Computational Identification and Prediction of
Tissue-Specific Alternative Splicing in H. Sapiens.
Eric Van Nostrand
CS229 Final Project

Introduction

RNA splicing is a critical step in eukaryotic gene expression provides the basis for the variety of proteins expressed by an organism. Splicing, the removal of intronic regions and the consequent joining of neighboring exons in a pre-mRNA, involves recognition of and cleavage at splice sites located at the exon-intron junctions by a complex of numerous RNAs and proteins, collectively denoted as the spliceosome. Splicing is often constitutive, that is, identical pre-mRNAs are always spliced in the same manner to form identical mature mRNAs, but can also be alternative, splicing identical pre-mRNAs differently to generate unique mature mRNAs. As many as 50-75% of human genes undergo some form of AS, and is not only directly responsible for protein diversity in eukaryotes, but also plays crucial roles in the regulation of gene expression. AS is a critical regulator of the physiology of the human heart, skeletal muscle, brain, and other tissues, and has been identified as crucial in fruit-fly development. Recent efforts have applied machine learning to computationally predict alternative splicing events from sequence features, and have shown that the presence of short *cis*-regulatory elements, most notably short 5-7 nucleotide RNA binding protein binding motifs, are crucial for determining the alternative splicing of a given exon.[1][2] It is therefore not surprising that the tissue specificity of AS events would be heavily regulated by RNA binding proteins through these intronic and exonic elements, and this has been verified to occur through the activity of a variety of RNA binding proteins. However, the lack of large-scale genome-wide tissue-specific AS information has limited the ability to computationally predict the tissue specificity of AS events.

At the present time, there have been three major attempts to determine AS events on a genome-wide scale. In 2003, Johnson *et al.* used exon junction micro-arrays in which array probes were localized to the junctions of neighboring exons to identify exon skipping events across 52 tissues and cell lines [3]. In addition, in 2006 *Sugnet, et al.* described a smaller set of exons specifically included or skipped in mouse brain and human tissues, as found by micro-arrays with probes targeting known or predicted tissue-specific splicing events [4]. However, the most high-quality data currently available was published by Affymetrix, who performed exon micro-arrays in 11 human tissues [5], which have been used to define AS events specific to different tissues [6]. Using these datasets, I set out to define features which would define these classes of AS events, and which could be used to predict the tissue specificity of AS events *in silico*.

Generation of Training Data & Feature Selection

To determine a set of tissue-specific alternative skipping events, exon-specific micro-array data for human brain and muscle was downloaded from Affymetrix [5]. Each exon in the human genome was defined as consisting of one or more probe set regions, each of which consists of multiple (up to 4) individual probes. For the set of probes in each probe set region, the normalized expression values in brain were compared to those in muscle, and this was then compared to the equivalent ratios for all exons in the gene. Thus, while constitutive exons will show an approximately equal ratio of expression in brain vs. muscle across an entire gene, an exons alternatively skipped in brain compared to muscle will show dramatically lower expression in brain compared to the other constitutive exons [6]. By performing this comparison across all exons in the genome, it is possible to obtain a score for each exon that describes the confidence with which it can be identified as skipped in muscle (included in brain), or skipped in brain (included in muscle).

Setting the cutoff for calling an exon alternatively spliced to a value of 3 standard deviations above or below the median expression value in the gene yielded 263 muscle skipped exons (the "+" training set) and 148 brain skipped exons (the "-" training set), or 1304 and 686 respectively for 2 standard deviations. As a first approach, I trained a Naïve Bayes model on these two training sets, using counts of all 1024 possible 5 nucleotide long elements (5-mers) in the exon, as well as in the

flanking 400nt of the upstream and downstream introns. In addition, using alignments of these exons to orthologous exons in the mouse, rat, and dog genomes [7], I counted the number of 5-mers in the exon and flanking introns that were conserved across all four of these genomes, giving a total of 6144 conserved and non-conserved 5-mer features.

