

Sequence Motifs Predictive of Tissue-Specific Skipping

by

Neha Soni

B.S. Engineering and Applied Science
California Institute of Technology, 2003

Submitted to the Department of Electrical Engineering and Computer
Science

in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2006

© Massachusetts Institute of Technology 2006. All rights reserved.

Author
Department of Electrical Engineering and Computer Science
September 15, 2005

Certified by
Tomaso Poggio
Eugene McDermott Professor
Thesis Supervisor

Accepted by
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Sequence Motifs Predictive of Tissue-Specific Skipping

by

Neha Soni

Submitted to the Department of Electrical Engineering and Computer Science
on September 15, 2005, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

Alternative splicing plays a major role in protein diversity and regulating gene expression. Motifs that regulate tissue-specific alternative splicing have been identified by groups studying small sets of genes. We introduce a tissue-specific skipping score for skipped exons using exon-exon junction microarray data. We compare these exons with known literature-verified EST skipped exons and exons predicted to be skipped in both human and mouse. After deriving tissue-specific skipped exon sets for brain, heart, muscle and testis, we find sequence features in the exon and flanking introns that distinguish these tissue-specific skipped exons from constitutive exons. Lastly, we use sequence-based scoring based on these features to predict tissue-specific skipped exons and compare these with EST data to demonstrate the tissue-specificity of the motifs.

Thesis Supervisor: Tomaso Poggio
Title: Eugene McDermott Professor

Acknowledgments

I would like to thank my advisors Dr. Tomaso Poggio and Dr. Chris Burge for this fantastic opportunity, and for their help, support and guidance in this project. Dr. Gene Yeo, who I admire greatly, has been a great mentor without whom this work would not have been possible. He has motivated me, guided me and has also been a great friend. I am grateful to the CBCL lab members for sharing their knowledge and for making CBCL a nice place to work at; Gabriel Kreiman for his help ; Stan Bileshci, Ryan Rifkin and Lior Wolf for being good office neighbors and for answering my queries; Bernd Heisele for proofreading the thesis; Masa, Jake, Ethan, Davide, Adlar, Sasha, Gadi, Ulf, Sanmay, Minjoon, Thomas, Giorgos, and Yuri for being good office buddies; Mary Pat for all her help, and Tony Ezzat and Jane Wu for being always supportive and very good friends. I want to thank the Burge lab members for their help; in particular, I want to thank Noam Shomron for his help, expertise and extreme patience with the wet lab experiments, Eric, Brad Friedman, Zefeng Wang and Jianming Zhang also for their help with the experiments, and Rickard Sandberg, Michael Stadler and Craig Rothman for their help with the data.

I am thankful to my undergraduate Caltech friends who have been so greatly supportive through the years and have become an integral part of my life; Mona Sheikh, Jim Pugh, Jon Chang, and Kevin Ko with whom I've had tons of useful, interesting and fun discussions online even while far away; and my Boston-based Caltech friends Elaine Ou, Ben Lee and Kaisey Mandel for keeping in touch regularly and being there. At MIT, I am so glad I got to live with a roommate as amazing as Diana who took such great care of me, and for all the fun kitchen chats with Hiba and Diana that helped us relax when we most needed it. I want to thank all my close friends in graduate school for making this a very intense, emotional and fun experience that I will cherish forever; my next roommate, Evelyn, who has been so sweet to me, and with whom I wish I had gotten to spend more time with; my fellow CSAIL friends; the Boston crowd Aarthi, Andy, Anoop, Ashish K, Ashish N, Charu, Chintan, Hetal, Joyce, Lavanya, Mahesh, Manish, Neha Singh, Rupa, Sapna, Vipin

and Vish who made Boston a home away from home; my Sidney-Pacific dorm friends; my hall councilor Lucy, and my fellow officers for a memorable graduate experience. Finally, I want to thank my brother who has always believed in me and whom I admire greatly, and above all, my parents for their unconditional love.

Contents

1	Introduction	15
1.1	Alternative Splicing and disease	15
1.2	Previous work on tissue-specific alternative splicing	17
1.3	Prediction of tissue-specific skipping	18
1.4	Thesis Outline	18
2	Microarray-based prediction of tissue-specific skipping	21
2.1	Microarray dataset	21
2.2	Prediction of skipping from microarray data by assigning scores . . .	22
2.2.1	Normalization of the microarray per gene	22
2.2.2	Scoring of the exons	23
2.2.3	Learning the weight parameters	26
2.3	Correlation with previous predictions	27
2.4	Tissue-specific skipping	28
2.5	Reading frame	28
3	Extraction of tissue-specific skipping motifs	31
3.1	Comparing TSS sets with the background set	31
3.2	TSS motifs in brain, muscle, heart and testis	32
4	Sequence-based prediction of tissue-specific skipping	35
4.1	Using TSS motifs to predict tissue-specific skipping	35
4.2	Tissue specificity of motifs	36

4.3	Classification using sequence-based scoring	41
5	Conclusions and Perspectives	45
A	Tables of TSS motifs	47

List of Figures

1-1	A. Regulatory elements that can affect alternative splicing B. Types of alternative splicing events [2].	16
2-1	Original hybridisation intensities for the EPB41L2 gene.	22
2-2	Normalised hybridisation intensities for the EPB41L2 gene. The darker cells correspond to lower intensities.	23
2-3	Binned intensity values for the EPB41L2 gene. The darkest bin correspond to the exons that are predicted to be skipped while the lighter bins correspond to exons predicted to be included.	24
2-4	Mean squared error versus gradient descent iterations. The parameters are the weights assigned to the different bins, and they determine the predicted skipping score for the exons. The parameters are learned by minimising the mean squared error between the known and the predicted labels on the training set using a gradient-descent approach. The solid curve depicts the mean squared error on the training set, and the dashed curve depicts the error on the validation set.	26

2-5 This figure shows the cumulative distribution curves for the M-NS scores assigned to exons. The M-NS scores depict the skipping of the exon from the microarray data. In each of the figures, the difference between the cumulative distribution curves for the skipped exons (dashed-dotted curve) and the constitutive exons (solid curve) provides a cutoff measure for the M-NS scored exons. (a) Cumulative distribution curves for exons with EST evidence for skipping versus constitutive exons. (b) Cumulative distribution curves for the ACEScan(+) exons with EST evidence versus constitutive exons. 29

4-1 Tissue-specificity of the brain motifs. The exons (21000 exons) were ordered from highest to lowest according to the S-TSS scores using the brain-specific sequence motifs, and these ordered exons were divided equally into partitions with the first partition containing the exons with the highest S-TSS scores. The number of exons in each partition predicted to be skipped in the brain according to the microarray (M-TSS) scores is shown in the figure. 37

4-2 Tissue-specificity of the heart motifs. The exons (21000 exons) were ordered from highest to lowest according to the S-TSS scores using the heart-specific sequence motifs, and these ordered exons were divided equally into partitions with the first partition containing the exons with the highest S-TSS scores. The number of exons in each partition predicted to be skipped in the heart according to the microarray (M-TSS) scores is shown in the figure. 38

4-3	Tissue-specificity of the muscle motifs. The exons (21000 exons) were ordered from highest to lowest according to the S-TSS scores using the muscle-specific sequence motifs, and these ordered exons were divided equally into partitions with the first partition containing the exons with the highest S-TSS scores. The number of exons in each partition predicted to be skipped in the muscle according to the microarray (M-TSS) scores is shown in the figure.	39
4-4	Tissue-specificity of the testis motifs. The exons (21000 exons) were ordered from highest to lowest according to the S-TSS scores using the testis-specific sequence motifs, and these ordered exons were divided equally into partitions with the first partition containing the exons with the highest S-TSS scores. The number of exons in each partition predicted to be skipped in the testis according to the microarray (M-TSS) scores is shown in the figure.	40
4-5	Cumulative distribution functions of the STSS scores. The dashed curves represent the cumulative brain-motif based STSS scores assigned to exons that have no EST evidence of skipping in any tissue. (a) The solid curve represents the cumulative distribution function of the STSS scores assigned to exons that have shown skipping from EST data. (b) The solid curve represents the cumulative distribution function of the STSS scores assigned to exons that have shown EST-based skipping in any tissue other than the brain.	42
4-6	Classification performance using STSS score cutoff values to classify exons skipped in the brain. The solid curve represents the classification performance between exons skipped in the brain according to EST data and the constitutive exons. The dashed curve represents the classification performance between exons skipped in non-brain tissues and the constitutive exons. The dotted curve represents classification by chance.	43

List of Tables

A.1	Brain-specific TSS motifs	48
A.2	Heart-specific TSS motifs	49
A.3	Testis-specific TSS motifs	50
A.4	Muscle-specific TSS motifs	51

Chapter 1

Introduction

Relative to transcription, a less well understood mechanism of gene expression control which has recently come under much attention, is via the regulated removal or inclusion of RNA sequence regions in premature messenger RNA by alternative splicing, thereby altering the function of a gene in a tissue-specific fashion, or by down-regulation of the gene by generating a premature termination stop codon, leading to degradation via the nonsense-mediated decay pathway. Alternative splicing plays a major role in protein diversity and regulating gene expression in higher eukaryotes [5, 4], and more than half the human genes have been estimated to be alternatively spliced. Motifs that regulate splicing (splicing enhancers or silencers) can be found on the exons as well as the flanking introns as shown in Figure 1-1 A. Figure 1-1 B shows the different elementary alternative splicing events that can occur [2].

