Tissue-Specific Classification of Alternatively Spliced Human Exons

by

Craig Jeremy Rothman

S.B. Electrical Engineering and Computer Science
S.B. Biology
Massachusetts Institute of Technology, 2005

SUBMITTED TO THE DEPARTMENT OF BIOLOGICAL ENGINEERING IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ENGINEERING IN BIOMEDICAL ENGINEERING
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2007

Signature of Author: ______________________________
Department of Biological Engineering
January 19, 2007

Certified by: ______________________________
Christopher Burge
Associate Professor of Biology and Biological Engineering
Thesis Supervisor

Accepted by: ______________________________
Bevin Engelward
Associate Professor of Biological Engineering
Program Director

Accepted by: ______________________________
Alan Grodzinsky
Professor of Biological Engineering
Chair of BE Graduate Committee

Tissue-Specific Classification of Alternatively Spliced Human Exons

by

Craig Jeremy Rothman

Submitted to the Department of Biological Engineering
on January 19, 2007 in Partial Fulfillment of the
Requirements for the Degree of Master of Engineering in
Biomedical Engineering

Abstract

Alternative splicing is involved in numerous cellular functions and is often disrupted and involved in disease. Previous research has identified methods to distinguish alternative conserved exons (ACEs) in human and mouse. However, the cellular machinery, the spliceosome, does not use comparative genomics to decide when to include and when to exclude an exon. Human RefSeq exons obtained from the University of California Santa Cruz (UCSC) genome browser were analyzed for tissue-specific skipping. Expressed sequence tags (ESTs) were aligned to exons and their tissue of origin and histology were identified. ACEs were also identified as a subset of the skipped exons. About 18% of the exons were identified as tissue-specifically skipped in one of sixteen different tissues at four stringency levels. The different datasets were analyzed for both general features such as exon and intron length, splice site strength, base composition, conservation, modularity, and susceptibility to nonsense-mediated mRNA decay caused by skipping. *Cis*-element motifs that might bind protein factors that affect splicing were identified using overrepresentation analysis and conserved occurrence rate between human and mouse. Tissue-specific skipped exons were then classified with both a decision-tree based classifier (Random Forests™) and a support vector machine. Classification results were better for tissue-specific skipped exons vs. constitutive exons than for tissue-specific skipped exons vs. exons skipped in other tissues.

Thesis Supervisor: Christopher Burge
Title: Associate Professor of Biology and Biological Engineering

Acknowledgements

I would like to thank my advisor, Dr. Christopher Burge for giving me the opportunity to perform this research, and for providing his counsel, advice, and guidance. I would also like to thank all members of the Burge laboratory who offered support and recommendations throughout this process. Specifically I would like to thank Rickard Sandberg and Xinshu (Grace) Xiao for their contributions to my data (mouse skipped exons and known tissue-specific binding motifs) and their methods for finding significant motifs. I would also like to thank Michael Stadler who was always willing to answer questions and help me with computer and cluster problems.

Spending five-and-a-half years at MIT has also allowed me to make great friendships with so many wonderful people. My friends from the crew team, Baker House, and the Class of 2005 were always there to provide a distraction from work. They provided the social relief I needed to get through the arduous work that makes MIT, MIT. I would also like to thank my friends at Theta Xi who have provided friendship and support throughout the last four-and-a-half years. And last, but certainly not least, I would also like to thank my parents and sister who were always there to encourage me through all the work, late nights, and always believed in me.

Table of Contents

Table of Figures

Table of Tables

1 Introduction

Since the discovery of the structure of DNA by Watson and Crick in the 1950s, scientists have been studying how cells function with regard to their genome and proteins. In the 1960s, biologists decoded the method by which DNA encodes for different amino acids and thus proteins. By latest accounts, there are over 20,000 genes in the human genome and many have not been studied in detail. With the advent of the human genome project, more computational research has become possible allowing for speedier analysis versus that which can be done on a laboratory bench. Areas of computational research include protein folding prediction and analysis of transcription and translation controls. A subset of this research is the study of alternative RNA splicing.

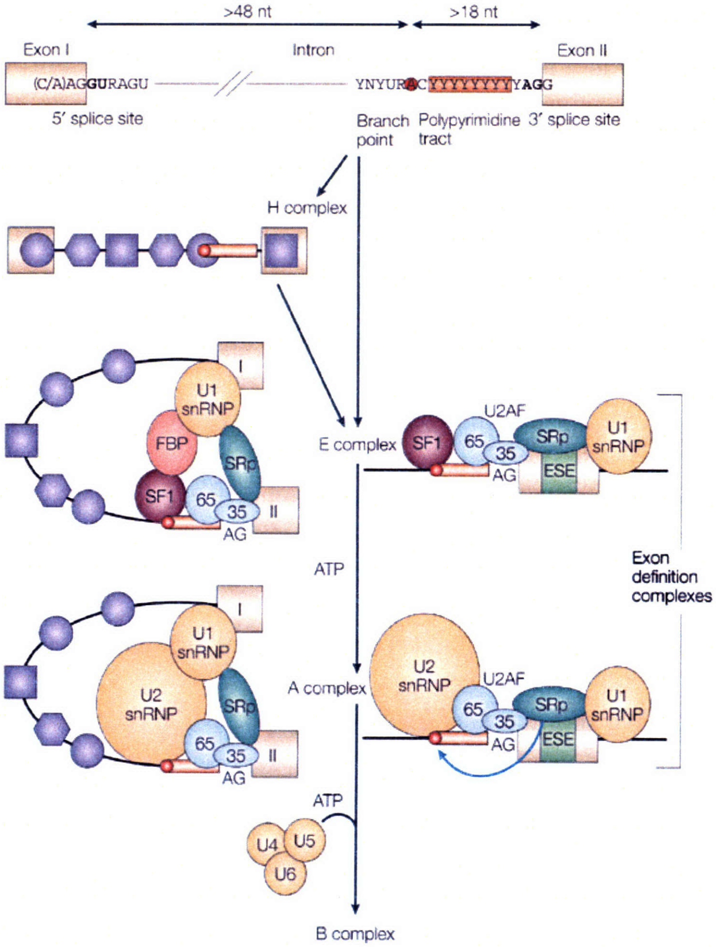
When the human genome project was finished, the fact that humans had only 20,000-30,000 genes instead of the previously thought 100,000 genes shocked many scientists. However, alternative pre-mature messenger ribonucleic acid (pre-mRNA) splicing, which results in different mature messenger ribonucleic acid (mRNA) products, and thus different proteins, may explain the small number of human genes. Yet, the details of how the cell decides to splice a pre-mRNA differently are not well understood. In this paper, I explore computational methods to predict whether an exon will be constitutive or alternatively spliced in certain tissues based solely on the sequence of a single species, human.

1.1 RNA Splicing

A living cell uses its genetic information, which is composed of deoxyribonucleic acid (DNA), to direct synthesis of specific proteins, the enzymes that control cellular function. The cell transcribes DNA to pre-mRNA, then processes pre-mRNA into mature mRNA, and then translates mature mRNA into proteins. The middle step, the processing of pre-mRNA into mature mRNA is a field of much current interest as it is involved in many diseases and the function of different tissues. The pre-mRNA is composed of exons and introns. Exons are the portions of DNA that exist in the mature RNA and code for amino acids, while the introns are the areas of DNA that do not exist in the mature RNA because they are 'spliced out'.

Currently, there exists a basic understanding of the locations in a gene where splicing occurs and it is possible to predict the exon-intron structure of a gene just from the sequences of chromosomes (3). Exon-intron structure prediction is an important step in understanding splicing and the basic cellular machinery and the process in which it splices out the introns and joins the exons to produce mRNA has been identified. At the 3′ end of the exon/5′ end of the intron, exists the 5′ splice site, and at the 5′ end of the exon/3′ end of the intron, exists the 3′ splice site. A branch site (an adenosine deoxyribonucleic acid) and a polypyrimidine tract exist in the intron within about fifty nucleotides from the 3′ splice site. Splicing occurs via two trans-esterification reactions. First, the 2′ OH of the adenosine at the branch site attacks the 5′ splice site. Then, the 3′ OH left behind by the attack on the 5′ splice site attacks the 3′ splice site. This leaves the exons joined and the intron excised as a lariat that is then de-branched and degraded. The catalysis of RNA splicing involves five small nuclear ribonucleoprotein (snRNP) particles, U1, U2, U4, U5, and U6. These particles recognize the branch and splice sites, bring them together in the proper order, and then catalyze the process of RNA cleavage and joining (see Figure 1) (1).

There are other protein factors and DNA sequences that help define the splice site and allow the snRNPs to recognize them. The many other protein factors include U2AF and SR proteins, which usually bind to pre-mRNA elements (see Figure 1) (1). The pre-mRNA elements are certain sequences of nucleotides called exonic splicing silencers (ESS), exonic splicing enhancers (ESE), intronic splicing silencers (ISS), and intronic splicing enhancers (ISE), the name defining where they are located in the gene and what function they play in splicing. These elements have recently received much attention and are discussed further below.



Nature Reviews | Molecular Cell Biology

Figure 1: Spliceosome Assembly. The splice sites and branch site are bound by small ribonucleoprotein particles (yellow circles) and other factors (blue and green ovals) that recognize the exon and splice out the intron. The blue circles, squares, and hexagons represent hnRNPs binding to the mRNA and affecting spliceosome assembly. Adapted from (1).

As discussed earlier, most genes code for more than one protein in human and other vertebrates through alternative splicing. Alternative splicing occurs when the splicing machinery combines fragments of the pre-mRNA in different combinations to produce different mature mRNAs from one gene.

1.2 Alternative RNA Splicing

There are five common types of alternative splicing: exon skipping, intron retention/inclusion, alternative 5′ and 3′ splice site usage, and mutually exclusive exon usage (see Figure 2). In each type except for intron retention, all or some part of an exon is not contained in the final mature mRNA product in the alternatively spliced version. In intron retention, what would normally be an intron is instead included in the mature mRNA to form part of an exon. It is currently believed that the genome contains about 20,000-30,000 genes and about 30-70% of these genes contain alternatively spliced exons. However, it is difficult to predict the exact number that contain alternatively spliced exons for the following three reasons: (i) it is hard to determine which splicing events are functional; (ii) some are specific to disease states or mutations; and (iii) the current annotation may be missing some exons. Most available data comes from expressed sequence tags and microarray expression data, both of which have deficiencies (4). There are many biological functions for alternative splicing and the most prevalent type of alternative splicing is exon skipping (5).

Exon skipping can cause three major changes to the final structure of the mRNA. After removing an exon it can: (i) keep the rest of the mRNA the same without disrupting the reading frame; (ii) introduce a frameshift and cause the spliceosome to reach an early stop codon, usually generating a substrate for nonsense-mediated mRNA decay (NMD); or (iii) remove an exon and expand the mRNA into a longer functional form. Each of these results can have a diverse range of effects on the mRNA product and translated protein (6-8). Alternative splicing has also been associated with disease due to an alteration of a needed protein and may be caused by mutations in the gene or by misregulation of a splicing factor (9).

