Evolutionary Divergence of Exon Flanks: A Dissection of Mutability and Selection

Yi Xing, Qi Wang, Christopher Lee*
Molecular Biology Institute
Center for Genomics and Proteomics
Dept. of Chemistry & Biochemistry
University of California, Los Angeles
Los Angeles, CA 90095-1570

EMAIL: leec@mbi.ucla.edu
TEL: 310-825-7374
FAX: 310-267-0248

*Author to whom correspondence should be addressed.

Draft 13
April 30, 2006

Running title: Evolution of exon flanks
Keywords: evolution, exon flanks, alternative splicing, selection pressure, mutation rate
Abstract

The intronic sequences flanking exon-intron junctions (i.e. exon flanks) are important for splice site recognition and pre-mRNA splicing. Recent studies show a higher degree of sequence conservation at flanks of alternative exons, compared to flanks of constitutive exons. In this manuscript we performed a detailed analysis on the evolutionary divergence of exon flanks between human and chimpanzee, aiming to dissect the impact of mutability and selection on their evolution. Inside exon flanks, sites that might reside in ancestral CpG dinucleotides evolved significantly faster than sites outside of ancestral CpG dinucleotides. This result reflects a systematic variation of mutation rates (mutability) at exon flanks, depending on the local CpG contexts. Remarkably, we observed a significant reduction of the nucleotide substitution rate in flanks of alternatively spliced exons, independent of the site-by-site variation in mutability due to different CpG contexts. Our data provide concrete evidence for increased purifying selection at exon flanks associated with regulation of alternative splicing.
Introduction

The intronic sequences that flank exon-intron junctions play a critical role in exon recognition and pre-mRNA splicing (FAIRBROTHER and CHASIN 2000; SUN and CHASIN 2000; ZHANG et al. 2003). These intronic sequences, referred to as “exon flanks” (ZHANG et al. 2005), contain important splicing regulatory signals such as intronic splicing enhancers and silencers. Zhang and colleagues performed a genome-wide study on the splicing signals in 50nt exon flanks. They showed that exon flanks displayed a bimodal distribution of GC content, and high-GC and low-GC exon flanks might be associated with different mechanisms of exon recognition (ZHANG et al. 2005). Some important splicing regulatory motifs are phylogenetically conserved in exon flanks (KOL et al. 2005; MINOVITSKY et al. 2005). Mutations at exon flanks frequently cause disruption of normal splicing patterns, such as the recognition of a cryptic splice site or the skipping of an entire exon (COSTA et al. 2005; COUTINHO et al. 2005; ENG et al. 2004). Such splicing mutations are one of the major causes for various human diseases (GARCIA-BLANCO et al. 2004). Therefore, although non-coding, exon flanks experience negative selection during evolution due to constraints of splicing regulation.

Recent comparative genomic analyses indicate a reduced nucleotide substitution rate at flanks of alternatively spliced exons. Sorek and Ast analyzed the 100nt flanks of exons that were alternatively spliced in human and mouse transcripts. They observed a nearly 20% increase of percent sequence identity at 30 nt exon flanks of alternatively spliced exons, compared to flanks of constitutive exons (SOREK and AST 2003). The average length of conserved sequences was 103nt for upstream flanks of alternative exons, and 94nt for downstream flanks of alternative exons (SOREK and AST 2003). Other studies reported similar results (SUGNET et al. 2004; SUGNET et al. 2006; XING and LEE 2005). In fact, the increased sequence identity at flanks of alternative exons has been successfully used to predict alternative splicing from raw genomic sequences (PHILIPPS et al. 2004; SOREK et al. 2004b; YEO et al. 2005). It has been suggested that such a higher sequence conservation reflects stronger negative selection associated with regulation of alternative splicing (ITOH et al. 2004; SOREK and AST 2003).
However, there are other factors that affect the nucleotide substitution rate in a genomic region. One important factor is the mutation rate. It is well known that the mutation rate varies significantly across the entire genome, leading to mutational hot-spots and cold-spots (Chuang and Li 2004). Even within the same gene, there can be remarkable site-by-site variations of the mutation rate. For instance, CpG dinucleotides are highly mutable due to methylation effects (Li et al. 2002). Keightley et al. analyzed the substitution rate in rodent genomes. They found that the substitution rate was significantly higher in CpG-susceptible sites (i.e. sites preceded by C or followed by G, thus were likely to be parts of ancestral CpG dinucleotides) compared to non-CpG-susceptible sites (Gaffney and Keightley 2005; Keightley and Gaffney 2003). A similar observation was made through genome-wide human-chimpanzee comparisons (Ebersberger et al. 2002; Helmann et al. 2003), and more recently at distant intronic sites and 4-fold redundant synonymous sites using human-chimpanzee genome alignments (Kondrashov et al. 2005). These data indicate the heterogeneity of mutation rate even within a small region of the gene, depending on CpG contexts.

