

Prediction of biologically important features related to intron retention events based on machine learning algorithms

Felipe E. Ciamponi*, Michael T. Lovci, Katlin B. Massirer

Abstract

The retention of introns represents a class of alternative splicing (AS) events frequently associated with diseases. Despite the recent development of many AS identification tools, most of the tools do not consider their relationship to relevant biological features. We developed a package capable of accessing many features to AS events and evaluated association between events. We identified that retained introns and nearby exons have lower GC content than their non-retained counterparts.

Key words:

Splicing, Bioinformatics, Machine learning

Introduction

The splicing process is one of the main events associated with the variability of protein isoforms found in higher eukaryotes. Failures in this event, which is responsible for altering the nucleotide sequence of the messenger RNA (mRNA), can lead to the appearance of diseases and other phenotypes harmful to the organism. Although there are numerous tools for identification of such alterations, few of these methodologies offer the interpretation of biological characteristics which might be associated with them, leaving this task to researchers, a process often is wearing and time-consuming. As a solution to this problem, our group developed a computational package based on Python that uses machine learning approaches to analyze dozens of biological features simultaneously, classifying them in order of importance and performing statistical tests to identify which features are the most influent for separating the analyzed events from the rest of the genome. Our package is capable of extracting information fast and quantitatively, allowing for a greater efficiency and easiness in the process of biological characteristics associated with the splicing, helping in the guidance of hypothesis and design of further experiments. This study aims to demonstrate the application of this algorithm in intron retention events, but the package is currently under development and can be applied to a wide range of splicing events.

Results and Discussion

Figure 1 – GC content of intronic and exonic regions are the most important variables in intron retention events.

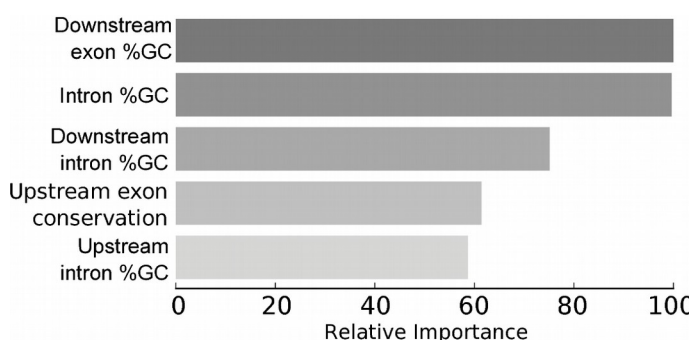


Figure 2 – Partial dependence of GC content of introns and downstream exons is highly correlated.

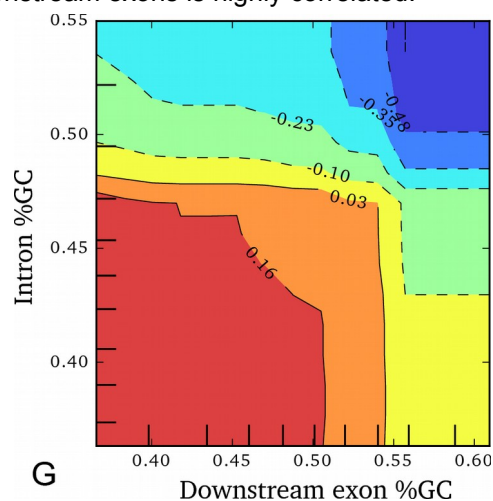
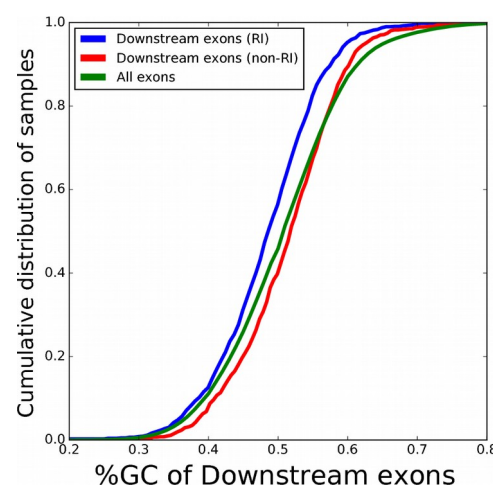


Figure 3 – Exons downstream of retained introns have a lower %GC ratio than non-retained or other exons.



Conclusions

Our package was capable of identifying that the lower GC content present in retained introns and their downstream exons are the most important features related with intron retention events in our model.