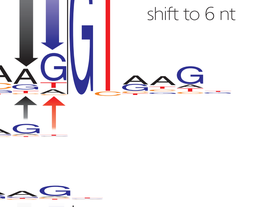
Overlapping alternative donor splice sites

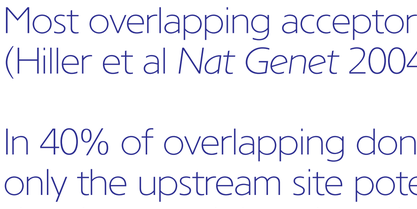
Ekaterina O. Ermakova[1], Ramil N. Nurtdinov[2], and Mikhail S. Gelfand[1,2]
ermakova@iitp.ru n_ramil@mail.ru gelfand@iitp.ru

[1]Institute for Information Transmission Problems (the Kharkevich Institute), Russian Academy of Sciences: Bolshoi Karetny per. 19, 127994, Moscow, Russia
[2]Department of Bioengineering and Bioinformatics, Moscow State University: Vorob'evy Gory, 1-73, 119992, Moscow, Russia



One genome: human

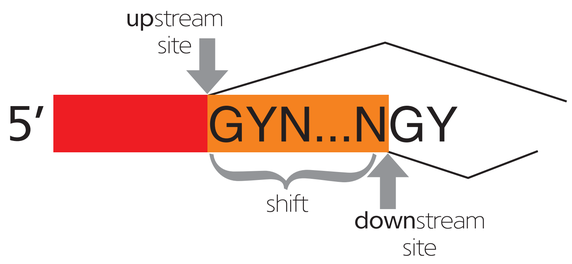
The structure of a **donor splice site**:



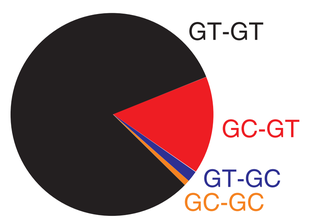
We assign a **potential donor splice site** function to a motif of 9 nucleotides numbered (3, 2, 1, +1, +2, +3, +4, +5, +6) with GY at positions (+1, +2).
Two overlapping potential donor sites form a pair. The distance (in nucleotides) between their splicing positions is the site shift.
The upstream 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 3 through 6 nucleotides from the active site. We call potential upstream sites **up6**, **up5**, **up4**, **up3** (with respect to the site shift), and potential downstream sites **down3**, **down4**, **down5**, **down6**.

In usability studies only sites with GT at positions (+1, +2) were considered.

An alternatively spliced **pair** of donor sites:



Frequencies of core combinations in alternatively spliced pairs of overlapping donor splice sites



Counts and frequences of potential alternative sites 3 through 6 nucleotides upstream or downstream of active donor sites

position of the potential site	up6	up5	up4	up3	down3 (min)	down4 (max)	down5	down6
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%

We considered 187725 human donor splicing sites. 96968 (52%) of them have GT dinucleotide at the position up6, up5, up4, up3, down3, down4, down5, or down6

Site preferences in alternatively spliced pairs

	shift to 3 nt	shift to 4 nt	shift to 5 nt	shift to 6 nt	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

max in the row max in the column both

major site (in a pair): used in ⩾**66%** of cases based on the EST data
minor site: used in <**33%** of cases

When two splice sites overlap, their consensi interact

Most overlapping donor splice sites shift the reading frame and yield major changes in proteins or untranslated isoforms.
Most overlapping acceptor splice sites are in-frame ones and yield minor changes in proteins (Hiller et al *Nat Genet* 2004).

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 in 15% of the donor pairs only the downstream site does, and thus the other isoform might be inducing nonsense-mediated decay.