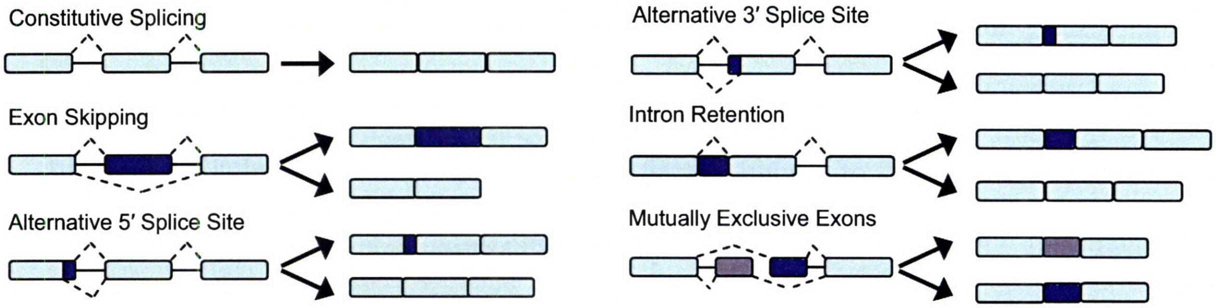


Figure 2: Types of Alternative Splicing. Constitutive splicing and five common types of alternative splicing are shown. Colored boxes represent exons, lines between the boxes represent introns, and dashed lines represent how the splice sites combine to form the mature mRNA. Dashed lines on the top and bottom represent alternative splice forms with the two mature mRNA products shown to the right of the arrows.

Many sequence features or *cis*-acting RNA elements in both exons and introns have been discovered that help regulate alternatives splicing. Proteins may bind to these elements in the cytoplasm and alter splicing. These features are called exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs), and intronic splicing silencers (ISSs). Their names derive from where they are located in a gene and what action they perform, to enhance splicing or to silence splicing (10-16). Many of the exonic splicing regulatory motifs are conserved among most vertebrates, but the intronic splicing regulatory motifs appear to differ between mammals and fish (16). In addition, many of these elements are specific to certain tissues and may be used by the cell to control tissue-specific alternative splicing (17-19). There are three possible mechanisms that may regulate tissue-specific alternatives splicing: (i) factors may be tissue-specifically expressed; (ii) factors may be expressed everywhere but at different levels in different tissues; and (iii) factors may be expressed everywhere but tissue-specifically spliced genes have evolved to only contain certain RNA elements (7).

1.3 Machine Learning

Machine learning is a subset of artificial intelligence and is the process by which a machine analyzes, learns, and separates a group of distinct classes from each other. Machine learning usually involves giving a classifier two sets composed of feature vectors: a training set, used to learn a discrimination task, and a test set, used to test the classifier and its performance.

There are two types of machine learning: unsupervised and supervised. Unsupervised learning occurs when the machine separates the data into different groups without having any knowledge of which data belongs to which set or class. Supervised learning occurs when the machine separates the data into different groups with the knowledge of which class each data vector belongs to. Unsupervised learning is generally harder and more complex because the machine has to determine which class each element belongs to and then create a classifier around that. Supervised learning is somewhat easier because the machine/algorithm has more information, and works by taking each element along with its features and class and trying to define a separating plane (in an appropriate feature space). There exist many algorithms for supervised learning that use different methods for trying to separate the classes.

1.3.1 Random Forests

Random Forests™ (RF) is a supervised learning method created by Leo Breiman and Adele Cutler that is based on decision trees (20). RF works by combining decision tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. During tree development, bagging is used along with random feature selection. Each new training set is drawn, with replacement, from the original training set, and then a tree is grown on the new set using random feature selection. The final tree is not pruned. Bootstrapping is used for each training set—about one-third of the training data is left out, and thus no cross-validation or separate test set is needed to obtain an unbiased error estimate. After a large number of trees are generated, each tree 'votes' for the most popular class. The error rate is based on the correlation between any two trees and the strength of each individual tree in the forest. Increasing the correlation increases the error rate while increased tree strength decreases the error rate. The correlation and strength are determined by how many variables are selected at random for each node, the mtry0 value; the more variables the greater the correlation and the strength, and thus there is some number of variables that produce a correlation and strength that give the best classifier (20).

RF has several advantages over other types of supervised learning algorithms when it comes to the dataset under question (21):

- It is unexcelled in accuracy among current tree-based algorithms, such as AdaBoost.
- It can handle thousands of input variables without variable deletion.
- It gives estimates of what variables are important in the classification.
- It generates an internal unbiased estimate of the generalization error as the forest building progresses.
- It has methods for balancing error in class population unbalanced data sets.
- Generated forests can be saved for future use on other data.
- Prototypes are computed that give information about the relation between the variables and the classification.
- It computes proximities between pairs of cases that can be used in clustering, locating outliers, or (by scaling) give interesting views of the data.

RF always converges and thus it does not overfit, as experienced in this research, due to the Strong Law of Large Numbers, proved on page 30 of "Random Forests" in Machine Learning, 2001. RF has also been shown to be an accurate classifier even when noise is introduced into the data (20). While the above advantages make RF look like a great tool for classification, Hastie *et al.* notes that although trees have emerged as the most popular learning method for data mining, they are seldom the ideal tool due to inaccuracy (22).

It also has been shown that RF works well with imbalanced datasets. The data set used here, constitutive and skipped exons, are imbalanced as there are about five times as many constitutive exons as skipped exons and even more of an imbalance when looking solely at tissue-specific skipped exons, for which current datasets number in the low hundreds. RF implements a class weights method to alter the weights during training so one can use different sized training sets for each class. The class weights are incorporated into the algorithm in the tree induction procedure (class weights are used to weight the Gini criterion for finding splits) and in the terminal nodes of each tree. Chen and co-workers tested this method on datasets such as oil, mammography, satellite images, hypothyroid, and eurothyroid where the minority class is 2-10% of the majority class. Good prediction results were obtained when compared against several other weighted methods (23).

1.3.2. Support Vector Machines

Support vector machines (SVMs) describe a type of classifier that has many implementations with different variations and features. SVMs work by taking a training set and trying to separate the data by transforming the features onto some higher dimensional hyperplane as specified by the inputs into one of the implementations of the algorithm (24). Most implementations of an SVM allow the user to specify the type of kernel used to separate the data: linear, radial, polynomial, etc. The algorithm then determines which vectors of the training set will act as support vectors for the classification. SVMs work well for many classification problems and depend largely on the type of kernel chosen. However, they are generally slow and do not work well with large datasets and feature spaces. It is also possible to overfit the classifier such that it cannot be generalized to a greater set of data than what the classifier was trained on, and thus it requires a test set separate from the training set for cross-validation (25). However, SVMs are good classifiers when the data classes overlap and they have good accuracy once the correct kernel is chosen and other parameters are tuned (22).

1.4 Previous work on predicting alternative splicing

Expressed sequence tags (ESTs), complementary DNA (cDNA), and microarray data have been used to identify alternatively spliced exons. There are many methods that use this data to detect alternatively spliced exons (26-32). However, ESTs and cDNAs are not always high quality and cDNAs that were spliced incorrectly or came from a damaged or diseased cell could lead to detection of false positive alternative splicing events. This method also does not allow detection of alternatively spliced exons that occur at low levels or in tissues that have not been adequately surveyed for ESTs and cDNAs. The human genome project has produced the genetic information necessary to discover a molecular code for how the splicing machinery decides which exons to splice together. The ability to determine whether an exon is alternatively or constitutively spliced based on the gene sequence would the prediction of alternative splicing in genes where it has not been experimentally observed.

There has been much progress in this field over the past few years, but only in specialized cases of exon skipping. Computational biologists have developed methods that predict conserved alternatively spliced exons (ACEs), exons that are alternatively spliced in more than one species, using SVMs (33-35). They predict ACEs instead of just skipped exons because it increases the likelihood that the skipping is functional. In addition, there are many sequence features that separate exons skipped in more than one species and exons that are skipped in only one species or not at all (36). These classifiers use sequence conservation of both exons and flanking introns between species such as human and mouse as a major feature to predict exon skipping as they differ greatly between ACEs and constitutive exons (34, 37). They also use other sequence features such as exon and intron length, exon divisibility by 3, counts of tetra- and penta-nucleotides in exons and flanking introns, poly-pyrimidine tract (PPT), and splice site scores, which differ between constitutive and alternatively spliced exons.

One issue with these classifiers is that they require the use of comparative genomics, the conservation of sequence between two related species. However, the cellular machinery does not use comparative information to make its decision when deciding how to splice pre-mRNA. Another issue is that since they use conservation, they only predict alternative splicing events that are common among several species; however, many exons may be alternatively spliced in only a single organism (and perhaps very closely related species). Thus, a method to predict exon skipping using the genome of just a single organism is needed. Ratsch and co-workers have made some headway in this area with the nematode (roundworm) genome *C. elegans* (38). The *C. elegans* exon classifier used an SVM and it reportedly had a true positive rate of 48.5% with a false positive rate of 1%. However, the *C. elegans* genome is much simpler than mammalian genomes and their method does not work for mammals.

The sequence and biological differences between alternatively spliced exons and constitutive exons have been explored and identified (39-41). For example, guanosine-rich motifs interact with heterogeneous nuclear ribonucleoprotein F (hnRNP F) and hnRNP A1 to affect exon skipping (13, 42). Iida and co-workers found that GC-ending codons were found more often in constitutive exons than alternative exons in both *Drosophila* and humans (43). Yeo and co-workers found that 70% of alternative conserved exons (ACEs) had a length that was a multiple of three compared to around 40% for non-conserved skipped exons and constitutive exons. These conserved skipped exons also were shorter than constitutive exons and they had greater conservation of their nucleotides in both the exon and first 150 nucleotides in the upstream and downstream introns (34). Zhuang and co-workers found that amino acid usage was nearly identical between constitutive and alternatively spliced exons and also observed that the average length of alternatively spliced exons was less than that of constitutive exons. Clark and Thanaraj found that a majority of skipped exons are modular, occur in low G+C regions, and have weaker splice signals than constitutive exons (44).

The biological differences between tissue-specific skipped exons and non-tissue-specific skipped exons have also been explored and identified. Xing and co-workers have identified that tissue-specific exons are more likely to be modular and have a length that is a multiple of 3 than constitutive exons by using microarray data (45). Others have identified certain *cis*-acting sequence elements that exist in tissue-specific skipped exons (17, 18). Yet, much more needs to be learned about tissue-specific skipping and it has not been compared to alternative conserved skipping events.

1.5 Prediction of alternative splicing and tissue-specific skipping

I have worked to create a program that would be able to identify an exon as alternatively or constitutively spliced based solely on the sequence of one mammalian genome. The initial version of the program is for human exons and is tissue-specific, i.e. there are multiple different versions of the classifier trained for different tissues, beginning with those tissues such as brain, liver, testis and skeletal muscle that have high rates of alternative splicing and large amounts of available transcript data (2). I have also identified different sequence features and motifs of tissue-specific exons and compared them to those from non-tissue-specific exon and alternative conserved exons.

Most previous programs used SVMs as the exon classifier (33, 34, 38). SVMs are good classifiers when the data classes overlap as they produce nonlinear boundaries by constructing a linear boundary in a large, transformed version of the feature space (22). However, they do not work as well with a large number of variables and are relatively slow to train. Thus, I used both RF and an SVM to perform three two-way classifications: all tissue-specific skipped exons vs. constitutive exons, tissue-specific skipped exons for each tissue vs. all other tissue-specific skipped exons, and tissue-specific skipped exons for each tissue vs. constitutive exons.

2 Methods

To create a classifier for the prediction of alternative splicing, the first step was to acquire the dataset, the second step was to analyze the different classes of data and determine separating features, and the third step was running the classifying algorithms.

