Classification of RNA Splice Junctions Based on Genomic Signals

Mavis Li
Yun (Sammy) Long
John Mu

Abstract
Classification of splice junctions is a critical component in the identification of new RNA isoforms in genetic annotations. Current high-throughput experimental techniques are unable to reliably identify splice junctions with high confidence. We propose to use a support vector machine (SVM) based model to improve the specificity of such experiments. Compared to existing schemes, our model performs competitively despite its simplicity. Various parameters such as the length of the sequence feature vector and the distance between nucleotides were tuned to optimal values. Overall, it was shown that the SVM is a capable tool for identification of splice junctions based on their contextual genomic signals.

1 Introduction

Proteins are an important class of molecules in living organisms and are a part of almost all biochemical processes. A central part in the production of proteins within a cell is the formation of messenger ribonucleic acid (mRNA) from the deoxyribonucleic acid (DNA) sequence. Formation and regulation of mRNA is not a well understood process. Hence, we aim to apply machine learning techniques to model the splicing step of the process. A brief overview of splicing and why it is possible to model is given below.

In the formation of a protein, DNA first transcribes to pre-mRNA, which is essentially the same sequence as the original DNA segment. Next, a number of post-processing steps are applied to the pre-mRNA by various enzymes and complexes to form the final mRNA. The mRNA is then transported to the cytoplasm and translated to a protein. One of the most important steps in the post-processing is known as RNA splicing. In this step, large parts of the pre-mRNA are removed by a specialized complex called the spliceosome. These removed segments are known as introns and the remaining segments are called exons. Only exons can potentially code for a protein as they are not removed. The location of the removal is known as a splice junction. A splice junction in mammals usually begins with GT and ends with AG. For practical reasons, we only consider splice sites with this highly conserved GT-AG sequence.

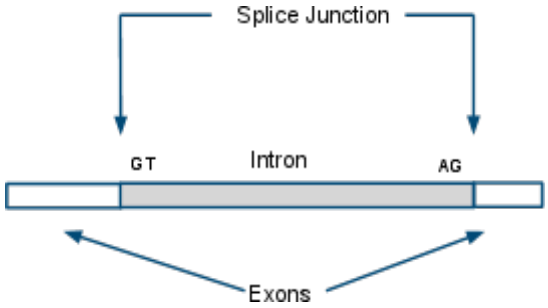

Figure 1 - Splice junction illustration

In this project, the contextual DNA sequence around each splice junction (apart from the GT-AG) is used to learn whether a given splice junction is real or not. We initially considered many of algorithms presented in class. However, we eventually settled on an SVM due to it's simple implementation and high classification accuracy. Section 2 will discuss why the identification of splice junctions is an important and interesting problem.

There is a long history of splice junction identification models. A prominent one is GENSCAN (Burge and Karlin,1997), which used a Maximal Dependence Decomposition (MDD) model to build a tree of the most

informative features. Yeo (Yeo, 2004) generalized this to a maximum entropy model. Later there were several attempts to use SVMs (Baten, 2006, 2008) and neural networks (Wang, 2006) to classify splice sites. Currently we are comparing our model with Yeo's model as that is the most popular model.

2 Motivation

A key aspect in the splicing process is that introns are much longer than exons (average 25k vs 100 nucleotides). Hence, the spliceosome must be able to recognise the precise junction locations from a comparatively long sequence. The precise location is important as a single base shift would change the entire protein structure. This suggests that there is a strong genomic signal near the splice junctions (apart from the GT-AG) that defines its location. The genomic signal is also confirmed by much experimental and biological evidence.

Traditionally, the detection of splice junctions from DNA sequence was used as a part of gene finding models (Burge and Karlin,1997). With the recent advances of high-throughput sequencing, there is great interest in high-throughput sequencing of RNA (known as RNA-seq). One issue with the RNA-seq protocol is that the reads generated are fairly short and frequently cannot be aligned reliably to the genome as a split read. As a result, it is hard to reliably call junctions by simply aligning RNA-seq reads to the genome. We would like to use machine learning combined with the contextual sequences to improve the reliability of RNA-seq alignments.

3 Method

RNA-seq reads from a wild type mouse cerebrum are aligned to the genome with SpliceMap (Au,2010). Junctions with more than three non-redundant reads are labelled as true junctions (positive examples). This has been experimentally shown to be a good criteria in the paper. Negative examples are generated from other sequences which contain GT-AG and are near the true junction. These sequences are the most likely to be mistaken as a splice site. The eventual goal is to apply this model to the junctions with less than or equal to three non-redundant reads in order to improve specificity of the overall alignment.

