

OVERLAPPING ALTERNATIVE DONOR SPLICE SITES IN THE HUMAN GENOME

EKATERINA O. ERMAKOVA^{*,†,§}, RAMIL N. NURTDINOV^{‡,¶}
and MIKHAIL S. GELFAND^{†,‡,||}

[†]*Institute for Information Transmission Problems (Kharkevich Institute)
Russian Academy of Sciences*

Bolshoi Karetny per. 19, 127994 Moscow, Russia

[‡]*Department of Bioengineering and Bioinformatics
Moscow State University, Vorob'evy Gory, 1-73*

119992 Moscow, Russia

[§]*ermakova@iitp.ru*

[¶]*n_ramil@mail.ru*

^{||}*gelfand@iitp.ru*

Received 16 January 2007

Revised 30 May 2007

Accepted 1 June 2007

Over 50% of donor splice sites in the human genome have a potential alternative donor site at a distance of three to six nucleotides. Conservation of these potential sites is determined by the consensus requirements and by its exonic or intronic location. Several hundred pairs of overlapping sites are confirmed to be alternatively spliced as both sites in a pair are supported by a protein, by a full-length mRNA, or by expressed sequence tags (ESTs) from at least two independent clone libraries. Overlapping sites may clash with consensus requirements. Pairs with a site shift of four nucleotides are the most abundant, despite the frameshift in the protein-coding region that they introduce. The site usage in pairs is usually uneven, and the major site is more frequently conserved in other mammalian genomes. Overlapping alternative donor sites and acceptor sites may have different functional roles: alternative splicing of overlapping acceptor sites leads mainly to microvariations in protein sequences; whereas alternative donor sites often lead to frameshifts and thus either yield major differences in the protein sequence and structure, or generate nonsense-mediated decay-inducing mRNA isoforms likely involved in regulated unproductive splicing pathways.

Keywords: Alternative splicing; splice sites; untranslated isoforms; nonsense-mediated decay.

1. Introduction

Alternative splicing is a major source of mRNA and protein diversity in an eukaryotic cell. Alternative donor sites contribute up to 15% of all alternative splicing events in human genes.¹ The nine-nucleotide consensus of human donor splice sites

*Corresponding author.

is MAG|GTRAGT (the vertical line marks the exon–intron boundary; we use the DNA notation throughout), and alternative sites tend to have weaker fit to the consensus compared to constitutive ones.² The closer a site is to the consensus (that proximity may be formalized in terms of free energy or positional weight matrix score), the more often it is used by the basic splicing machinery when additional regulation is absent.

Several recent studies considered alternatively spliced NAG|NAG| pairs of acceptor sites^{3–6} and |GYN|GYN pairs of donor sites.⁷ These alternative splicing events do not disrupt the reading frame, and in most cases they introduce microvariations in the protein structure, influencing one or two amino acids: only ten alternatively spliced NAG|NAG| pairs of acceptor sites³ and four |GYN|GYN pairs of donor sites⁷ create or destroy a stop codon.

Bioinformatics analysis shows that NAG|NAG or NAGNAG| acceptor sites occur in 30% of human genes, and 5% of human genes contain at least one alternatively spliced NAG|NAG| pair of acceptor sites.³ Tadokoro *et al.*⁵ confirmed alternative splicing for 236 NAG|NAG| pairs of acceptor sites by reverse transcription–polymerase chain reaction (RT-PCR). It has been experimentally verified that in the *ITGAM*, *SMARCA4*, and *BTNL2* genes, splicing of the NAG|NAG| acceptor sites is tissue-specific.³ In *IGF1R*, selective usage of tandem acceptor sites yields two protein isoforms of the receptor having different signaling activities and internalization rates⁸; while in *DRPLA*, two such isoforms have different subcellular localizations.⁵ Other alternatively spliced NAG|NAG| acceptor sites are also likely to be functionally important, as they represent nearly half of all human/mouse-conserved alternative acceptor splice sites.^{6,9}

In NAG|NAG and NAGNAG| acceptor sites, most of the active (i.e. used in splicing) NAGs are HAGs (N = any nucleotide, H = not G). GAG acceptor sites are rare, usually inactive,³ and accumulate more SNPs than HAGs.⁴ Since CAG and TAG are the preferred acceptor sites, alternatively spliced YAG|YAG| acceptor sites are almost as frequent as YAG|YAG and YAGYAG| singlets taken together.³ The upstream NAG tends to be the major one.³

The functional role of NAGNAG acceptors is obscure. Although it has been proposed that the spliceosome can bind to an acceptor site with a probability that depends on the site score computed using a positional weight matrix (with correction for possible action of nonsense-mediated decay),¹⁰ this model does not fully account for the observed tissue-specific patterns of NAG|NAG| usage.¹¹