2.1 Dataset

The genomic data used in this research came from the University of California Santa Cruz (UCSC) genome browser, available online at http://genome.ucsc.edu/ (46). Human reference sequence (RefSeq) exons were obtained and matched to homologous mouse exons. Then human expressed sequence tags (ESTs) and mRNAs (referred to hereafter solely as transcripts) aligned to the human genome by UCSC were matched to each exon. Finally, the tissue and histology of each transcript were identified and each skipped exon was categorized as tissue-specific or not at four different probability values using Fisher's exact test.

2.1.1 Exons

The hg18 (human) and mm8 (mouse) RefSeq exons were downloaded from the UCSC Genome Browser Database in mid-September 2006 (47-49). All transcripts that contained one or two exons were discarded since only internal exons were needed to study exon skipping. Then all internal coding exons were identified and separated from first, last, and UTR exons in the remaining transcripts.

2.1.2 Homology

The internal coding exons were filtered based on homology to mouse exons to ensure only high quality exons were studied because homologous exons are more likely to be functional. Homology was determined using the multiz 17-way alignment from UCSC. A mouse exon was specified as homologous to a human exon if their RefSeq transcript coordinates overlapped at least 90% with each other.

2.1.3 Classification of alternative splicing

Transcripts aligned to the genome by UCSC were obtained and matched to each exon as evidence of skipping or inclusion events. The transcripts that were aligned to more than one place in the genome were discarded. To classify a transcript as evidence of a skipped exon (referred to hereafter as 'skipped transcript'), the transcript needed to include the last 50 nucleotides of any upstream exon and the last 50 nucleotides of any downstream exon, spliced together, and with no other nucleotides in between. To classify a transcript as evidence of an included exon (referred to hereafter as 'included transcript'), the transcript needed to include the middle 100 nucleotides of the exon (see Figure 3). For quality, the exon-junctions needed to be the canonical and non-canonical exon-junctions of 'GT-AG' or 'GC-AG' as they make up 99.27% of splice junctions (50). Then all exons with at least one skipped transcript were classified as skipped exons while all others were classified as constitutive exons.

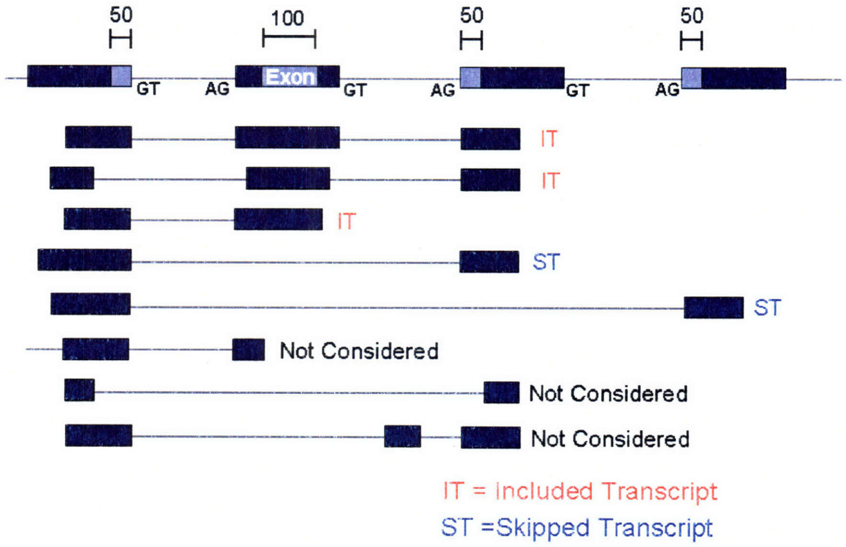


Figure 3: Classifying transcripts as skipped or included. Transcripts that contained the middle 100 nucleotides of an exon were considered an 'included transcripts.' Transcripts that contained the last 50 nucleotides of an upstream exon and the first 50 nucleotides of a downstream exon, with nothing in between, were considered 'skipped transcripts.' Examples are shown in the figure.

2.1.4 Library information

The UCSC genome browser provided the name of the library from which each transcript came. Using this information, the histology and tissue type of each transcript was identified using library data obtained August 1, 2006 from the Cancer Genome Anatomy Project (CGAP). CGAP provided a file that contained the histology, preparation, tissue type, etc. for 8,858 libraries. The library name from UCSC was matched to the library name from CGAP to obtain each transcript's tissue type and histology. Some tissue types were combined into more general categories: brain included cerebrum and cerebellum; eye included retina; lymph included lymph node, lymphoreticular, spleen, thymus, and bone marrow; and endocrine included pineal gland, pituitary gland, thyroid, parathyroid, adrenal cortex, adrenal medulla, and pancreatic islet. The tissue type and histology of some transcripts could not be identified because some transcript library names were not contained in UCSC's data or the library was not contained in CGAP's data.

2.1.5 Classification of tissue specificity

Fisher's exact test, a statistical test to determine if there are non-random associations between two variables, was used to identify whether a skipped exon was tissue-specifically skipped. It utilizes a matrix representing the observations of each of two states and involves calculating the conditional probability of seeing that matrix or a matrix with more biased distribution of entries given its column and row sums compared to all other possible matrices that would give the same column and row sums (without using negative numbers). The cutoff probability value for each matrix is a multivariate generalization of the hypergeometric probability function (51). Fisher's exact test was used instead of other methods such as the chi-squared test because there are often very few transcript counts per tissue per exon.

In this research, 2 x 2 matrices were created for each exon for each tissue: the first row contained the number of skipped transcripts; the second row contained the number of included transcripts; the first column contained the number of transcripts from one tissue; and the second column contained the sum of the number of transcripts from all other tissues. Before running Fisher's exact test, all transcripts classified as coming from cancerous tissues were eliminated because only normal alternative splicing was studied and not splicing caused by mutation or disease. All transcripts classified with the broad tissue types of 'head and neck,' 'pooled tissue,' and 'whole body' were also discarded. The data was then imported into MATLAB (The Mathworks, Natick, MA) and four different probability value cutoffs (0.05, 0.10, 0.20, and 0.25) were used to test for tissue specific skipping using a right-sided Fisher's exact test (52). An exon was identified as tissue-specific skipped in a given tissue if that tissue and only that tissue was found significant at the given probability value or the exon was classified as tissue-specific skipped in the given tissue and only that tissue at a more stringent probability value.

2.1.6 Classification of Alternatively Conserved Exons (ACEs)

To ensure the quality of the tissue-specific and non-tissue-specific skipped exons and their features, an alternatively conserved exon (ACE) subset of the tissue-specific skipped exons was created. ACEs are skipped exons whose skipping is conserved in more than one species. Each skipped exon whose homologous mouse exon was also a skipped exon based on available mouse transcript data was identified as an ACE (R. Sandberg, personal communications).

2.2 Features

The classifiers required features from each exon to separate the different classes. Analysis of each exon included five regions: the first 200 nucleotides or first ⅓ of the upstream intron, the last 400 nucleotides or last ⅔ of the upstream intron, the exon, the first 400 nucleotides or first ⅔ of the downstream intron, and the last 200 nucleotides or last ⅓ of the downstream intron. The last twenty-five nucleotides at the 3′ end of the introns were excluded because it includes the 3′ splice site and polypyrimidine tract. The first ten nucleotides at the 5′ end of the introns were excluded because it includes the 5′ splice site. The fractions of an intron were used when the intron length was less than 636 nucleotides (see Figure 4).

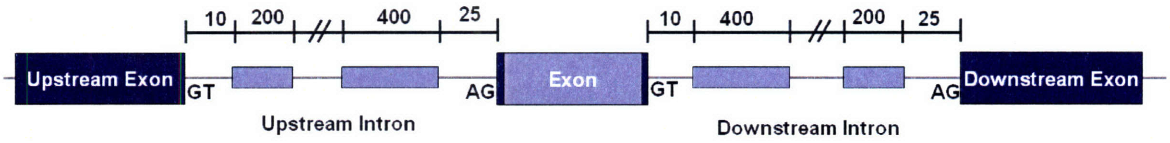


Figure 4: Exon and Intron Regions for Analysis. Light blue regions represent the exon regions for analysis. Lengths of the regions are displayed above the image (note: picture not drawn to scale). The dark blue regions represent exon regions not included in analysis and the grey bar represents intron regions not included in analysis. For introns less than 635 nucleotides, the 200 nucleotide region was 1/3 the nucleotides of the intron and the 400 nucleotide region was 2/3 the nucleotides of the intron. Splice site junctions were required to be either the canonical GT-AG or GC-AG.

All exons with a length less than 13 nucleotides and all exons with an adjacent intron with a length less than 53 nucleotides were discarded as they did not provide enough sequence to analyze once the discarded regions were removed. In addition, all exons with a length greater than 1000 nucleotides were discarded because they are likely to be regulated differently. Each RefSeq transcript from which an exon came was tested to ensure that its annotation was correct—only one stop codon occurs in the mature mRNA at the end of the last coding exon—otherwise the exon was discarded. Both general (length, reading frame conservation, splice site score, etc.) and motif sequence features were obtained from the five regions.

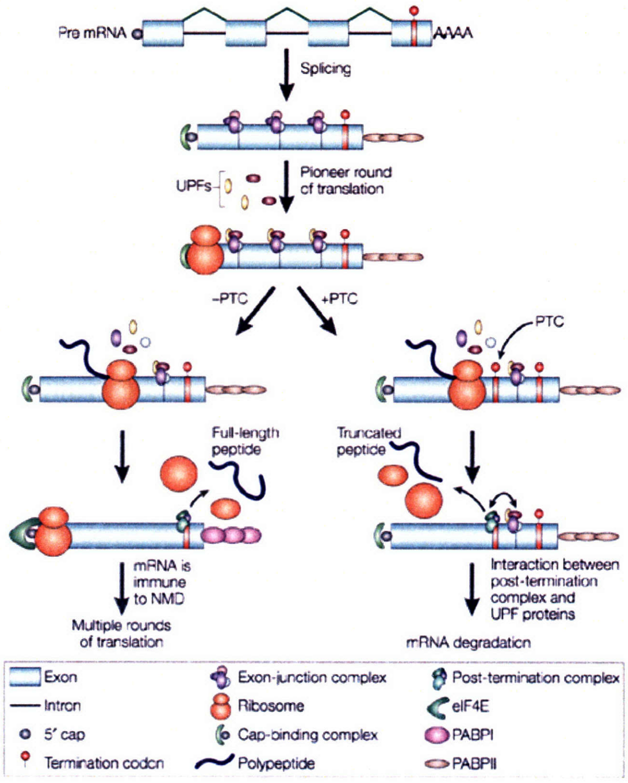
2.2.1 General Sequence Features

Several general sequence features were identified from the tissue-specific skipped exons, non-tissue-specific skipped exons, and constitutive exons to look for differences between the exon sets. The sequence features identified and used in classification were: exon length, upstream intron length, downstream intron length, 5′ splice site score, 3′ splice site score, exon length divisibility by three, and the percent G+C of each region described above. Other features identified but not used in classification were: exon modularity, percent generating a substrate for NMD if the exon is skipped, percent extension of open reading frame (ORF) if the exon is skipped, percent conservation of the exon, and percent conservation of the 150 nucleotides of each intron most adjacent to the exon.

The scores of 5′ and 3′ splice sites were calculated using a maximum entropy model (53). The 5′ splice site score was calculated from the 3 nucleotides before and 4 nucleotides after the ‘GT’ or ‘GC’ of the 5′ splice site. The 3′ splice site score was calculated from the 18 nucleotides before and 3 nucleotides after the canonical ‘AG’ of the 3′ splice site.