Using filtering-based feature selection, I next turned to using only those k-mers which differed significantly between the two categories. I initially implemented two approaches to identify these informative features: first, the mutual information (MI) of each 5-mer in each location was calculated with respect to the positive and negative training examples, and second, three different χ^2 enrichment values were calculated. For each 5-mer, I calculated the χ^2 value from the 2x2 table of each specific 5-mer against all other 5-mers in the human sequence of the "+1" vs. "-1" class, the similar 2x2 table for counts of conserved occurrences across the 4 mammalian genomes of a specific 5mer against all other 5-mers in "+1" vs. "-1", and the enrichment of conservation of the 5-mer relative to its frequency in the human sequences as compared to the typical conservation rate for all other 5-mers. For this study, only 5-mers were used due to the relative scarcity of positive training examples, which would make it difficult to obtain statistically significant results for sequences larger than 5-mers. Both approaches were used to calculate information both for 5-mers in the human sequence, as well as for the set of 5-mers that are not only present in the human sequence but are also conserved in sequence in the mouse, rat, and dog orthologous sequence. As the MI strategy gave a similar list of 5-mers as did the simpler χ^2 test for enrichment, I chose to focus only on using features based upon the χ^2 tests, which also have the advantage of easier biological interpretation. To avoid over-fitting, I chose my set of 5-mer features to consist of 5-mer counts in the human sequence for the subset of 5-mers which were significantly enriched ($p < 0.001$, corrected for multiple hypothesis testing) either in the human sequence or for conservation relative to the human sequence, and 5-mer counts in conserved aligned sequence for those which were significantly enriched for conservation or significantly enriched within conserved sequence overall.

I also used additional non-k-mer features in developing the classifiers described below. Specifically, conservation of the exon, upstream intron, and downstream intron were each used as features, as these have been previously found to be characteristic of alternative splicing events. In addition, I chose to use splice site strength [8], as well as overall G/C content of the exon and flanking introns as features which should help lower the generalization error of this model. The overall expression of the gene containing each exon was also tested as a feature for both classifiers described below, but did not provide enough additional classification power to make its inclusion valuable (after considering that such data was only available for <50% of transcripts).

Prediction of novel tissue-specific skipping events

Using the feature selection described above, I first considered the problem of classifying brain skipping events as compared to exons not tissue-specifically alternatively spliced in brain. Using the 289 5-mer features that were significantly enriched or depleted ($p<0.00001$, corrected for multiple hypothesis testing) in the set of 686 brain skipped exons, I first trained a Naïve Bayes model using a randomly selected set of 686 non-skipped exons as the "-1" control set. Using 10-fold cross-validation, I found that either with or without the use of the additional non-k-mer features, the Naïve Bayes model had ~45% generalization error. With such high error, I next turned to a SVM approach (which typically has greater success in classification problems with reasonably highly sized set of training examples). As the SVM implementation developed in class took far too long to train on the set of ~1300 training examples, I turned to utilizing the SVM[light] publicly available SVM implementation, which not only had the benefit of faster run-time, but also easily allows differential weighting of inaccuracy on "+1" vs. "-1" training examples.

Using 10-fold cross-validation, the training set was used to build an SVM model using typical parameters (C=1 and termination criteria of 0.001), and it was found to have ~58% accuracy. However, there are two main flaws in only considering generalization error for this problem: first, it is possible that there are improperly labeled "+1" examples in the "-1" set that were simply missed by the micro-array analysis, and second, it is more important to have a low false positive rate than to accurately identify all positive training examples. Thus, by setting a cutoff such that examples are only classified as "+1" if they

