Evidence for the regulation of alternative splicing via complementary DNA sequence repeats

Yun Lian and Harold R. Garner*

Departments of Biochemistry and Internal Medicine, McDermott Center for Human Growth and Development and Center for Biomedical Inventions, The University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390, USA

Received on July 13, 2004; revised on November 8, 2004; accepted on November 23, 2004

ABSTRACT
Motivation: While the mechanism for regulating alternative splicing is poorly understood, secondary structure has been shown to be integral to this process. Due to their propensity for forming complementary hairpin loops and their elevated mutation rates, tandem repeated sequences have the potential to influence splicing regulation.

Results: An analysis of human intronic sequences reveals a strong correlation between alternative splicing and the prevalence of mono- through hexanucleotide tandem repeats that may engage in complementary pairing in introns that flank alternatively spliced exons. While only 44% of the 18 173 genes in the Human Alternative Splicing Database are known to be alternatively spliced, they contain 84% of the 694 237 intronic complementary repeat pairs. Significantly, the normalized frequency and distribution of repeat sequences, independent of their potential for pairing, are indistinguishable between alternatively spliced and non-alternatively spliced genes. Thus, the increased prevalence of repeats with pairing potential in alternatively spliced genes is not merely a consequence of more repeats or repeat composition bias. These results suggest that complementary repeats may play a role in the regulation of alternative splicing.

Contact: harold.garner@utsouthwestern.edu

INTRODUCTION
Alternative splicing of pre-mRNA is a key mechanism for achieving genetic diversity in higher eukaryotes (Black, 2003). Alternative splicing generates functionally distinct spliceforms that are critical to many cellular and developmental processes such as tissue differentiation (Hou and Conboy, 2001), apoptosis (Black, 2003) and immune and nervous responses (Modrek and Lee, 2002). Errors in the regulation of alternative splicing can result in truncated or unstable proteins, some of which are responsible for human diseases such as prostate cancer (Carstens et al., 1997) and schizophrenia (Huntsman et al., 1998). Because recent analyses of expressed sequence tags (EST) suggest that alternative splicing occurs in 30–60% of human genes (Black, 2003; Modrek and Lee, 2002; Mironov et al., 1999; Kan et al., 2001), an understanding of the regulatory mechanisms of RNA splicing is fundamentally important to almost all biological issues.

Many patterns of alternative splicing have been reported, but the regulatory mechanisms of alternative splicing that modulate these patterns are still poorly understood (Black, 2003; Maniatis, 1991; Nasim et al., 2002). In recent years, the discoveries of many regulatory elements within introns have increased the interest in their role in regulating alternative splicing (Cartegni et al., 2003; Miriami et al., 2003). Repeat sequences, especially complementary repeats, possess sequence motifs and symmetry elements that potentially allow the formation of secondary structures on single strand pre-mRNAs, such as hairpins and triplexes (Mitas, 1997). These repetitive elements are frequently polymorphic, and their potential for polymorphism can be predicted (Fondon et al., 1998; Wren et al., 2000). Because it has been suggested that the formation of secondary structure can participate in the regulation of splice site selection (Black, 2003; Muro et al., 1999; Tu et al., 2000), the potential role of complementary simple sequence repeats (mono- through hexanucleotide repeat units) in the regulation of alternative splicing was investigated. The results showed a correlation between alternative splicing and frequency of DNA complementary repeat pairs in intronic sequences that flank alternative spliced sites.

IMPLEMENTATION
Data source
The Human Alternative Splicing Database (http://www.bioinformatics.ucla.edu/HASDB) used in this study currently provides the most comprehensive information for alternative splicing in the human genome (Modrek et al., 2001; Lee et al., 2003). This database is based on genome-wide analyses of alternative splicing in humans, and it includes 30 793 alternative splicing relationships derived from alignment of UniGene clusters of expressed sequences to the draft human genome sequence (Lee et al., 2003). This release of the database contains significantly more data than its 2000 edition; for example there are 3.5 times as many alternatively spliced genes noted now. Our analysis is based on the 2002 database and the state of knowledge then, so it should be noted that our analysis therefore reflects that state and the inherent limitations of a rapidly changing dataset. The Human Alternative Splicing Database and the Mouse Alternative Splicing Database, which were released in January 2002, were downloaded from UCLA-ASAP (the Alternative Splicing Annotation Project) website and imported into an Oracle database. Each mapped UniGene cluster of genomic sequence was split into an intron cluster and an exon cluster on the basis of the intron and exon tables from the database. A summary of the analyzed human data set is shown in Table 1.
Regulation of alternative splicing via complementary DNA sequence repeats