The G+C percent of each region was calculated by counting the number of G, C, A, and T nucleotides and then dividing the number of G and C nucleotides by the total number of nucleotides. The divisibility of each exon by three was calculated to determine whether or not removing the exon preserves the reading frame of downstream exons. An exon was considered modular if its length was divisible by three or if all the RefSeq annotations that skipped the exon preserved the stop codon in the last exon (other exons are skipped when this exon is skipped).

The human exon and intron sequence conservation with homologous mouse exon and intron sequence was calculated using CLUSTALW (54). Susceptibility to NMD was determined by skipping the exon in the mature mRNA or looking at RefSeq transcripts that skipped the exon, and assessing whether the first stop codon reached in frame did not occur in the last exon and occurred at least 55 nucleotides before the end of the penultimate exon (see Figure 5) (55, 56). Open reading frame (ORF) extension was calculated by skipping the exon in the mature mRNA or looking at RefSeq transcripts that skipped the exon and finding whether any stop codons occurred in frame.



Nature Reviews | Genetics

Figure 5: Nonsense-mediated mRNA decay (NMD). NMD is an mRNA surveillance mechanism that ensures mRNA quality by selectively targeting mRNAs that harbor premature termination codons (PTCs) for rapid degradation. PTCs caused by exon skipping can lead to non-functional or deleterious proteins. PTCs in higher eukaryotes are only recognized as such when they occur upstream of a 'boundary' on the spliced mRNA that is situated ~55 nucleotides before the last exon–exon junction. As summarized in the accompanying figure, the prevalent view of the NMD mechanism is that the splicing process leaves a ‘mark’ ~20 nucleotides upstream of each exon–exon boundary, in the form of an exon-junction complex (EJC), which in turn provides an anchor for up-frameshift suppressor proteins (UPFs). During the first round of translation of a normal mRNA, the stop codon is located downstream of the last mark, and all EJCs are displaced by elongating ribosomes. During subsequent rounds of translation, the cap-binding complex is replaced by eIF4E (eukaryotic initiation factor 4E) and PABPII (poly(A)-binding protein II) is replaced by PABPI, new ribosomes no longer encounter EJCs, and the mRNA is immune to NMD. However, when a PTC is present, ribosomes stop and fail to displace the downstream EJCs from the transcript. Interactions between the marking factors and components of the post-termination complex trigger mRNA decay. Adapted from (55).

For each tissue, each general sequence feature was analyzed to determine if there was a statistical difference between the non-ACE skipped exons, the ACE skipped exons, and constitutive skipped exons. The binomial test was used to test the significance of the following features: exon length divisibility by three, exon modularity, percent generating a substrate for NMD if the exon is skipped, and percent extension of ORF if the exon is skipped. A two-tailed Kolmogorov-Smirnov (KS) test was used to test the significance of the following features: 5′

splice site score; 3′ splice site score; exon length; upstream intron length; downstream intron length; percent G+C of each region; percent conservation of the exon; and percent conservation of the 150 nucleotides of each intron most adjacent to the exon. Only those comparisons found to be significant with a p-value < 0.01 were considered. To test the significant difference of a feature between one tissue's tissue-specific skipped exons and all other tissue-specific and non-tissue-specific skipped exons, the feature value in the tissue under study was compared to the average of the feature values in the other tissues.

2.2.2 Motifs

A major feature for classification of ACEs is the increased conservation of the exon and upstream and downstream intron sequences over constitutive exons (34, 37). However, the purpose of this research was to classify exon skipping using a single genome; therefore, conservation could not be used. However, since ACEs have greater sequence conservation in the exon and flanking introns, it is believed that the conservation is due evolutionary constraints on certain sequence elements or motifs that regulate splicing. These sequence elements are likely bound by factors that either enhance or silence splicing. Because of the size of the dataset, and the sizes of motifs typically bound by splicing factors, five-nucleotide motifs, pentamers, were used to identify regulatory sequences that control tissue-specific exon skipping; thus, 1024 motifs were analyzed in each region described above. Two methods were employed to identify candidate functional motifs: overrepresentation and conservation. Motifs that are conserved and/or overrepresented in the five regions described above may have regulatory influences on splicing.

The first method to find important motifs was the overrepresentation of certain sequences of DNA over the background frequency or expectation in an exon and the 200 nucleotides of the intron regions adjacent to the exon. For each region, the number of occurrences of each pentamer was counted. The expected frequencies were estimated using a 1st order Markov model for which the single nucleotides and dimers are counted in each sequence and those counts are stratified according to the GC content of the whole sequence into four equally-spaced bins (0-25, 25-50,50-75, and 75-100 % GC). The expected counts were then summed over each GC-bin into the total number of expected counts. The log odds ratios, $\log(observed/expected)$, were then calculated for each pentamer.

The other method to find important motifs was conservation. The conserved occurrence rate (COR) involves testing the conservation of certain sequence motifs between homologous segments of human and mouse. The 150 nucleotides of the intron regions most adjacent to an exon were used in the algorithm described by Wang and co-workers (15). The COR calculation for pentamers began by obtaining counts of each pentamer in each of the human and mouse orthologous regions, in this example, exons. Denote this count as $pentamer_j^{Hexon_i}$ and $pentamer_j^{Mexon_i}$ for the j^{th} pentamer in the i^{th} exon. Next, the difference of pentamer counts in human and mouse orthologs were calculated along with the ratios COR_H and COR_M. COR_H is defined as follows, with the sums taken over all ($pentamer_j^{Hexon_i} - pentamer_j^{Mexon_i} > 0$):

$$COR_H = 1 - \frac{\sum_{i,j}\left(pentamer_j^{Hexon_i} - pentamer_j^{Mexon_i}\right)}{\sum_{i,j} pentamer_j^{Hexon_i}}$$

COR_M is defined analogously, with the sums taken over all exons j such that ($pentamer_j^{Hexon_i} - pentamer_j^{Mexon_i} < 0$).

$$COR_M = 1 - \frac{\sum_{i,j}\left(pentamer_j^{Mexon_i} - pentamer_j^{Hexon_i}\right)}{\sum_{i,j} pentamer_j^{Mexon_i}}$$

Then, the COR of the i^{th} pentamer is: $COR = \left(COR_H + COR_M\right)/2$

The next step was calculating the p-value for each pentamer. To do this, other pentamers that had the same (or similar) total counts as pentamer *i* were found in the human sequence (e.g. exons). Suppose *N* such pentamers were found. Now *N* vectors of counts each with length *L* (*L* is the total number of exons we are analyzing) were constructed. For pentamer *i* in exon *j*, one element from all elements of the *N* vectors was randomly picked so that it had the same count in human as pentamer *i* in exon *j* (if the same count could not be found, then the closest count was chosen). This was performed for all exons. This resulted in a control vector *V* that was the same (or similar) as the count vector of pentamer *i*. For this control vector, the COR value was calculated as described previously. This random process was repeated 2000 times, giving 2000 control COR values. A p-value for the observed COR value was then calculated by fitting a normal distribution to the distribution of control COR values.

The *cis*-element motifs were then analyzed to determine if they matched known ISEs, ISSs, ESEs, or ESSs. A list of these elements was obtained from RESCUE-ESE, FAS-ESS-cut2 and unpublished data on RESCUE-based ISEs, RESCUE-based ISSs, and FAS-ISS-cut3 (unpublished data from X. Xiao and N. Shomron) (11, 14). Since the *cis*-elements were hexamers, six nucleotides, and the motifs analyzed in this dataset were pentamers, five nucleotides, a pentamer set was created from the hexamer set by taking those pentamers that existed as both the first five nucleotides in one hexamer and the last five nucleotides in another hexamer. This resulted in a set of *cis*-element pentamers that contained 50-57% the number of *cis*-element hexamers. A list of known tissue-specific splicing factors, muscleblind-like (MBNL) [9], CELF [17], PTB [5], FOX [2], hnRNP A1 [2], hnRNP H,F [1], Nova [8], and SF1 [5], and their *cis*-elements (the number of *cis*-elements that match each factor is in brackets) was identified from the literature (1, 57-65). The *cis*-elements that the factors bound to ranged from four to six nucleotides in length. Therefore, if the pentamer was contained in the factor's *cis*-element or the factor's *cis*-element was contained in the pentamer, then the pentamer was considered as being bound by the factor.

2.3 Classification

A set of general features and counts of significant pentamers from each region were assembled for classification. Three two-way separations were attempted: tissue-specific skipped exons from one tissue vs. all other tissue-specific skipped exons, tissue-specific skipped exons from one tissue vs. constitutive exons, and all tissue-specific skipped exons vs. constitutive exons. The significant pentamers for the feature set were chosen by combining the top two most significant overrepresented and the top two most significant conserved pentamers in each region for each tissue and taking the symmetric difference for those pentamers found significant for the constitutive exons. The binary classifiers utilized a training and test set to identify the weights on each feature needed for separating the two classes. The dataset was randomly separated into training and test sets and classified using two types of classifiers: RF and SVM.

2.3.1 Random Forests

Version 5.1 of the Random Forests™ software is available from Adele Cutler's site at http:/www.math.usu.edu/~adele/forests/cc_software.htm or Leo Breiman's site at http://www.stat.berkeley.edu/~breiman/RandomForests/cc_software.htm (20). The training set was composed of 75% of the tissue-specific skipped exons and 2.5% of the constitutive exons (because so many are available). An mtry0 value of twenty, the number of random features selected at each node of a tree, was used for the data containing all general features and the counts of the most significant pentamers for each of the five regions. For each binary classifier, 2500 trees were generated, except for the all tissue-specific skipped vs. constitutive classifier, for which only 500 trees were generated as they took much longer to generate. The program was compiled using the GNU g77 fortran compiler using gcc version 3.3.3 (http://gcc.gnu.org/) and run under Fedora Core Linux release 2.

2.3.2 Support Vector Machine

Version 2.83 of the support vector machine LibSVM is available online at http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (66, 67). It was chosen because it supports weights for unbalanced data and has a tool to make choosing parameters for an SVM with a radial kernel easy, which was important because SVM parameters are difficult to pick and the parameters control the quality of the classification. The training set was composed of 75% of the tissue-specific skipped exons and the number of constitutive exons equaled 1.25x the number of the tissue-specific skipped exons. The python tool easy.py, included with LibSVM, was used to scale the features and tune the parameters of the radial kernel for the many binary classifiers (68). The program was run under Fedora Core Linux release 2.

3 Results and Discussion

The UCSC genome data were filtered and classified to create a dataset that could be used for training and testing a tissue-specific exon skipping classifier. The data were analyzed for both general features and sequence motifs and then the RF and SVM classifiers were used to classify the datasets. For RF, the classification was skewed to a 5% error for the constitutive or other tissue-specific skipped exons by using weights, because a low false positive rate is useful for making predictions about novel tissue-specific spliced exons. This resulted in various error percentages for the tissue-specific skipped exons. For SVM, the classification was skewed to a low error for the constitutive exons by including more constitutive exons than tissue-specific skipped exons in the training set. The classification was run using the easy.py tool, resulting in several good classifications and several very poor classifications.

3.1 Dataset

The UCSC hg18 revision (human) included 25,165 RefSeq transcripts and the mm8 revision (mouse) included 19,877 RefSeq transcripts. After one- or two-exon RefSeq transcripts were removed, 22,333 hg18 RefSeq transcripts and 16,441 mm8 RefSeq transcripts remained. After the exons from each of these transcripts were filtered for only internal coding exons, 146,826 unique hg18 exons and 126,221 unique mm8 exons remained. Then the hg18 exons were matched to homologous mm8 exons, creating 115,739 hg18-mm8 homologous exon pairs.