are at least *l* above the hyper-plane defined by the support vectors, I was able to check whether the algorithm was able to classify a smaller subset of positive training examples with higher confidence. Unfortunately, this approach was met with limited success; using a cutoff of 0 (a default SVM implementation), there were 385 accurate positive predictions as compared to 281 inaccurate positive predictions (57.8% accurate), whereas at a cutoff of 1.7 (corresponding to an overall false-positive rate of 10%), there were 148 accurate as compared to 68 inaccurate predictions (68.5%). Thus, I was only able to gain ~10% accuracy at a tradeoff of losing over 60% of the positive data.
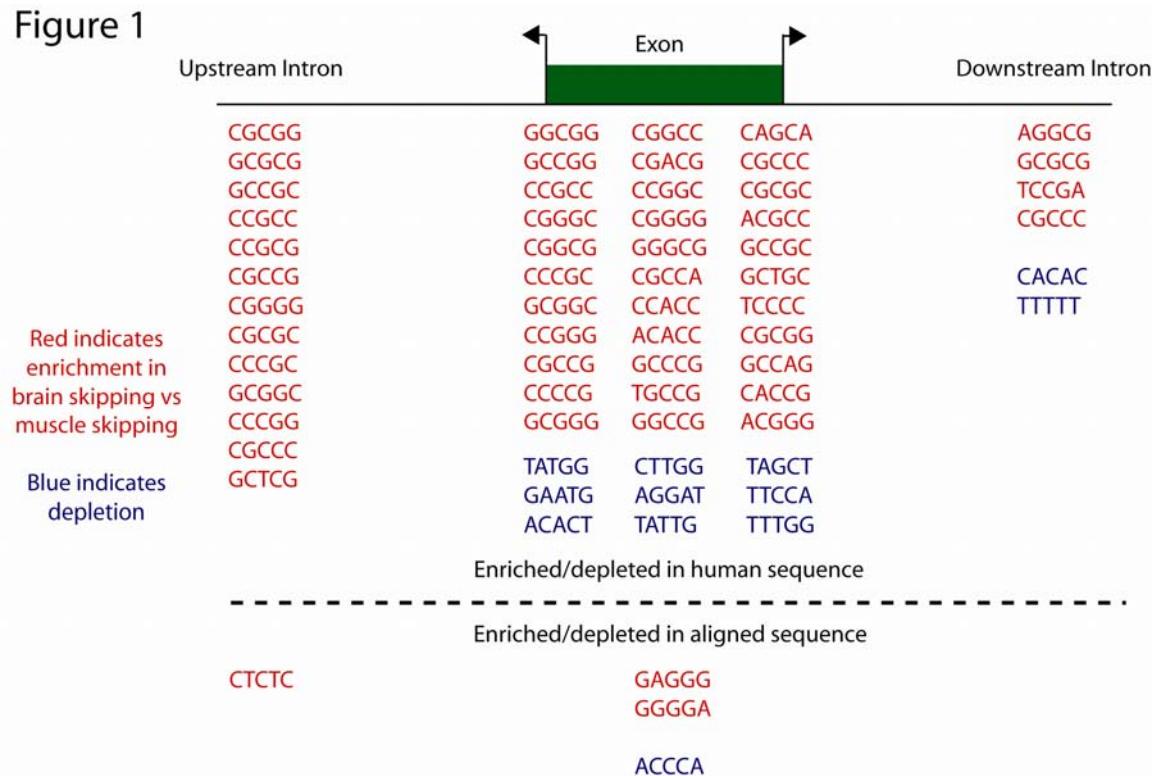
By implementing the corresponding procedure for muscle-specific skipping events, using the 309 *k*-mer features as described above, and training an SVM using similar parameters, it was found to have only ~55% accuracy. Similar to the brain-specific classification attempt, at a cutoff of 0 there were 720 accurate positive predictions as compared to 579 inaccurate predictions, whereas using a cutoff of 1.4 gave a ~7% increase in accuracy to 219 to 135 accurate to inaccurate positive predictions.

Although these results are discouraging, it is clear that there is some amount of predictive power in the features being used, and furthermore that the most trustworthy predictions (those furthest from the separating hyper-plane) are more likely to be accurate than those which are less clearly separated. As the tissue-specific training examples are from the first generation of exon-arrays, and the algorithms to identify these events are currently very basic, with better array technology and more accurate algorithms it should become possible to generate a larger set of more confident tissue-specific events. By using these events, it should be possible to implement this procedure with even greater success.

Classification of brain- vs. muscle-specific skipping events

As a secondary question, I set out to develop a classifier that, when given a list of exons that are known to be tissue-specific AS events in either brain or muscle, could accurately predict which tissue the AS was specific to. Although this is perhaps less informative for genome-wide identification of novel AS events, it is nonetheless interesting from a biological perspective to both be able to identify the features common to each class of events, as well as to identify tissue-specificity for AS events *de novo*, and to help annotate the specificity of events that were either not tested or are experimentally difficult to test by the current high-throughput approaches. Using the set of 686 brain-specific and 1304 muscle-specific skipping events, I extracted the set of 69 5-mer features that were significant at *p<0.001* after correcting for multiple hypothesis testing (Figure 1), along with the other features (such as G/C content, conservation, etc) as described above. Using these features, I again trained an SVM using SVM[light], with the parameter that training errors on brain-specific events ("+1") would outweigh errors on the muscle-specific events ("-1") by a factor of 1304/686 = 1.90. This was done in an effort to utilize all possible training data, while avoiding an algorithm that simply predicted "-1" for all feature vectors.