Contextual sequences extracted from the 5'(donor) and 3'(acceptor) sites of the splice junctions are used as features. For each end, we varied the length from the exon and the intron (with the canonical GT-AG sequence excluded) in order to determine the optimal sequence length for identifying splice sites. See Figure 2 below for illustration.

Figure 2 - Contextual sequences

In addition to looking at the 5' and 3' sequences individually, we also concatenated the 5' and 3' sequences that correspond to the same splice site. Further, we fixed the exon length at both ends to its optimal value and tuned the intron length to optimize the combined sequence length.

Each of the four nucleotides A, C, T and G was represented by real vectors `[a 0 0 0]`, `[0 c 0 0]`, `[0 0 t 0]` and `[0 0 0 g]`, respectively. That is, each *L*-nt training example is converted to a length *4L* real vector so that it can be directly inputted into a standard SVM package. Assuming equal distances `(a = c = t = g = 1)` yields acceptable results. However, it was found that altering the nucleotide vectors slightly improved accuracy.

We initially trained an SVM model using the LIBLINEAR library (http://www.csie.ntu.edu.tw/~cjlin/liblinear/). LIBLINEAR implements an SVM with a linear kernel, which allows for fast training on large data sets. We set LIBLINEAR to solve support vector classification with L1-regularized logistic regression. We further trained an SVM model with a Gaussian kernel by using the LIBSVM library (http://www.csie.ntu.edu.tw/~cjlin/libsvm/). The

training time of the Gaussian kernel on 200k examples takes about 10 minutes, compared to 30 seconds on the linear kernel.

In order to assess significance of appending paired donor and acceptor sites, we compared the ROC curve generated from training with appended sequences from the same splice sites versus random permuted pairings of 5' (donor) and 3' (acceptor) sequences.

4 Results

Using data from Gene Yeo's paper (Yeo, 2004), we compared our algorithm to a linear SVM and their model. Our algorithm appears to perform very similarly to their maximum entropy model (MEM) algorithm with the 5-prime data and is competitive with the 3-prime data (Figure 3).

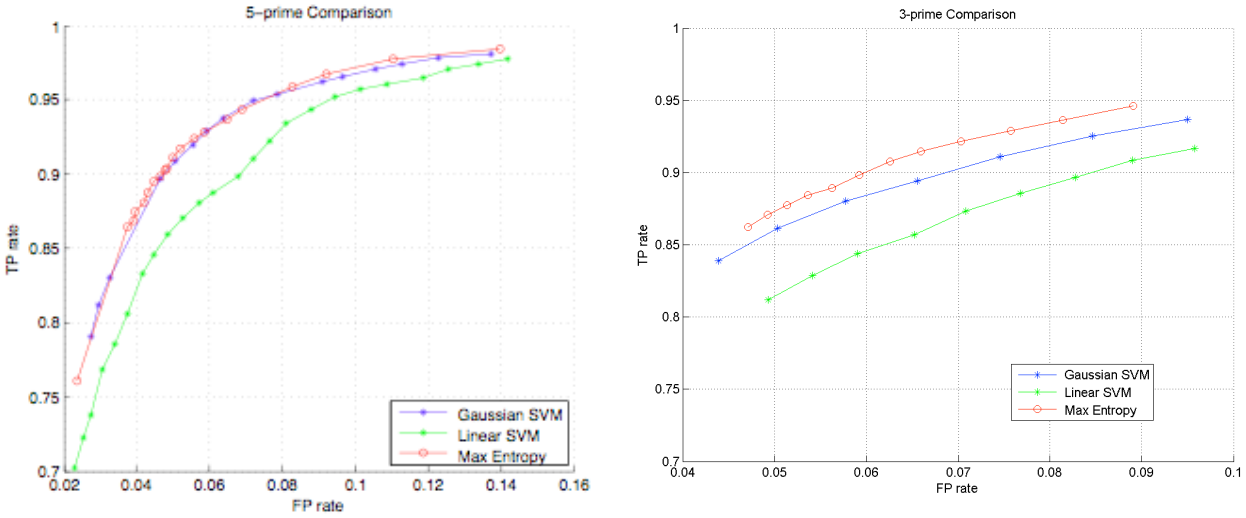

Figure 3 - Comparison between alternative models

The rest of our results used the RNA-Seq data from the previously mentioned wild type mouse cerebrum. Simple cross validation with 70% training and 30% test is used to evaluate the model's performance.

