Discovering candidate intron motifs regulating tissue-specific alternative splicing patterns in exons as identified by exon junction microarrays

Elaine Gee and Mike Lin
MIT 6.895 (Manolis Kellis, Piotr Indyk, Pouya Kheradpour)

Abstract

Alternative splicing is an important mechanism by which eukaryotic species create abundant proteomic variation from a given set of protein coding genes. More than 74% of human multi-exon genes are estimated to be subject to alternative splicing, in most cases in characteristic tissue-specific patterns (Johnson et al., 2003). While the cellular machinery that splices introns out of pre-mRNA has been thoroughly studied, the regulatory mechanisms that govern its action in specific tissues are not yet well understood. To apply these methods for the identification of intronic splicing regulatory motifs by identifying sets of “co-regulated” exons, i.e. those with similar splicing patterns in different tissues. We hypothesize that different alternatively spliced transcripts exhibiting correlated expression levels in specific exons may share similar regulatory mechanisms that control their pattern of alternative splicing. These regulatory mechanisms recognize their exons through sites located in adjacent, upstream introns. From analysis of the Johnson et al. (2003) dataset of tissue-specific alternative splicing events, we clustered these exons from over 10,000 genes in 54 different tissues to organize them into clusters of similar expression patterns using complete link clustering. By applying Gibbs sampling techniques by using in the form of BioProspector (Liu et al., 2001), we identified statistically significant motif candidates within the intron sequences preceding the excised exon pattern. These intronic motifs exhibit strand bias by occurring with a higher frequency on the forward strand of the DNA that encodes the mRNA transcript.
**Introduction**

Alternative RNA splicing is a mechanism that adds diversity to the proteome for a given set of genes. Genes at the pre-RNA level contain protein coding and non-protein coding regions called exons and introns, respectively. In order to generate a functional mRNA transcript, constitutive RNA splicing must occur to remove the interspersed introns nestled between the exons. However, it is possible to also remove targeted exon sequences through methods of alternative splicing, which may occur in one of four ways: deletion of an optional exon, inclusion of an optional intron, sorting of mutually exclusive exons, and splicing at interior intronic splice sites. Based on the total number of exons in the pre-RNA transcript, the possible combinatorial patterns expand the diversity of the proteins isoforms to be translated from the post-transcriptional processing of the pre-RNA (Figure 1).

**Figure 1.** (a) Constitutive splicing splices out introns from the pre-RNA while (b) alternative splicing also removes targeted exons. Images courtesy of Alberts et al. (2002).

Alternative splicing is time dependent, as seen by its selective action during organismal development. For example, the sex of the fruit fly is determined by the alternative splicing pattern of the tra gene, which encodes the Tra regulatory protein. Tra is involved in alternatively splicing the doublesex gene, and depending on the pattern of alternative splicing the fruit fly may be male or female. Alternative splicing is also a tissue specific process to facilitate cell differentiation (Alberts et al., 2002). The alpha-tropomyosin gene in rats regulates muscle contraction, and we see that the gene is alternatively spliced into various RNA transcripts (Figure 2) to give protein isoforms that are different across the different tissues types such as striated versus smooth muscle tissue (Alberts et al, 2002). It is estimated that almost three-quarters of human genes are involved in alternative splicing (Johnson et al., 2003).

**Figure 2.** The alpha-tropomyosin gene in the rat regulates muscle contraction and undergoes different patterns of alternative splicing in a tissue-specific manner. Images courtesy of Alberts et al. (2002).
Within the human genome, splicing motifs at the pre-RNA level are used to recruit the spliceosome machinery to bind and perform the two transesterification reactions required to excise the intron sequence in the form of a lariat to be targeted for degradation in the nucleus. Three types of sequences are required for the recognition of the targeted intron by the spliceosome machinery: the 5’ splice site (AG(GURAGU), branch site (GTRAYY), and polypyrimidine tract (YYYYYYYYNCAG|G) (Figure 3) (the vertical line indicates the intron-exon junction). Together, the branch site and the polypyrimidine tract are known as the 3’ splice site. These consensus motifs are highly variable, which makes their discovery difficult on the computational level, but there are a few consistent characteristics. Exon1 usually ends in AG, the branch point is an A (which is where the lariat forms), and Exon2 begins with a G to give the final spliced product an intersection of AGG. Based on the existence of motifs for constitutive splicing, we are interested in looking for motifs that are specific to alternative splicing that generates similar splicing patterns across different human tissues. We want to analyze intron sequences from genes with similar alternative splicing patterns across different human tissues based on the RNA exon junction microarray data available from previous studies by Johnson et al. (2003).

