TRIFID: determining functional isoforms

Fernando Pozo Ocampo[1], Laura Martínez-Gómez[1], Michael L. Tress[1]

[1] Bioinformatics Unit, Spanish National Cancer Research Center (CNIO)

Introduction

Background

- Alternative Splicing (AS) of messenger RNA can generate a wide variety of mature RNA transcripts and this expression is confirmed by experimental transcript evidence. However, it is not clear how many alternative transcripts will code for functional proteins. (Tress et al. 2017, Blencowe et al. 2017).
- Proteomics analyses have shown that most coding genes have a single main splice isoform (Ezkurdia et al. 2015).
- Human population variation data indicate that most AS transcripts are evolving neutrally (Tress et al. 2017).
- Ever more splice isoforms are annotated every day, but we do not know their function.

Objectives

1. Grading what proportion of Alternative Splicing is functional.

2. Developing a ML based tool for predicting splice isoform functional importance.

Methods

`TRIFID` (Tool to Reliable Identification of Functional Isoform Data) is a Random Forest based **predictor of the relative functional importance of splice isoforms**.

It has been trained on reliable peptide evidence of 497 genes from the largest tissue-based proteomics analysis to date (79 experiments) and 47 features categorized in 5 groups (*genome annotation*, *structural*, *splicing impact*, *cross-species conservation* and *RNA-seq expression*).

Code development available here.

Results

SHAP feature importances and model interpretation

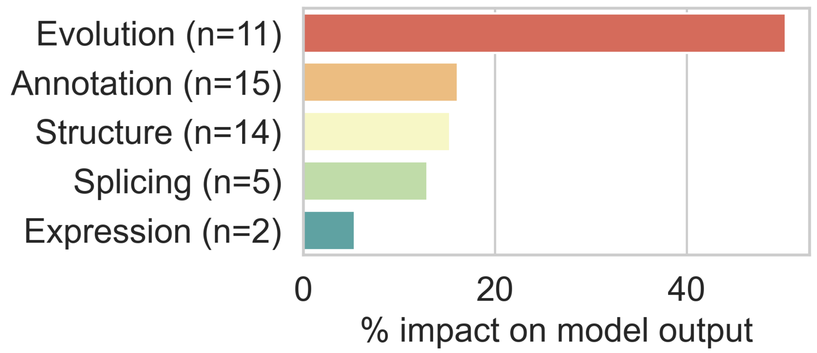

Figure 1: SHAP values (Lundberg et al. 2020) added by category. Features that best distinguish functional isoforms in the training set are conservation-based. Other important features include the length difference between the alternative isoform and the longest isoform, whether or not the transcript has a CCDS (Pruitt et al. 2009), and the conservation of Pfam functional domains. This predictions also can provide clues to the influence of features on individual predictions (see example: Fibroblast growth factor receptor 1 (TRIFID Scores)

Functional importance in the human genome

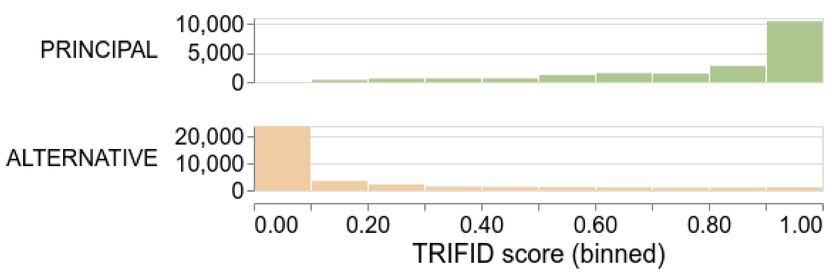

Figure 2: Non-redundant isoforms divided into PRINCIPAL or ALTERNATIVE according to their annotation in APPRIS (Rodriguez et al. 2017). Most AS isoforms have TRIFID scores below 0.05. Most PRINCIPAL isoforms have predictor scores above 0.9.

Validating the model against an external source of information

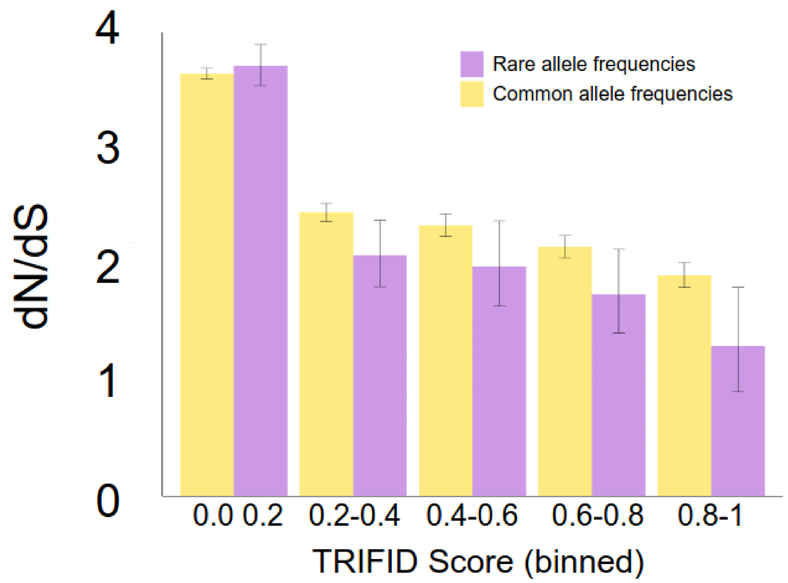

Figure 3: It shows non-synonymous to synonymous ratios for alternative exons (exons that are exclusively present in alternative isoforms). Exons under selective pressure should have significantly lower dn/ds ratios for common than for rare allele frequencies.

Conclusions

- It discriminates functionally important isoforms with high confidence (MCC=0.89, AUPRC=0.98 over 5-folds CV of the training set).
- It predicts that a large majority of splice variants (85-90%) in the human genome are likely to not be functionally important at the protein level.
- *dn/ds* ratios from show that exons from the highest scoring of alternative transcripts are under selective pressure, while low scoring exons have little or no evidence of selection.
- The model can be succesfully exported to different genome species and genome annotation databases.