In mid-September 2006, the hg18 genome had 7,737,713 ESTs and 214,057 mRNAs aligned to it. After the ESTs and mRNAs were filtered for those that aligned to only one place in the genome, 6,816,156 ESTs and 192,677 mRNAs remained. The ESTs and mRNAs allowed the identification of 18,824 (16.26%) skipped exons and 96,915 (83.74%) non-skipped exons out of the 115,739 hg18-mm8 homologous exon pairs.

There were 11,046 'exon-skipping' transcripts and 111,693 'exon-including' transcripts that were not characterized because their library names were not contained in UCSC's data. There were 3,753 'exon-skipping' transcripts and 37,371 'exon-including' transcripts that were not characterized because 1,191 library names contained in UCSC's data were not contained in CGAP's data. The tissue type and histology of some transcripts also could not be identified because the histology and/or tissue type of the library were not identified in CGAP's data. There were 45,673 'exon-skipping' transcripts and 730,382 'exon-including' transcripts that were not characterized in CGAP's data. Since only normal alternative splicing was studied, and not splicing caused by mutation or disease, all transcripts classified as coming from cancerous tissues were eliminated. This discarded 56,520 'exon-skipping' transcripts and 1,255,102 'exon-including' transcripts. Thus, 219,721 'exon-skipping' transcripts and 3,766,497 'exon-including' transcripts remained that came from normal tissue and whose tissue type was identified.

About 10-30% of the skipped exons were identified as tissue-specific skipped after Fisher's exact test was run using all the normal (non-disease associated) transcripts for skipped exons. Those tissues with the most tissue-specific skipping are listed in Table 1 at the four different p-value cutoffs used with Fisher's exact test.

Table 1: Tissue-specific Homologous Skipped Exons using Normal Tissue ESTs at four p-values

Tissue	$p < 0.05$		$p < 0.10$		$p < 0.20$		$P < 0.25$	
	Non-ACE/ACE	ACE	non-ACE/ACE	ACE	non-ACE/ACE	ACE	non-ACE/ACE	ACE
non-specific	15788	1831	14461	1626	12459	1349	11839	1255
Brain	181	38	328	56	696	106	865	135
Nervous	161	35	313	58	558	97	645	109
Placenta	132	13	249	27	455	56	523	71
Testis	127	19	217	34	362	53	400	58
Eye	119	23	208	37	331	53	359	58
Endocrine	105	14	153	24	211	30	227	33
Liver	98	24	169	31	214	38	226	38
Kidney	93	11	150	19	207	24	222	25
Lung	87	12	180	25	286	42	302	43
Lymph	80	13	149	26	254	39	286	41
Embryonic	76	19	165	37	429	73	506	84
Skin	58	6	86	11	124	14	130	14
Prostate	58	11	97	21	135	26	140	27
Heart	51	11	71	15	84	17	89	18
Muscle	41	8	80	14	114	18	124	20
Vascular	30	5	48	9	76	15	80	18

The number of tissue-specifically skipped exons for each tissue correlates well with the tissues that have the highest proportion of genes with skipped exons (see Table 1 and Figure 6). Yeo and co-workers found that the top tissues whose genes have the greatest proportion of skipped exons were brain, testis, lung, eye-retina, and placenta. Consistently, I found that the most tissue-specific exons were found in brain, nervous, placenta, testis, and eye.

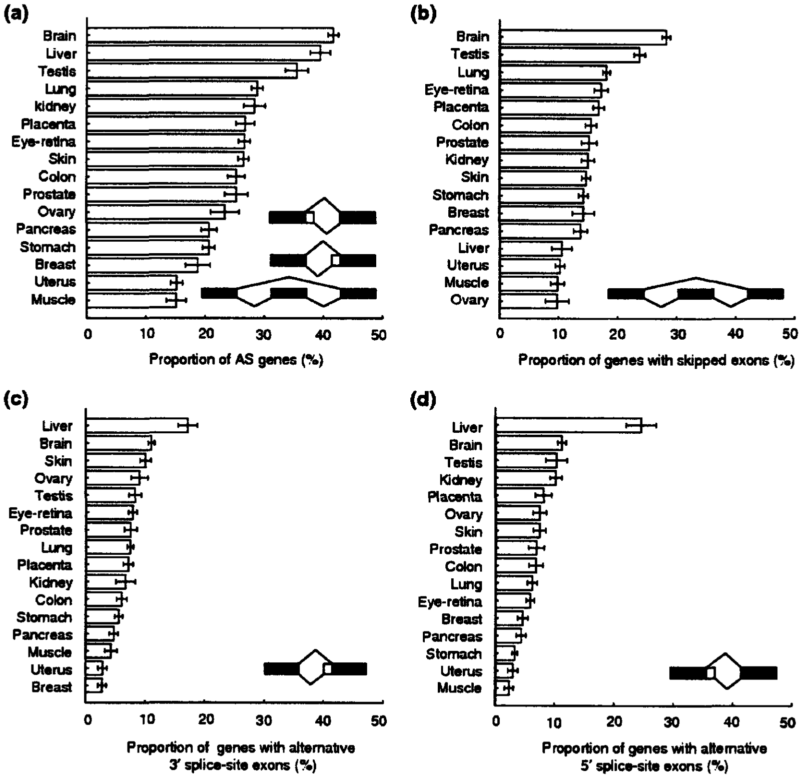


Figure 6: Levels of alternative splicing in 16 human tissues with moderate or high EST sequence coverage. Horizontal bars show the average fraction of alternatively spliced (AS) genes of each splicing type (and estimated standard deviation) for random samplings of 20 ESTs per gene from each gene with ≥ 20 aligned EST sequences derived from a given human tissue. The different splicing types are schematically illustrated in each subplot. (a) Fraction of AS genes containing skipped exons, alternative 3′ splice site exons (A3Es) or 5′ splice site exons (A5Es), (b) fraction of AS genes containing skipped exons, (c) fraction of AS genes containing A3Es, (d) fraction of AS genes containing A5Es. Adapted from (2).

3.2 Features

Both general and motif sequence features were analyzed to find those features that may help a classifier separate skipped exons from constitutive exons, tissue-specific skipped exons from non-tissue-specific skipped exons, and ACE skipped exons from non-ACE skipped exons.

3.2.1 General Sequence Features

For each tissue, each general sequence feature was analyzed to determine if there was a statistical difference between a feature for a specific tissue within all non-ACE skipped exons and for a specific tissue within all ACE skipped exons. Only those features with a p-value < 0.01 by the KS or binomial test were considered (see Figure 7).

Fewer ACE features were found to be significant from the other ACEs than for non-ACEs, suggesting that either most tissue-specific ACEs have the same features, or there just were not enough ACEs to sustain significance amongst the different tissue-specific skipped sets. A few generalizations can be made from these results: non-ACE kidney and non-ACE liver skipped exons had significantly more or significantly less G+C nucleotides, respectively, in all their regions. Non-ACE prostate exons also had significantly more G+C nucleotides in three out of the five regions. The significant high G+C content of the non-ACE liver skipped exons and their flanking introns probably results in the significant greater amount of percent extension of ORF if the exon is skipped. This probably results from stop codons being composed mostly of A+T nucleotides, making a stop codon less likely to be encountered in high G+C regions, and thus less likely for a frameshift caused by exon skipping to introduce an early stop codon (44).

The high percent of substrates susceptible to NMD when an exon is skipped witnessed in muscle exons, and non-ACE skipped exons in general, may exist to ensure only the right isoform of a protein is present in a cell. Pulak and Anderson showed that NMD is used to degrade protein fragment alleles of myosin globular head domain in *C. elegans* (69). The same may be the case for human muscle genes. This may also explain why more gene substrates susceptible to NMD are seen when exons are skipped in non-ACE and constitutive exons (see Figure 8). Most isoforms generated when a non-ACE exon is skipped may be aberrant and not functional, and thus, the cell would want to target them for degradation by NMD so they do not interfere with the good isoform (70, 71). It seems likely, as shown in Figure 8, that a lesser percentage of ACE skipped exons are substrates for NMD because the conservation of skipping between two species over 90 million years of evolution is likely due to the cell keeping a protein isoform that has some biological function.

It is also worth noting that non-ACE liver tissue-specific-skipped exons, non-ACE prostate tissue-specific skipped exons, non-ACE embryonic tissue-specific skipped exons and non-ACE kidney tissue-specific skipped exons had many features that were significant, 5/17, 5/17, 5/17, and 7/17, respectively, compared to non-ACE tissue-specific skipped exons in other tissues. The tissue-specific differences in general sequence features may represent novel mechanisms of splicing and/or regulation in the tissues with significantly different features.

In addition to the tissue-specific analysis, all tissues were combined into three distinct sets and then each general sequence feature was analyzed to determine if there was a statistical difference between the non-ACE skipped exons, the ACE skipped exons, and constitutive skipped exons. Only those features with a p-value < 0.01 by the KS or binomial test were considered (see Figure 8).