When our study was completed, a paper about one particular type of overlapping donor splice sites, nonframeshifting |GYN|GYN, was published.⁷ In that study, 110 alternatively spliced overlapping pairs of human donor sites with the |GYN|GYN motif were analyzed. The |GTN|GTN pairs appear to be the most common ones (89%); half of them are in phase 0 and cause a valine indel. The downstream donor site (“e donor” in Ref. 7) in alternatively spliced |GTN|GTN pairs and the GT donor site in |GTN|GCN and |GCN|GTN pairs are usually the major ones. The number of expressed sequence tags (ESTs) using the site was negatively

correlated with the free energy of U1 snRNA binding. Similar levels of expression of both isoforms in different tissues were confirmed experimentally for seven human genes: *ANAPC4*, *ANGPT1*, *SEMA5B*, *RBM10*, *TOM1*, *STAT3*, and *Cxorf44*. It was also shown that unconfirmed potential alternative donor sites in the |GYNGYN or GYN|GYN motifs were usually weaker (in terms of free energy) than the annotated ones. A control experiment was performed with nine |GYN|GYN motifs with low free energy of both overlapping sites and only one of them known to be active, and it showed that the potential alternative site was indeed not functional.

Alternatively spliced NAG|NAG| pairs of acceptor sites and |GYN|GYN pairs of donor sites are not specific to the human genome. They are abundant in other mammals, fruitflies, worms, and plants.^{3,7,12,13} Two isoforms of the murine *Pax-3* gene resulting from alternative splicing of a NAG|NAG| pair of acceptor sites differ by the presence of a single glutamine residue, but demonstrate dramatically different DNA-binding activity.¹⁴ *STAT3* orthologs in various mammals retain the GTNGTN motif, and the distributions of the isoforms are similar to the one in human cells.⁷

Nonetheless, a site shift to three nucleotides is not the only possible one. The |GYN|GYN motif is nonframeshifting, but it does not fit the consensus, with severe disagreement for the upstream site. Thus, we considered other types of overlapping donor sites. Of these, |GYNN|GY has the best overall match to the consensus, but it introduces a frameshift. Although that motif is the most abundant one,^{5,6} it has not been previously considered in detail. The site choice in a |GYNN|GY pair can have severe consequences: the use of the upstream site in a |GTNN|GT pair (caused by a single nucleotide polymorphism [SNP]) in the human *BTNL2* gene yields a truncated protein lacking the C-terminal IgC domain and the transmembrane helix, and results in predisposition to sarcoidosis.¹⁵

Here, we consider pairs of alternative donor splice sites at a distance of three to six nucleotides and potential splice sites at the same distances from active sites. We demonstrate that pairs with a site shift of four nucleotides are the most abundant despite the frameshift. We also consider the conservation of potential and active splice sites in the mouse and dog genomes. Intronic potential sites are less conserved than exonic ones, except potential GT donor sites shifted to four nucleotides from the active site that perfectly fit the splice site consensus of the latter. Major donor splice sites are more frequently conserved than minor ones, and frame-preserving pairs are more frequently conserved than frame-shifting ones. In 55% of alternatively spliced pairs of donor sites, one isoform is translatable while the other (usually the minor one) is not. Those untranslatable isoforms are likely targets to nonsense-mediated mRNA decay (NMD) and may have a regulatory role.

2. Definitions

The consensus of human donor splice sites is MAG|GTRAGT.¹⁶ We consider a donor splice site to be nine nucleotides long, and the nucleotides to be enumerated

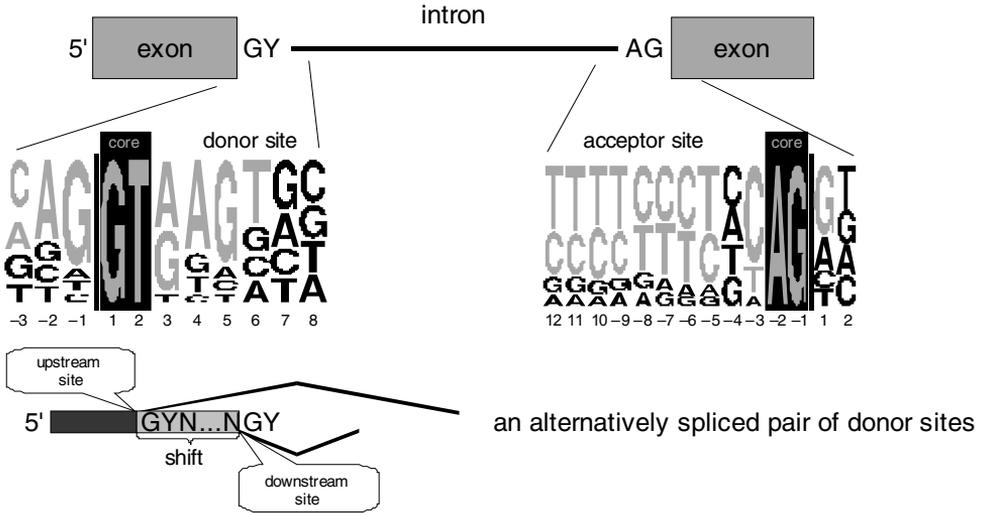


Fig. 1. Definitions.

as shown in Fig. 1. We call the dinucleotide (+1,+2) the core of a donor splice site.