Table 1. Summary of analyzed dataset downloaded from Human Alternative Splicing Database

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>AS</th>
<th>NAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starting number of UniGene clusters</td>
<td>96,109</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mapped to draft genome</td>
<td>68,177</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mapped gene sequences (clusters)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>18,173</td>
<td>7,991</td>
<td>10,182</td>
</tr>
<tr>
<td>Mapped introns</td>
<td>133,369</td>
<td>93,446</td>
<td>39,923</td>
</tr>
<tr>
<td>Mapped exons</td>
<td>96,589</td>
<td>68,189</td>
<td>28,400</td>
</tr>
<tr>
<td>Alternative splicing relationships&lt;sup&gt;b&lt;/sup&gt;</td>
<td>30,793</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Tabulation of the number of introns or exons and number of distinct UniGene clusters in which they were observed from the total dataset (Total), and clusters in alternatively spliced genes (AS) and non-alternatively spliced genes (NAS).

<sup>b</sup>Alternative splicing relationship includes alternative 3′, alternative 5′, exon skipping and mutually exclusive exons, but not other events such as intron retention.

RESULTS AND DISCUSSION

Limited information exists regarding the relationship between the secondary structures of pre-mRNA and alternative splicing. This study has shown that complementary repeat pairs that flank alternative splicing sites potentially allow secondary structure formation that may play a role in the regulation of alternative splicing. For example, in the ACOX1 gene (UniGene Cluster I.D. HS.100099), the second exon is alternatively spliced (exon skipped), and the flanking introns contain five complementary repeat pairs (for example, A23...6175nt...exon 2...1261nt...T22, A23...6175nt...exon 2...1556nt...T29).

Complementary repeat pairs are unevenly distributed

To investigate the potential role of complementary repeats in alternative splicing, the distributions of complementary repeat pairs in alternatively and non-alternatively spliced genes were compared. For alternatively spliced genes, the analysis was further sub-divided into introns containing complementary repeats that either flank or do not flank known alternative splice sites. To do this, the human gene dataset in the ASAP database, which contains 18,173 genes (clusters), was divided into two subsets: alternatively spliced genes and non-alternative splicing genes (Table 1). The alternatively spliced gene subset contained 7991 genes, each containing at least one type of alternative splicing (alternative 5′, alternative 3′, exon skipping, or mutually exclusive exons). The non-alternatively spliced gene subset contained 10,182 genes currently designated as not alternatively spliced. Alternatively spliced genes were found to have a higher incidence of complementary repeat pairs than non-alternatively spliced genes. A total of 585,179 complementary repeat pairs were found in the introns of alternatively spliced genes (84% of total detected complementary repeat pairs) in comparison to 109,058 (16%) in the introns of non-alternatively spliced genes (Table 2). Conversely, 53% of alternatively spliced genes (4255 genes) were identified as containing intronic complementary repeat pairs. In contrast, only 21% of non-alternatively spliced genes (2109 genes) contained intronic complementary repeat pairs. The top five repeat units that formed these complementary repeats (in descending order: U, A, UA, UA and CA) make up 95% of the complementary repeat pairs in both the alternatively spliced and non-alternatively spliced genes.

Because alternative spliced genes are usually larger and contain more introns and exons than non-alternatively spliced genes (Zhuang et al., 2003), distributions of complementary repeat pairs were also normalized by number of introns and by the size of the introns (in nucleotides) for both subsets (Table 2). The difference in frequency of intronic complementary repeat pairs in alternatively spliced genes and non-alternatively spliced genes was still significant after the normalization: 73 complementary repeat pairs per alternatively spliced gene compared to an average of 11 such repeat pairs in non-alternatively spliced genes; 6.3 complementary repeat pairs per intron in alternatively spliced genes in contrast to 3 complementary repeat pairs per intron in non-alternatively spliced genes.

