracts features from DNA sequences ranging from raw sequence near splice sites to counts of known RNA binding protein motifs. The deep learning model is trained on RNA-seq data from the Illumina Human Body Map 2.0 project and then used to predict changes in splicing caused by 658,420 documented single nucleotide variations (SNVs). By overlapping the set of SNVs causing large changes in splicing with the set of RNAs known to be associated with diseases, the authors identify SNVs that likely cause disease by disrupting splicing. It is also evaluated across three diverse disorders - Spinal Muscular Atrophy, Nonpolyposis Colorectal Cancer, and Autism.

When applied to explore misregulation of SMN1/2 for Spinal Muscular Atrophy, the model is evaluated by using it to identify nucleotides that can increase likelihood of exon-7 skipping (which may lead to loss of function of SMN2) and to predict change in Ψ for SNVs in three exonic regulatory regions, as well as explore mutations which result in gain of SM2 function. In nonpolyposis colorectal cancer, predictions were made for mapping SNVs to $\Delta\Psi$ values for splicing in MLH1 and MSH2. The model is also used to identify genes with SNVs that may cause splicing misregulation in ASD cases, using brain same data from Autism Tissue Program, both to verify existing candidate genes and to suggest new genes in the autism function regulation process. In general, the method demonstrated its success in inferring SNVs related to disruptive splicing and loss of function for these diseases, and suggests possible unknown factors which can lead to these disruptions.

1 Computational Model

An ensemble of neural networks is used to predict the percentage of transcripts with the central exon spliced in (ψ), from the extracted DNA sequence features. The base neural network has a single layer of 30 hidden units, with sigmoidal activation functions and a softmax function to predict the output. In order to prevent overfitting, a prior distribution is put over the models that shuts off connections between the input and hidden units with Bernoulli probability $1 - \alpha$ and then draws the number of hidden units to be used to predict output from a Poisson distribution with an expectation of λ. Gibbs sampling is used to sample from the posterior distribution of models. In practice, $\alpha = 0.1$ (encouraging sparse usage of DNA sequence features) and $\lambda = 10$. The researchers experimented with models that jointly predicted for different tissues (i.e. tissues share the hidden units) and models that picked different tissues separately (i.e. a model was trained for each tissue) and found that jointly predicting tissues performed slightly higher on the test set [1]. Nearest neighbors, multinomial regression, support vector machines, and non-Bayesian neural networks were also experimented with, and found to have inferior performance.

1.1 Weaknesses and Extensions

One of the most interesting aspects of the model is the size: in expectation the neural network has one layer of 10 hidden units. Given the small number of examples (10,698 cassette exons) and large number of features (extracted 1393 DNA sequence features), overfitting is a large concern, which motivates the use of a prior distribution that encourages few connections in the network. This leads to three questions about the procedure: Would a different regularization technique lead to better results? Is a neural network an appropriate model? Would a model trained jointly among different tissue types be more effective?

1.1.1 Regularization

The ensembling of a family of neural networks with stochastically removed connections brings to mind the dropout technique, which has been shown to be an effective method for regularization of neural networks, and has been applied to computational biology and other domains. Dropout randomly drops hidden units during train time, effectively training an exponentially large number of smaller networks, and then at test time takes the average prediction of the ensemble [2]. It would be interesting to see if a network regularized with dropout leads to different results.

1.1.2 Different Models

Neural networks have gained popularity recently in large part because of their effectiveness on large datasets [3]. For splicing prediction, it was found that on some tissue types, a simple regularized regression was nearly as effective as the Bayesian neural network that was chosen [1]. This raises the question of whether a completely different model might be more effective.

1.1.3 Jointly Trained Model

The current model takes DNA sequences as input, and predicts the degree to which the variant affects splicing in human tissues. According to the authors, potential sources of prediction error include unaccounted-for RNA features, inaccuracies in computed features, imperfect modeling of splicing levels, and limitations from placing a heavy emphasis on cassette splicing. To address this shortcoming, one approach could be to jointly model processes that are known to affect transcription levels within the cell in a highly integrated manner. Examples of such steps in gene regulation include protein stabilization, protein synthesis, chromatin dynamics, polyadenylation, and mRNA turnover. Furthermore, joint modeling could be useful since DNA elements previously thought to be pertinent to only a single regulatory process actually take part in multiple steps in the regulatory chain [4].

2 Features

For each cassette exon, the model is fed 1393 features characterizing the corresponding sequence. These features were selected from sequence characteristics that were shown to modulate splicing in previous studies. The features range from fundamental properties of the sequence, such as intron lengths and raw sequence near splice sites to curated features such as known RNA-binding protein (RBP) binding sites and protein coding potential.

Two features that we found particularly interesting were the nucleosome positioning and RNA secondary structure metrics. We find these interesting because previous studies have shown that they influence splicing, but it is unclear in what way. Intuitively, these features likely exert context dependent effects, making deep learning the best option for characterizing their influence. We would appreciate if the authors attempted to interpret how their model is using these features and include these findings in the manuscript.