1.1 Alternative Splicing and disease

It is widely believed that alternative splicing is of particular importance in the nervous system. Defects in the splicing machinery have been known to cause a substantial fraction of human genetic diseases. Errors in alternative splicing regulation directly cause a large number of diseases, including several types of cancer, dementia, genetic defects and lethal infant degenerative conditions such as spinal atrophy [14, 8]. Correction of erroneous splicing is an important goal of molecular therapies [13]. This is

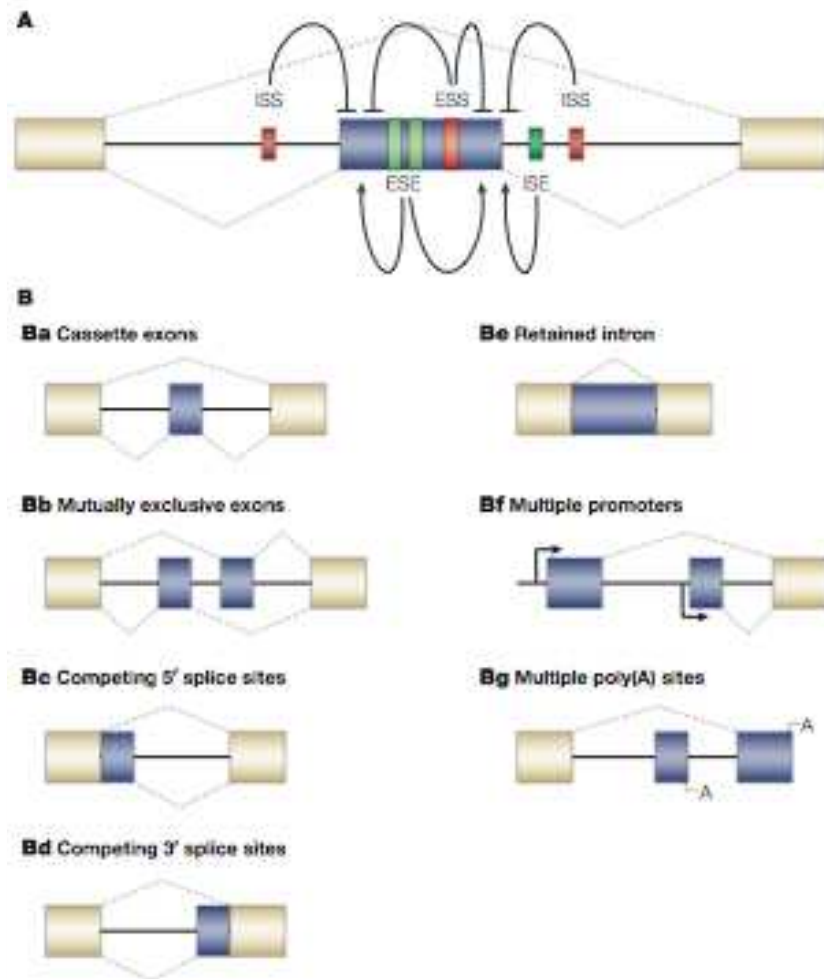


Figure 1-1: A. Regulatory elements that can affect alternative splicing B. Types of alternative splicing events [2].

the main motivation for studying alternative splicing in greater depth and using computational approaches to understand the signals that affect tissue-specific alternative splicing, which shall be explored in this thesis.

1.2 Previous work on tissue-specific alternative splicing

Cell type and tissue-specific regulation of gene expression is of utmost importance in specifying normal growth and development, metabolism, and other fundamental processes in living organisms. Control of tissue-specific expression of a gene can be achieved by various different mechanisms, such as binding of tissue-specific transcription factors to the promoter regions of genes, and more recently, complementary base-pairing of tissue-specifically expressed microRNAs to the 3'-UTR of target genes. A common theme in understanding such specific regulation is to identify sequences that act as cis-regulatory binding sites to tissue-specific trans-acting elements (proteins or RNA sequences). A basis for most methods depends on extracting motifs that distinguish sets of transcripts enriched in one tissue versus other tissues. To systematically and efficiently perform such computational searches of sequence motifs, large experimental sources of data across a wide variety of tissues of interest are necessary. Such data can come from expressed sequence tags (EST's), SAGE tags or microarray data. For example, tissue-specific binding sites (particular to muscle and liver) have been identified by Smith, Sumazin and Zhang [1].

Splicing is regulated by the interactions of protein factors with the splicing machinery, where the factors bind to cis-elements in the regulated exons or the flanking introns. Careful studies of the binding affinities of particular splicing trans-factors such as ETR3 [15] and NOVA-1 [16, 10], or mutation of cis-regulatory elements, or a combination, have identified tissue-specific cis and trans relationships. Systematic larger-scale searches using experimental and computational methods [22, 18, 19, 12] have uncovered additional splicing cis-elements. In addition, motifs that regulate

tissue-specific alternative splicing have been identified by groups studying small sets of genes [17]. Efforts involving larger sets of tissue-specifically regulated exons have come from utilizing expressed sequence tags (EST's) [6]. Even after controlling for the non-uniformity of tissue libraries, ESTs retain significant limitations because of bias in transcript coverage [9, 7]. Fortunately, more recent large-scale experimental approaches utilizing splicing-specific microarrays have provided a more comprehensive set of tissue-specific alternative splicing exons[9, 3].

1.3 Prediction of tissue-specific skipping

In this study, a binning approach is described that assigns a potential microarray-based tissue-specific skipping (M-TSS) score for exon skipping to each exon in the Rosetta microarray dataset [9]. The scores were utilized to compartmentalize exons into tissue-specific skipping (TSS) sets. Predicted skipped exons were compared to exons with a positive ACEScan score (ACEScan[+] exons) [20] predicted using a comparative genomics and statistical learning method. A simple statistical motif-finding approach was used to identify tissue-specific motifs particular to exon skipping events. These motifs were used to predict tissue-specific skipped exons from the sequence data by assigning to them a sequence-based tissue-specific skipping (S-TSS) score. In combination with methods such as ACEScan, we thus suggest the potential to predict tissue-specific exon skipping from sequence data alone and compare the skipping scores assigned to exons with EST skipping evidence. Such cis-elements are likely to be important in specifying tissue-specific regulation of alternative splicing in vivo.

1.4 Thesis Outline

Chapter one introduces the motivation behind studying alternative splicing, and looks at some previous relevant work in this area. Chapter two describes the methods for using microarrays to predict tissue-specific alternative splicing in human tissues.

Chapter three describes the extraction of motifs that distinguish between tissue-specific skipped exons and constitutive exons. Chapter four describes the prediction of tissue-specific skipping from sequences using the motifs found with the help of the microarray data and demonstrates the tissue-specificity of the motifs. Chapter five concludes the work, and offers some perspectives.

Chapter 2

Microarray-based prediction of tissue-specific skipping

2.1 Microarray dataset

The microarray data consists of exon-exon junction microarrays for more than 10,000 human genes in 52 tissues and cell lines [9]. The tissue set consists of some diseased cell-lines in addition to normal tissues. Microarray data tends to be noisy, and there are many factors that cause variation within this data. The tissue-specific expression levels and the binding affinities of the probes are estimated from the data itself. In addition, there are a number of probes that do not hybridize target RNA well or at all. All these factors have to be taken into account when determining exon skipping in a tissue.

Each gene is characterized by a matrix of hybridization intensities of exon-exon junction probes across splice junctions for the tissues and cell lines as shown in Figure 2-1 for the EPB41L2 gene. The tissues and cell-lines as ordered on the microarray are listed in Table ???. Log-transformed data is often used in microarray analysis to do statistical tests. But untransformed data can sometimes be more useful for detecting global correlations or finding significant changes in intensity levels [21].