The justification for looking for intronic motifs that may govern patterns of alternative splicing is based on the existence of positive and negative regulatory factors that can bind to intronic sequences that help to stabilize or repel the spliceosome from a nearby exon. Alternative splicing regulatory motifs may be analogous to promoter motifs that influence transcription initiation by recruiting transcription factors. Transcriptional regulatory motifs have traditionally been identified by applying methods such as Gibbs sampling to the promoters of co-regulated genes, and we look to perform a similar analysis on intron sequences to look for alternative splicing regulatory motifs. Here we operate under the assumption that the motif lies only within the intron sequences that precedes the first missing exon in a given splicing pattern, so we will only examine one intron sequence per gene. A database of known transcript isoforms, such as the Alternative Splicing Database (http://www.ebi.ac.uk/asd/altsplice/index.html), is not sufficient for this purpose of intronic motif discovery, since detailed information about the tissue-specificity of each alternative transcript is required.

*Exon junction microarrays for identification of tissue-specific alternative splicing events*
Recently, gene expression microarray technology has been extended to probe the presence or absence of individual exons in whole transcripts. For example, exon junction arrays, described in Johnson et al., place a probe targeting the junction between each pair of adjacent exons in a transcript. Adjacent exon junction probes indicate that the flanked exon is absent in the sample when their intensity is low in relation to other probes targeting the same processed RNA transcript (Figure 4).

![Figure 4. Conceptual schematic of an exon junction microarray experiment. The low intensity of probes targeting the junction of exons 2 and 3 and exons 3 and 4, in relation to the other probes targeting this transcript, indicate that exon 3 is removed through splicing.](image)

From exon junction arrays targeting some 100,000 exon junctions in more than 10,000 human genes, the Johnson et al. (2003) study identified alternative splicing events in mRNA samples from 54 different human tissues. Thus, a consistent, near-genome-wide database of tissue-specific alternative splicing events is now available.

**Methods**

Our strategy to systematically identify candidate intronic alternative splicing regulatory motifs is as follows. First, we will process the Johnson et al. (2003) microarray dataset to produce an expression vector across different tissues for each exon measured. We will then identify sets of “coregulated” exons by performing hierarchical clustering on these expression vectors. For each set of coregulated exons, we will extract the intronic sequence upstream of the first exon in the set and use existing motif finding tools such as BioProspector to identify candidate motifs. Finally, we will assess the significance of the motifs we identify through several different statistically- and biologically-motivated tests.

**Microarray data processing**

The full details of the microarray data processing strategy are reported in Johnson et al (2003). Briefly, linear models are fit to each probe based on its intensity across all tissue samples and of all probes targeting the same transcript, to predict the probe intensity in each tissue, under the assumption that the targeted exons are included in the transcript in that tissue. Each exon junction probe in each tissue is then assigned an integer value between 0 and 3 indicating how far below the predicted intensity the actual observed intensity was; for example, a value of 0 means that the observed intensity was mostly explained by the model, suggesting that both exons targeted by the probe are present in the sample. A value of 3 means that the observed intensity is far below expectation, suggesting that one or both exons targeted by the probe are not present. There may also be missing values; for example, if an entire transcript did not seem to be expressed in a certain tissue, then the probes targeting it will be assigned missing values.

**Generation of exon expression vectors**
We chose to represent the alternative splicing pattern of each exon as a binary vector with an entry for each of the 54 tissues containing 1 if the exon is expressed in that tissue, 0 if it is spliced out, or the missing value (“NA”) if there is not sufficient evidence to decide. For each exon targeted by the microarray, we must fill in each value in this vector based on the values of the two exon junction probes flanking the exon in the corresponding tissue. We assigned the entry to 0 if the value of one of the junction probes was 3, indicating the strongest prediction of an alternative splicing event, and the value of both junction probes was at least 1. We assigned the entry to NA if either probe value was missing and 1 otherwise.

