Biophysics, Vol. 48, Suppl. 1, 2003, pp. S30–S35.
Original Russian Text Copyright © 2003 by Denisov, Gelfand.

Conservedness of the Alternative Splicing Signal UGCAUG in the Human and Mouse Genomes

S. Denisov[1] and M. S. Gelfand[2,3,4]

[1]*Department of Biology, Moscow State University, Vorobievy Gory, Moscow, 119992 Russia*
[2]*State Scientific Center GosNIIGenetika, 1 1st Dorozhny proezd, Moscow, 117545 Russia; E-mail: gelfand@ig-msk.ru*
[3]*Institute for Problems of Information Transmission,*
Russian Academy of Sciences, 19 Bolshoi Karetny per., Moscow, 127994 Russia
[4]*Department of Bioengineering and Bioinformatics, Moscow State University,*
1-73 Vorobievy Gory, Moscow, 119992 Russia

Received November 7, 2003; in final form, December 26, 2003

Abstract—It has recently been shown that the sequence UGCAUG occurred in introns downstream of brain-specific cassette exons more often than in a control sample of introns following constitutive exons. Here we analyzed conservation of all UGCAUG hexanucleotides observed in 1000 nucleotide segments following cassette exons. Some 50% of UGCAUG sites in the human introns and 70% sites from mouse introns were conserved in both genomes. Since conservation is a feature of functionally important genomic regions (e.g., protein-coding or regulatory regions), it is likely that the conserved sites are functional and regulate alternative splicing. Besides, analysis of newly available mRNA and EST data demonstrated that the considered exons are not strictly brain-specific, and thus UGCAUG appears to be an enhancer of exon inclusion not only in the brain, but also in other tissues.

Key words: human genome, alternative splicing, regulation, cassette exon

INTRODUCTION

Recently published sequences of the human and mouse genome open new possibilities for studying cellular processes. The human (*Homo sapiens*) and mouse (*Mus musculus*) genomes diverged about 75 million years ago [2]. Genomic regions not subject to selection (mainly noncoding regions) changed considerably, with about 0.52 mutations per site [3]. On the contrary, protein-coding regions that were subject to stabilizing selection are more conserved. Indeed, homologous human and mouse exons can be aligned with average identity of about 85%, whereas 60% of introns cannot be aligned at all, and the average identity of alignable introns is 69% [3].

Nevertheless, many studies (e.g., [2]) show that introns and intergenic regions contain segments of considerable similarity. These segments do not encode proteins, but are conserved, and a natural explanation is that they have a regulatory function. In particular, conserved intronic sites may be involved in regulation of alternative splicing.

Alternative splicing, leading to generation of multiple mRNA isoforms corresponding to one gene, is an important mechanism for generation of protein identity. Current estimates of the fraction of alternatively spliced genes in the human genome are 30% through 60% [4, 5]. Not much is known about regulation of alternative splicing. Cis-regulatory sites may reside both in exons and introns [6]. These sites are bound by proteins (SR-proteins, hnRNP, etc.) that regulate splicgin events, e.g. inclusion of cassette exons. If a regulatory element promotes exon inclusion, it is called an enhancer; if it leads to exon skipping, a silencer. Some studies also show that splicing may be regulated by RNA secondary structure [7, 8].

Brudno et al. [1] studied a sample of brain-specific cassette exons and adjacent intronic fragments. It was shown that introns downstream of such exons are enriched in the UGCAUG motifs compared to a control sample containing constitutive exons. Previously it was assumed that UGCAUG regulates splicing of brain-specific exons, but the same study [1] showed that this hexanucleotide is also found in introns downstream of muscle-specific exons.