Figure 1

Upstream Intron   Exon   Downstream Intron

Upstream Intron	Exon			Downstream Intron
CGCGG	GGCGG	CGGCC	CAGCA	AGGCG
GCGCG	GCCGG	CGACG	CGCCC	GCGCG
GCCGC	CCGCC	CCGGC	CGCGC	TCCGA
CCGCC	CGGGC	CGGGG	ACGCC	CGCCC
CCGCG	CGGCG	GGGCG	GCCGC	
CGCCG	CCCGC	CGCCA	GCTGC	CACAC
CGGGG	GCGGC	CCACC	TCCCC	TTTTT
CGCGC	CCGGG	ACACC	CGCGG	
CCCGC	CGCCG	GCCCG	GCCAG	
GCGGC	CCCCG	TGCCG	CACCG	
CCCGG	GCGGG	GGCCG	ACGGG	
CGCCC				
GCTCG	TATGG	CTTGG	TAGCT	
	GAATG	AGGAT	TTCCA	
	ACACT	TATTG	TTTGG	

Red indicates enrichment in brain skipping vs muscle skipping

Blue indicates depletion

Enriched/depleted in human sequence

- -

Enriched/depleted in aligned sequence

CTCTC   GAGGG
                         GGGGA

   ACCCA

Figure 1: 5-mer features used for the SVM. Elements are grouped by their location relative to exons, with elements enriched in brain (as compared to muscle) in red, and depleted elements in blue.

Using 10-fold cross-validation to train and test the classifier, it obtained ~64% accuracy with a C value of 1 (modification of this C value did not significantly alter the accuracy of the classifier). In the hopes of increasing classification accuracy, I again tested the effect of requiring that the margin of an example be more than a certain cutoff. Testing different cutoffs, I determined that at an optimal cutoff of ~1.15, the examples predicted to be in the +1 or -1 class were over 3.2-fold more likely to be accurate than inaccurate, corresponding to an accuracy over 75%, whereas without such stringent cutoffs there was only a ~1.7-fold enrichment for accuracy (Figure 2, Full SVM line). To test the effect of different feature sets, I completed the same analysis for an SVM using only k-mer features (i.e., not using conservation, G/C content, etc), which achieved fairly similar accuracy with low cutoff values, but was not able to achieve over 75% accuracy at any cutoff (Figure 2, SVM with only k-mer features line). After retraining this model over the entire training set, I predicted the tissue-specific splicing for all ~90,000 exons in the full exon micro-array dataset, and obtained ~7600 predicted brain-specific and ~14000 predicted muscle-specific events. Although most of these events are likely to be false-positives (since they are likely not alternatively spliced in any tissue), it would be interesting to experimentally test a subset of these exons using more stringent methods than the exon micro-array to determine if many of these exons do in fact show a tissue-specific splicing pattern.

As an additional test, I attempted to use the ratio of overall expression of the gene in brain vs. muscle as an additional feature, based upon the GeneAtlas publicly available gene-level tissue expression [10]. Using only the subset of training examples for which this data was available, I trained an SVM on this ~1/3 of the full training set with the additional gene expression feature. Perhaps as a result of this smaller training set, training an SVM using this additional feature resulted in far lower accuracy (Figure 2, Full SVM with gene tissue specificity added line). Retraining the algorithm using different parameters (C, stopping criteria) as well as using different kernels (including polynomial kernels of different degree) gave similar or worse accuracy, indicating that the results in Figure 2 are approaching the limit of discriminatory power within this training set.