These rules were chosen somewhat arbitrarily, and future research should investigate the most meaningful ways to represent the alternative splicing pattern of each exon based on the microarray data.

*Clustering expression vectors to identify coregulated exons*

We next applied hierarchical clustering to the exon expression vectors in order to identify sets of exons that appear to be coregulated across different tissues. Hamming distance is a natural choice of distance metric for binary vectors; however, we required a metric that is able to deal with missing values. We used the following distance function on exon expression vectors, intended to lessen the importance of “mismatches” due to missing values where vectors \( u \) and \( v \) are the expression vectors to be compared:

\[
d(u, v) = \sum_i \begin{cases} 
0 & \text{if } u_i \text{ or } v_i \text{ absent} \\
+1 & \text{if } u_i \neq v_i \\
-1 & \text{if } u_i = v_i 
\end{cases}
\]

Using the complete link method and this distance metric to our expression vectors, we defined clusters by choosing a “cut height” for the tree (see Results). Finally, we performed two post-processing steps on the clusters: first, we filtered out clusters containing only a small number of exons; second, in cases where multiple adjacent exons appeared in a cluster (e.g. the 3rd and 4th exon of a certain transcript), we discarded all but the first.

*Intron sequence extraction*

Having identified sets of coregulated exons, we next extracted the sequence of the intron upstream of each exon. This is conceptually straightforward, but proved to be a significant bioinformatics challenge. From the Johnson et al. (2003) dataset, we have for each exon a gene name, a RefSeq ID, an exon number, and the sequence of the flanking probes. We downloaded the full transcript mRNA sequence corresponding to each RefSeq ID in our clusters and aligned them to the human genome using BLAT (Kent 2002). For each exon and the corresponding alignment for its transcript, we searched for a pair of adjacent alignment blocks with ends matching the corresponding sequence in the 5’ exon junction probe. We then recorded the sequence occurring in between the alignment blocks as the desired intron. This approach ensured that we extracted the correct sequence whenever possible, without having to rely on existing databases of alignments for these (often outdated) RefSeq transcripts.

We extracted at most 10,000 bases upstream of each exon, in order to limit the search space and data set size. We will later discuss the potential implications of having chosen this cutoff.
Motif Finding

A probabilistic approach towards motif finding includes the method of Gibbs sampling that calculates possible motifs based upon their frequency of occurrence within aligned sequences. Given a sample of n strings $S_1, S_2, ..., S_c, ... , S_n$ that contain a common motif pattern of length $m$, we create a list of one substring of length $m$ from each of the strings whose starting position within a given string is randomly chosen. Then at random we choose one string $S_c$ to exclude, and calculate a motif matrix based on the $n-1$ motifs available for the other $S_1, S_2, ..., S_{c-1}, S_{c+1}, ..., S_n$ strings. Based on this motif matrix, we then score all possible substrings of length $m$ from the string $S_c$ to calculation the new starting position within $S_c$ based on a probability that is proportional to that position’s score. By iterating through these steps, we can continue until we reach convergence.

For the intron motif search, we implement BioProspector (Liu et al., 2001), which uses the Gibbs sampling algorithm implemented with a 3rd order Markov model. Under the assumption that the presence of a nucleotide base within the intron sequences is dependent on a local neighborhood of size three (for example, to model the size of codons as length three), we can calculate the probability of a sequences based on the three bases preceding it. For example, to calculate the probability of having a sequence AUGUA, we have the following:

$P(AUGUA) = P(A) * P(U | \text{ previous base is } A) * P(G | \text{ previous 2 bases are AU}) * P(U | \text{ previous 3 bases are AUG}) * P(A | \text{ previous 3 bases are UGU})$

The BioProspector will return the top ten scoring motifs from each cluster. The motif score is based on a relative entropy calculation:

$$\text{Motif Score} = \sum_{w} \sum_{all \ positions \ j} q_{ij} \log \left( \frac{q_{ij}}{p_{ij}} \right)$$

where $j$ denotes the nucleotide, $\Theta$ is the motif matrix, and $B$ is the background probability model. We found that as motif width increases, the motif score decreased linearly (Figure 5).