Sorec and Ast [2] demonstrated that intron regions adjacent to cassette exons contain segments conserved between human and mouse, whereas the degree of conservation around constitutive exons is lower. They selected 132 alternative exons containing upstream and downstream conserved segments of length at least 50 nt, and analyzed the most frequent hexanucleotides in 100 nt intron fragments adjacent to such exons. The most frequent hexanucleotide in downstream exons was UGCAUG. It was over-represented compared both to downstream introns of constitutive exons and to upstream introns of alternative exons.

Among 25 genes studied in [1], the influence of the UGCAUG site on exon inclusion was experimentally observed only for exon N1 of murine gene c-src [9]. The regulation involves binding to this site by protein KSRP and FBP [6]. The latter study identified a longer sequence (pos. 37–70 relative to the donor site of the cassette exon), named DCS (downstream control sequence). In addition to UGCAUG, it contains two sites, one bound by PTB/nPTB, and the other, by hnRNP H and hnRNP F [6]. PTB is a well-known splicing silencer [6]; among genes in our sample, it regulates alternative splicing c-src, $\gamma 2$ subunit of $GABA_A$ receptor, light chain of clatrin B(CLCB), and NMDA-receptor NR1. Notably, PTB has a brain-specific homolog nPTB, that binds to the same site as PTB, but is less effective as a silencer, thus promoting exon inclusion in the brain-specific isoform [8, 10]. One more protein involved in alternative splicing, Fox-1, binds to GCAUG [11]. It is expressed in heart and skeletal muscle.

Some occurrences of the UGCAUG hexanucleotides identified by Brudno *et al.* [1] could be random. Our aim was to use human-mouse genomic comparisons in order to distinguish conserved (and thus likely functional) UGCAUG site from non-conserved (spurious) ones. In addition, we used newly available data to check whether all exons from [1] are indeed brain-specific.

DATA AND METHODS

For each exon from [1] we (1) analyzed conservation of UGCAUG in the human and mouse genomes; (2) based on that, made a counclusion about functionality of UGCAU; (3) considered tissue- and organ-specificity of the cassette exon.

Genomic sequences were analyzed using Human/Mouse/Rat Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway) and program BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat). The following version of genomic sequences were used: human of July 2003, mouse of February 2003, rat of June 2003. Analyzed exons were taken from http:// //www-gsb.lbl.gov/~dubchak/splicing-data.

It should be noted that the initial sample contained both human and mouse exons. For each such exon the orthologous gene was identified in the remaining genome, and conservation of UGCAUG was considered. Thus the initial inhomogeneity could not influence the result.

For each exon, the following steps were done:

(1) localization of the exon in the genome using BLAT;

(2) identification of the corresponding mRNA and protein sequence in the Human/Mouse Genome Browser.If such mRNA was not found, an isoform containing adjacent exons was selected, the studied exon was artificially introduced into this mRNA in the correct position, and then the results was translated into protein;

(3) identification of the orthologous gene by spliced alignment of the human protein with the mouse genome or vice versa using BLAT;

(4) mapping of the exon corresponding to the initial cassette exon using the BLAT alignment and analysis of the corresponding mRNA/EST in the Human Genome Browser;

(5) analysis of all UGCAUG sites in 1000 bp in downstream introns, separately for the human and mouse genes. For introns shorter than 1000 bp, the analyzed fragment terminated at the acceptor site;

(6) alignment of introns in order to determine whether the observed UGCAUG sites are conserved in the two genomes. This was done by alignment of UGCAUG sites from one genome with 100 bp flanks on both sides with the intronic fragment of the orthologous gene from the other genome.

RESULTS

Conservation of Exons

As exons are subject to evolutionary changes (e.g. loss or birth, see [12]), existence of each exon in the human and mouse genome was verified using mRNA/EST data. In some cases, no mRNA/EST containing the exon could not be found. This could mean that either the exon does exist at all, or that the database does not contain the corresponding isoform, although it exists. All such cases are listed in Table 1. High level of sequence conservation of tentative exons and existence of mRNA-confirmed exons in related organisms confirms that in all but two cases these exons are real. The exclusions are (1) ankyrin B, large insert (mouse) where the exon alignment is rather weak and (2) KOR-3a (mouse) where the corresponding exon could not be found at all. These cases were excluded from consideration. Mouse exon of gene NMDA-R1 is considerably longer than the homologous human exon. This can be explained by a mutation in the donor site (GT to GC). This leads to the loss of recognition by the spliceosome, that inseted uses a cryptic GT site downstream.