Figure 2
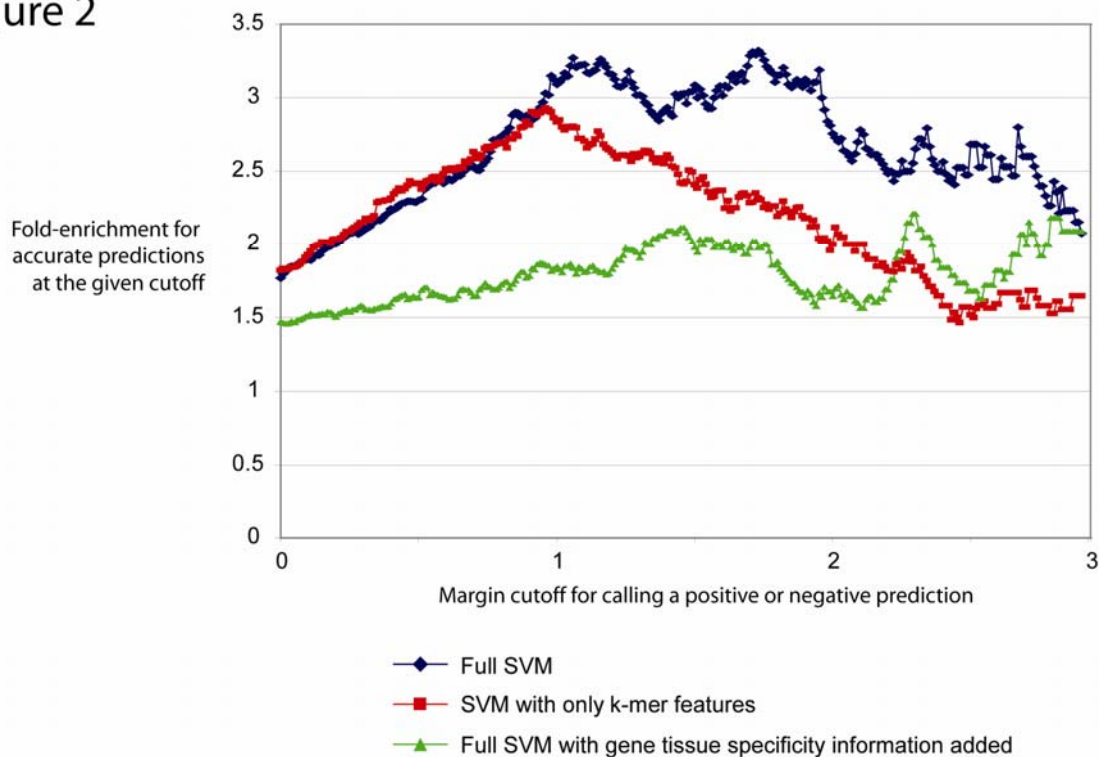
Figure 2: SVM accuracy with various margin cutoffs. Multiple variants of the SVM were trained and compared for their ability to make correct predictions above different margin cutoff values.

Conclusion

Gaining knowledge about the mechanisms behind tissue-specific alternative splicing is a critical step in the understanding of what distinguishes different tissues in mammalian organisms. Using genome-wide tissue-specific AS information from recent exon micro-array analyses in human brain and muscle tissue, I was able to achieve moderate success in classifying exons according to their tissue-specific AS pattern. Using sequence information alone, I was able to achieve over 75% accuracy in distinguishing between brain-specific and muscle-specific events, a level of success that could enable further experimental analysis to verify predictions of novel brain-specific or muscle-specific events that are either not covered by the current micro-array technology, or are simply differences that are too small to identify using current micro-array analysis tools. With the continual availability of splicing micro-arrays from additional tissues, it will be interesting to apply this approach to events observed in other tissues, in the hopes that the use of additional data will allow better identification of AS events, and thus make possible better classification tools.

[1] Yeo, G.W., *et al. Proc Natl Acad Sci U S A*. **102:** 2850-5.
[2] Sorek, R., *et al. Genome Res.* **14:** 1617-23
[3] Johnson, J.M., *et al. Science*. **302**. 2141-4
[4] Sugnet, C.W., *et al. PLoS Computational Biology*. **2**. e4.
[5] *http://www.affymetrix.com/products/arrays/exon_application.affx*
[6] Yeo, G. *Personal communication*
[7] Blanchette, M. *et al. Genome Res* **14**, 708-15.
[8] Yeo, G., *et al. J Comput Biol*. **11**. 377-94
[9] Joachims, T. Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.
[10] Su, A.I. *et al. Proc Natl Acad Sci U S A*. **99**: 4465-70.