The consensus for human acceptor splice sites is $(Y_n)NYAG|G$. We enumerate the nucleotide positions as shown in Fig. 1. We call the dinucleotide (-2,-1) the core of an acceptor splice site.

We assign a potential donor splice site function to a motif of nine nucleotides (-3,-2,-1,+1,+2,+3,+4,+5,+6) enumerated with GT at positions (+1,+2).

In 11% of alternative donor site pairs, the two sites overlap.¹⁷ Two overlapping potential donor sites form a pair. The distance (in nucleotides) between their splicing positions is the site shift. The upstream site and the downstream site in a pair may be active splicing sites, or they may be silent.

We consider only potential sites with site shifts of three to six nucleotides from the active site. We call potential upstream sites u6, u5, u4, u3 (with respect to the site shift); and potential downstream sites, d3, d4, d5, d6.

We call a splicing event confirmed if it is supported by a protein, by a full-length mRNA, or by ESTs from at least two independent clone libraries.

A site in a pair is called major if it is used in $\geq 66\%$ of cases based on the EST data. A site in a pair is called minor if it is used in $< 33\%$ of cases. In some alternatively spliced pairs, there is no strong bias in the site usage, so the major site cannot be defined. We used EST data from the EDAS (EST-Derived Alternative Splicing) database¹⁷ to ascertain relative site usage in the pairs.

When we write “ GTN_kGT ,” we mean all three possibilities: alternative site $|GTN_k|GT$, upstream active site $|GTN_kGT$, and downstream active site $GTN_k|GT$.

3. Data

Splicing annotations for human genes were taken from the EDAS database.¹⁷ EDAS database annotations are based on protein, mRNA, and EST sequences mapped to the human genomic sequence.

Orthologous triples of human, mouse, and dog genes were taken from supplemental information to Ref. 18 at the Broad Institute (http://www.broad.mit.edu/ftp/pub/papers/dog_genome/suppinfo).

Only donor sites confirmed by a protein, by a full-length mRNA, or by ESTs from at least two independent clone libraries were considered. We considered only canonical sites with the GT core.

4. Results

We considered 187,725 human donor splicing sites. Of these, 96,968 (52%) had a GT dinucleotide at the position u6, u5, u4, u3, d3, d4, d5, or d6 (see Table 1). Potential sites of the d3 type were the least frequent (0.6%) while the d4 ones were the most frequent (39.4%), as GT is the consensus for the positions (+5, +6) of the human donor site.

Sometimes, there were several potential alternative sites near the active site. Such cases were considered independently for all possible sites.

We obtained 385 confirmed pairs of alternatively spliced human overlapping donor sites with a site shift of three to six nucleotides mapped to orthologous triples of human, mouse, and dog genes. Pairs with a site shift of four nucleotides were the most abundant. Site preferences with respect to the site shifts are shown in Table 2.

The scores of the upstream (w_u) and downstream (w_d) sites were calculated for all alternatively spliced pairs of donor sites, as described in Sec. 6. The joint

Table 1. Counts and frequencies of potential alternative sites three to six nucleotides upstream or downstream of the active donor site.

Position of the potential site	u6	u5	u4	u3	d3	d4	d5	d6
Count	8841	5555	3379	3895	1182	74019	7181	12034
Frequency	4.7%	3.0%	1.8%	2.1%	0.6%	39.4%	3.8%	6.4%

Table 2. Site preferences in alternatively spliced pairs.

	Site shift (nt)				
	3	4	5	6	Total
Upstream major	9	148	26	31	214
No major	6	21	4	15	46
Downstream major	37	45	16	27	125
Total	52	214	46	73	385

nt: nucleotide.

distribution of w_u and w_d for the alternatively spliced pairs with the upstream site being the major one, with the downstream site being the major one, and without distinct preference for any of the two sites is shown in Fig. 2. In the $|GTN|GT$ pairs, the two sites cannot be strong simultaneously because of their overlap inducing the conflict between their consensi. In the $|GTN_2|GT$ pairs, the upstream sites are on average stronger and preferred. For the $|GTN_3|GT$ and $|GTN_4|GT$ pairs, the distributions for w_u and w_d are rather similar. For all site shifts, the site strength