Conservation of UGCAUG Sites in Introns

We analyzed all occurrences of UGCAUG in 1000 nt intronic segments downstream of cassette exons. In 24 intron pairs, 28 motifs were found in human introns, and 20 in mouse introns. The data about conservation of these sites is given in Table 2. Fourteen sites are conserved in the two genomes. Thus the degree of conservation is 14/28 = 50% for human sites and 14/20 = 70% for mouse sites.

In one case (B-KSR1) UGCAUG (H. sapiens) aligned with UGCgUG (M. musculus), although this site was embedded into highly conserved region. In PMCA4, the region of possible occurrence of UGCAUG in the mouse intron was not sequenced. In both cases we identified the corresponding exons in the rat genome. The UGCAUG site was conserved there.

Tissue and Organ Distribution of Isoforms Containing Cassette Exons

It was assumed that all exons in the studied sample are brain-specific [1]. We analyzed the tissue and organ distribution of isoforms containing these exons and found that that at least 10 human exons (nos. 2, 3, 5, 7, 8, 11, 13, 17, 19, 25 in Table 2) and 4 mouse exons (nos. 2, 8, 10, 24 in Table 2) are not strictly brain-specific. Thus UGCAUG does not seem to be a brain-specific enhancer of alternative splicing. It should be noted that Brudno et al. [1] observed increased frequency of UGCAUG in a smaller sample

Table 1. Exons not confirmed by EST/mRNA

Gene name	Alignment	AT	Genscan	Genomes containing a homologous exon
		Human		
Ca-channel of N-type / CACNA1B	+	ag/gt	–	rat
NF1 exon 9a / NF1	+	ag/gt	–	rat
Activin receptor of type II / ACVR2	+	ag/gt	–	rat
c-src exon N / SRC	+	ag/gt	–	mouse, rat
		Mouse		
Ankyrin B, large insert / ANK2	–	ag/gt	+	human
Ñà-pump PMCA4 / ATP2B2	+	ag/gt	–	human, rat, cow, dog
Ca-channel of N-type / CACNA1B	+	ag/gt	–	human, rat
NMDA-R1 exon 5 / GRIN1	+	ag/gt	5′-ext.	rat
B-KSR1 / KSR	+	ag/gt	+	human
HDlg / DLG1	+	ag/gt	–	human
MHC-B / MYH10	+	ag/gC	–	human, rat, cow (not extended)
NF1 exon 9a / NF1	+	ag/gt	–	rat
Activin receptor of type II / ACVR2	+	ag/gt	–	rat

* Alignment: "+" – good , "–" – bad (many insertions/deletions, frameshifts, etc.). Genscan: "+" – exon predicted by GenScan [13], "–" – exon not found, "5′-ext." – predicted exon is extended at 5′-end compared to the human homolog. In gene MHC-B the exon is extended at 5′-end compared to homologous exon from human, rat and cow.