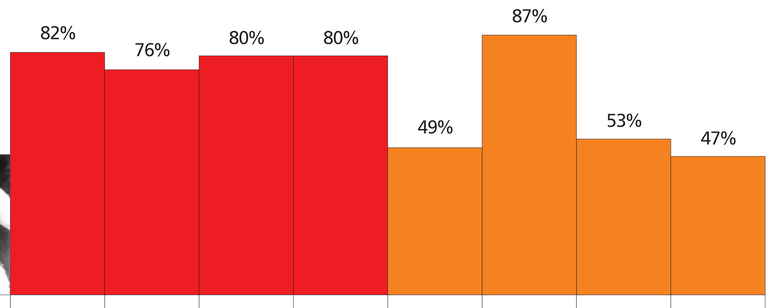
(upstream translatable, downstream translatable)	shift to 3 nt	shift to 4 nt	shift to 5 nt	shift to 6 nt	total
(+, +)	14	31	20	52	117
(+, –)	7	121	15	10	153
(–, +)	28	23	5	3	59
(–, –)	3	39	6	8	56
total	52	214	46	73	385

Usability of the overlapping alternatively spliced donor splice sites in proteins: predicted translatability of the isoforms

(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

Majority and translatability

Two genomes: + dog



position of the potential site	up6	up5	up4	up3	down3	down4	down5	down6
count in human near active sites conserved in dog	4209	2464	1250	1739	557	35984	3385	5810
conserved in dog	3464	1880	998	1388	274	31282	1786	2703

exonic potential donor splice sites intronic potential donor splice sites

	shift to 3 nt		shift to 4 nt		shift to 5 nt		shift to 6 nt		total	
	upstream conserved	dowstream conserved	upstream conserved	dowstream conserved	upstream conserved	dowstream conserved	upstream conserved	dowstream conserved	upstream conserved	dowstream conserved
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%)	24/26 (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%)	25/46 (50%)	57/73 (80%)	49/73 (70%)	278/385 (70%)	244/385 (60%)

Conservation of potential donor splice sites. Exonic potential sites are more frequently conserved than intronic ones except down4 potential sites. In the latter case the intronic sites are more conserved due to the donor splice site consensus.

Conservation of alternatively spliced pairs of overlapping donor splice sites. Major sites are more frequently conserved than minor ones. The minor sites exonic relative to the major site are more frequently conserved than intronic minor sites.

Methods:

Splicing annotations for human genes: **EDAS** (EST-Derived Alternative Splicing) database; Neverov et al, *BMC Bioinformatics* (2005); http://www.belozersky.msu.ru/edas/

Confirmation: Only donor sites confirmed by a **protein**, by a **full-length mRNA**, or by **ESTs** from at least **two** independent clone **libraries** were considered

Orthologous genes: Linblad-Toh et al, *Nature* (2005); the Broad Institute; http://www.broad.mit.edu/ftp/pub/papers/dog_genome/suppinfo/

Orthologous splice **sites** identification: BLAT, Kent, *Genome Res.* (2002); Pro-Gen, Novichkov et al, *Bioinformatics* (2001)

Translated isoforms prediction: IsoformCounter; Neverov et al, *BMC Bioinformatics* (2005) Logos: WebLogo; Crooks et al, Genome Res. (2004); http://weblogo.berkley.edu

E. O. Ermakova, R. N. Nurtdinov, and M. S. Gelfand,
Overlapping alternative donor splice sites in the human genome, *J Bioinform Comput Biol*, (2007), in press

M. Hiller, K. Huse, K. Szafranski, P. Rosenstiel, S. Schreiber, R. Backofen and M. Platzer,
Phylogenetically widespread alternative splicing at unusual GYNGYN donors, *Genome Biol.* 7, R65 (2006)

M. Akerman and Y. Mandel-Gutfreund,
Alternative splicing regulation at tandem 3' splice sites, *Nucleic Acids Res.* 34(1), 2331 (2006).

We are grateful to A. A. Mironov for useful discussions **and to A. D. Neverov** for fine-tuning his program IsoformCounter

This study was partially supported by grants from HHMI (55001056), INTAS (05-8028), RFBR (07-04-00343), and RAS (program "Molecular and Cellular Biology")

This presentation was supported by Sixth Framework Programme of the European Union, through the EMBRACE project