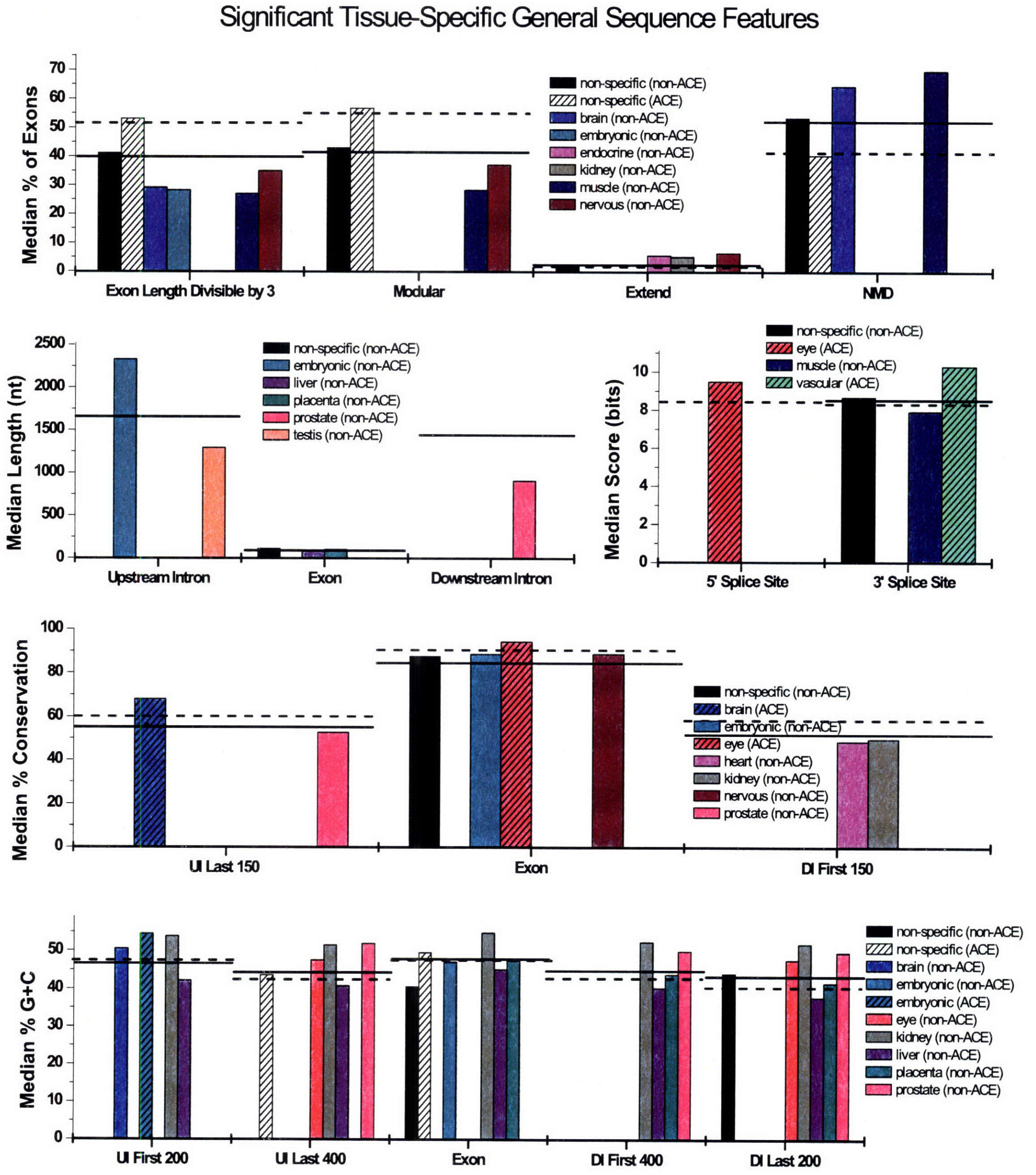


Figure 7: Significant General Sequence Features within Tissue-Specific ACE Exons and within Tissue-Specific non-ACE exons. The average values of each feature in the upper leftmost graph were compared using the binomial test and significant difference was established for p < 0.01. The distribution of each feature in all other graphs were compared using the Kolmogorov-Smirnov (KS) test and significant difference was established for p < 0.01. The tissues in the graph for each feature are the tissues that are significantly different from all other tissue-specific skipped exon tissues of the same type (ACE or non-ACE). The non-ACE features are solid and the ACE features are solid with slashes through the bar. Solid lines (——) represent median non-ACE values and dotted lines (- - - -) represent median ACE values. UI = Upstream Intron, DI = Downstream Intron.

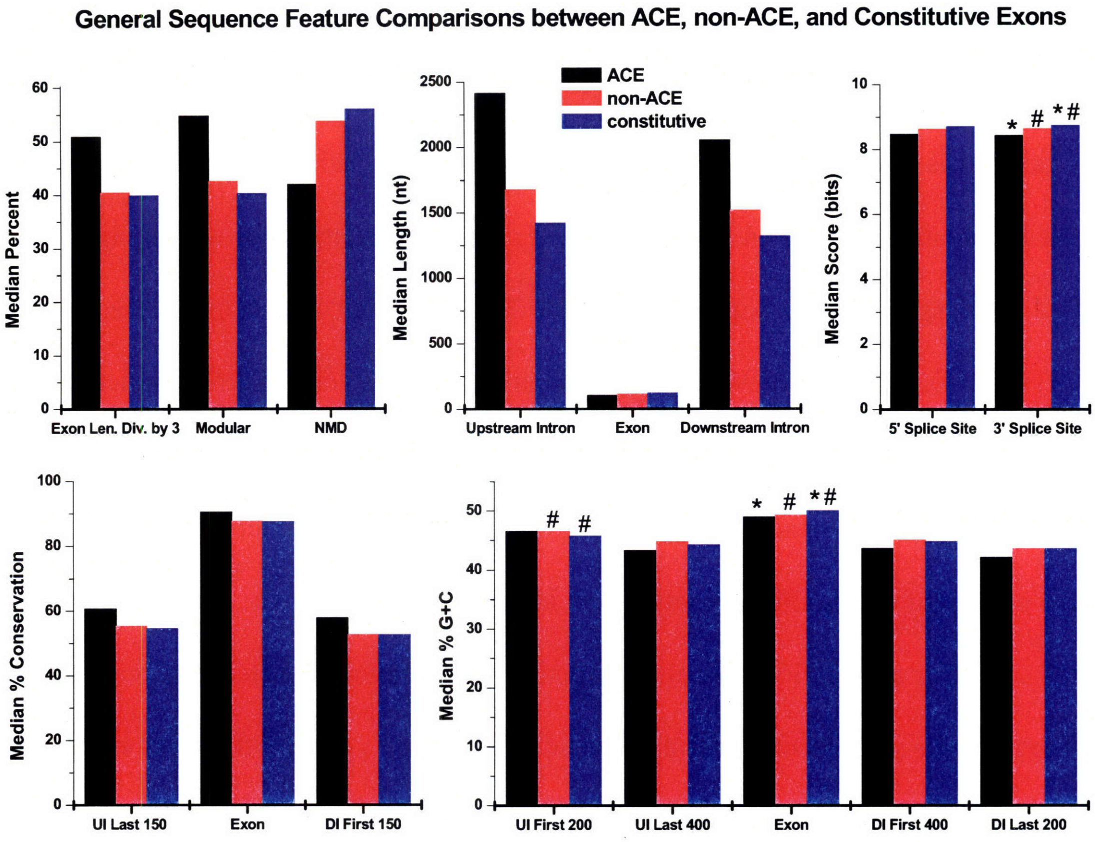


Figure 8: General Sequence Feature Comparisons between ACE Skipped Exons, non-ACE Skipped Exons, and Constitutive Exons. The average values of each feature in the upper leftmost graph were compared using the binomial test and significant difference was established for $p < 0.01$. The distribution of each feature in all other graphs were compared using the Kolmogorov-Smirnov (KS) test and significant difference was established for $p < 0.01$. Within each group, all three comparisons (non-ace/constitutive, ace/constitutive, and ace/non-ace) are significant unless symbols (*,#) appear above any of the bars in a group, in which case, only those bars within the same group that have the same symbol are significantly different. UI = Upstream Intron, DI = Downstream Intron.

For every feature but one (median % G+C for upstream intron first 200 nucleotides), the ACE skipped exons have greater separation from constitutive exons than the non-ACE skipped exons have from constitutive exons. In general, the ACE skipped exons and non-ACE skipped exons are more divisible by 3, more modular, have less NMD, have longer introns, have shorter exons, have weaker splice site scores, have greater sequence conservation between human and mouse, and have less G+C nucleotides than the constitutive exons. These results are consistent with the results by Yeo and co-workers and Clark and Thanaraj (34, 44).

3.2.2 Motifs

The motifs found by overrepresentation and COR were combined for both the tissue-specific skipped vs. constitutive classifications and tissue specific-skipped vs. other tissue-specific skipped as described in the Methods. The motifs from each set were then analyzed to determine if they matched known ESEs, ESSs, ISEs, and ISSs elements or were motifs associated with tissue-specific skipping (see Table 2) (unpublished data from X. Xiao and N. Shomron) (11, 14). There were 94 pentamers identified as ESSs (out of 176 hexamers), 118 pentamers identified as ESEs (out of 237 hexamers), 281 pentamers identified as ISSs (out of 491 hexamers), and 227 pentamers identified as ISEs (out of 459 hexamers). Forty-three to forty-five percent of the significant motifs were identified as known ISSs, ISEs, ESSs, or ESEs.

The number of significant motifs matched to known regulatory elements (see Table 2) was not found to be significantly overrepresented using Fisher's exact test, except known ISEs in the upstream intron first 200, which were significantly underrepresented. However, in most cases, the number of motifs matched to known regulatory elements was greater than the expectation. For example, 38% of the regulatory elements in the upstream intron first 200 were identified as known ISSs while the expectation is for only 27% to be identified as known ISSs.

Table 2: Comparison of Significant Motifs to Known Regulatory Elements
(UI = Upstream Intron, DI = Downstream Intron, parentheses value is # of motifs that are an enhancer & silencer, ISS/ISE for intron regions, ESS/ESE for exon region, * represents significance at p < 0.01)

Region	Comparison Type	Fisher Exact Cutoff (p < 0.10)			Fisher Exact Cutoff (p < 0.20)		
		Total Motifs	Known ISE/ESE	Known ISS/ESS	Total Motifs	Known ISE/ESE	Known ISS/ESS
UI First 200	Skipped vs. constitutive	50	4	19	49	1 *	18
UI Last 400		52	12 (2)	13 (2)	53	10 (1)	19 (1)
Exon		54	8 (1)	3 (1)	49	9	4
DI First 400		53	10 (1)	16 (1)	47	9 (2)	14 (2)
DI Last 200		52	8 (2)	15 (2)	51	9 (1)	14 (1)
Total		261	42 (6)	66 (6)	249	38 (4)	69 (4)
UI First 200	Tissue-specific vs. other skipped	49	4	19	44	1 *	16
UI Last 400		53	11 (2)	11 (2)	50	8	19
Exon		49	8 (1)	3 (1)	49	9	4
DI First 200		47	15	10	45	8 (2)	13 (2)
DI Last 400		51	7 (2)	14 (2)	48	7	14
Total		249	45 (5)	57 (5)	236	33 (2)	66 (2)

Motifs were then analyzed for similarity to binding motifs of known tissue-specific splicing factors. Eight factors were identified from literature but only five appeared in the significant motifs (see Table 3). Those that did not appear only had 1, 2, or 5 motifs that represented the binding factor. To test the significance of the number of binding factors found in the significant motifs, the binding factor motifs were shuffled, preserving dinucleotide frequency where possible. When the analysis was re-run using the shuffled motifs, at most, one motif was identified as PTB or CELF in a region. Therefore, it is likely that the presence of the MBNL, CELF, and hnRNP A1, but not PTB, motifs in the significant set is due to them being tissue-specific splicing factor *cis*-elements.

Several generalizations can be made about which tissues the significant motifs that matched the known splicing factors came from. The MBNL motifs in both upstream and downstream introns appear to be a top significant motif for almost every tissue, including muscle and eye. However, the MBNL motif in the exon came only from muscle tissue. The CELF motifs were highly significant in many tissues, but more so in the exon than in the introns. The hnRNP A1 motif in the upstream intron came from lymph tissue while the hnRNP A1 motif in the exon came from brain and nervous tissues. The PTB motif was significant in only a few tissues and the two Nova motifs encountered came from heart tissue. However, nova is known to control splicing in the brain, but it only came out as a top significant motif in one region of heart tissue (63). MBNL is named for its regulation of muscle and eye splicing, but it also regulatory function in other tissues, at least in introns. Recent research suggests that MBNL and CELF regulate splicing in an antagonistic manner, which likely explains many more tissues having CELF in the exon but MBNL in the introns (57). The hnRNP A1 factor is thought to bind ESS motifs in exons and thus play a role in alternative splicing (72). Yeo and co-workers also found hnRNP A1 to be overrepresented in the brain (2). It has also been proposed that hnRNP A1 plays a role in introns as well and thus significant motifs appearing in lymph tissue may be due to its role there (73).