Table 2. Conservation of UGCAUG sites in human and mouse exons

No.	Gene name	Gene in the sample	Position of UGCAUG relative to donor sites		Alignment	Predicted functionality
			Human	Mouse		
01	Ankyrin B, large isnert / ANK2	human	none	1. 226	yes	false
02	FHL1B / FHL1	human	none	none	n/a	n/a
03	Ca-pump PMCA4 / ATP2B2	human	1. 22	none	no	false
04	Na-channel SCN8 / SCN8A	human	1. 30	1. 30	yes	true
05	Amphyphysin II (region I) / BIN1	human	none	none	n/a	n/a
06	Ca-channel of N-type / CACNA1B	human	none	none	n/a	n/a
07	NMDA-R1 exon 5 / GRIN1	human	1. 8	1. 8	yes	true
			2. 257	2. 262	yes	true
			3. 494	3. 497	yes	true
08	CLCB / CLTB	human	none	none	n/a	n/a
09	Myelin-associated glycoprotein exon 12 / MAG	human	1. 63	1. 63	yes	true
				2. 599	no	false
10	4.1R exon 15 / EPB41	human	none	none	n/a	n/a
11	B-KSR1 / KSR	human	1. 12	none	yes	?
12	4.1N / EPB41L1	human	1. 5	1. 5	yes	true
13	4.1B exon 15 / EPB41L3	human	1. 286	1. 271	yes	true
			2. 344	2. 328	yes	true
14	HDlg / DLG1	human	1. 243	1. 243	yes	true
			2. 275	2. 274	yes	true
15	KOR-3a / OPRK1	mouse		1. 304	Excluded from the sample	
16	Agrin exon 33 / AGRN	mouse	none	none	n/a	n/a
17	MHC-B / MYH10	human	1. 244	none	no	false
18	NF1 exon 9a / NF1	human	1. 57	1. 57	yes	true
			2. 117	2. 118	yes	true
19	Tyrosine phosphatase LAR / PTPRF	human	1. 884	1. 912	no	false
			2. 888	none	no	false
20	Agrin exon 32 / AGRN	mouse	1. 247	1. 257	no	false
			2. 251	2. 265	no	false
			3. 365			false
			4. 445			false
			5. 620			false
			6. 642			false
			7. 734			false
			8. 885			false
21	Activin receptor of type II / ACVR2	human	none	none	n/a	n/a
22	GABA gamma2 / GABRG2	human	1. 31	1. 31	yes	true
23	c-src exon N / SRC	human	1. 66	1. 57	yes	true
			2. 289	2. 622	no	false
24	Agrin exon 28 / AGRN	human	none	1. 875	no	false
25	FE65 / APBB1	human	none	none	n/a	n/a

```
Human    uugaguggcuuucuaggcuaaaaagagguaaugaaaguauauacug—--ucucucaccau
         ||| |||    ||||| ||| |||||||| ||||||||||||||||   ||||||||| |
Mouse    uuggguguuggucuagacuacaaagaggucaugaaaguauauacugucuucucucaccgu

Human    uUGCAUGaauuauuaaggugugaaauauucaaauUGCAUGuguuugcuaguauguaagcc
         |||||||| |||||||| | ||  ||| || || ||||||| ||| | ||||||||| ||
Mouse    uUGCAUGaguuauuaaguuuugggaua-uccaacUGCAUGuuuuuac-aguauguaaacc
```

Fig. 1. Two UGCAUG sites of gene HDIg (DLG1) in conserved environment.

```
Human    gaguaaaaaaaggaacuaugaaaaccucgaccaacuguccuaugacaacaagcgcggacc
         ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Mouse    gaguaaaaaaaggaacuaugaaaaccucgaccaacuguccuaugacaacaagcgcggacc

Human    caagguauauaUGCAUGgacgugcacgccaccca-cggcuagggagcccuggccu-cggc
         |||||||||||||||||||||||||||||||||| || |||| |||||| ||||| |||
Mouse    caagguauauaUGCAUGgacgugcacgccacccaccgacuagcgagccccggccugcggu
```

Fig. 2. Exon 5 of gene NMDA-R1 (GRIN1) (underlined) and a part of the conserved sequence containing UGCAUG.

of muscle-specific exons. When more mRNAs/ESTs are sequenced, it is likely that the number of purportedly brain-specific exons will further decrease.