To illustrate that the average number of complementary repeat pairs was not dominated by any particular portion of the gene size distribution as measured by its total intron length, a histogram of the ratio of numbers of complementary repeat pairs for the two groups of genes was computed (Fig. 1A). The genes were divided into 500 nucleotide bins and, within each bin, an equal number of

---

Identification of interspersed repeats in the human genome

Simple repeat elements and low complexity DNA sequences were identified from the genomic sequences with RepeatMasker using the default settings for repeat purity minima (Smit, 1999, http://ftp.genome.washington.edu/RM/RM_details.html). By default, only mono- to pentameric including some hexameric repeat units were scanned by RepeatMasker and any contiguous sequence elements shorter than 20 nucleotides constructed from these simple repeat units were ignored. All identified repetitive elements, their genomic locations and sequence purity were imported into the Oracle database in preparation for complementary repeat analysis.

Identification of complementary repeat pairs

We define complementary repeat pairs as those simple sequence repeats found by RepeatMasker on the same strand of a gene’s pre-mRNA molecule which can loop back to form stable Watson–Crick basepairs over an extent of 20 basepairs or more. To identify the complementary repeat pairs among all introns and exons, a parser and analysis system was developed using PERL and PL/SQL languages. Of the 16 possible base-pairings, only six (AU, GU, GC, UA, UG and CG) make up 95% of the complementary repeat units that formed these complementary repeats (in descending order: UA, UG, UA and CA). Approximately 0.8% of all complementary repeats had greater than a 30% mismatch in finding the complements, but it should be noted that only 0.8% of all complementary repeats had greater than a 30% mismatch. All results were imported into the Oracle database for further analysis. The differences between the distribution of complementary repeat pairs in alternatively spliced genes and the distribution in non-alternatively spliced genes were analyzed using the Fisher Exact test.
both types of genes were compared. The number of genes compared was determined by the total number of the two gene types in the bin, equaling the smaller of the total in each bin. A random selection equaling this number was made from the gene type with the larger total. As the figure displays, the number of intronic complementary repeat pairs in alternatively spliced genes was 2.6 times higher than in non-alternatively spliced genes on average (see also Table 2). Statistical analysis (Wilcoxon test) showed a significant difference in the distribution of complementary repeat pairs between alternatively and non-alternatively spliced genes (P < 0.00000001).

A Fisher Exact test was used to examine the statistical significance of the difference in the distribution of complementary repeat pairs between alternatively spliced genes and non-alternatively spliced genes. The results showed that there are significant differences in the distributions of ‘intron–intron’ complementary repeat pairs (P < 0.00000001) and ‘intron–exon’ complementary repeat pairs (P = 0.0036) between the two groups of genes. The significance of the difference in the distribution of ‘exon–exon’ complementary repeat pairs could not be determined due to sample size.

The statistical analysis revealed a strong correlation between the high frequency of intronic complementary repeat pairs and alternatively spliced genes. A mechanism by which complementary repeats may influence alternative splicing in a manner similar to hnRNP A1-mediated splicing repression (Black, 2003) is suggested in Figure 2. The complementary repeats could combine together and form hairpins or other secondary structures, resulting in ‘looping out’ of the exon and exon skipping. Several studies have shown the potential for secondary structure to influence splicing choices in cellular pre-mRNAs and some viral transcripts (Solnick and Lee, 1987; Vamvakopoulos et al., 2002). The secondary structures of pre-mRNA formed by the complementary repeats could also interfere with the binding of spliceosome components or some regulatory factors, blocking the exon bridging interactions during exon definition, or enhancing splicing by decreasing physical distances between splice sites (Black, 2003; Jacquent et al., 2001). It should be noted that the presence of repeats that span intron–exon boundaries (220 in 211 genes) and the existence of complementary repeat pairs between repeats internal to exons and other exons or introns (Table 2) suggest that this same ‘looping out’ mechanism could potentially play a role in alternative 5′ and 3′ type splicing. We observed an anti-correlation between alternative splicing relationships and complementary repeat pairs found within the same intron (Table 2), indicating that either the shortening of an intron by splicing out does not play a major role in this process or there still remain many more alternative splicing events to be discovered for this class of introns.