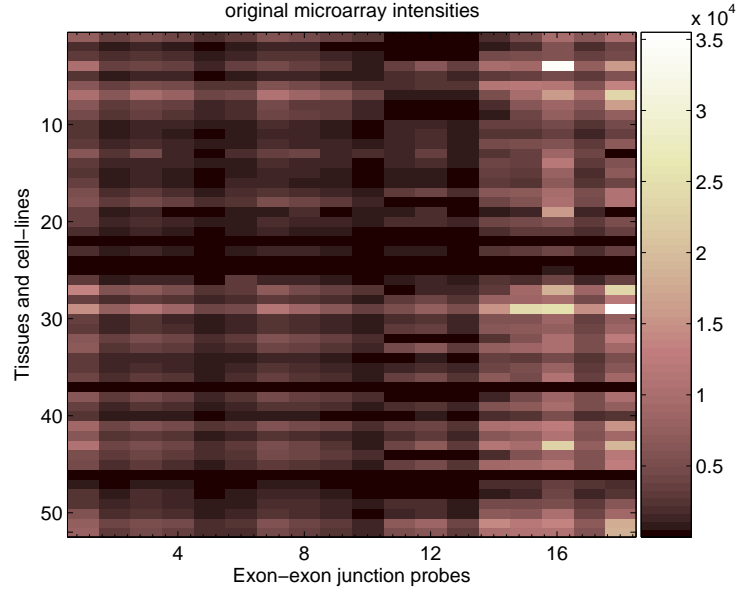


Figure 2-1: Original hybridisation intensities for the EPB41L2 gene.

2.2 Prediction of skipping from microarray data by assigning scores

2.2.1 Normalization of the microarray per gene

For each gene, the probe binding affinity for a particular junction probe was approximated by the median of the original intensities across all tissues for that probe. Similarly, the tissue-specific gene expression level was approximated by the median of the intensities across all junction probes in the gene for that tissue. The intensities for the gene were thus normalized in the following way:

$$x'_{p,t} = \frac{x_{p,t} - \mu_t}{\sigma_t}$$

$$x''_{p,t} = \frac{x'_{p,t} - \mu'_p}{\sigma'_p}$$

where $x_{p,t}$ is the original intensity value for the p^{th} probe in the t^{th} tissue, $x'_{p,t}$ and $x''_{p,t}$ are the intensity values after normalizing across tissues and then across

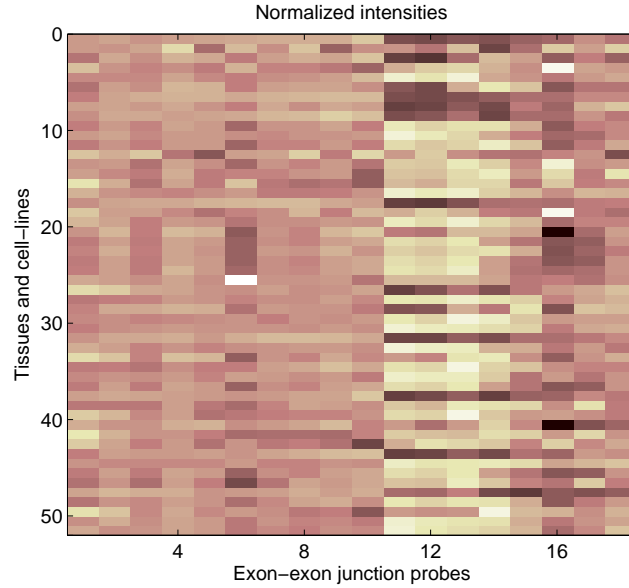


Figure 2-2: Normalised hybridisation intensities for the EPB41L2 gene. The darker cells correspond to lower intensities.

junction probes respectively, (μ_t, σ_t) is the median and standard deviation of the original intensity values across all junction probes for that gene in tissue t within that gene, and (μ'_p, σ'_p) is the median and standard deviation of the normalized intensity values of junction probe p across all tissues after the first normalization. Figure 2-2 displays the normalized intensity values for the EPB41L2 gene.

2.2.2 Scoring of the exons

After normalization, all the normalized intensity values within a gene were binned into multi-variant bins irrespective of tissue or junction, using a euclidean k -means approach. In this approach, the intensities are partitioned into k disjoint subsets or bins $b_j, j \in [1, k]$, such that each bin contains intensities $x_{j,i}, i \in [1, |b_j|]$ where μ_j is the mean intensity value for that bin and $|b_j|$ is the size of the j^{th} bin.

Initially, the bin centers μ_j are chosen randomly. In each iteration, the intensities are reassigned to the closest bin so as to minimize the sum of distances of each intensity from its new bin center. Thus the following sum of squares criterion is minimized at each step, to get an optimal assignment of the intensities to the bins:

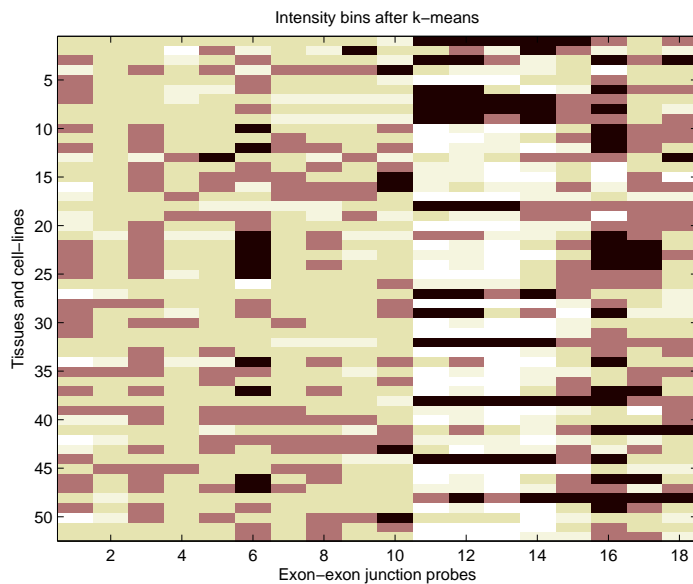


Figure 2-3: Binned intensity values for the EPB41L2 gene. The darkest bin correspond to the exons that are predicted to be skipped while the lighter bins correspond to exons predicted to be included.

$$\sum_{j=1}^k \sum_{i=1}^{|b_j|} |x_{j,i} - \mu_j|^2$$

The iterations are stopped when the stopping criteria is met, i.e. when there is no further reassignment of the intensities to the bins. To prevent getting stuck in a local minimum, this was repeated a few times for each gene and the bin assignment that resulted in the minimum sum of squares was chosen. Figure 2-3 shows the bin assignments of all the intensity values after applying k-means for $k = 5$. Thus, the intensities that are assigned to the first bin have the lowest intensities and are predicted to correspond to skipped exons.

To assign a skipping score to an exon in a particular tissue t , each of the bins above was given a weight. These bin weights were estimated from the training data with a gradient descent approach. The microarray-based tissue-specific skipping (M-TSS) score for exon e in tissue t is thus calculated by summing the bin weights for

the intensities corresponding to that exon.

$$\text{M-TSS}(e, t) = w[\text{bin}(p_e, t)] + w[\text{bin}(p_{e+1}, t)]$$

where p_e is the junction probe overlapping with exon($e-1$) and exon (e), $\text{bin}(p_e, t)$ is the bin assignment for the intensity of the junction probe p_e in tissue t , and $w[\text{bin}(p_e, t)]$ is the weight of that bin. High weights correspond to an affinity for exon skipping, while low weights correspond to an affinity for exon inclusion. Hence, the lower intensity bins get assigned a higher weight so that a high M-TSS score corresponds to the likelihood of the exon to be skipped in that tissue. These weights are determined via a learning algorithm using gradient descent. To determine the optimum bin weights, a training set of about 40 positive (alternatively spliced) and 40 negative (constitutive) exons was used. The set consisted of non tissue-specific skipped exons. A validation set of similar size is used to prevent over-fitting to the training set. All these training exons are either literature-verified or verified in previous studies by reverse transcription polymerase chain reaction (RT-PCR) and sequencing.

In order to compare the M-TSS predicted skipped exons with the training set of skipped exons, a microarray-based net skipping (M-NS) score was calculated for each exon e .

$$\text{M-NS}(e) = \frac{\sum_t \text{MTSS}(e, t) - \mu(\forall e, \sum_t \text{MTSS}(e, t))}{\sigma(\forall e, \sum_t \text{MTSS}(e, t))}$$

where $\sum_t \text{MTSS}(e, t)$ is the net skipping score for the exon, i.e., the sum of the M-TSS scores of the exon over all tissues, $\mu(\forall e, \sum_t \text{MTSS}(e, t))$ is the median of the net skipping scores of all exons in the gene, and $\sigma(\forall e, \sum_t \text{MTSS}(e, t))$ is the standard deviation of the net skipping scores of all exons in the gene. Under the assumption that a majority of the exons within a gene are constitutive, this transforms the M-TSS scores to have zero median and unit standard deviation.