DISCUSSION

Efficacy of the Approach

We applied the assumption that conservation of a site implies its functional significance. The question arising is whether one could expect such conservation by chance. A back-of-an-envelope calculation is as follows. Denote the average number of nucleotide substitutions in introns by q. Then the probability of nucleotide conservation is $1 - q$, and the probability of hexanucleotide conservation is $(1 - q)^6$, and of at least one mutation, $1 - (1 - q)^6$. The average level of conservation in human and mouse non-coding regions is 0.52 substitutions per nucleotide [3]. This might serve as a reasonable estimate of intronic conservation as well. If $q = 0.52$, then $1 - (1 - q)^6 = 98{,}8\%$, and thus most hexanucleotides would contain mismatches.

On the other hand, one cannot exclude the possibility that some non-conserved sites are functional. Regulation of alternative splicing could change in course of evolution, and some sites could lose its function in one of the lineages. Thus we do not claim that we have identified all functional sites, but we believe that those that have been selected are much more likely to be functional that the remaining ones.

Participation of UGCAUG in Regulation of Alternative Splicing

The identified hexanucleotides fall within larger conserved regions on both sides (an example is shown in Fig. 1). Often the conserved regions are proximal to the exon compared to the UGCAUG site (e.g. Fig. 2). This shows that regulation of alternative splicing is a complicated process involving many factors, whereas the studies motif is only a part of the system of cys-regulatory elements. This has been demonstrated in experiments with c-src, discussed in the introduction [6].

As noted, the enhancer function of UGCAUG has been demonstrated for c-src (exon N1) [9]. Assuming that conservation implies functional significance, we identified 14 sites that are likely to be functional (Table 2). In gene4site was not sequenced in mouse, but was observed in rat, and we believe that it also is functional. In one case (gene B-KSR10) a mouse-specific mutation was observed, whereas the site is retained in rat, and no conclusion about functionality of this element could be made.

ACKNOWLEDGMENTS

We are grateful to A.A. Mironov, R. Nurtdinov, and I. Kosmodemianskii for useful discussion.

This study was supported by grants from HHMI (55000309) and LICR (CRDF RBO-1268)

REFERENCES

1. Sorec, R. and Ast, G., *Genome Research*, 2003, vol. 13, pp. 1631–1637.
2. Waterson, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.*, *Nature*, 2002, vol. 420, pp. 520–562.
3. Mironov, A.A., Fickett, J.W., and Gelfand, M.S., *Genome Reseach*, 1999, vol. 9, pp. 1288–1293.
4. Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J., and Bork, P., *FEBS Lett.*, 2000, vol. 474, pp. 83–86.
5. Ladd, A.N. and Cooper, T.A., *Genome Biology*, 2002, vol. 3, P. REVIEWS0008.
6. Mirami, E., Margalit, H., and Sperling, R., *Nucleic Acids Research*, 2003, vol. 31, pp. 1974–1983.
7. Blanchette, M. and Chabot, B., *RNA*, 1997, vol. 3, pp. 405–419.
8. Brudno, M., Gelfand, M.S., Spengler, S., Zorn, M., Dubchak, I., and Conboy, J.G., *Nucleic Acids Research*, 2001, vol. 29, pp. 2338–2348.
9. Modafferi, E.F. and Black, D.L., *Molecular Cell Biology*, 1997, vol. 17, pp. 6537–6545.
10. Markovtsov, V., Nikolis, J.M., Goldman, J.A., Turck, C.W., Chou, M.Y., and Black, D.L., *Molecular Cell Biology*, 2000, vol. 20, pp. 7463–7479.
11. Jin, Y., Suzuki, H., Maegawa, S., Endo, H., Sugano, S., Hashimoto, K., Yasuda, K., and Inouve, K., *EMBO J.*, 2003, vol. 22, pp. 905–912.
12. Modrek, B. and Lee, C.J., *Nature Genetics*, 2003, vol. 34, pp. 177–180.
13. Burge, C. and Karlin, S., *J. Mol. Biol.*, 1997, vol. 268, pp. 78–94.