Splice Junction Identification Using Long Short-Term Memory Neural Networks.

Authors:

Kevin Regan, Abolfazl Saghafi and Zhijun Li

Affiliation:

Department of Mathematics, Physics and Statistics, University of the Sciences, Philadelphia, PA

Abstract:

Background: Splice junctions are the key to going from pre-messenger RNA to mature messenger RNA in many multi-exon genes due to alternative splicing. Since the percentage of multi-exon genes that undergo alternative splicing is very high, identifying splice junctions is an attractive research topic with important implications.

Objective: The aim is to develop a deep learning model capable of identifying splice junctions in RNA sequences using 13,666 unique sequences of primate RNA.

Method: A Long Short-Term Memory (LSTM) Neural Network model is developed that classifies a given sequence as EI (Exon-Intron splice), IE (Intron-Exon splice), or N (No splice). The model is trained with groups of trinucleotides and its performance is tested using validation and test data to prevent bias.

Results: Model performance was measured using accuracy and f-score in test data. The finalized model achieved an average accuracy of 91.34% with an average f-score of 91.36% over 50 runs.

Conclusion: Comparisons show a highly competitive model to recent Convolutional Neural Network structures. The proposed LSTM model achieves the highest accuracy and f-score among published alternative LSTM structures.