In alternatively spliced genes, introns that contain complementary repeat pairs and also flank alternative splicing sites were designated as ‘companion introns’. Sequence elements in these introns are potentially more likely involved in regulating alternative splicing, thus the distribution of complementary repeat pairs among these introns within alternatively spliced genes were further analyzed (Table 3). Companion introns contained 17.2 complementary repeat pairs per intron (or alternative splice site) on average. Also, these complementary repeat pairs in companion introns spanned on average 41 bp of complementarity. Among these complementary repeat pairs, 78 665 (54%) spanned skipped exons (10 350 of the total 30 793 alternative splicing relationships), suggesting that they may be involved in exon skipping. The remaining 68 132 complementary repeat pairs are associated with the other forms of alternative splicing, including alternative 3′, alternative 5′ and mutually exclusive exons (the remaining 20 443 alternative splicing relationships). We further define ‘non-companion’ introns as those introns in alternatively spliced genes that do not flank a known alternative splice site. In alternatively spliced genes, 38 372 (29%) of introns were ‘non-companion’ and contained 11.4 complementary repeat pairs per intron on average. Those complementary repeat pairs had only 38 bp of complementarity, precisely the same level of complementarity as seen in complementary repeat pairs of non-alternatively spliced genes. Furthermore, the introns that contain complementary repeat pairs within non-alternatively spliced genes have only 6.9 complementary repeat pairs per intron.

Averaged over all the companion intron splice sites, the distance between the start of the exon and the upstream repeat of a complementary pair is 6750 nt and the distance between the end of the exon and the other flanking repeat of a complementary pair is 3528 nt. So,
the loops are on average over 10,000 nt in size, which is sufficiently long that disfavored entropic constraints are not a factor. Another significant difference is also noted in the average distance between complementary repeat pairs in all introns of alternatively spliced and non-alternatively spliced genes, and not just those flanking the splice site, where the averages are 59,174 and 81,416 nt, respectively. Therefore, alternatively spliced genes contain more complementary repeat pairs, and those complementary repeat pairs span longer regions of complementarily, are in closer proximity to one another and are predominantly distributed among introns flanking alternative splicing sites. These results strongly support our suggested mechanism model. Since there are alternatively spliced genes that do not contain any complementary repeat pairs, the molecular mechanism of alternative splicing must involve other regulatory factors.

Interestingly, since there are significantly more complementary repeat pairs in non-companion introns of known alternatively spliced genes than in the introns of non-alternatively spliced genes, it may suggest that only a fraction of the potential alternative splicing sites have been found within those known alternative spliced genes. In support, we searched the literature for a small set of ‘cancer’ genes which contain an abundance of complementary repeat pairs indicative of genes that should be alternatively spliced but were not indicated
Fig. 2. Models for mechanism of complementary repeat mediated alternative splicing. Exons are represented by black boxes if they are not alternatively spliced, dark gray if they are alternatively spliced, and introns are represented by lines. Companion introns are designated by brackets. Repeats are designated by small, light gray boxes, with complementary repeat pairs coupled by lines. (A) It is proposed that, in some tissues, specific proteins bind to the complementary repeat pairs and block the formation of a hairpin so that ‘normal’ intron splicing is induced. (B) Hairpin model: in other tissues, the complementary repeat binding proteins are absent or lacking, permitting ‘looping out’ of the exon and exon skipping.

Table 3. Distribution of complementary repeat pairs in introns that are or are not associated with alternatively spliced exons

<table>
<thead>
<tr>
<th>Types of gene</th>
<th>Types of intron</th>
<th>Number of introns with CR</th>
<th>Number of CR</th>
<th>CR number/intron</th>
<th>CR number/nt</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS</td>
<td>Companion introns</td>
<td>8530</td>
<td>146797</td>
<td>17.2</td>
<td>0.0008</td>
</tr>
<tr>
<td></td>
<td>Non-companion introns</td>
<td>38372</td>
<td>438382</td>
<td>11.4</td>
<td>0.0011</td>
</tr>
<tr>
<td>NAS</td>
<td>NAS introns</td>
<td>15708</td>
<td>109058</td>
<td>6.9</td>
<td>0.0005</td>
</tr>
</tbody>
</table>

Companion introns contain more complementary repeat pairs on average. CR: complementary repeat pairs; AS: alternatively spliced genes; NAS: non-alternatively spliced genes; Companion introns: the introns in genes known to be alternatively spliced that flank alternative splicing sites; Non-companion introns: the introns in genes known to be alternatively spliced that do not flank any alternative splicing sites; NAS introns: the introns in non-alternatively spliced genes; nt: nucleotides in introns.

as such in the January 2002 release of the ASAP database. We found that, indeed, 8 out of 11 have since been described as now alternatively spliced in humans (5) or in mice or rats (3).