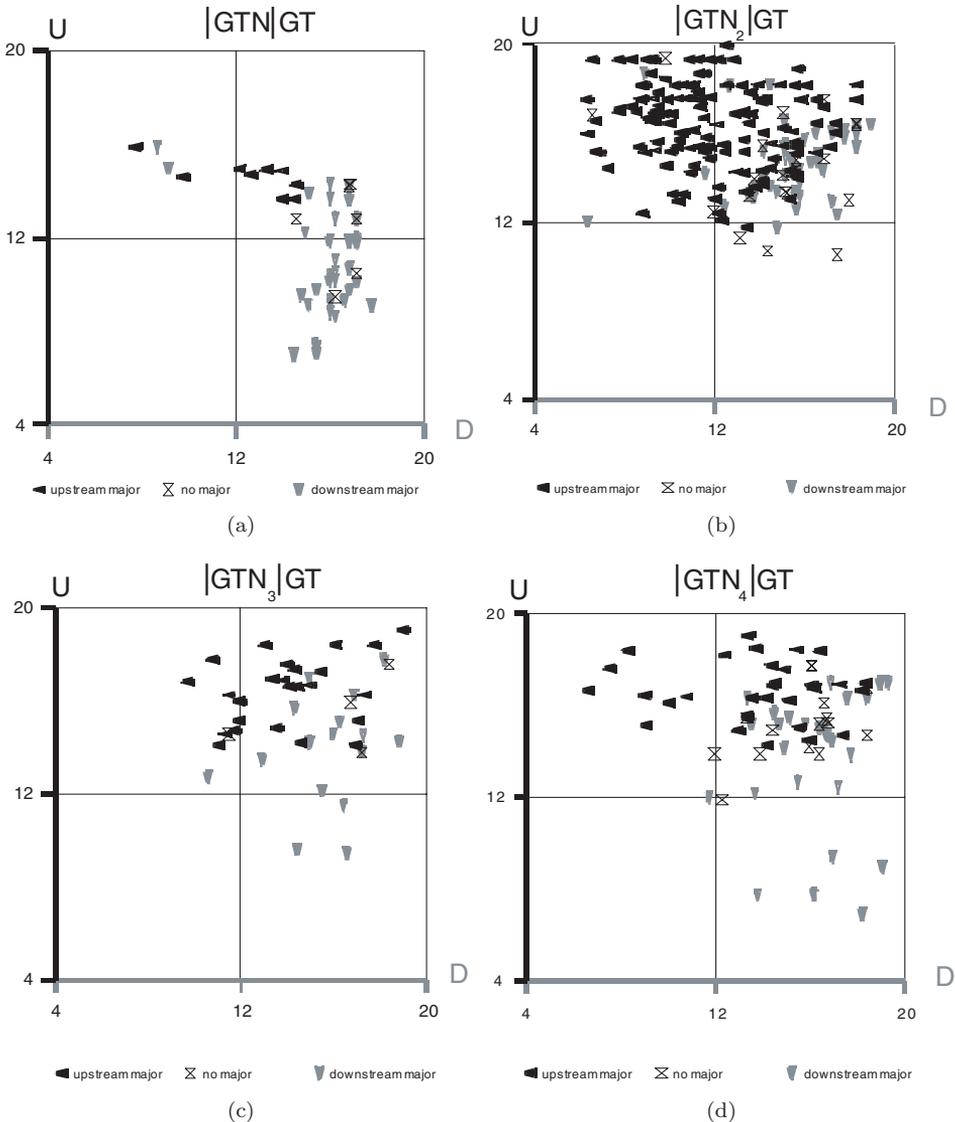


Fig. 2. Correlation of the site score with the site usage in alternatively spliced pairs of donor sites. Horizontal axis: score of the downstream site. Vertical axis: score of the upstream site.

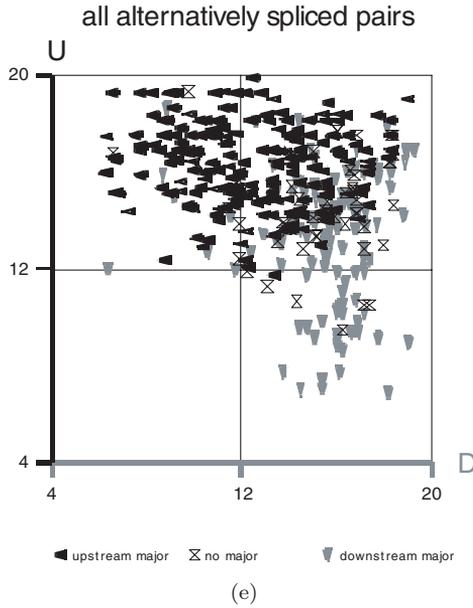


Fig. 2. (Continued)

Table 3. Usability of the overlapping alternatively spliced donor sites in proteins.

Upstream translatable	Downstream translatable	Site shift (nt)				
		3	4	5	6	Total
+	+	14	31	20	52	117
+	-	7	121	15	10	153
-	+	28	23	5	3	59
-	-	3	39	6	8	56
	Total	52	214	46	73	385

nt: nucleotide.

usually (but not always) determines whether the site would be a major or minor one.

