

Extracting biological knowledge from SVM models capable of distinguishing alternative and constitutive splicing

T. Murlidharan Nair¹
mnair@iusb.edu

Michael Gribskov²
gribskov@purdue.edu

¹ Departments of Biological Sciences and Computer Science/Informatics, Indiana University South Bend, 1700 Mishawaka Ave, South Bend, Indiana 46634

² Department of Biological Sciences, Purdue University, Lilly Hall of Life Sciences, 915 W. State Street, West Lafayette, Indiana 47907

Abstract

Support Vector Machine based learning methods have been applied to a difficult classification problem involving biological sequences that of distinguishing alternatively spliced junctions from constitutively spliced junctions. While achieving correct classification only solves half the problem and main problem is in understanding the basis that was used to classify the data. Here we discuss a simple approach that is capable of revealing the biological knowledge that was used in creating the soft margin between the two classes of data.

Keywords: Alternative splicing, Constitutive splicing, SVM

1 Introduction

An important observation that resulted from the genome data is the lack of correlation between the complexity of the organism and the number of genes its genome contains. The human genome for instance contains far fewer genes than expected. Precursor-mRNAs of genes in the human genome are split into regions that code called exons and those that do not code called introns [1]. There is an efficient usage of those few genes in a highly complex and combinatorial usage of its parts. This efficient usage involves combinatorial editing where in two or more exons are joined together in several different fashion. In some cases the splicing process may include a particular exon in every transcript (constitutively spliced), while in other the exons may be skipped or alternatively spliced. The region being spliced is recognized by the splicing machinery which in part is defined by the 5' and 3' splice sites and the branch points at the ends of the intervening segments (introns). One of the most important questions is to understand the signals that could be involved in the regulation of alternatively spliced products.

The problem that was addressed here is to characterize the sequence elements that could potentially distinguish alternatively and constitutively spliced junctions. A machine learning approach was used to develop an algorithm that was capable of delineating regions that was considered important in the internal representation of the models built to classify the regions.

2 Method and Results

SVM models were built to distinguish between alternatively spliced junction and constitutively spliced junction of Human and mouse acceptor (AA/CS) and donor regions (AD/CSD). A maximum accuracy of 70% was achieved for class separation. The data for building the model was mined from the Manually Annotated Alternatively Spliced Events database (MAASE) [2]. Since SVM models are not capable of

revealing the biological knowledge that it considered important in the construction of the soft margin, we have developed an algorithm that circumvents this and helps delineate the regions in the input space that are critical for the SVM in building the maximum margin hyper plane. The approach works by introducing noise into a window of a particular size and then presenting the window-randomized input set to the SVM classifier. The details of the algorithm will be discussed in the poster. The interpretable SVM algorithm was implemented in the R-programming language. Figure 1 gives the relative importance of features as a function of the position dependent noise introduction in a hexamer window along the sequence.

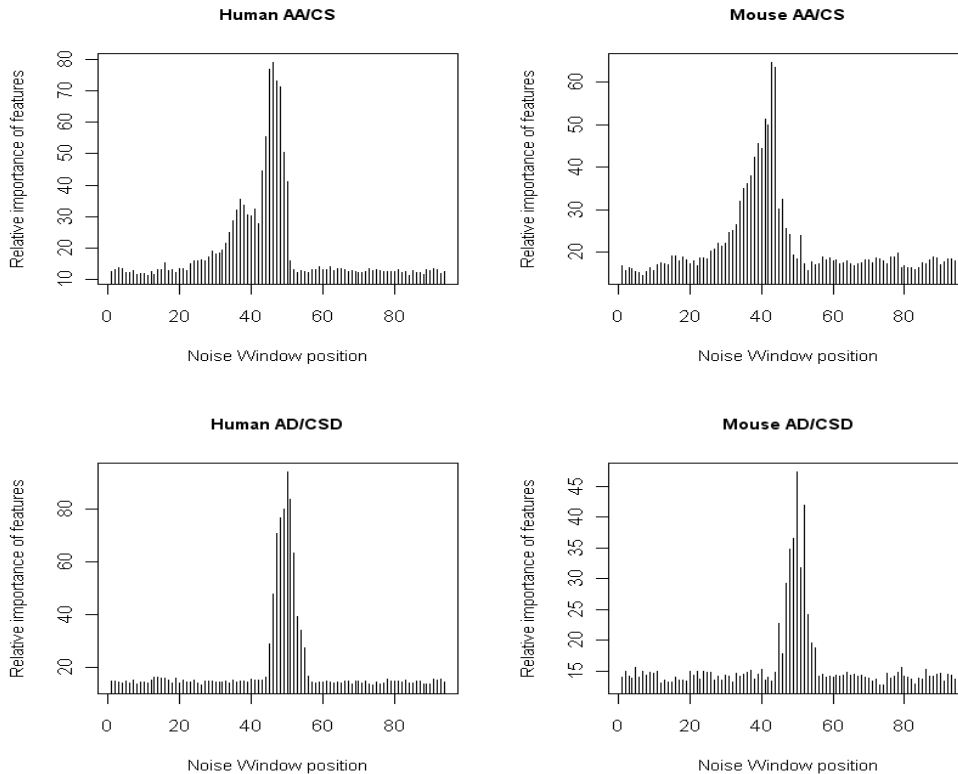


Figure 1: Interpretation of the SVM model based on performance as a function of positional noise introduced into the window of the test sequences.

3 Discussions

The results of provide evidence to the complexity of the features and signals that distinguish an alternatively spliced acceptor/donor junction from that of a constitutively splice acceptor/donor junction. The simulation results also point to need for larger data sets as learning space for building models. The interpretable SVM approach seems to yield biologically relevant information from the SVM model and is dependent on the accuracy of the model.

References

- [1] Sharp, P. A. Split genes and RNA splicing. *Cell* 77, 805-815, 1994.
- [2] Zheng CL, Kwon YS, Li HR, Zhang K, Coutinho-Mansfield G, Yang C, Nair TM, Gribskov M, Fu XD. MAASE: an alternative splicing database designed for supporting splicing microarray applications. *RNA*. 11(12):1767-1776, 2005