Paper Review

Paper: **Integrative Deep Models For Alternative Splicing**

Authors: **Anupama Jha, Matthew R Gazzara, Yoseph Barash**

Reviewers: **Sebastian Astiz Le Bras, Ana-Maria Istrate, Michael Painter**

Paper Summary

Splicing codes are used to predict percent splicing index (PSI) of exons during alternative splicing. This work improves upon existing Bayesian Neural Network (BNN) and Deep Neural Network (DNN) implementations of splicing codes using two modifications. Firstly, new target variables are introduced to directly model PSI with probabilistic regression, rather than quantizing the problem into a classification task. Secondly, the network architecture is adapted to integrate further data sources, and furthermore allows integration of knockdown or knockout data.

Significant Contributions and Differences From Existing Work

The new target function described allows for a continuous value for PSI to be calculated directly. Previous work discretized PSI values into three buckets representing Low, Medium, or High Values. The introduction of the new function is largely responsible for the increase in accuracies over previous models and is perhaps the most significant contribution of the work. The function improves the variance explained by the PSI prediction from an average of ~36% to an average of~72% across different Tissues.

The introduction of auxiliary data sources such as CLIP-seq data and knockout, knockdown and over expression condition data. The introduction of these sources not only improves the accuracy of the models described but provides opportunity for more scalable and functional results. Integrating experimental data into the splice code models is another contribution the authors bring to the field and that has potential to be explored more in future work. Comparison between the BNN and DNN models helped make a better assessment between the two, as it normalized their findings by using the same datasets and target functions.

Criticism, Suggestions and Further Work

In the results section, a phrase *cassettization* is introduced but defined ambiguously in section 2.1, three pages prior. This could be improved upon by explicitly defining the term earlier in the paper.

It’s mentioned that deeper networks were trained and that regularization was also explored, however, no data was published. We believe that it would have been beneficial to include the architectures and results in the supplementary material as guidance for future work.

We noticed that in the cross-validation algorithms that the entire dataset it shuffled before selecting folds, including the fold for testing, else it is poorly described. This is problematic because it allows every example in the dataset to contribute to hyperparameter selection, allowing overfitting to the test data and suggests that any model chosen may not generalize well.

Two potential areas of exploration are: to use RNNs for automatic RNA feature extraction rather than manually designing features, and to try different DNN architectures, such as CNNs. Moreover, further work should also include generalizing the findings in this paper to other conditions and datasets.