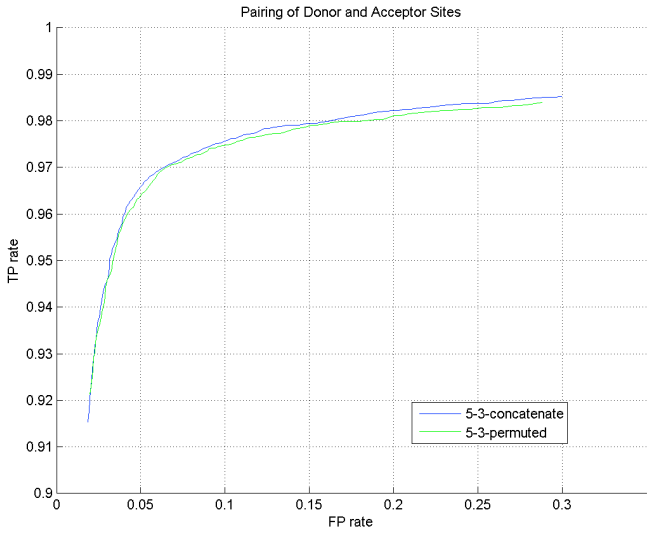

Figure 4 - Pairing of donor and acceptor sites

In Figure 4, we see that appending sequences from the donor and acceptor of the same splice site tells us slightly more information than if the two sites were randomly paired.

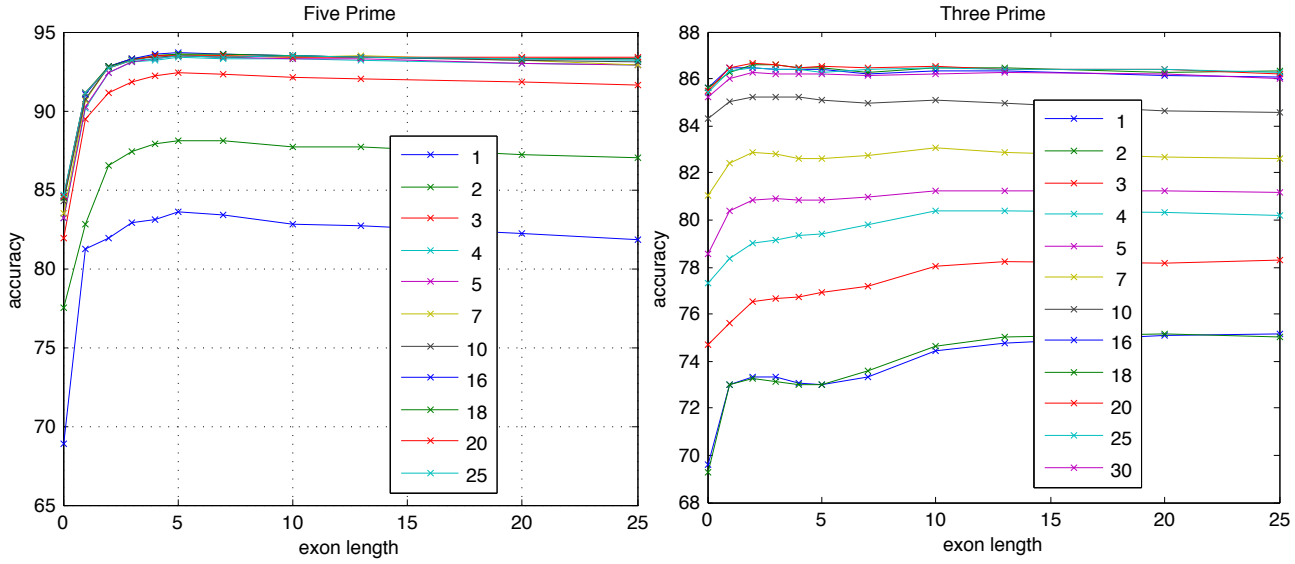

Figure 5 - 5' and 3' feature length versus accuracy

Based on our results learning from 5' and 3' separately (Figure 5), we concluded that including 5 bp from the 5' exon and 3 bp from the 3' exon should provide adequate information. In Figure 6, we concatenated the 5' and 3' sites. The 5' intron was set to a fixed length, as the length of the 3' intron was varied. As shown in Figure 6, accuracy increased as the length of the 3' intron increased. However, little improvement was observed after including 15 bp from the 3' intron. The different curves in figure 5 corresponds to a different length of the 5' intron. We also observed increased accuracy with increases in the length of the 5' intron. An acceptable outcome was observed after inclusion of 10 bp from the 5' intron.