Table 3: Comparison of Significant Motifs to Known Splicing Factors
(UI = Upstream Intron, DI = Dowstream Intron)

Region	Comparison Type	Fisher Exact Cutoff (p < 0.10)					Fisher Exact Cutoff (p < 0.20)				
		MBNL	CELF	PTB	hnRNP A1	Nova	MBNL	CELF	PTB	hnRNP A1	Nova
UI First 200	Skipped vs. constitutive	2	4	2	0	0	1	3	0	0	0
UI Last 400		2	1	0	0	0	2	1	1	1	0
Exon		1	6	0	1	0	1	3	0	0	0
DI First 400		2	1	2	0	0	1	2	2	0	0
DI Last 200		2	2	0	0	1	2	2	0	0	0
Total		9	14	4	1	1	7	11	3	1	0
UI First 200	Tissue-specific vs. other skipped	2	4	2	0	0	1	4	0	0	0
UI Last 400		2	1	0	0	0	2	1	1	1	0
Exon		1	6	0	1	0	1	3	0	0	0
DI First 400		2	1	2	0	0	1	2	2	0	0
DI Last 200		2	2	0	0	1	2	2	0	0	0
Total		9	14	4	1	1	7	12	3	1	0

3.3 Classification

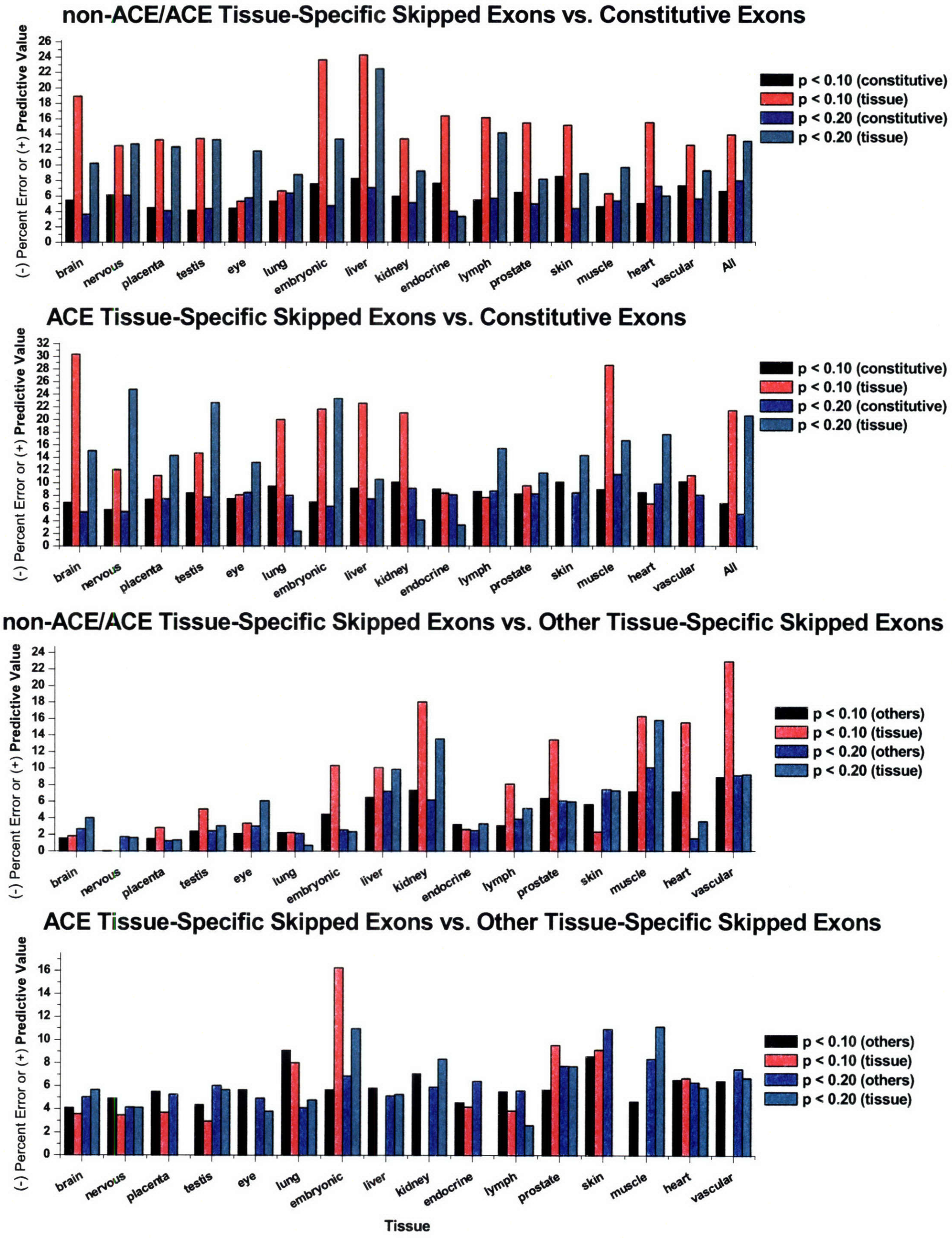
Three two-way separations were attempted: tissue-specific skipped exons from one tissue vs. all other tissue-specific skipped exons, tissue-specific skipped exons from one tissue vs. constitutive exons, and all tissue-specific skipped exons vs. constitutive exons. Most of the RF classifiers had about the same results while the SVM results varied greatly. Only those exons identified as tissue-specific at p-value cutoffs of $p < 0.10$ and $p < 0.20$ were used.

3.3.1 Random Forests

The relative error rates for each class in the RF classifier were easily controlled and they were set to produce around a 5% error rate for the constitutive exon or other tissue-specific skipped sets. This provided a low false positive rate for possible discovery purposes. However, a low false positive rate produced a generally high false negative rate for most tissues. The RF classifier did not overtrain; therefore, the results shown on the next page are the combination of the training and test sets. Generally, the tissue-specific skipped exons were easier to classify against the constitutive exons, and the ACE exons were easier to classify than the non-ACE/ACE exons (see Figure 9). Brain, embryonic, liver, and muscle were the tissues that classified the best. Classification was difficult because many of the features overlap between the two sets and there is probably substantial noise in the datasets, the features, and the tissue-specificity of skipping.

3.3.2 Support Vector Machine

The SVM classifier was much more difficult to control and there were very few support vectors that could be defined for the tissue-specific set as there were few tissue-specific skipped exons compared to the many constitutive exons. This often led to overtraining and all members of both classes being classified as one class or the other. Therefore, the results reflected in Figure 10 are only from the test set in order to display a more real presentation of the accuracy of the classifier (see Figure 10). The SVM performed generally worse than the RF classifier. The tissue-specific skipped exons vs. all other tissue-specific skipped exons performed poorly with the SVM (data not shown). It is likely that finer tuning and the use of different kernels could produce better SVM classifiers.



non-ACE/ACE Tissue-Specific Skipped Exons vs. Constitutive Exons

ACE Tissue-Specific Skipped Exons vs. Constitutive Exons

non-ACE/ACE Tissue-Specific Skipped Exons vs. Other Tissue-Specific Skipped Exons

ACE Tissue-Specific Skipped Exons vs. Other Tissue-Specific Skipped Exons

Figure 9: Random Forest Classification of Tissue-Specific Skipped Exons. The top two graphs display the classification of non-ACE/ACE and ACE tissue-specific skipped exons against constitutive exons. The bottom two graphs display the classification of non-ACE/ACE and ACE tissue-specific skipped exons for each tissue against other tissue-specific skipped exons. Black and blue bars represent the negative percent error = FN/(TN+FN) of constitutive or other tissue-specific skipped exons of $p < 0.10$ and $p < 0.20$ datasets, respectively. Red and dark cyan bars represent the positive predictive value = TP/(TP+FP) of tissue-specific skipped exons of $p < 0.10$ and $p < 0.20$ datasets, respectively. Low negative percent error and high positive predictive value is better.

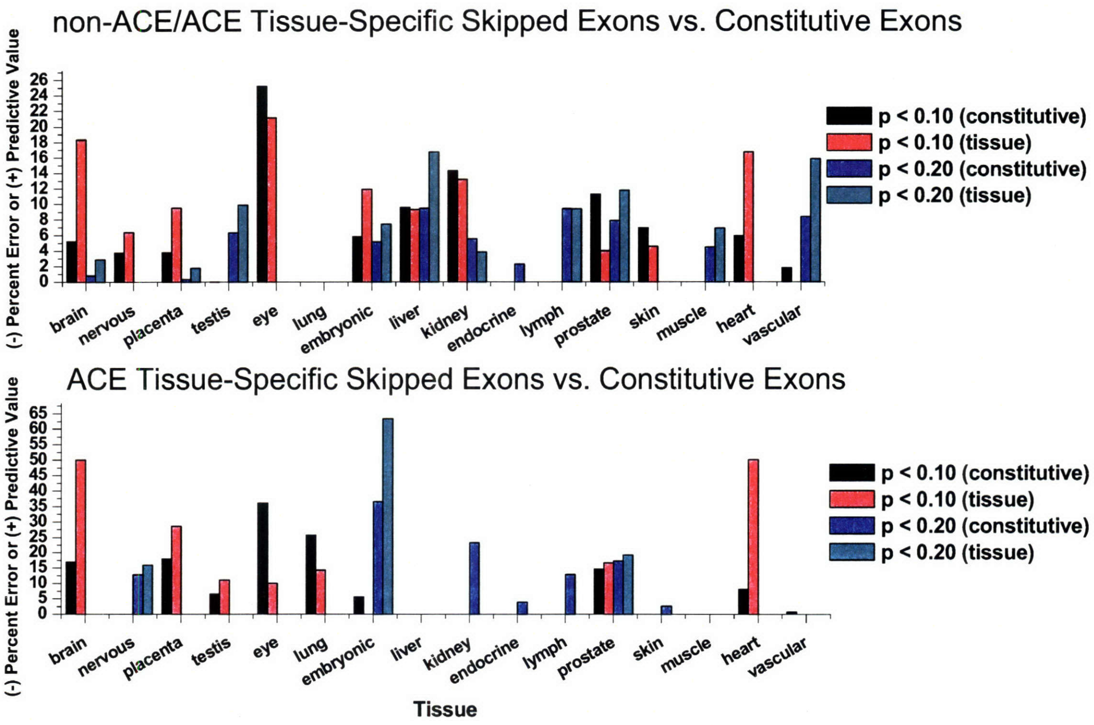


Figure 10: Support Vector Machine Classification of Tissue-Specific Skipped Exons. The graphs display the classification of non-ACE/ACE and ACE tissue-specific skipped exons against constitutive exons. Black and blue bars represent the negative percent error = FN/(TN+FN) of constitutive skipped exons of $p < 0.10$ and $p < 0.20$ datasets, respectively. Red and dark cyan bars represent the positive predictive value = TP/(TP+FP) of the tissue-specific skipped exons of $p < 0.10$ and $p < 0.20$ datasets, respectively.

4 Conclusion

This thesis identifies tissue-specific skipping datasets at different stringencies using Fisher's exact test. The identification was performed using only ESTs that came from normal tissues, but it is possible to perform the same classification using ESTs that solely come from cancerous tissues or a combination of both, thereby allowing analysis of cancer-specific, tissue-specific skipping.

Features that separate tissue-specific skipped exons in sixteen different tissues from each other were identified. In addition, features that separate non-ACE skipped exons, ACE skipped exons, and constitutive exons were identified and confirmed with previous studies. Tissue-specific motifs were found using overrepresentation and conserved occurrence rate (COR). Several known tissue-specific binding elements and factors were found within the set of motifs identified. These methods also identified novel tissue-specific *cis*-elements that could be studied for further analysis.

Classification using both the general sequence features and significant motifs provided some good results and some poor results. Generally, both RF and SVM were not able to separate tissue-specific skipped exons from constitutive exons enough to be useful for biological analysis. Finding more feature differences between the sets of exons and tuning the parameters and kernel of the support vector machine should allow for even better classification. In addition, it would be interesting to do an analysis comparing tissue-specific exon skipping with tissue-specific microarray protein expression data. This analysis could help define certain regulatory factors that are crucial for tissue-specific skipping. This thesis has shown that tissue-specific skipping is an area of research that is both very complex and not heavily studied, but is biologically relevant to cellular function in health and potentially disease as well.

5 References

1. Matlin AJ, Clark F, Smith CW. Understanding alternative splicing: Towards a cellular code. Nat Rev Mol Cell Biol. 2005 May;6(5):386-98.

