DI
ML
BI

MRNA ISOFORM PREDICTION

Allocation: Innovation and Exploration/84.9 Knh
PI: Julie Dickerson[1]
Co-PI: Gaurav Kandoi[1]

[1]Iowa State University

EXECUTIVE SUMMARY

Genes perform multiple functions, in part owing to their multiple messenger RNA isoforms. Messenger RNA (mRNA) is a large family of RNA molecules that conveys genetic information from DNA to the ribosome, where they specify the amino acid sequence of the protein products of gene expression. Alternative splicing produces multiple mRNA isoforms of genes that have important diverse roles such as regulation of gene expression, human heritable diseases, and response to environmental stresses. Alternative splicing is one mechanism that allows genes to perform multiple functions. Despite the diverse role of alternative splicing, very little has been done to assign functions at the mRNA isoform level. Therefore, differentiating the functions of mRNA isoforms is vital for understanding the underlying mechanisms of biological processes.

The goal of this study is to develop a functional network and recommendation (recommender) system to predict tissue-specific mRNA isoform function and understand implications of such alternate isoforms on metabolic pathways for the mouse and *Arabidopsis thaliana* model systems. (*Arabidopsis* is a small flowering plant that is widely used as a model organism in plant biology.)

Highlights from the research team's outcomes include: (1) the processing of over 100 tissue-specific *Arabidopsis* RNA–Seq data sets as well as mRNA and protein sequence characterizations; (2) developing a random forest-based framework for mRNA-level functional network prediction; and (3) optimizing hyperparameters for the recommender system.

RESEARCH CHALLENGE

This research considered three major challenges in mRNA isoform function prediction. The first is the unavailability of mRNA isoform-level functional data, which is required to develop machine learning tools. However, the available data, even at the gene level, does not include all genes, further complicating the matter. The second challenge is the lack of information about tissue specificity in functional databases such as Gene Ontology, Kyoto Encyclopedia of Genes and Genomes, and UniProt. The third challenge is the lack of mRNA isoform-level "ground truth" functional annotation data. Therefore, the research team's work includes using mRNA isoform and protein sequences, high-throughput RNA-sequencing data, and functional annotations at the gene level to develop computational methods for predicting functions for alternative spliced mRNA isoforms.

Some limitations of previous studies that have been overcome in the current work include: (1) predicting novel mRNA isoform interactions with no gene-level interaction information in current biological databases, (2) predicting tissue-specific mRNA isoform-level functional networks and mRNA isoform function, (3) limiting bias in the machine learning model by using a more biologically sound way of defining nonfunctional (negative pairs) mRNA isoform pairs, (4) formulating the task of mRNA isoform-level functional network prediction as a simple supervised learning task and formulating the task of mRNA isoform function prediction as a recommendation system, and (5) incorporating the relations between the Gene Ontology terms apart from the obvious hierarchical relations.

METHODS & CODES

The team has used STAR [1] and StringTie [2], which are codes for high-performance alignment and assembly of RNA–Seq reads and expression analysis for transcripts, without compromising mapping accuracy. These scale almost linearly with an increasing number of processing cores with a minimal increase in the memory requirement. Both tools are written entirely in C++ for higher efficiency and faster performance. All mRNA and protein sequence properties were calculated using R Bioconductor packages. These provide the means to calculate several diverse types of sequence properties. In addition, the team used TensorFlow and scikit-learn to build the machine learning systems.

RESULTS & IMPACT

One of the outcomes of this project was the evaluation and validation of the team's supervised learning-based machine learning framework for predicting tissue-specific mRNA isoform functional networks in *Arabidopsis*. Tissue-spEcific mrNa iSoform functIonal Networks (TENSION) makes use of single mRNA-producing gene annotations and gene annotations tagged with "NOT" to create high-quality mRNA isoform-level functional data. The team uses these to train random forest algorithms to develop mRNA isoform functional network prediction models. By using a leave-one-tissue-out approach and incorporating tissue-specific mRNA isoform-level predictors along with those obtained from mRNA isoform and protein sequences, the team has developed mRNA isoform-level functional networks for *Arabidopsis* tissues.

Another outcome is the evaluation of different combinations of hyperparameters of the mRNA function recommendation system (mFRecSys) for making tissue-specific function recommendations for mRNA isoforms. In mFRecSys, the team considers mRNA isoforms as "users" and Gene Ontology biological process terms as "items." By using explicit contexts for mRNA isoforms, Gene Ontology biological process terms, and tissue-specific mRNA isoform expression, mFRecSys is able to make tissue-specific mRNA isoform function recommendations.

This work emphasizes the significance of incorporating diverse biological context to develop better machine learning tools for biology. It also highlights the use of simplified supervised learning methods for biological network prediction. The machine learning models and recommendation systems developed as part of this work also draw attention to the power of simple mRNA isoform sequence-based predictors to improve mRNA isoform function prediction. The methods developed have potential practical applications, for instance, as predictive models for distinguishing the functions of different mRNA isoforms of the same gene or identifying tissue-specific functions of mRNA isoforms.

WHY BLUE WATERS

Blue Waters was essential to the two research outcomes of this project, as they both involved extensive validation and optimization. The computational resources of Blue Waters allowed the research team to perform these experiments at scale, enabling the optimization of many different experimental settings. Furthermore, conversations with the staff have been instrumental in improving job efficiency.

PUBLICATIONS & DATA SETS

G. Kandoi and J. A. Dickerson, "Tissue-spEcific mrNa iSoform functIOnal Networks (TENSION) Collection," 2018, doi: 10.25380/iastate.c.4275191.

G. Kandoi and J. A. Dickerson, "Tissue-specific mouse mRNA isoform networks," *BioRxiv*, 2019, doi: 10.1101/558361.

G. Kandoi and J. A. Dickerson, "Differential alternative splicing patterns with differential expression to computationally extract plant molecular pathways," in *2017 IEEE International Conf. on Bioinformatics and Biomedicine*, doi: 10.1109/BIBM.2017.8217993.