![Motif Score as a Function of Motif Width](image)

Figure 5. The general trend for the motif score as a function of motif width decreases linearly for the twelve clusters analyzed.
Analysis: Statistical Significance

In order to evaluate the significance of these generated motifs, we are interested in how these motifs score against motifs found in a cluster of randomly generated sequences similar in the number, length, and nucleotide distribution to the input cluster sequences. For our studies, we generated 25 Monte Carlo simulations of these random sequences and scored them. From this null distribution, we can calculate the mean $\mu$ and standard deviation $\sigma$. To calculate the p-values for the motif scores of the sequences, we compare the sample to the extreme value distribution categorized by the parameters $a$ and $b$ (Figure 6), with the extreme value as the

$$P(x) = \frac{1}{b} e^{-\frac{(x-c)}{b}}$$

$$CDF(x) = e^{-\frac{(x-a)}{b}}$$

$$\mu = a + b \gamma$$

$$\sigma = \frac{1}{\pi^2 b^2}$$

$$\gamma = \text{Euler-Mascheroni constant} = \lim_{n \to \infty} \left( \sum_{k=1}^{n} \frac{1}{k} - \ln n \right) \approx 0.577...$$

Figure 6. The extreme value distribution depends on parameters $a$ and $b$.

the maximum scoring motifs. We can integrate the probability density function $P(x)d(x)$ to get the cumulative distribution function as a function of the motif score $x$.

Analysis: Strand Bias

We pursued a second avenue through which to assess the potential significance of our motifs based on testing for enrichment of the motif on the sense strand only. We presume that alternative splicing regulatory motifs are biologically functional (for example, as binding sites for splicing factor proteins) post-transcriptionally in the pre-mRNA, rather than in the genomic DNA. If this is the case, then it is likely that whatever recognizes these sequences can only recognize those sequences, and not their reverse complement. This is in contrast to genomic transcription factor binding sites, where the double-stranded nature of DNA implies that if the reverse complement of a motif occurs on one strand, then the motif itself occurs on the other strand.

Thus, (some or most) post-transcriptional factor-binding motifs should be enriched only in one direction in the sequence. In contrast, if a certain motif was just occurring by chance, then one would expect it to occur forwards and backwards on both strands of the DNA with approximately equal frequency. We formulated a test for strand bias as follows. Given a total number of motif occurrences, we compute the expected number of occurrences on the forward
and reverse strands under some background model. We then use a chi-square test to compare the actual number of occurrences on the forward and reverse strand to the expected value. This assigns a p-value to each motif that indicates the confidence that it is biased to occur only on one strand.

We used a simple GC content background model, implying that there is always an expected even split between the forward and reverse strands. Future research should enrich this background model to capture more complex properties of the intron sequences, against which we wish to find motifs.

**Results**

*Clustering Results*

We identified 5,038 exon expression vectors that indicated alternative splicing in at least 5 of the 54 tissues, a cutoff we chose to ensure high clustering specificity. When we performed hierarchical clustering on these vectors, we observed the presence of seemingly unusually well-defined clusters in the hierarchical clustering tree (see Figure 7), and chose a low cutoff height intended to target these clusters. This resulted in 12 clusters containing at least 15 exons. An example is shown in Figure 8. We successfully extracted the intronic sequence upstream of each of the exons in these clusters, except for a very small number of cases in which apparently a low-quality transcript sequence could not be confidently aligned to the genome.

![Figure 7. Hierarchical clustering tree of ~5,000 exon expression vectors. Note the appearance of localized structures towards the right-hand side of the magnified view. To define clusters, we chose a tree cut height intended to target these structures.](image-url)
Figure 8. Example of a cluster of coregulated exons. Each row in the matrix represents an exon, and each column represents a tissue; each entry is colored green if the expression vector indicates the exon is included in that tissue, and red if it is spliced out.

Motif Results

We recorded the top ten highest scoring motifs for twelve clusters for motif widths of 6, 8, 10 bases. Based on 120 total scores at each cluster, we found that for motif width of length 6, only 68% of the motifs scored above random, but for longer lengths above size 8, at least 98% of the scores were statistically significant at the p = 0.05 cut off. We also applied our strand bias test and found that, with higher motif widths, an increasing proportion of the motifs found exhibited significant strand bias (Table 1).