In this manuscript, we performed an in-depth analysis on the evolutionary divergence of exon flanks. We are interested in the following questions. First, how does the local CpG context affect substitution rates at flanks of constitutive and alternative exons? Second, what’s the major cause for the increased sequence conservation at flanks of alternative exons? Is it reduced mutability or increased selection or both? Since multiple substitutions at single sites can obscure the impact of ancestral CpG dinucleotides, it’s important to compare closely related species, where multiple substitutions are unlikely to occur (Ebersberger et al. 2002). The human and chimpanzee exon flanks are the most suitable for our comparison, because these two species diverged only about 5 million years ago (Hedges 2002). The goal of our analysis is to dissect the contribution of mutability and selection to the evolutionary rate of exon flanks. We divided sites within exon flanks into distinct types according to their local CpG contexts. Windows ranging from 50nt to 350nt were used in previous analyses of exon flanks (Kaufmann et al. 2004; Sorek and Ast 2003; Sugnet et al. 2004; Sugnet et al. 2006; Xing and Lee 2005; Yeo et al. 2005). Throughout this manuscript, following the consensus choice in
the current literature (SOREK and AST 2003; SUGNET et al. 2004; SUGNET et al. 2006), we defined our exon flanks as the 100nt intronic sequences immediately upstream (upstream exon flanks) or downstream (downstream exon flanks) of the exon-intron junctions. A different definition (e.g. 50nt exon flanks) produced very similar results (data not shown).

Methods and Materials

We identified alternatively spliced exons in the human and mouse genome, by aligning expressed sequences (mRNA/EST) to the genomic sequences (MODREK et al. 2001). We matched orthologous exons as previously described (MODREK and LEE 2003). We compiled a set of ancestral alternatively spliced exons between human and mouse, i.e. exons whose inclusion and skipping forms were observed in both human and mouse transcripts (XING and LEE 2005). The independent observation of alternative splicing in two species is widely adopted as a criterion for functional alternative splicing events (SOREK et al. 2004a; YEO et al. 2005). For such an exon, alternative splicing is likely to be present in the common ancestor of human and mouse, therefore affecting the rate of nucleotide substitutions throughout subsequent evolution. We also used the data of Pan and colleagues to compile an independent set of ancestral alternatively spliced exons (PAN et al. 2004). Pan and colleagues identified mouse alternative exons that were also alternatively spliced in human, using EST-genome alignments and microarray data. For each mouse exon, we mapped it to the human genome (hg17, http://hgdownload.cse.ucsc.edu/goldenPath/hg17/bigZips/) using BLAST (ALTSCHUL et al. 1997), requiring a match to the exon over its entire length and an expectation value of less than 1e-4. Lastly, we compiled a set of exons that were constitutively spliced in human and mouse.

For each ancestral alternative or constitutive exon, we calculated the evolutionary divergence of its flanks between human and chimpanzee. We extracted the alignment of its flanks between human and chimpanzee using the UCSC human-chimpanzee (http://hgdownload.cse.ucsc.edu/goldenPath/hg17/vsPanTro1/axtNet/) pairwise genome alignments (KENT et al. 2003; SCHWARTZ et al. 2003). We divided the nucleotide sites at exon flanks into four distinct types based on the local CpG context: sites preceded by a C
and followed by a G (postCpreG), sites followed by a G but not preceded by a C (preG), sites preceded by a C but not followed by a G (postC), and sites neither preceded by a C nor followed by a G (nonCpG) (Keightley and Gaffney 2003; Kondrashov et al. 2005). The first three types of sites were CpG-susceptible according to the definition of Keightley and Gaffney, i.e. they were likely to be parts of ancestral CpG dinucleotides (Keightley and Gaffney 2003). We required each site to be flanked by two conserved nucleotides in the alignment, and excluded sites that corresponded to insertions or deletions. For each nucleotide site, we checked the alignment to see whether it was conserved or substituted between human and chimpanzee. For each type of CpG sites, we counted the number of substituted sites and the total numbers of sites. The alignment data can be accessed at http://bioinfo.mbi.ucla.edu/yxing/cpg/.