2. Yeo G, Holste D, Kreiman G, Burge CB. Variation in alternative splicing across human tissues. Genome Biol. 2004;5(10):R74.

3. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997 Apr 25;268(1):78-94.

4. Blencowe BJ. Alternative splicing: New insights from global analyses. Cell. 2006;126:37-47.

5. Ast G. How did alternative splicing evolve? Nat Rev Genet. 2004 Oct;5(10):773-82.

6. Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H. Function of alternative splicing. Gene. 2005 Jan 3;344:1-20.

7. Black DL. Mechanisms of Alternative Pre-Messenger RNA Splicing. Annu Rev Biochem. 2003;72(1):291-336.

8. Hiller M, Huse K, Platzer M, Backofen R. Creation and disruption of protein features by alternative splicing -- a novel mechanism to modulate function. Genome Biol. 2005;6(7):R58.

9. Garcia-Blanco MA, Baraniak AP, Lasda EL. Alternative splicing in disease and therapy. Nat Biotechnol. 2004 May;22(5):535-46.

10. Fairbrother WG, Yeo GW, Yeh R, Goldstein P, Mawson M, Sharp PA, Burge CB. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. Nucleic Acids Res. 2004 2004 Jul 1;32:W187-90.

11. Fairbrother WG, Yeh RF, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. Science. 2002 Aug 9;297(5583):1007-13.

12. Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR. ESEfinder: A web resource to identify exonic splicing enhancers. Nucleic Acids Res. 2003 Jul 1;31(13):3568-71.

13. Grabowski PJ. A molecular code for splicing silencing: Configurations of guanosine-rich motifs. Biochem Soc Trans. 2004 Dec;32(Pt 6):924-7.

14. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. Systematic identification and analysis of exonic splicing silencers. Cell. 2004;119(6):831-45.

15. Wang Z, Xiao X, Van Nostrand E, Burge CB. General and specific functions of exonic splicing silencers in splicing control. Mol Cell. 2006 Jul 7;23(1):61-70.

16. Yeo G, Hoon S, Venkatesh B, Burge CB. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. Proc Natl Acad Sci U S A. 2004;101(44):15700-5.

17. Minovitsky S, Gee SL, Schokrpur S, Dubchak I, Conboy JG. The splicing regulatory element, UGCAUG, is phylogenetically and spatially conserved in introns that flank tissue-specific alternative exons. Nucleic Acids Res. 2005;33(2):714-24.

18. Elliott DJ, Grellscheid SN. Alternative RNA splicing regulation in the testis. Reproduction. 2006;132(6):811-9.

19. Brudno M, Gelfand MS, Spengler S, Zorn M, Dubchak I, Conboy JG. Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. Nucleic Acids Res. 2001;29(11):2338-48.

20. Breiman L. Random forests. Mach Learning. 2001;45(1):5-32.

21. Breiman L., Cutler A. Random Forests - Classification Description [Internet]. ; 2006 [December 2006] . Available from: http://www.math.usu.edu/~adele/forests/cc_home.htm.

22. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Bickel, P.; Diggle, P.; Fienberg, S.; Krickberg, K.; Olkin, I.; Wermuth, N.; Zeger, S., editors. New York: Springer; 2001.

23. Chen, Chao; Liaw, Andy; Breiman, Leo. Using random forest to learn imbalanced data. 2004 July 2004:1-12.

24. Cristianini, Nello; Shawe-Taylor, John. An Introduction to Support Vector Machines and other kernel-based learning methods. 1st ed. Cambridge, UK: Cambridge University Press; 2000.

25. Burges CJC. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery. 1998;2(2):121-67.

26. Gupta S, Zink D, Korn B, Vingron M, Haas SA. Genome wide identification and classification of alternative splicing based on EST data. Bioinformatics. 2004 Nov 1;20(16):2579-85.

27. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science. 2003 Dec 19;302(5653):2141-4.

28. Kim N, Shin S, Lee S. ECgene: Genome-based EST clustering and gene modeling for alternative splicing. Genome Res. 2005 Apr;15(4):566-76.

29. Kim P, Kim N, Lee Y, Kim B, Shin Y, Lee S. ECgene: Genome annotation for alternative splicing. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D75-9.

30. Leipzig J, Pevzner P, Heber S. The alternative splicing gallery (ASG): Bridging the gap between genome and transcriptome. Nucleic Acids Res. 2004 Aug 3;32(13):3977-83.

31. Modrek B, Resch A, Grasso C, Lee C. Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res. 2001 Jul 1;29(13):2850-9.

32. Xu Q, Modrek B, Lee C. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. Nucleic Acids Res. 2002 Sep 1;30(17):3754-66.

33. Dror G, Sorek R, Shamir R. Accurate identification of alternatively spliced exons using support vector machine. Bioinformatics. 2005 Apr 1;21(7):897-901.

34. Yeo GW, Van Nostrand E, Holste D, Poggio T, Burge CB. Identification and analysis of alternative splicing events conserved in human and mouse. Proc Natl Acad Sci U S A. 2005 Feb 22;102(8):2850-5.

35. Sorek R, Shemesh R, Cohen Y, Basechess O, Ast G, Shamir R. A non-EST-based method for exon-skipping prediction. Genome Res. 2004;14(8):1617.

36. Sorek R, Shamir R, Ast G. How prevalent is functional alternative splicing in the human genome? Trends Genet. 2004 Feb;20(2):68-71.

37. Sorek R, Ast G. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. Genome Res. 2003 Jul;13(7):1631-7.

38. Ratsch G, Sonnenburg S, Scholkopf B. RASE: Recognition of alternatively spliced exons in C.elegans. Bioinformatics. 2005 Jun 1;21 Suppl 1:i369-77.

39. Itoh H, Washio T, Tomita M. Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes. RNA. 2004 Jul;10(7):1005-18.

40. Ladd AN, Cooper TA. Finding signals that regulate alternative splicing in the post-genomic era. Genome Biol. 2002 Oct 23;3(11):reviews0008.

41. Pan Q, Shai O, Misquitta C, Zhang W, Saltzman AL, Mohammad N, Babak T, Siu H, Hughes TR, Morris QD. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Mol Cell. 2004;16(6):929–941.

42. Han K, Yeo G, An P, Burge CB, Grabowski PJ. A combinatorial code for splicing silencing: UAGG and GGGG motifs. PLoS Biol. 2005 May;3(5):e158.

43. Iida K, Akashi H. A test of translational selection at 'silent' sites in the human genome: Base composition comparisons in alternatively spliced genes. Gene. 2000 Dec 30;261(1):93-105.

44. Clark F, Thanaraj TA. Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. Hum Mol Genet. 2002 Feb 15;11(4):451-64.

45. Xing Y, Lee CJ. Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. PLoS Genetics. 2005;1(3):0323-8.

46. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002;12(6):996-1006.

47. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ. The UCSC genome browser database. Nucleic Acids Res. 2003;31(1):51-4.

48. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, Hillman-Jackson J, Kuhn RM, Pedersen JS, Pohl A, Raney BJ, Rosenbloom KR, Siepel A, Smith KE, Sugnet CW, Sultan-Qurraie A, Thomas DJ, Trumbower H, Weber RJ, Weirauch M, Zweig AS, Haussler D, Kent WJ. The UCSC genome browser database: Update 2006. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D590-8.

49. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.

50. Burset M, Seledtsov IA, Solovyev VV. SpliceDB: Database of canonical and non-canonical mammalian splice sites. Nucleic Acids Res. 2001 Jan 1;29(1):255-9.

51. Weisstein Eric W. Fisher's Exact Test [Internet]. MathWorld–A Wolfram Web Resource; 1999. Available from: http://mathworld.wolfram.com/FishersExactTest.html.

52. Duembgen, Lutz. Fisher's exact test. 2002 August 6, 2002.

53. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol. 2004;11(2-3):377-94.

54. Higgins DG, Thompson JD, Gibson TJ. CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673-80.

55. Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: Exonic mutations that affect splicing. Nat Rev Genet. 2002 Apr;3(4):285-98.

56. Lewis BP, Green RE, Brenner SE. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. PNAS. 2003;100(1):189-92.

57. Ho TH, Charlet-B N, Poulos MG, Singh G, Swanson MS, Cooper TA. Muscleblind proteins regulate alternative splicing. EMBO J. 2004 Aug 4;23(15):3103-12.

58. Adereth Y, Dammai V, Kose N, Li R, Hsu T. RNA-dependent integrin alpha3 protein localization regulated by the muscleblind-like protein MLP1. Nat Cell Biol. 2005 Dec;7(12):1240-7.

59. AMIR-AHMADY B, BOUTZ PL, MARKOVTSOV V, PHILLIPS ML, BLACK DL. Exon repression by polypyrimidine tract binding protein. RNA. 2005;11(5):699-716.

60. Underwood JG, Boutz PL, Dougherty JD, Stoilov P, Black DL. Homologues of the caenorhabditis elegans fox-1 protein are neuronal splicing regulators in mammals. Mol Cell Biol. 2005 Nov;25(22):10005-16.

61. Blanchette M, Chabot B. Modulation of exon skipping by high-affinity hnRNP A1-binding sites and by intron elements that repress splice site utilization. EMBO J. 1999;18:1939-52.

62. Sugnet CW, Srinivasan K, Clark TA, O'Brien G, Cline MS, Wang H, Williams A, Kulp D, Blume JE, Haussler D, Ares M,Jr. Unusual intron conservation near tissue-regulated exons found by splicing microarrays. PLoS Comput Biol. 2006 Jan;2(1):e4.

63. Ule J, Stefani G, Mele A, Ruggiu M, Wang X, Taneri B, Gaasterland T, Blencowe BJ, Darnell RB. An RNA map predicting nova-dependent splicing regulation. Nature. 2006 Nov 30;444(7119):580-6.

64. Faustino NA, Cooper TA. Identification of putative new splicing targets for ETR-3 using sequences identified by systematic evolution of ligands by exponential enrichment. Mol Cell Biol. 2005 Feb;25(3):879-87.

65. Cooper TA. Muscle-specific splicing of a heterologous exon mediated by a single muscle-specific splicing enhancer from the cardiac troponin T gene. Mol Cell Biol. 1998 Aug;18(8):4519-25.

66. Chang C. C., Lin C. J. LIBSVM: a library for support vector machines [Internet]. ; 2007 January 2, 2007 [January 10, 2007] . Available from: http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf.

67. Fan RE, Chen PH, Lin CJ. Working set selection using second order information for training support vector machines. Journal of Machine Learning Research. 2005;6:1889-918.

68. Hsu C. W., Chang C. C., Lin C. J. A practical guide to support vector classification [Internet]. ; 2003. Available from: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

69. Pulak R, Anderson P. mRNA surveillance by the caenorhabditis elegans smg genes. Genes Dev. 1993 Oct;7(10):1885-97.

70. Frischmeyer PA, Dietz HC. Nonsense-mediated mRNA decay in health and disease. Hum Mol Genet. 1999;8(10):1893-900.

71. Baker KE, Parker R. Nonsense-mediated mRNA decay: Terminating erroneous gene expression. Curr Opin Cell Biol. 2004 Jun;16(3):293-9.

72. Del Gatto-Konczak F, Olive M, Gesnel MC, Breathnach R. hnRNP A1 recruited to an exon in vivo can function as an exon splicing silencer. Mol Cell Biol. 1999 Jan;19(1):251-60.

73. Martinez-Contreras R, Fisette JF, Nasim FU, Madden R, Cordeau M, Chabot B. Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. PLoS Biol. 2006 Feb;4(2):e21.