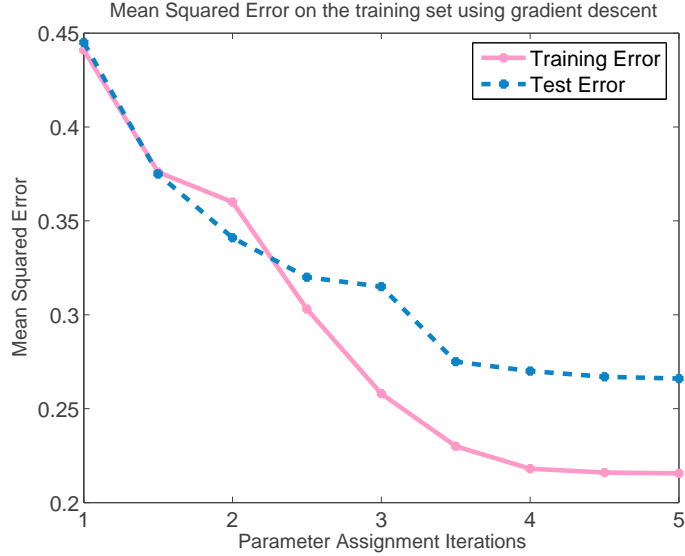


Figure 2-4: Mean squared error versus gradient descent iterations. The parameters are the weights assigned to the different bins, and they determine the predicted skipping score for the exons. The parameters are learned by minimising the mean squared error between the known and the predicted labels on the training set using a gradient-descent approach. The solid curve depicts the mean squared error on the training set, and the dashed curve depicts the error on the validation set.

2.2.3 Learning the weight parameters

The weight parameters are obtained using a gradient descent approach to minimize the root mean squared prediction error $F(\vec{w}_n)$ which is the error between the known and predicted labels of the exons. \vec{w}_n is the vector of bin weights at the n^{th} iteration of the algorithm, the predicted label $\hat{y}_e(\vec{w}_n)$ for the e^{th} exon is the normalized M-NS score lying between 0 and 1 ($\hat{y}_e = 1$ for the strongest skipping score, $\hat{y}_e = 0$ for the weakest skipping score) calculated using the n^{th} iteration weights, and y_e is the known binary label for that exon ($y_e = 1$ for alternatively spliced exons, $y_e = 0$ for constitutive exons). The prediction error at the n^{th} iteration is given by:

$$F(\vec{w}_n) = \frac{1}{n} \sqrt{\sum_e (y_e - \hat{y}_e(\vec{w}_n))^2}$$

The weights are updated in the direction of the negative of the gradient of the mean

error on the training set, such that the step size, γ , is allowed to decrease at every iteration. To avoid overfitting to the training set, the training data was partitioned into the training set and the validation set with 40 positive and 40 negative exons each. The stopping criteria is defined by a change in mean-square error per iteration of less than 0.01. In Figure 2-4, the dashed curve depicts the mean squared error on the training set, and the solid curve depicts the error on the validation set, confirming that the parameters have not been overfitted to the training set.

2.3 Correlation with previous predictions

To compare the exons predicted to be skipped from microarray data with ACEscan(+) exons, each exon was given a Microarray-based Net Skipping (M-NS) score. This net score is the number of standard deviations away that the total score of the exon in all tissues is from the median net score given to all exons in that particular gene.

The predicted M-NS scores of exons with EST skipping evidence (SE exons) were compared with the M-NS scores of the background set of constitutive exons (CE exons). These exons are literature-verified and RT-PCR verified. The SE set had 1400 exons, and the CE set had about 30,000 exons. Random sets of 500 exons in the skipped set and the background constitutive set were selected over 20 iterations and the corresponding M-NS scores were found. The figures show a correlation between known skipped exons and high M-NS scores, i.e. between EST data evidence and microarray data predictions. We also compared ACEscan predictions with the M-NS predictions. ACEscan uses conservation in human and mouse to make predictions about exon skipping from sequence data.

Figure 2-5 shows the cumulative distribution curves for the M-NS scores assigned to exons. The M-NS scores depict the skipping of the exon from the microarray data. In each of the figures, the difference between the cumulative distribution curves for the skipped exons depicted by dashed curves and the constitutive exons depicted by solid curves provides a rough cutoff measure for the M-NS scored exons. The EST skipped exons have a higher cumulative M-NS score distribution than the constitutive

exons.

Figure 2-5(a) shows the cumulative distribution curves for the EST-based skipped exons versus the constitutive exons. Interestingly, the cumulative distribution curves for the ACEscan(+) EST skipped exons versus the constitutive exons in Figure 2-5(b) are further apart. The significantly larger difference between the two distributions could be due to the SE set of exons containing exons being skipped in diseased tissues or non-functional isoforms in addition to the regulated evolutionarily conserved alternatively spliced ones (ACEscan[+] exons). An alternative explanation is that skipping events that are conserved in human and mouse are skipped in a greater number of tissues and are thus detected more easily across tissues.

could be skipped in more tissues and hence is detected more easily across tissues.

2.4 Tissue-specific skipping

According to the predictions, the brain exons exhibited most skipping (973 exons), followed by the heart (850 exons), muscle (440 exons) and testis (390 exons). Exons predicted to be skipped in a particular tissue, like the brain, were compared with exons that were included in all the tissues. Several motifs were found to be over-represented in the exons and the flanking intron regions. We considered regions of 150bp in the upstream and downstream introns avoiding splice sites.

2.5 Reading frame

Out of roughly 21,000 exons that were skipped in at least one normal tissue, about 8200 preserved the reading frame (i.e., the number of base pairs in their exon sequences was a multiple of three). Out of 2276 exons that were skipped in at least 4 normal tissues, 877 exon sequences were a multiple of 3, and 1399 were not.

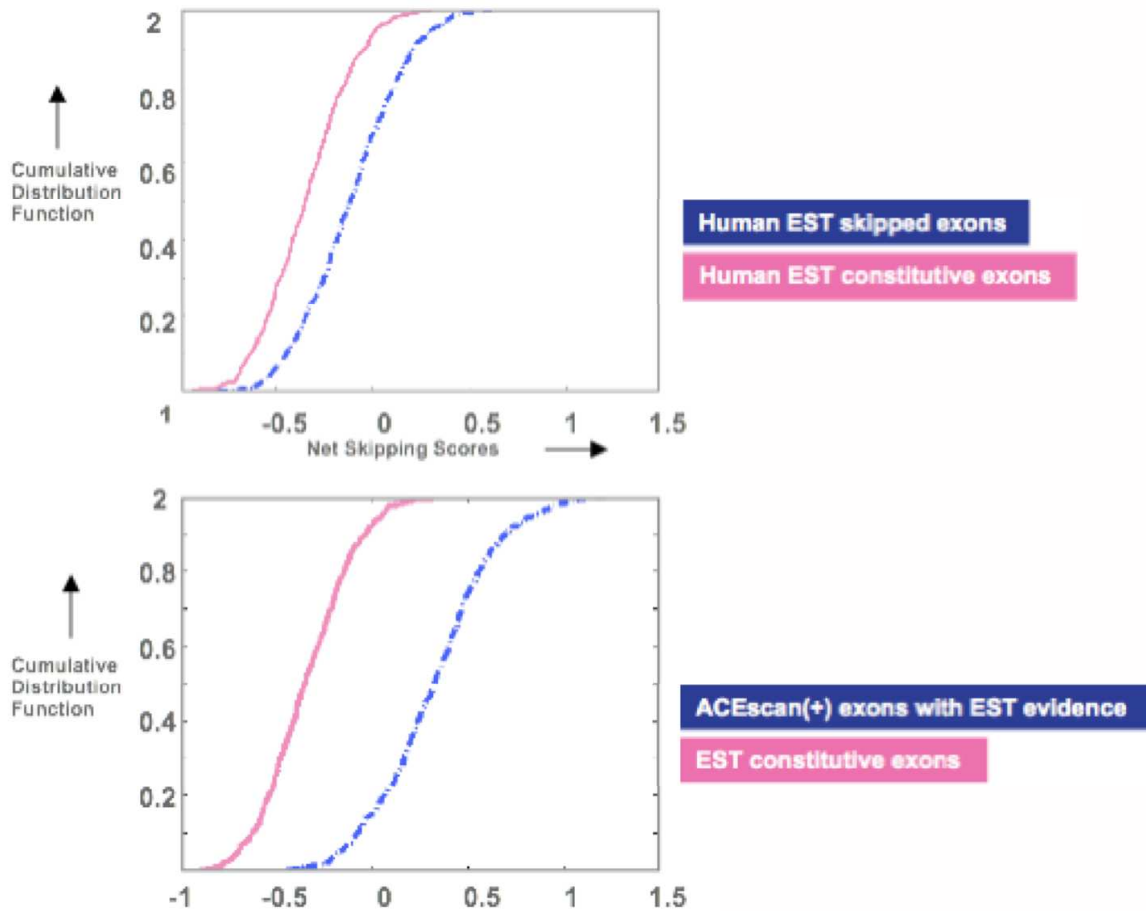


Figure 2-5: This figure shows the cumulative distribution curves for the M-NS scores assigned to exons. The M-NS scores depict the skipping of the exon from the microarray data. In each of the figures, the difference between the cumulative distribution curves for the skipped exons (dashed-dotted curve) and the constitutive exons (solid curve) provides a cutoff measure for the M-NS scored exons. (a) Cumulative distribution curves for exons with EST evidence for skipping versus constitutive exons. (b) Cumulative distribution curves for the ACEscan(+) exons with EST evidence versus constitutive exons.