Results and Discussion

We investigated the evolutionary divergence of exon flanks between human and chimpanzee. The complication of multiple substitutions at single sites is minimal because of the close relationship of human and chimpanzee (Ebersberger et al. 2002; Hedges 2002), allowing an accurate inference for the ancestral CpG dinucleotides. We extracted the alignments of 141 upstream flanks of ancestral alternatively spliced exons from the UCSC human-chimpanzee pairwise alignment. We also extracted the alignments of 40638 upstream flanks of constitutive exons. We divided all nucleotide sites into four distinct types, according to the local CpG contexts (see Methods and Materials).

We observed a systematic difference of nucleotide substitution rates among different types of sites. In upstream flanks of the constitutive exons, the nucleotide substitution rate averaged over all four types of sites was 0.0113. postCpreG sites had the highest rate of nucleotide substitution at 0.0172, followed by postC (0.0134) and preG (0.0124) sites. nonCpG sites had the lowest rate of nucleotide substitution at 0.0097, a nearly two fold reduction compared to postCpreG sites (see Table 1). The nucleotide composition at different types of sites also showed a systematic difference: postC sites had a very low frequency of G; preG sites had a very low frequency of C; while postCpreG sites were strongly depleted of both C and G. We observed a similar trend at upstream flanks of
ancestral alternatively spliced exons. The overall nucleotide substitution rate was 0.0075. The rate was the lowest (0.0063) for non-CpG-susceptible sites (nonCpG sites, compared to CpG-susceptible sites including postCpreG, postC and preG sites) (see Table 1). Analyses of downstream exon flanks yielded very similar results (see Table 2). Our results were consistent with human-chimpanzee analyses on distant intronic sites or 4-fold redundant synonymous sites in exons (Kondrashov et al. 2005). These data indicate increased mutability at CpG-susceptible sites compared to non-CpG-susceptible sites in exon flanks, regardless of constitutive splicing or alternative splicing.

Comparing upstream flanks of alternative and constitutive exons, we observed a significant reduction of nucleotide substitution rate associated with alternative splicing (see Table 1). The overall nucleotide substitution rate at upstream flanks was 0.0075 for alternative exons, a 33.6% reduction compared to constitutive exons (0.0113). Remarkably, all four types of sites experienced a reduced substitution rate at flanks of alternative exons (see Table 1), ranging from a reduction of 45.9% (at postCpreG sites) to 26.6% (at preG sites). We also compared the substitution rate at downstream flanks of alternative and constitutive exons, and observed a similar trend. The overall substitution rate had a 38.3% reduction at downstream flanks of alternative exons. All four types of CpG sites had reduced substitution rates (see Table 2). These data suggest that alternative splicing is associated with a universal increase in the strength of negative selection at exon flanks, independent of the site-by-site variation of mutation rates due to different CpG contexts (see Fig. 1, black bars vs white bars).

To rule out the possibility of artifacts due to potential sampling biases in our alternative splicing dataset, we analyzed an independent set of ancestral alternatively spliced exons. This dataset included 289 exons, which were identified using EST-genome alignments and microarray data (Pan et al. 2004). We observed the same trend in this dataset (see Fig. 1). All four types of CpG sites had a statistically significant reduction in their substitution rates, compared to constitutive exons (see Fig. 1, gray bars vs white bars). We also observed the same trend when we restricted our analysis to 50nt exon flanks. These data demonstrate that our reported finding is a confident result.
Our study extended previous comparative analyses of intronic sites flanking alternatively spliced exons (SOREK and AST 2003; SUGNET et al. 2004; XING and LEE 2005). Instead of analyzing the evolutionary divergence of the entire exon flanks, we divided all nucleotide sites at exon flanks into distinct types, according to their local CpG contexts. Such treatment allowed us to separate the contribution of mutability and selection, two most important factors that affect the divergence rate of exon flanks. Our study indicates significant site-by-site variations of mutability in the exon flanks, depending on the local CpG context. Such a phenomenon was observed in flanks of both alternative and constitutive exons. However, the site-by-site variation of mutability is not the major contributor to the increased conservation at flanks of alternative exons. The higher sequence conservation is not due to a reduced fraction of hyper-mutable sites (i.e. CpG-susceptible sites). In fact, at all types of CpG sites (with different propensities for mutation) inside flanks of alternative exons, we observed a significant reduction of the nucleotide substitution rate (see Fig. 1). This consistent pattern, independent of the variation in mutability due to different CpG contexts, strongly argues for an increased selection at flanks of alternative exons. Such a strong purifying selection might reflect RNA sequence motif (ITOH et al. 2004) or secondary structure (MEYER and MIKLOS 2005) requirements for proper controls of alternative splicing. Our study provides concrete evidence for widespread selection pressure in mammalian genomes that is associated with alternative splicing (BAEK and GREEN 2005; SOREK and AST 2003; XING and LEE 2005). Understanding selection pressure at exon flanks will shed light on the regulation of alternative splicing and how it evolves.
Acknowledgements