As alternatives might be confirmed only by ESTs, we used the IsoformCounter algorithm¹ to decide whether an isoform was translatable (see Table 3).

“Both nontranslatable” pairs occurred in untranslated regions. In |GTN₃|GT pairs with a single translatable site, the downstream site was usually the translatable one; while in |GTN₄|GT pairs, the upstream site was. Expectedly, the |GTN₆|GT pairs had the largest fraction of the “both in protein” annotations (71%). First, the two sites do not interfere much. Second, the change is an in-frame one, so it may have less drastic consequences for the protein.

A translated isoform is usually the major one according to the EDAS annotation (see Table 4). However, the untranslated isoforms might be subject to nonsense-mediated decay or other regulated decay processes, and thus be

Table 4. Majority and translatability.

Upstream translatable	Downstream translatable	Upstream major	No major	Downstream major	Total
+	+	49	22	46	117
+	-	146	5	2	153
-	+	0	3	56	59
-	-	19	16	21	56
	Total	204	46	125	385

underrepresented in EST libraries. When both or none of the sites in a pair can be used in a protein, there is no bias toward the use of the upstream or the downstream site.

A single human donor site was considered conserved in the mouse (dog) genome if it could be located in it using BLAT¹⁹ and Pro-Gen,²⁰ and if GT at positions (+1, +2) was conserved. If a donor splice site was conserved, a potential site was considered conserved if GT was retained at the orthologous positions.

Of 126,326 donor splice sites in the human genes mapped to human, mouse, and dog ortholog triples, 88,696 (70%) were conserved in the mouse genome and 89,280 (71%) in the dog genome. For counts and conservation of potential sites near a conserved donor splice site, see Tables 5(a) and 5(b). Expectedly, in general, intronic potential sites were less conserved than exonic ones, except d4 potential GT donor sites that perfectly fit the splice site consensus of the active site. The least conserved GTs were those at the d3 position, as they contradicted the consensus.

For conservation of upstream and downstream sites in alternatively spliced pairs with respect to majority and site shift, see Tables 6(a) and 6(b). Expectedly, major sites were more frequently conserved than minor ones, and frame-preserving pairs were more frequently conserved than frame-shifting ones.

Table 5(a). Conservation of the potential sites near active human donor sites conserved in the mouse genome.

Position of the potential site	u6	u5	u4	u3	d3	d4	d5	d6
Count in human near active sites conserved in mouse	4171	2462	1286	1744	530	35833	3388	5758
Conserved in mouse	3247 (78%)	1731 (70%)	957 (74%)	1335 (77%)	190 (36%)	29823 (83%)	1309 (39%)	1950 (34%)

Table 5(b). Conservation of the potential sites near active human donor sites conserved in the dog genome.

Position of the potential site	u6	u5	u4	u3	d3	d4	d5	d6
Count in human near active sites conserved in dog	4209	2464	1250	1739	557	35984	3385	5810
Conserved in dog	3464 (82%)	1880 (76%)	998 (80%)	1388 (80%)	274 (49%)	31282 (87%)	1786 (53%)	2703 (47%)

Table 6(a). Conservation in the mouse genome of upstream (u) and downstream (d) sites in 385 alternatively spliced pairs of overlapping human donor sites.

Site shift	3		4		5		6		Total	
	u	d	u	d	u	d	u	d	u	d
Upstream major	8/9 (90%)	5/9 (60%)	120/148 (80%)	97/148 (70%)	22/26 (80%)	12/26 (50%)	25/31 (80%)	13/31 (40%)	175/214 (80%)	127/214 (60%)
No major	4/6 (70%)	3/6 (50%)	10/21 (50%)	8/21 (40%)	1/4 (30%)	2/4 (50%)	10/15 (70%)	11/15 (70%)	25/46 (50%)	24/46 (50%)
Downstream major	22/37 (60%)	28/37 (80%)	24/45 (50%)	29/45 (60%)	11/16 (70%)	12/16 (80%)	16/27 (60%)	24/27 (90%)	73/125 (60%)	93/125 (70%)
Total	34/52 (70%)	36/52 (70%)	152/214 (70%)	128/214 (60%)	35/46 (80%)	22/46 (50%)	51/73 (70%)	45/73 (60%)	272/385 (70%)	231/385 (60%)

Table 6(b). Conservation in the dog genome of upstream (u) and downstream (d) sites in 385 alternatively spliced pairs of overlapping human donor sites.