<table>
<thead>
<tr>
<th>Motif Width</th>
<th>Proportion Scoring Above Random</th>
<th>Proportion Scoring with a p-value &lt; .05</th>
<th>Proportion with Strand Bias p-value &lt; .05</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.68</td>
<td>0.31</td>
<td>0.19</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0.98</td>
<td>0.28</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>0.38</td>
</tr>
</tbody>
</table>

Table 1. Out of 120 scores, this is the distribution of scores for 3 different motif widths scoring above what is randomly expected, and also the fraction of those with statistically significant motif scores and strand bias.

Therefore, we will only consider 8-mer and 10-mer motifs in our subsequent analysis. Within this sample subset, we counted approximately one-third of these motifs to exhibit statistically significant strand bias at the p = 0.05 level. However, due to repeated motif results for a given cluster based on multiple BioProspector results, this number is an overestimate of the true distribution of strand biased motifs. From our collection of top ten scoring motifs from the twelve clusters, three motif candidates have been shown below:
Figure 9. Three candidate motifs selected from the sample size of twelve clusters based on their statistical significance in score and strand bias. Within each introgram, the introns for one cluster are represented as black horizontal lines whose lengths scale with the number of bases in the intron. The red boxes distributed on top of the lines represent motif occurrences on that intron. Note that the widths of the red boxes are not to scale. The third motif candidate contains the consensus sequence for the branch site as marked by the red box.

Note that the dotted boxed added to the sequence in the third candidate motif matches the branch site consensus sequence. For all the widths examined, the branch site consensus sequences was identified in a few clusters, but only the occurrences of the consensus sequence in the larger motifs of size 8 or 10 were statistically significant (Table 1):

<table>
<thead>
<tr>
<th>Motif Width</th>
<th>Number of occurrences of Consensus Motif</th>
<th>Number of Clusters Containing Consensus</th>
<th>Motif Score p-value &lt; .05?</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>No (Did not Score Above Random Sequences)</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>1</td>
<td>Yes</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>3</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 2. As the motif width increases, we can capture more occurrences of the consensus motif for the branch site.

Discussion

Our results offers initial glimpses into alternative splicing regulatory motifs involved in tissue-specific exon manipulation. By combining both results from exon junction microarrays and motif discovery algorithms, we have generated candidate motif sequences that several methods of analysis suggest may be biologically functional.

A weakness to our approach is that we have used many arbitrarily chosen cutoffs to define exon clusters and analyze motifs, and we did not systematically try different combinations of these cutoffs. Future analysis should explore different cutoffs for generating exon expression vectors, hierarchical clustering tree cut height, intron sequence length, and other various parameters. For example, transcriptional regulatory motifs typically occur within only a few hundred bases of the transcription start site (Xie et al. 2005), while we explored up to 10,000 bases of the introns. Also, varying the hierarchical tree cut height may lead to many more clusters than the 12 we analyzed with a fairly low cutoff.

One of the assumptions made in this analysis is that only one intron, the one preceding the first missing exon, is involved in coordinating a pattern of alternative splicing. It is plausible to imagine that the intron sequences that precede each spliced out exon may act in a coordinated fashion with each other, but we leave this for future analysis. Potentially, we can examine more than one intron sequence per gene to see if multiple signals are required for orchestrating
alternative splicing in a tissue (or time) specific manner. Also, it is known that sequence motifs within the exons themselves can have a splicing regulatory function (Fairbrother et al. 2002), so our approach might even be applied to exon sequences.

From our small selection of model motifs presented in our results, it was encouraging to find that the branch site consensus motif appeared despite the high variability in the consensus motif. Although not all the consensus motifs were found in all the clusters, BioProspector could identify this highly variable motifs in a select number of clusters. It is also possible that the candidate motifs that do not match the consensus motifs for constitutive splicing suggests that the motifs associated with alternative signalling may be stronger within the introns. The Gibbs Sampling approach depends highly on whether the input sequences all exhibit the given motif, and it may be possible that many more motifs were not identified in our motif search.

Future analysis may include comparison of the candidate motifs against randomly selected intron sequences from the human genome in order to distinguish alternative splicing regulatory motifs from other motifs that may happen within introns. Analysis of their conservation with other genomes may also help to validate them. Ultimately, however, experimental validation such as site-directed mutagenesis experiments would be necessary to prove the biological function of these motifs.

References