Chapter 3

Extraction of tissue-specific skipping motifs

3.1 Comparing TSS sets with the background set

Using the microarray data to assign skipping scores to exons in each tissue, we attain tissue-specific skipping (TSS) sets for brain, muscle, heart and testis. We consider exons with the strongest evidence for skipping in that particular tissue, i.e. the exons for which both the corresponding probe intensity values belonged to the lowest intensity bins.

Given a TSS set of exon sequences, we search for common sequence patterns that occur more frequently than chance in constitutive exons and their flanking introns, similar to the method used in [11]. Statistically significant hexamer and pentamer motifs enriched in the exonic and intronic regions of the sequences exhibiting tissue-specific alternative splicing are found as described below.

The frequency of occurrence of motif m in the background set is $p_m = \frac{e_{b,m}}{n_b}$, where $e_{b,m}$ is the number of samples containing motif m in the set of n_b samples analyzed. For the enrichment of a motif in tissue t with respect to all exon samples, $n_b = 4900$ for the background set that was considered. n_t was the size of the TSS set. We assume as a null hypothesis that the motif was randomly distributed across all exons or introns. So the enrichment is defined as the probability that the number of samples

e_t in which the motif m is present, is larger than or equal to the observed value x assuming the frequency p_m in all exons. This follows a hypergeometric distribution which converges to the binomial distribution when $n_b \gg n_t$. The probability of enrichment is expressed as

$$P(e_t \geq x) = \sum_{i=x}^{n_t} \binom{n_t}{i} p_m^i (1 - p_m)^{n_t-i}$$

The enrichment probability is computed for all the motifs. P-values for the motifs are calculated by comparing their enrichment probabilities with motifs from randomly distributed sets. Motifs with p-value < 0.01 are selected as TSS motifs. The motif weight assigned to motif m for a particular tissue t is based on the frequency of occurrence of the motif in the tissue-specific set and is given by

$$w = \log\left(\frac{f_{m,t}}{f_{m,b}}\right)$$

where $f_{m,t}$ is the frequency of occurrence of motif m in the TSS set of exons skipped in tissue t , and $f_{m,b}$ is the frequency of occurrence of the motif m in the background set of constitutive exons.

3.2 TSS motifs in brain, muscle, heart and testis

Statistically significant motifs over-represented or under-represented in the exonic and intronic regions of the sequences exhibiting alternative splicing were found as described in Section 3.1. The p-values for all the motifs were calculated and those with p-values < 0.01 are listed in Appendix A. The motifs are listed for the TSS sets that are skipped in the brain, muscle, heart and testis.

The tables in Appendix A list the motifs that were found over-represented and under-represented in the exonic and intronic regions in four tissues: brain, muscle, heart and testis. GU-rich sequences present in the intronic regions were shown to

enhance inclusion of the exon in that gene, i.e. they were under-represented in exons skipped in that particular tissue. This is consistent with findings that show that ETR-3 binding sites enriched by SELEX were GU rich and enhanced exon inclusion [15]. Splicing regulation by ETR-3 was stated to be relevant in muscle, heart and brain since ETR-3 is abundant in these tissues.

The motifs over-represented in sequences that exhibit greater exon skipping in the brain were found to be high in G-rich and C-rich sequences. G-rich and C-rich sequences were identified earlier as potential intronic splicing enhancers [6], found more frequently in introns flanking weak exons than in introns flanking strong exons. It is likely that these ISEs acted to enhance the inclusion of these exons in non-brain tissues. This is also consistent with findings that alternative exons included in brain had a deficiency of G-rich sequences in the flanking introns [17].

Chapter 4

Sequence-based prediction of tissue-specific skipping

4.1 Using TSS motifs to predict tissue-specific skipping

Leave-one-out cross-validation was used to find tissue specific motifs and test these motifs on the data. During a single validation experiment, the gene that the validation exon belonged to was left out and the TSS motifs were found for the rest of the data. The TSS motifs in Appendix A were weighted according to the log of the ratio of frequency of occurrence in the tissue-specific set of alternative spliced exons versus the background set of constitutive exons. These weights were used to calculate the score for the validation exon sequence by considering all hexamers and pentamers in the exon as well as those in the flanking intronic regions of 150 bp, avoiding the splice sites.

During each cross-validation, the weighted motifs found were used to predict tissue-specific splicing in the exon. The test exon was given a sequence-based tissue-specific skipping score (S-TSS score) by summing up the weight of each motif in its sequence as well as the flanking intronic sequences. The S-TSS score is thus given by

$$\theta = \sum_{m \in M} \log\left(\frac{f_{m,br}}{f_{m,b}}\right)$$

where $M = (M_{exon} \cup M_{dnint} \cup M_{upint})$ is the set of the motifs over-represented in the TSS sets of exons and the corresponding flanking downstream and upstream introns respectively, $f_{m,br}$ is the frequency of motif m in the brain-specific splicing set of exons, and $f_{m,b}$ is the frequency of motif m in the background set of constitutive exons.

Leave-one-out cross-validation was carried out about 21,000 times where the sequence-based tissue specific skipping (S-TSS) score was calculated for the validation exon each time. The test S-TSS scores were ranked and divided into partitions. In the figure, the x-axis shows 42 partitions of 500 exons each, where the first partition contains the exons with the highest scores, and the last partition contains the exons with the lowest score. The y-axis depicts the number of exons in that bin, which were predicted to be skipped from the microarray data. The figures thus show the correlation between the predictions from the microarray data and those from the sequence data using the weighted TSS motifs.

4.2 Tissue specificity of motifs

The next few figures are used to demonstrate the tissue-specificity the motifs for each of the four tissues in Appendix A. In Figure 4-1, the exons are scored separately according to brain motifs, muscle motifs, heart motifs and testis TSS motifs. The four plots show the correlation of S-TSS scoring for all tissues with the M-TSS scoring for brain skipped exons. The S-TSS scoring for brain shows the strongest correlation with the M-TSS scores, thus validating the brain-specificity of the brain TSS motifs. In a similar fashion, the rest of the Figures 4-2, 4-3 and 4-4 demonstrate the tissue-specificity of the motifs for the heart, muscle and testis respectively.

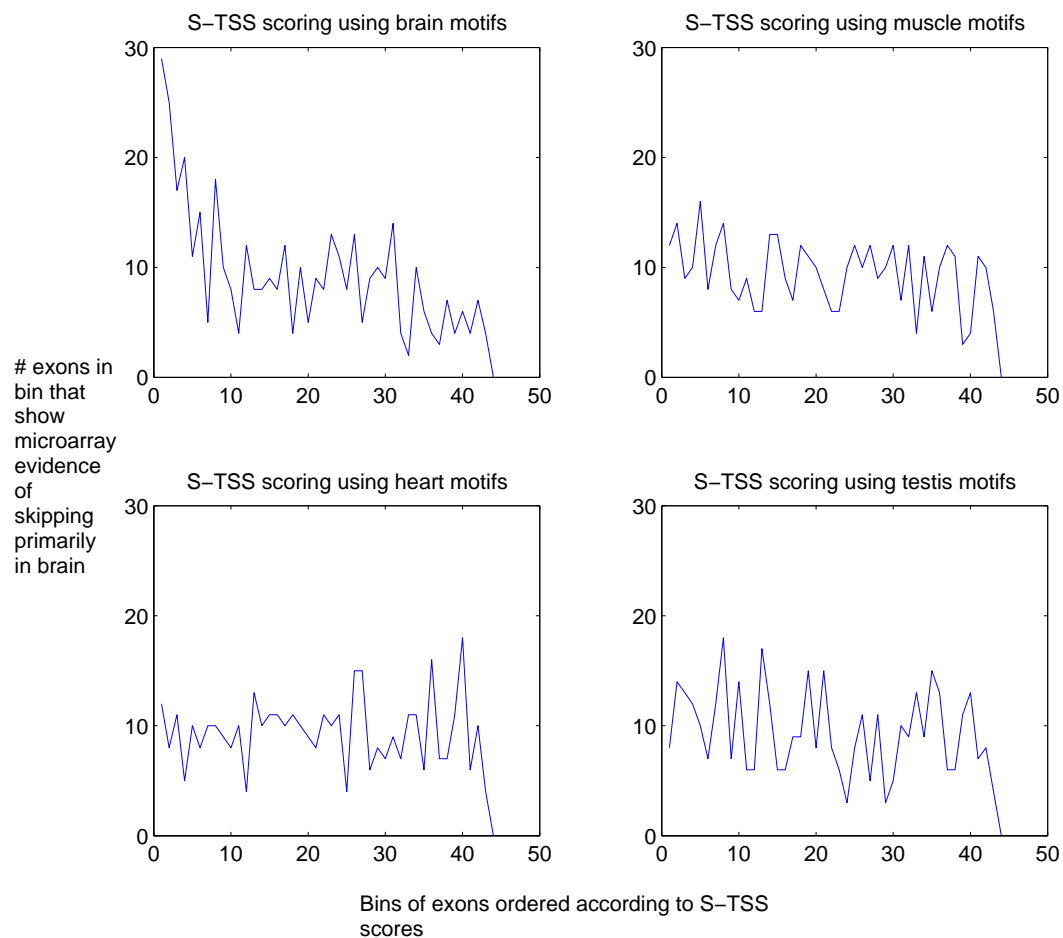


Figure 4-1: Tissue-specificity of the brain motifs. The exons (21000 exons) were ordered from highest to lowest according to the S-TSS scores using the brain-specific sequence motifs, and these ordered exons were divided equally into partitions with the first partition containing the exons with the highest S-TSS scores. The number of exons in each partition predicted to be skipped in the brain according to the microarray (M-TSS) scores is shown in the figure.