Although most complementary repeat pairs were intron–intron complementary repeats, some were found to be intron–exon (85) or exon–exon (6) complementary repeats. Further analyses revealed that 72 of the intron–exon complementary repeat pairs were in alternatively spliced genes (85%) and that four of the exon–exon complementary repeat pairs were in alternatively spliced genes (66%).

Simple repeats are evenly distributed

It is possible that the high frequency of complementary repeat pairs in alternatively spliced genes is due to an elevated frequency of simple repeats (all repeats, whether in complementary pairs or not) in these genes. Therefore, a systematic analysis of the distributions of frequency and length of simple repeats in all genes was performed. A total of 117123 intronic simple repeats were found in alternatively spliced genes (7991 genes) and 55113 in non-alternatively spliced genes (10182 genes). The frequencies of different repeat lengths in alternative splicing genes were not significantly different from those in non-alternatively spliced genes (Fig. 3). Only 478 simple repeats were found in exons: 327 in alternatively spliced genes (68%) and 151 in non-alternatively spliced genes (34%). Even though many more simple repeats were found in alternatively spliced genes than in non-alternatively spliced genes, there is no significant difference in frequencies of simple repeats between the two subsets after normalization by the number of introns: 1.25 simple repeats per intron in alternatively spliced genes compared to 1.38 simple repeats per intron in non-alternatively spliced genes. After normalization by intron length for the two groups of genes, the frequencies of intronic simple repeats in the two groups of genes were almost the same (Fig. 1B). The result of Wilcoxon test showed
This study suggests the possibility of a mechanism of alternative splicing via complementary DNA sequence repeats. The characteristics of these complementary repeat pairs are also different. Many of the complementary repeat pairs that were selected for cannot be the result of a random process, but must over time have been selected for.

Intronic and exonic repeat unit length distributions in both types of genes were also compared. There was a significant difference between genes that were alternatively and non-alternatively spliced sized-matched genes with bin size = 500 bp ($P = 0.23$). Thus, the high frequency of complementary repeat pairs in alternatively spliced genes does not exhibit a direct relationship to the intron size or the frequency of simple repeats. Since the frequencies of simple repeats in alternatively and non-alternatively spliced genes are almost the same, the high frequency of complementary repeat pairs in alternatively spliced genes cannot be the result of a random process, but must over time have been selected for.

Intronic and exonic repeat unit length distributions in both types of genes were also compared. There was a significant difference between intronic and exonic repeat unit length distributions. For intronic repeat units, there is a very large population of dimer and multiple-dimer repeat unit lengths. There was a much higher proportion of trimer and multiple-trimer repeat unit lengths in exonic repeats (Fig. 4). Overall, both alternatively and non-alternatively spliced genes had very similar distributions of lengths of repeat units.

**CONCLUSION**

The analysis of human (and mouse, see Supplemental Information) tandem repeat sequences presented here has revealed a strong correlation between alternative splicing and high frequency of complementary repeat pairs in intron sequences in alternatively spliced genes relative to non-alternatively spliced genes. Many of the characteristics of these complementary repeat pairs are also different. This study suggests the possibility of a mechanism of alternative splicing mediated through complementary repeats via their elevated polymorphism potential (see Supplemental Information). Further biological experiments will allow detailed analysis of the effect of complementary repeats on specific alternatively spliced genes.

**ACKNOWLEDGEMENTS**

We thank Dr John Fondon III, Dr Elizabeth Mary (Lena) Flood and Dr Michael Huebschman for helpful comments on the manuscript. We also thank Dr Tracy Xu for technical help. This work was supported by NIH/NCI Grant R01-CA096901, the Hudson Foundation and the NIH/NHLBI Training Grant 1 ROI CA096901 (Y.L.).

**REFERENCES**


**Fig. 4.** The abundance and distribution signatures of repeats in exons and introns are very different. (A) Distributions of intronic repeats in AS and NAS genes. (B) Distributions of exonic repeats in AS and NAS genes.
Y.Lian and H.R.Garner