Site shift	3		4		5		6		Total	
	u	d	u	d	u	d	u	d	u	d
Upstream major	8/9 (90%)	5/9 (60%)	118/148 (80%)	91/148 (60%)	23/26 (90%)	8/26 (30%)	25/31 (80%)	10/31 (30%)	174/214 (80%)	114/214 (50%)
No major	6/6 (100%)	5/6 (80%)	8/21 (40%)	6/21 (30%)	0/4 (0%)	1/4 (30%)	11/15 (70%)	12/15 (80%)	25/46 (50%)	24/46 (50%)
Downstream major	23/37 (60%)	29/37 (80%)	23/45 (50%)	28/45 (60%)	11/16 (70%)	12/16 (80%)	21/27 (80%)	24/27 (90%)	78/125 (60%)	93/125 (70%)
Total	37/52 (70%)	39/52 (80%)	151/214 (70%)	131/214 (60%)	33/46 (70%)	252/46 (50%)	57/73 (80%)	49/73 (70%)	278/385 (70%)	244/385 (60%)

5. Discussion

When two sites overlap, their consensi interact. For example, in alternatively spliced pairs of donor sites, the core and shift preferences result from a tradeoff between the requirements of the two sites, and this tradeoff is reflected in the consensi of the pairs. The consensi for constitutive donor sites and for overlapping alternative sites (including ones with the GC core) are shown in Fig. 3. When the consensus requirements of the upstream and downstream sites coincide, the consensus position of a pair becomes stronger and the corresponding positional weights predict the site usage well. For example, AG at positions (+4, +5) of the upstream site coincides with AG at positions (-2, -1) of the downstream site in the |GYN₃|GY pairs, and the resulting consensus in a pair is stronger than each individual consensus. In the |GYN₂|GY pairs, C at position -3 of the downstream site increases the fraction of C in position +2 of the upstream site. In the |GYN₄|GY pairs, GT at positions (+5, +6) of the downstream sites does not displace AG as the leader, but it replaces TA, the second most frequent nucleotides at these positions.

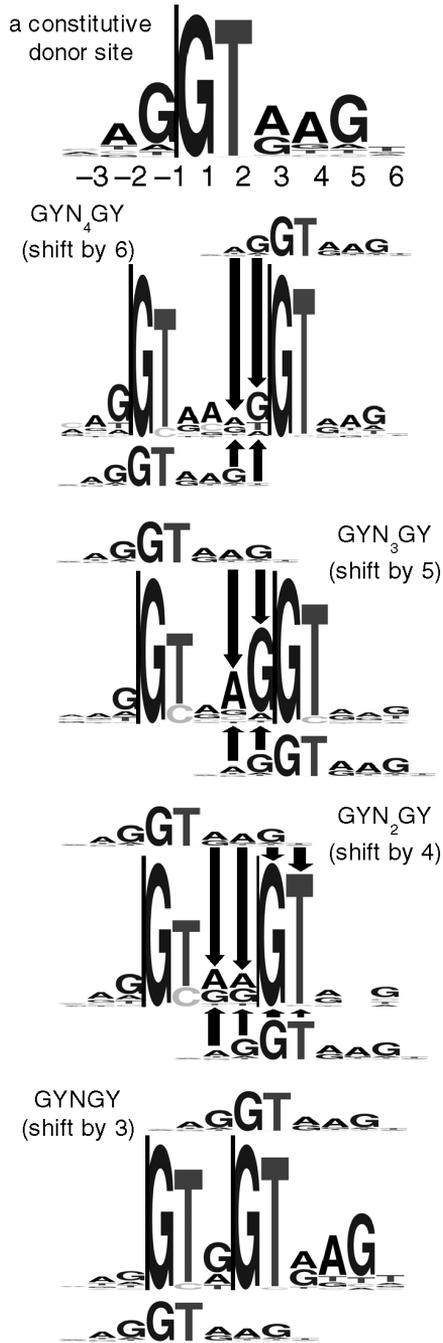


Fig. 3. Site consensi for a nonoverlapping constitutive donor splice site and for alternatively spliced pairs of overlapping donor sites with shifts by three to six nucleotides.

The structure of the consensus might determine functional features of overlapping donor sites. The consensus for the donor splice site contains a perfect core for a second site four nucleotides downstream, and so the upstream site is stronger and in alternatively spliced pairs it tends to be the major one (Table 2). It was previously shown that the distances between alternative donor sites are biased toward frameshifting events mostly because of the site shift to four nucleotides.^{5,6} Moreover, this frameshift is not compensated at the intron acceptor site; on the contrary, site shifts between the alternative acceptors are biased toward frame-preserving events.⁶ We show that in 40% of overlapping donor site pairs confirmed by a protein or by ESTs from at least two independent clone libraries, only the upstream site potentially produces a translated isoform; and that in 15% of the donor pairs, only the downstream site does — thus, the other isoform might be NMD-inducing. Hiller *et al.*⁷ confirmed the usage of both of the overlapping |GYN|GYN donor sites for seven human genes, but detected no difference among the analyzed tissues. Moreover, for the *STAT3* gene, different genotypes were studied and no difference in the expression of *STAT3* isoforms was detected either. When both overlapping donor sites produce translated isoforms, the selective usage of overlapping donor sites may influence the protein-binding properties¹⁴ or the subcellular localization of the isoforms.⁵ Extensive screening of the published data did not yield any experimental reports for overlapping donor sites producing tissue-specific isoforms, as opposed to overlapping acceptor sites that are often tissue-specific.⁵ These observations indicate that overlapping donor sites may regulate protein concentrations uniformly in different tissues rather than provide tissue-specific regulation.

