Application and Comparison of Deep Learning Methods in the Prediction of RNA Sequence Degradation and Stability

Singhal, Ankit (School: International School of Basel)

Synthesis and efficient implementation mRNA strands has been shown to have wide utility, especially recently in the development of COVID vaccines. However, the intrinsic chemical stability of mRNA poses a challenge due to the presence of 2'-hydroxyl groups in ribose sugars. As expected for in-line hydrolytic cleavage reactions, the chemical stability of mRNA strands is highly dependent on external environmental factors, e.g. pH, temperature, oxidizers, etc. Predicting this chemical instability using a computational model will reduce the number of sequences synthesized and tested through identifying the most promising candidates, aiding the development of mRNA related therapies. This paper adapts and evaluates three deep learning models (Long Short Term Memory, Gated Recurrent Unit, and Graph Convolutional Networks) as methods to predict the reactivity and degradation of mRNA sequences. The Stanford Open Vaccine dataset of 6034 mRNA sequences was used in this study. The training set consisted of 3029 of these sequences (length of 107 nucleotide bases) while the testing dataset consisted of 3005 sequences (length of 130 nucleotide bases), in structured (Lowest Entropy Base Pair Probability Matrix) and unstructured (Nodes and Edges) forms. The stability of mRNA strands was accurately generated, with the Graph Convolutional Network being the best predictor of reactivity (RMSE = 0.249) while the Gated Recurrent Unit Network was the best at predicting risks of degradation (RMSE = 0.266). Combining all target variables, the GRU performed the best with 76% accuracy. Results suggest these models can be applied to understand and predict the chemical stability of mRNA in the near future.