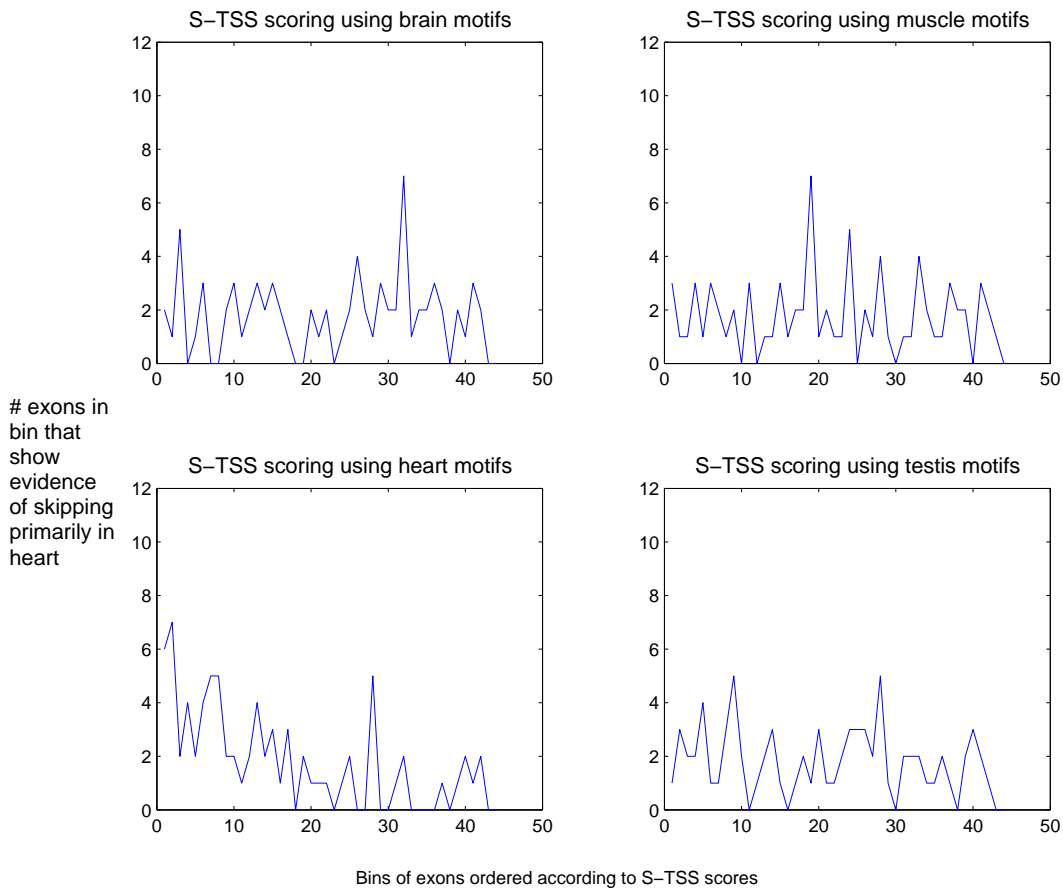


Figure 4-2: Tissue-specificity of the heart motifs. The exons (21000 exons) were ordered from highest to lowest according to the S-TSS scores using the heart-specific sequence motifs, and these ordered exons were divided equally into partitions with the first partition containing the exons with the highest S-TSS scores. The number of exons in each partition predicted to be skipped in the heart according to the microarray (M-TSS) scores is shown in the figure.

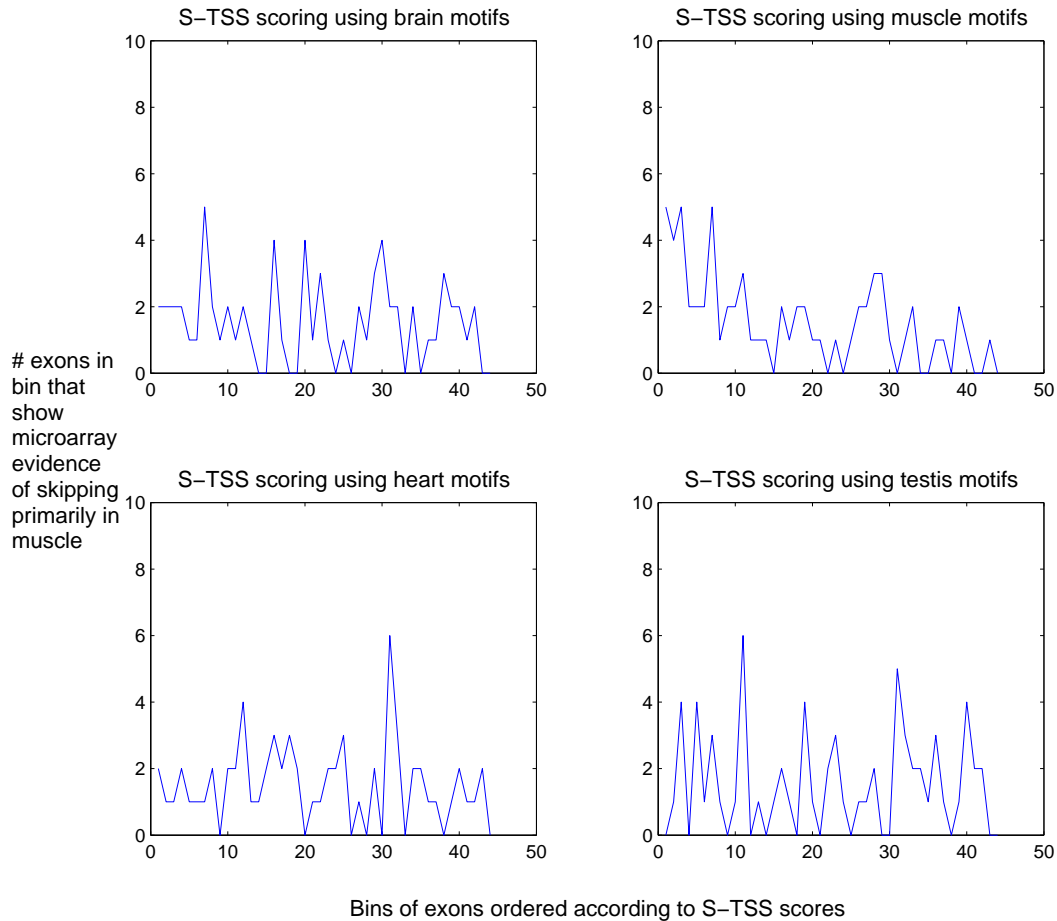


Figure 4-3: Tissue-specificity of the muscle motifs. The exons (21000 exons) were ordered from highest to lowest according to the S-TSS scores using the muscle-specific sequence motifs, and these ordered exons were divided equally into partitions with the first partition containing the exons with the highest S-TSS scores. The number of exons in each partition predicted to be skipped in the muscle according to the microarray (M-TSS) scores is shown in the figure.

