PREDICTION OF FULL-LENGTH INTERNAL RIBOSOME ENTRY SITES USING DEEP LEARNING

LI YP

RVAC Medicines Inc, Waltham, MA, USA

Internal ribosome entry sites (IRESs) are RNA elements that can recruit ribosomes and allow cap-independent translation. Analyzing IRESs is important for advancing our understanding of many biological functions and developing RNA-based therapeutics. However, IRESs do not have highly conserved primary sequences, secondary structures, or mechanisms to recruit ribosome subunit, which makes studying IRES and identify new ones challenging. A few traditional machine learning models have been developed to identify IRES and related sequence and structure features using a data-driven approach. However, it is noteworthy that most models were trained using the same oligonucleotide library with 550k 174-nt length sequences, many of which were truncated IRESs. Because full-length IRES is necessary for strong translation, we focused on developing a deep learning model to identify full-length IRES. Additionally, a deep learning model that learns directly from sequence information can also reduce the bias from feature engineering using in traditional machine learning models.

We downloaded all cis-regulatory RNA elements from Rfam, a database of RNA sequence families of structural RNAs. Since IRESs are longer than many other cis-regulatory RNA elements, we only kept the RNA families with a median length greater than 100 nt to reduce the bias when training the models. After further removing sequences with non-standard nucleotides, we had 1,443 full-length IRES sequences as positive samples and 166,138 non-IRES cis-regulatory RNA elements as negative samples.

We evaluated multiple deep learning architectures to build a binary classification model for identifying full-length IRES sequences. Since the data is highly imbalanced, we introduced a new oversampling technique to overcome the issue. In the 10-fold cross-validation (CV), only the positive samples in the training sets were oversampled. The reported measurements, e.g., AUC and AUPRC, were based on original validation sets. Data were split by RNA families during the CV to avoid data leakage due to the sequence similarity within a family. Among all the architectures we evaluated, a simple convolutional neural network (CNN) model with one-hot encoding of the nucleotides as input outperformed many other complicated architectures, such as nucleotide embedding and transformer, and those using secondary structure information, such as CNN with both nucleotide and secondary structure encodings as input, graph attention, and graph convolution. On average, the CNN model achieved 0.87 AUC and 0.39 AUPRC in the 10-fold CV. With the highly accurate classification model, we can further identify motifs that are important to the IRES function and discover new IRESs.