We thank Fyodor Kondrashov for key discussions that led to this analysis, and for sharing of one manuscript prior to its publication. We thank Sandy Pan and Benjamin Blencowe for sharing of mouse splicing microarray data; Xiang Zhang and Alex Alekseyenko for helpful comments and suggestions. This work was supported by NIH Grant U54-RR021813, a Teacher-Scholar award to C.J.L. from the Dreyfus Foundation, a DOE grant DE-FC02-02ER63421, and a Ph.D. dissertation fellowship to Y.X. from UCLA.
Literature Cited

ALTSCUL, S. F., T. L. MADDEN, A. A. SCHÄFFER, J. ZHANG, Z. ZHANG et al., 1997


HELLMANN, I., S. ZOLLNER, W. ENARD, I. EBERSBERGER, B. NICKEL et al., 2003
Selection on human genes as revealed by comparisons to chimpanzee cDNA.

ITOH, H., T. WASHIO and M. TOMITA, 2004 Computational comparative analyses of
alternative splicing regulation using full-length cDNA of various eukaryotes. Rna
10: 1005-1018.

KAUFMANN, D., O. KENNER, P. NURNBERG, W. VOGEL and B. BARTELT, 2004 In NF1,
CFTR, PER3, CARS and SYT7, alternatively included exons show higher
conservation of surrounding intron sequences than constitutive exons. Eur J Hum
Genet 12: 139-149.

KEIGHTLEY, P. D., and D. J. GAFFNEY, 2003 Functional constraints and frequency of
deleterious mutations in noncoding DNA of rodents. Proc Natl Acad Sci U S A
100: 13402-13406.

Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and

KOL, G., G. LEV-MAOR and G. AST, 2005 Human-mouse comparative analysis reveals
that branch-site plasticity contributes to splicing regulation. Hum Mol Genet 14:
1559-1568.

KONDRASHOV, F. A., A. Y. OGURTSOV and A. S. KONDRASHOV, 2005 Selection in favor
of nucleotides G and C diversifies evolution rates and levels of polymorphism at

650-656.

MEYER, I. M., and I. MIKLÓS, 2005 Statistical evidence for conserved, local secondary
structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. Nucleic
Acids Res 33: 6338-6348.

MINOVITSKY, S., S. L. GEE, S. SCHOKRPUR, I. DUBCHAK and J. G. CONBOY, 2005 The
splicing regulatory element, UGCAUG, is phylogenetically and spatially
conserved in introns that flank tissue-specific alternative exons. Nucleic Acids


YEO, G. W., E. VAN NOSTRAND, D. HOLSTE, T. POGGIO and C. B. BURGE, 2005
Identification and analysis of alternative splicing events conserved in human and

information for the splicing of human pre-mRNA identified by support vector

ZHANG, X. H., C. S. LESLIE and L. A. CHASIN, 2005 Dichotomous splicing signals in
Table 1: The nucleotide substitution rates at upstream exon flanks.

<table>
<thead>
<tr>
<th></th>
<th>Constitutive</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All Sites</td>
<td>postCpreG</td>
<td>PostC</td>
<td>preG</td>
<td>nonCpG</td>
</tr>
<tr>
<td># of sites</td>
<td>3825522</td>
<td>231695</td>
<td>701611</td>
<td>620871</td>
<td>2271345</td>
</tr>
<tr>
<td># of substituted sites</td>
<td>43207</td>
<td>3996</td>
<td>9412</td>
<td>7685</td>
<td>22114</td>
</tr>
<tr>
<td>Frequencies</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>22.9%</td>
<td>41.1%</td>
<td>21.6%</td>
<td>25.3%</td>
<td>20.8%</td>
</tr>
<tr>
<td>T</td>
<td>31.3%</td>
<td>44.7%</td>
<td>33.3%</td>
<td>32.5%</td>
<td>29.0%</td>
</tr>
<tr>
<td>C</td>
<td>24.2%</td>
<td>7.4%</td>
<td>40.4%</td>
<td>4.9%</td>
<td>26.1%</td>
</tr>
<tr>
<td>G</td>
<td>21.5%</td>
<td>6.7%</td>
<td>4.7%</td>
<td>37.3%</td>
<td>24.0%</td>
</tr>
<tr>
<td>Proportion of nucleotide difference</td>
<td>0.0113</td>
<td>0.0172</td>
<td>0.0134</td>
<td>0.0124</td>
<td>0.0097</td>
</tr>
</tbody>
</table>