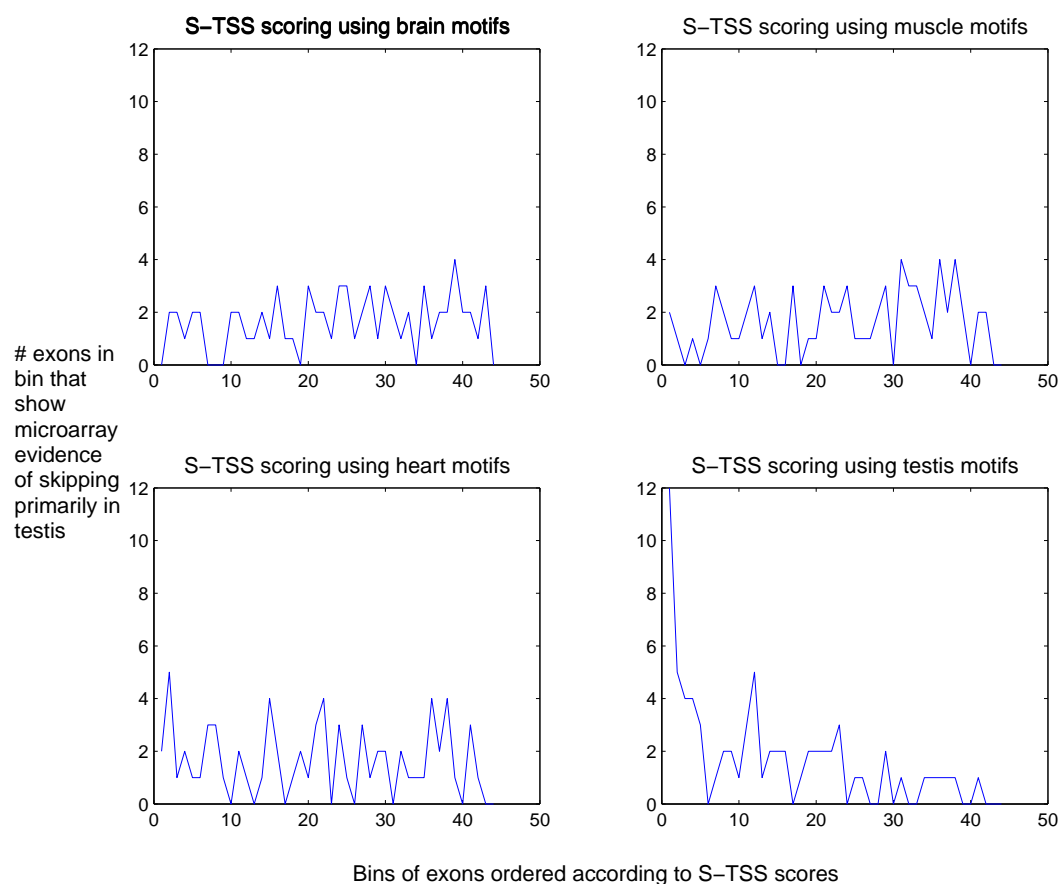


Figure 4-4: Tissue-specificity of the testis motifs. The exons (21000 exons) were ordered from highest to lowest according to the S-TSS scores using the testis-specific sequence motifs, and these ordered exons were divided equally into partitions with the first partition containing the exons with the highest S-TSS scores. The number of exons in each partition predicted to be skipped in the testis according to the microarray (M-TSS) scores is shown in the figure.

4.3 Classification using sequence-based scoring

To further demonstrate the tissue-specificity of the motifs and the usefulness of the STSS scores in predicting novel TSS exons, we used a test set of exons with mid-high brain MTSS scores which did not overlap with the TSS sets of exons used for detecting the motifs. The true labels for the exons are derived from EST data and the S-TSS scores are used to make predictions. The cumulative distribution curves in Figure 4-5 show the differences in STSS scores assigned to the exons skipped in the brain tissues, exons skipped in non-brain tissues and constitutive exons according to EST data. The scores of exons skipped in the brain appear higher, demonstrating that the brain motifs can be used to predict brain-specific skipping and not skipping in general, and hence showing the tissue-specificity of these motifs. Prediction of tissue-specific exon skipping can also be approached from a classification perspective as in Figure 4-6. Each point on the curve represents a specific cutoff value of the STSS score. The noticeable difference between the ROC curves for exons skipped in the brain and exons skipped in non-brain tissues also supports the tissue-specific feature of the brain motifs.

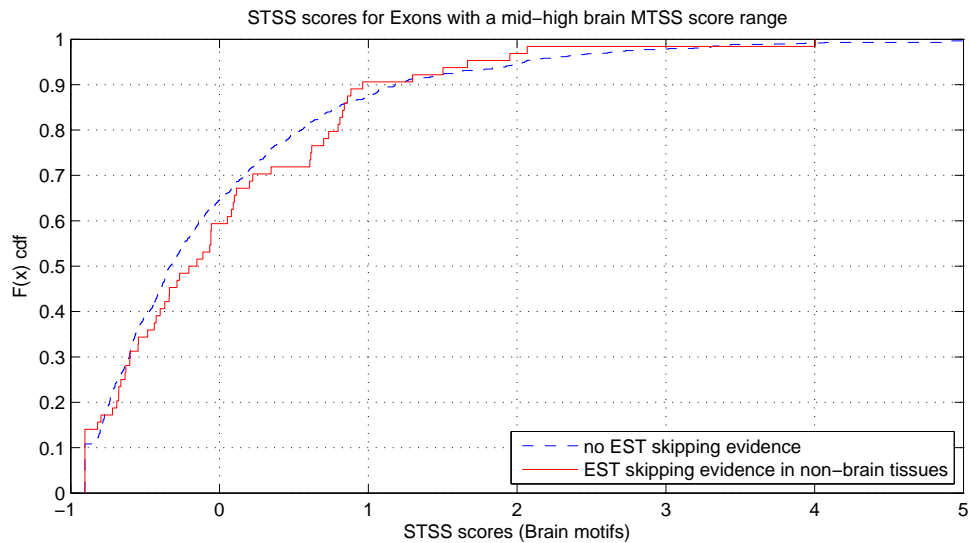
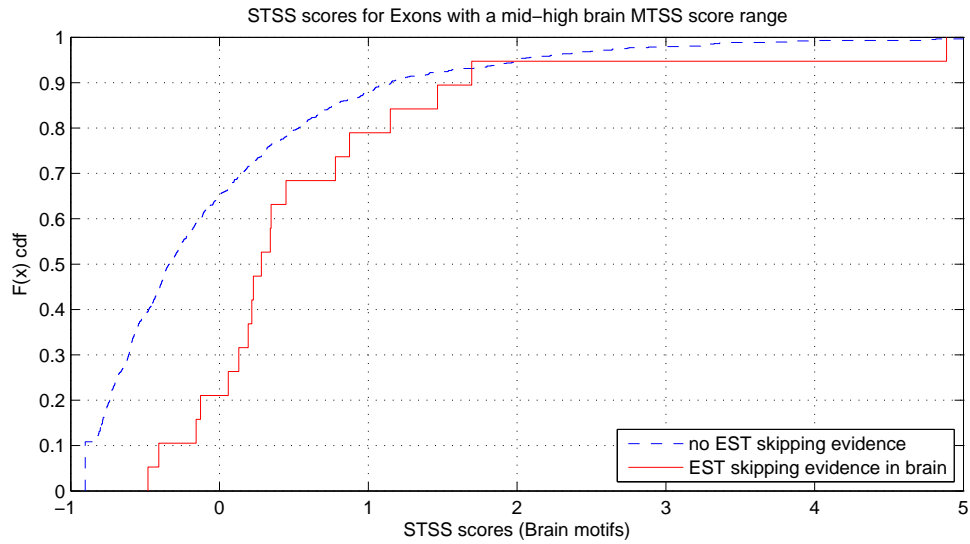


Figure 4-5: Cumulative distribution functions of the STSS scores. The dashed curves represent the cumulative brain-motif based STSS scores assigned to exons that have no EST evidence of skipping in any tissue. (a) The solid curve represents the cumulative distribution function of the STSS scores assigned to exons that have shown skipping from EST data. (b) The solid curve represents the cumulative distribution function of the STSS scores assigned to exons that have shown EST-based skipping in any tissue other than the brain.

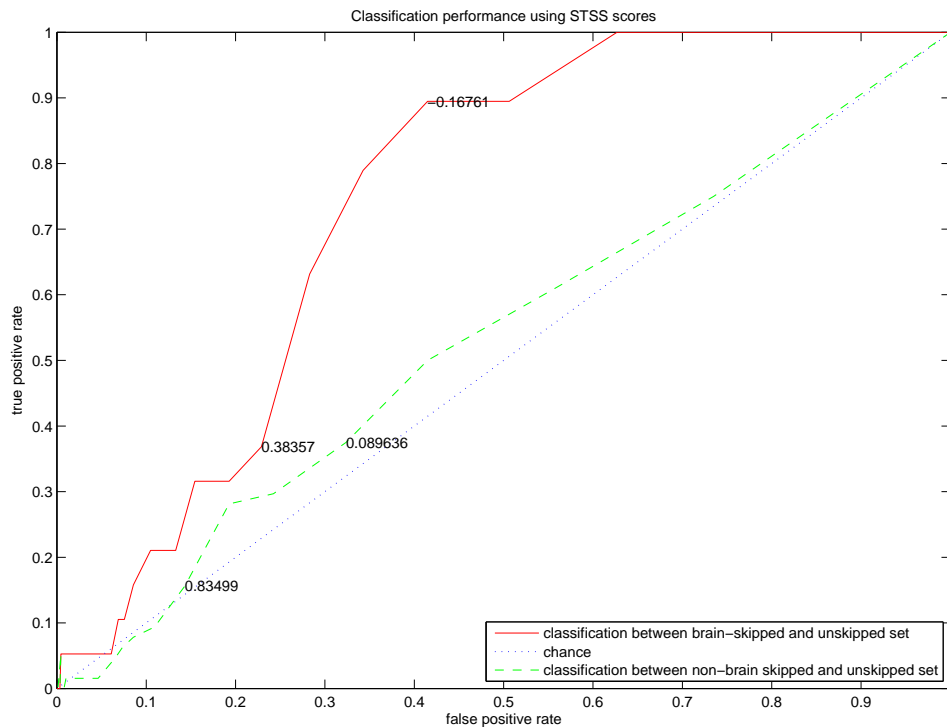


Figure 4-6: Classification performance using STSS score cutoff values to classify exons skipped in the brain. The solid curve represents the classification performance between exons skipped in the brain according to EST data and the constitutive exons. The dashed curve represents the classification performance between exons skipped in non-brain tissues and the constitutive exons. The dotted curve represents classification by chance.