A remaining unresolved question is the functional role of isoforms that would produce severely truncated proteins or are likely targets to NMD. The fact that the sites yielding such isoforms are used quite often argues for their importance. An intriguing possibility is that they have a regulatory role, channeling transcripts toward destruction in some specific conditions as suggested in general for NMD-inducing alternative splicing.²¹ Indeed, regulated unproductive splicing is widespread: it was shown that 45% of alternatively spliced human genes might produce an NMD-prone isoform with a premature termination codon.²²

6. Methods

6.1. Site scores

We used a sample of 85,798 confirmed constitutive donor sites to make a weight matrix that included splice site positions m from -3 to $+6$. The positional nucleotide weights were calculated as in Ref. 23:

$$W(b, m) = \log[N(b, m) + 0.5] - 0.25 \cdot \sum_{i=A,C,G,T} \log[N(i, m) + 0.5], \quad (1)$$

where $N(b, m)$ is the count of nucleotide b in position m in the training sample.

Table 7. The weight matrix for the human donor sites used in the analysis.

	-3	-2	-1	1	2	3	4	5	6
A	0.3945	1.2554	-0.1238	-1.0455	-2.5929	1.6810	1.4464	-0.3671	-0.2059
C	0.4488	-0.5347	-1.3751	-1.7430	0.4388	-1.3981	-0.7729	-0.8412	-0.3878
G	-0.2227	-0.4793	1.9448	5.5628	-2.9786	1.0544	-0.3427	1.7347	-0.1666
T	-0.6207	-0.2414	-0.4459	-2.7743	5.1327	-1.3372	-0.3307	-0.5264	0.7603

The $W(b, m)$ matrix is given in Table 7. The score of a site (b_{-3}, \dots, b_9) , where b_j are nucleotides, was then calculated as a sum of positional weights:

$$w(b_{-3}, \dots, b_9) = W(b_{-3}, -3) + \dots + W(b_9, 9). \quad (2)$$

6.2. Software

Orthologous splicing sites were identified using programs BLAT¹⁹ and Pro-Gen.²⁰ The IsoformCounter program¹ was used to predict translated isoforms. Logos for Fig. 2 were made using the WebLogo program.²⁴ Statistical tests were performed using R .²⁵

Acknowledgments

We are grateful to A. A. Mironov for useful discussions and to A. D. Neverov for finetuning his program IsoformCounter.

This study was partially supported by grants from the Howard Hughes Medical Institute (55001056), INTAS (05-8028), the Russian Foundation for Basic Research (07-04-00343), and the Russian Academy of Sciences (“Molecular and Cellular Biology” program).

References

1. Neverov AD, Artamonova II, Nurtdinov RN, Frishman D, Gelfand MS, Mironov AA, Alternative splicing and protein function, *BMC Bioinformatics* **6**:266, 2005.
2. Clark F, Thanaraj TA, Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human, *Hum Mol Genet* **11**(4):451–464, 2002.
3. Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Platzer M, Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity, *Nat Genet* **36**(12):1255–1257, 2004.
4. Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Platzer M, Single-nucleotide polymorphisms in NAGNAG acceptors are highly predictive for variations of alternative splicing, *Am J Hum Genet* **78**(2):291–302, 2006.
5. Tadokoro K, Yamazaki-Inoue M, Tachibana M, Fujishiro M, Nagao K, Toyoda M, Ozaki M, Ono M, Miki N, Miyashita T, Yamada M, Frequent occurrence of protein isoforms with or without a single amino acid residue by subtle alternative splicing: The case of Gln in DRPLA affects subcellular localization of the products, *J Hum Genet* **50**(8):382–394, 2005.