One included feature that we find controversial is the translatability of exon combinations. There is not (at least as far as we can intuit) a mechanistic reason why translatability would affect splicing. Of course, the model should simply learn to ignore features uncorrelated with splicing. However, in this case, translatability can influence the experimentally determined Ψ values through processes distinct from splicing, such as nonsense mediated decay. We discuss potential issues with including this feature in the next section.

3 Bias from Nonsense Mediated Decay

We are suspect of the high proportion of nonsense mutations among exonic SNVs that influence splicing on protein sequence. The authors report, "we found that significant deviations are induced by 9525 nonsense SNVs and 1273 missense SNVs but also by 579 synonymous SNVs", implying that 83% of these SNVs are nonsense mutations. If we assume that: 1) any given indel has a 2/3 chance of resulting in a frameshift, 2) all the significant exonic SNVs are indels (the best case for a high proportion of nonsense), and 3) every frameshift results in a nonsense mutation (again best case for high proportion of nonsense), then we conclude that around 66% of exonic SNVs should be nonsense mutations.

Experimentally measured Ψ values are influenced by nonsense mediated decay. Specifically, poison cassette exons have lower measured Ψ values than actual Ψ values (at the time of splicing) because the inclusion isoform is preferentially degraded. Based on this information, we suspect the model learned that untranslatable cassette exons tend to be excluded.

While including isoform-preferential degradation in the model provides more accurate prediction of the experimental Ψ values, their model ceases to be a model solely of splicing, but also of NMD. Furthermore, it implies that a large fraction of their set of SNVs that cause disease by disrupting splicing are not actually affecting splicing. We ask that the authors retrain their model while withholding poison cassette exons to assess how widespread a problem this has caused. [Through personal communication with the authors, we confirmed that they are aware of this issue and are okay with the fact that their model is a splicing and NMD predictor]

4 Evaluation Criteria

In general, the results cited by the authors are impressive in demonstrating the effectiveness of their method both as an isolated predictor and as an additional consideration to known predictive methods for splicing. However, some evaluation criteria that are used we think are questionable choices, and perhaps the results would be more informative if presented under different metrics.

For example, the choice of using the area under receiver operator curve (auROC) is perhaps misleading when the data is imbalanced - a better choice than the false negative rate would be the precision. Further, in this evaluation method, where the SNVs are binned into "low" and "high", precision and false negative rate are both probably inappropriate, because a "misclassification" isn't just when low is labeled as high, but should include the middle bin. A more informative measure in general might be the distribution of predicted Ψ against experimental Ψ under something like a squared-error loss function, instead of bin classification. The cited results are useful in showing that the method is good at identifying "important" SNVs with high Ψ, however.

Furthermore, in the experiment with the 4 individuals' Ψ levels for differing SNPs, where the method predicted the direction of change correctly in 73 percent of cases, we feel that the low P-value cited is misleading because all of these cases were significant examples where the change were above some noise threshold to begin with, so the direction of change should be easier to predict (very unbalanced prior). Thus, a figure of 73 percent seems to be a lot lower than is suggested by the binomial test.

5 Conclusion

Altogether, we believe the computational approach the paper takes is successful by many metrics, and has strong implications for applications in identifying candidate disease-causing SNVs, and in predicting splicing as a whole. We are interested in some extensions and have some criticisms for the methods used, however, and believe clarifying some points would yield a better understanding of both the model and what factors lead it to perform well.

Among extensions include regularization using dropout training (which seems to be similarly motivated to the removed-connections ensemble) and experimenting with additional regression models. Additionally, we think that it makes sense to jointly train the given prediction model with other processes that share mechanisms and are related to transcription. The features that have effects which we would like the authors to describe further include nucleosome positioning, RNA secondary structure, and the translatability of exon combinations. In particular, we believe using translatability as a feature causes bias towards nonsense mediated decay (NMD) which is a separate effect from detetrmining Ψ values that depends on the translatability, i.e. the predictions will be influenced in poison cassette exons by non-splicing factors. Through correspondence we found that the authors are aware of this, and explaining this property of the predictor would better the clarity of the paper. Finally, there were a few evaluation criteria that the paper cited that we felt were not the best choice, or for which there was a better alternative metric.

References

[1] Hui Yuan Xiong, Yoseph Barash, and Brendan J Frey. Bayesian prediction of tissue-regulated splicing using rna sequence and cellular context. *Bioinformatics*, 27(18):2554–2562, 2011.

[2] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[4] Ulrich Braunschweig, Serge Gueroussov, Alex M Plocik, Brenton R Graveley, and Benjamin J Blencowe. Dynamic integration of splicing within gene regulatory pathways. *Cell*, 152:1252–1269, 2013.