Chapter 5

Conclusions and Perspectives

This thesis finds tissue-specific sequence motifs that contribute to exon skipping. It uses microarray data to assign microarray based tissue-specific skipping scores to exons to predict tissue-specific skipping with a novel binning approach. The parameters for the algorithm were learned from a small set of literature-verified and RT-PCR verified training samples of known alternatively spliced exons. We then compared the skipped exons predicted using the microarray-based scores with skipped exons from the EST data. This helps to provide a rough cutoff measure for microarray-based scores in order to predict exon skipping.

Tissue-specific motifs that enhance skipping were found with a leave-one-out cross-validation method. The skipped exons predicted from the microarray data were divided into tissue-specific sets, and a simple statistical approach was used to discover over-represented motifs in their sequences and the sequences of the flanking introns. These motifs were weighted and then used to assign sequence-based tissue-specific skipping scores to the validation exon sequence to predict skipping. The tissue-specificity of the motifs was demonstrated, and the correlation between sequence-based and microarray-based skipping predictions was also shown.

The biggest contribution of this work is the demonstration of the ability to use short motif sequences in the exon and flanking introns to help predict alternative splicing. In conjunction with methods like ACEscan, the sequence-based tissue-specific scoring approach can be used on the subset of ACEscan predicted alternatively spliced

exons to provide a more accurate prediction for tissue-specific skipped exons.

Appendix A

Tables of TSS motifs

Table A.1: Brain-specific TSS motifs

Upstream intron	Exon	Downstream intron
Over-represented	Over-represented	Over-represented
ggcagc	ggcccc	ggaggg
gcagcg	acggcc	gccggc
cagccg	cgggcg	cccgtt
gcaggt	gcgcgc	acggcc
tggacg	ggggcc	ccccct
atcatc	gggggc	ggcctg
cctgac	tgggcc	cagggg
ccccgc	ggcacc	gcggag
ctgacc	gcagcc	gaagtt
gccctt	cagccc	tccggg
cccctg	tcgccg	gggagg
ctgacg	tggggg	ggggcc
caccgc	gtgcag	gcctgg
ccgact	cccacc	ggcagc
tcccag	cgcgcg	ctgccc
tgacac	tgccgc	tgcctt
ccaccg	cgccgc	gggacc
ccctgg	cgcgga	gggcgg
	agcagc	ggctgg
Under-represented	ggcctc	gacagg
aaataa	agccca	
aataaa	gggccc	Under-represented
	ggtgcc	atthttg
	cccctg	attaat
	aggagg	
	ggctcg	
	gccccg	
	cagcag	
	gccgct	
	Under-represented	
	tcaatg	
	ctcatt	

Table A.2: Heart-specific TSS motifs

Upstream intron	Exon	Downstream intron
Over-represented	Over-represented	Over-represented
accggt	cccca	aaacag
gacgcc	cccagg	cgagta
ccggac	cagtgg	gacaca
gcgaca	ccccag	agatta
cattga	cggctg	gcgtat
gcggc	cccag	agtctt
atgcgg		agcgta
gcggcc	Under-represented	ggtagt
tatttt	ttgaa	tagctg
accaaa		tgcaag
cgcggc		gcggag
cccgtg		cggga
tacaga		aaaaac
ggcgac		cgccg
ccgga		tcccg
ttcacc		gttggt
acgcc		ttggtg
cccgt		
cggcg		Under-represented
gacca		ttcca
tgtcct		
atfff		
Under-represented		
accat		
cctcaa		

Table A.3: Testis-specific TSS motifs

Upstream intron	Exon	Downstream intron
Over-represented	Over-represented	Over-represented
atgta	cgtcat	ctccgc
aacatt	gccgcg	cgggat
atacac	gtctgc	atcgg
ccgcgg	caccg	gcctag
ttcgcc		tgcgcc
taaagc	Under-represented	ctttaa
aactat	caatt	atcggg
aggatt		gtgaga
ctagat		cctagt
tcacgc		tgagcg
tttgtc		gttcg
ttaggt		atcatt
aaact		gtcggg
tagatt		ttatcg
cgcggc		cggga
gcggc		agttac
cgttc		tagtta
Under-represented		Under-represented
cctgg		agaact
		gctga
		acaca

Table A.4: Muscle-specific TSS motifs

Upstream intron	Exon	Downstream intron
Over-represented	Over-represented	Over-represented
aatgac	tgtggg	cggttg
gatctg	cctccc	gtgtcg
caggtt	aggcac	tgccct
ggtagg	gccagc	gcgagg
cgctct	gagtcg	
ggggta		Under-represented
gcaaat		gtgtgt
tggtc		
acccat		
Under-represented		
gtgtg		
gtgtg		
gtgtgt		

Bibliography

- [1] Smith AD, Sumazin P, and Zhang MQ. Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc Natl Acad Sci USA*, 102(5):1560–5, February 2005.
- [2] Matlin AJ, Clark F, and Smith C. Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology*, 6:386–398, 2005.
- [3] Blencowe BJ. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci.*, 25(3):106–10, March 2000.
- [4] D. L. Black. *Cell*, 13:367–370, 2000.
- [5] D. L. Black. *Annu. Rev. Genet.*, 32:279–305, 2003.
- [6] Yeo GW, Van Nostrand E, Holste D, Poggio T, and Burge CB. Identification and analysis of alternative splicing events conserved in human and mouse. *Proc Natl Acad Sci USA*, 102(8):2850–5, February 2005.
- [7] Castle J, Garrett-Engle P, Armour CD, Duenwald SJ, Loerch PM, Meyer MR, Schadt EE, Stoughton R, Parrish ML, Shoemaker DD, and Johnson JM. Optimization of oligonucleotide arrays and rna amplification protocols for analysis of transcript structure and alternative splicing. *Genome Biology*, 4(10), September 2003.
- [8] Caceres JG and Kornblihtt AR. Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet.*, 18(4):186–93, April 2002.

- [9] Johnson JM, Castle J, Garrett-Engle P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, and Shoemaker DD. Genome-wide survey of human alternative pre-mrna splicing with exon junction microarrays. *Science*, 302(5653):2141–4, December 2003.
- [10] Jensen KB, Dredge BK, Stefani G, Zhong R, Buckanovich RJ, Okano HJ, Yang YY, and Darnell RB. Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron*, 25(2):359–71, February 2000.
- [11] Kreiman. Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. *Nucleic Acids Research*, 32(9):2889–2900, May 2004.
- [12] Cartegni L, Wang J, Zhu Z, Zhang MQ, and Krainer AR. ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res.*, 31(13):3568–71, July 2003.
- [13] Garcia-Blanco MA, Baraniak AP, and Lasda EL. Alternative splicing in disease and therapy. *Nature Biotechnol.*, 22(5):535–46, May 2004.
- [14] Faustino NA and Cooper TA. Pre-mrna splicing and human disease. *Genes Dev.*, 17(4):419–37, February 2003.
- [15] Faustino NA and Cooper TA. Identification of putative new splicing targets for *etr-3* using sequences identified by systematic evolution of ligands by exponential enrichment. *Molecular and Cellular Biology*, February 2005.
- [16] Buckanovich RJ, Posner JB, and Darnell RB. Nova, the paraneoplastic ri antigen, is homologous to an rna-binding protein and is specifically expressed in the developing motor system. *Neuron*, 11(4):657–72, October 1993.
- [17] Minovitsky S, Gee SL, Schokrpur S, Dubchak I, and Conboy JG. The splicing regulatory element, *ugcaug*, is phylogenetically and spatially conserved in introns that flank tissue-specific alternative exons. *Nucleic Acids Res.*, 33(2):714–24, February 2005.

- [18] Fairbrother WG, Yeh RF, Sharp PA, and Burge CB. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297(5583):1007–13, August 2002.
- [19] Zhang XH, Kangsamaksin T, Chao MS, Banerjee JK, and Chasin LA. Exon inclusion is dependent on predictable exonic splicing enhancers. *Mol Cell Biol.*, 25(16):7323–32, August 2005.
- [20] Gene W. Yeo. *Identification, Improved Modeling and Integration of Signals to Predict Constitutive and Alternative Splicing*. PhD thesis, Massachusetts Institute of Technology, 2004.
- [21] Sample pooling for microarray analysis: A statistical assessment of risks and biases. Affymetrix.
- [22] Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, and Burge CB. Systematic identification and analysis of exonic splicing silencers. *Cell*, 119(6):831–45, December 2004.