6. Akerman M, Mandel-Gutfreund Y, Alternative splicing regulation at tandem 3' splice sites, *Nucleic Acids Res* **34**(1):23–31, 2006.
7. Hiller M, Huse K, Szafranski K, Rosenstiel P, Schreiber S, Backofen R, Platzer M, Phylogenetically widespread alternative splicing at unusual GYNGYN donors, *Genome Biol* **7**:R65, 2006.
8. Condorelli G, Bueno R, Smith RJ, Two alternatively spliced forms of the human insulin-like growth factor I receptor have distinct biological activities and internalization kinetics, *J Biol Chem* **269**(11):8510–8516, 1994.
9. Sugnet CW, Kent WJ, Ares M Jr, Haussler D, Transcriptome and genome conservation of alternative splicing events in humans and mice, *Pac Symp Biocomput* **9**:66–77, 2004.
10. Chern TM, van Nimwegen E, Kai C, Kawai J, Carninci P, Hayashizaki Y, Zavolan M, A simple physical model predicts small exon length variations, *PLoS Genet* **2**(4):e45, 2006.
11. Hiller M, Szafranski K, Backofen R, Platzer M, Alternative splicing at NAGNAG acceptors: Simply noise or noise and more?, *PLoS Genet* **2**(11):e207, 2006.
12. Ferranti P, Lilla S, Chianese L, Addeo F, Alternative nonallelic deletion is constitutive of ruminant alpha(s1)-casein, *J Protein Chem* **18**(5):595–602, 1999.
13. Li L, Howe GA, Alternative splicing of prosystemin pre-mRNA produces two isoforms that are active as signals in the wound response pathway, *Plant Mol Biol* **46**(4):409–419, 2001.
14. Vogan KJ, Underhill DA, Gros P, An alternative splicing event in the Pax-3 paired domain identifies the linker region as a key determinant of paired domain DNA-binding activity, *Mol Cell Biol* **16**(12):6677–6686, 1996.
15. Valentonyte R, Hampe J, Huse K, Rosenstiel P, Albrecht M, Stenzel A, Nagy M, Gaede KI, Franke A, Haesler R, Koch A, Lengauer T, Seeger D, Reiling N, Ehlers S, Schwinger E, Platzer M, Krawczak M, Muller-Quernheim J, Schurmann M, Schreiber S, Sarcoidosis is associated with a truncating splice site mutation in BTNL2, *Nat Genet* **37**(4):357–364, 2005.
16. Gelfand MS, Statistical analysis of mammalian pre-mRNA splicing sites, *Nucleic Acids Res* **17**(15):6369–6382, 1989.
17. Nurtdinov RN, Neverov AD, Mal'ko DB, Kosmodem'ianskii IA, Ermakova EO, Ramenskii VE, Mironov AA, Gelfand MS, EDAS, database of alternatively spliced human genes, *Biofizika* **51**(4):589–592, 2006.
18. Lindblad-Toh K, Wade CM, Mikkelsen TS *et al.*, Genome sequence, comparative analysis and haplotype structure of the domestic dog, *Nature* **438**(7069):803–819, 2005.
19. Kent WJ, BLAT — The BLAST-like alignment tool, *Genome Res* **12**(4):656–664, 2002.
20. Novichkov PS, Gelfand MS, Mironov AA, Gene recognition in eukaryotic DNA by comparison of genomic sequences, *Bioinformatics* **17**(11):1011–1018, 2001.
21. Lareau LF, Green RE, Bhatnagar RS, Brenner SE, The evolving roles of alternative splicing, *Curr Opin Struct Biol* **14**(3):273–282, 2004.
22. Lewis BP, Green RE, Brenner SE, Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans, *Proc Natl Acad Sci USA* **100**(1):189–192, 2003.
23. Gelfand MS, Koonin EV, Mironov AA, Prediction of transcription regulatory sites in Archaea by a comparative genomic approach, *Nucleic Acids Res* **28**(3):695–705, 2000.
24. WebLogo [<http://weblogo.berkeley.edu/>].
25. R [<http://www.r-project.org/>].

Ekaterina O. Ermakova is a junior researcher at the Research and Training Center in Bioinformatics of the Institute for Information Transmission Problems, RAS, in Moscow, Russia. She graduated from the Department of Mathematics and Mechanics of the Moscow State University in 2003 (M.S. in Mathematics, *magna cum laude*) and completed her graduate program at the Department of Bioengineering and Bioinformatics of the Moscow State University. Her research interests include alternative splicing and molecular evolution of eukaryotes.

Ramil N. Nurtdinov received his Master's degree in Physics from the Moscow State University in 2003. He is currently working on his Ph.D. at the Department of Bioengineering and Bioinformatics at Moscow State University. His research is centered on the evolution of mammalian alternative splicing.

Mikhail S. Gelfand is Head of the Research and Training Center in Bioinformatics of the Institute for Information Transmission Problems, RAS, in Moscow, Russia, and a Professor at the Department of Bioengineering and Bioinformatics of the Moscow State University. He graduated from the Department of Mathematics of the Moscow State University, received his Ph.D. (Mathematics) from the Institute of Theoretical and Experimental Biophysics, RAS (Pushchino), and the Doctor of Sciences from the State Research Institute for Genetics and Selection of Industrial Microorganisms (Moscow). He is an editorial board member of several journals, in particular, *PLoS Biology*, *Bioinformatics*, *BMC Bioinformatics*, *Journal of Bioinformatics and Computational Biology*, and *Journal of Computational Biology*. He received the A. A. Baev Prize (1999) from the Russian State Human Genome Council, and The Best Scientist of the Russian Academy of Sciences Award (2004). His research interests include comparative genomics, metabolic reconstruction and modeling, evolution of metabolic pathways and regulatory systems, function and evolution of alternative splicing, and functional annotation of genes and regulatory signals.