|                  | Alternative  |        |        |        |        |
|                  | All Sites    | postCpreG | PostC  | preG   | nonCpG |
| # of sites       | 13412        | 749     | 2801   | 2078   | 7784   |
| # of substituted sites | 101          | 7       | 26     | 19     | 49     |
| Frequencies      |              |         |        |        |        |
| A                | 21.1%        | 36.8%   | 20.1%  | 24.8%  | 19.1%  |
| T                | 32.4%        | 44.3%   | 33.5%  | 34.0%  | 30.4%  |
| C                | 26.2%        | 9.6%    | 41.7%  | 6.1%   | 27.5%  |
| G                | 20.3%        | 9.2%    | 4.7%   | 35.1%  | 23.0%  |
| Proportion of nucleotide difference | 0.0075 | 0.0093 | 0.0093 | 0.0091 | 0.0063 |
Table 2: The nucleotide substitution rates at downstream exon flanks.

<table>
<thead>
<tr>
<th>Constitutive</th>
<th>All Sites</th>
<th>postCpreG</th>
<th>postC</th>
<th>preG</th>
<th>nonCpG</th>
</tr>
</thead>
<tbody>
<tr>
<td># of sites</td>
<td>3831232</td>
<td>222572</td>
<td>614689</td>
<td>759625</td>
<td>2234346</td>
</tr>
<tr>
<td># of substituted sites</td>
<td>44107</td>
<td>4501</td>
<td>8806</td>
<td>9185</td>
<td>21615</td>
</tr>
<tr>
<td>Frequencies</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>24.2%</td>
<td>39.5%</td>
<td>24.2%</td>
<td>26.4%</td>
<td>21.9%</td>
</tr>
<tr>
<td>T</td>
<td>28.3%</td>
<td>44.3%</td>
<td>30.0%</td>
<td>28.6%</td>
<td>26.1%</td>
</tr>
<tr>
<td>C</td>
<td>22.1%</td>
<td>7.6%</td>
<td>40.6%</td>
<td>4.6%</td>
<td>24.4%</td>
</tr>
<tr>
<td>G</td>
<td>25.4%</td>
<td>8.6%</td>
<td>5.2%</td>
<td>40.4%</td>
<td>27.6%</td>
</tr>
<tr>
<td>Proportion of nucleotide difference</td>
<td>0.0115</td>
<td>0.0202</td>
<td>0.0143</td>
<td>0.0121</td>
<td>0.0097</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Alternative</th>
<th>All Sites</th>
<th>postCpreG</th>
<th>postC</th>
<th>preG</th>
<th>nonCpG</th>
</tr>
</thead>
<tbody>
<tr>
<td># of sites</td>
<td>13682</td>
<td>830</td>
<td>2394</td>
<td>2640</td>
<td>7818</td>
</tr>
<tr>
<td># of substituted sites</td>
<td>97</td>
<td>16</td>
<td>19</td>
<td>19</td>
<td>43</td>
</tr>
<tr>
<td>Frequencies</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>22.5%</td>
<td>35.8%</td>
<td>22.3%</td>
<td>24.7%</td>
<td>20.3%</td>
</tr>
<tr>
<td>T</td>
<td>28.6%</td>
<td>45.2%</td>
<td>30.7%</td>
<td>29.3%</td>
<td>26.0%</td>
</tr>
<tr>
<td>C</td>
<td>23.8%</td>
<td>9.2%</td>
<td>41.4%</td>
<td>5.2%</td>
<td>26.2%</td>
</tr>
<tr>
<td>G</td>
<td>25.1%</td>
<td>9.9%</td>
<td>5.6%</td>
<td>40.8%</td>
<td>27.5%</td>
</tr>
<tr>
<td>Proportion of nucleotide difference</td>
<td>0.0071</td>
<td>0.0193</td>
<td>0.0079</td>
<td>0.0072</td>
<td>0.0055</td>
</tr>
</tbody>
</table>
Figure 1: The nucleotide substitution rates over the 100nt upstream and 100nt downstream exon flanks of constitutive and alternative exons.
**Figure legends**

**Figure 1:** The nucleotide substitution rates over the 100nt upstream and 100nt downstream exon flanks of constitutive and alternative exons. Error bars represent 95% confidence intervals (calculated by assuming a binomial distribution). Alternative (Set 1): ancestral alternatively spliced exons from Xing and Lee (Xing and Lee, 2005). Alternative (Set 2): ancestral alternatively spliced exons from Pan and colleagues (Pan, et al., 2004). See details of these datasets in Methods and Materials.