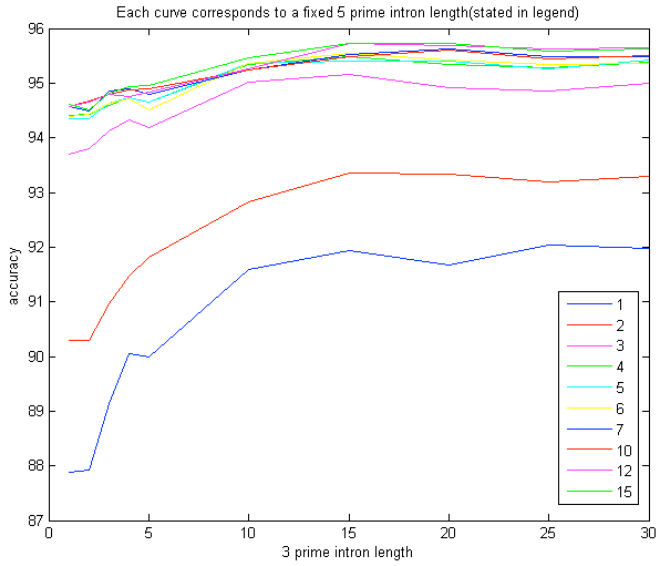

Figure 6 - Concatenated sequence accuracy

The distance between nucleotides is not expected to be symmetric since G-C bonds are slightly stronger than A-T bonds. However, it is not clear how this would should relate to the nucleotide vectors. Hence, we systematically tried some combinations nucleotide vectors, shown in Figure 7. The alternative nucleotide vectors provide better classification performance.

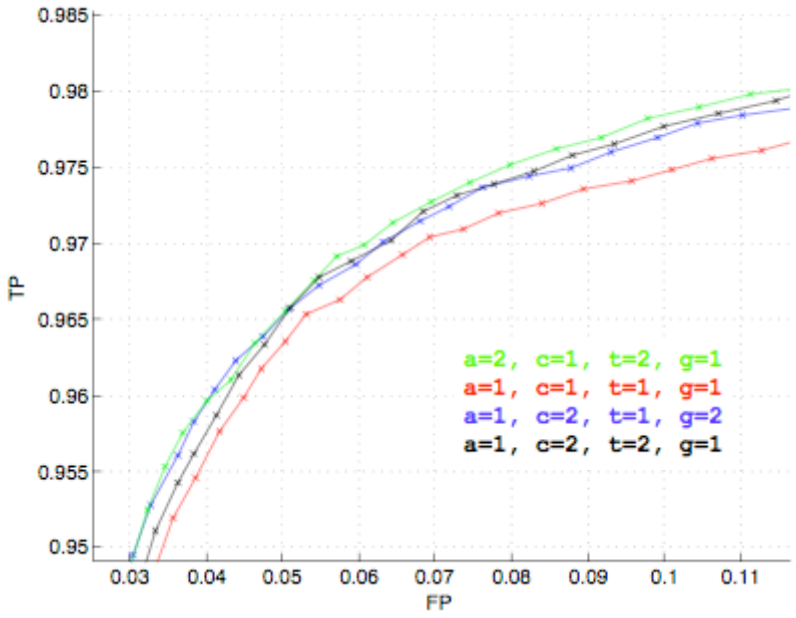

Figure 7 - Distance between nucleotides

5 Conclusion

Overall, an SVM model was implemented for classification of splice junctions. It has been shown to perform competitively compared to current models. The pairing of 3' and 5' sites was shown to provide additional information for classification. The optimal feature length for classification was examined. Finally, a non-uniform distance between nucleotides was demonstrated to improve classification accuracy. This model will be used as part of the pipeline to improve specificity of split read alignments.

6 Acknowledgements

Thanks to Quoc Le for his excellent advice, and Prof. Susan Ackerman's and Prof. Wing H Wong's labs for the mouse cerebrum RNA-Seq data. Thanks also to Prof. Andrew Ng for his fantastic teaching.

7 Appendix

Data files may be downloaded from: http://www-stat.stanford.edu/~johnmu/mouse/

References

Au, KF, Jiang, H, Lin, L, Xing, Y, and Wong, WH, **Detection of splice junctions from paired-end RNA-seq data by SpliceMap,** *Nucleic Acids Research*, 2010 doi:10.1093/nar/gkq211

Baten, A, Halgamuge, S.K. and Chang, B.C.H., **Fast splice site detection using information content and feature reduction**, *BMC Bioinformatics* **9** (Suppl. 12) (2008), p. S8.

Baten A, Chang B, Halgamuge S, Li J: **Splice site identification using probabilistic parameters and SVM classification.** *BMC Bioinformatics* 2006 , **7**(Suppl 5)**:**S15.

Burge and Karlin, 1997 C. Burge and S. Karlin, **Prediction of complete gene structures in human genomic DNA**, *J. Mol. Biol.* **268** (1997), pp. 78–94.

Wang M, Marin A. **Characterization and prediction of alternative splice sites**. *Gene* 2006; **366**: 219–227.

Yeo, G. W. and Burge, C. B. **Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals**. J. Computational Biol., 11, 475–494 (2004).