Introns and Reading Frames: Correlation Between Splicing Sites and Their Codon Positions

Masaru Tomita, Nobuyoshi Shimizu,† and Douglas L. Brutlag‡*

*Department of Environmental Information and Department of Molecular Biology, Keio University, Japan; †Department of Molecular Biology, Keio University; and ‡Department of Biochemistry, Stanford University

Computer analyses of the entire GenBank database were conducted to examine correlation between splicing sites and codon positions in reading frames. Intron insertion patterns (i.e., splicing site locations with respect to codon positions) have been analyzed for all of the 64 codons of all the eukaryote taxonomic groups: primates, rodents, mammals, vertebrates, invertebrates, and plants. We found that reading frames are interrupted by an intron at a codon boundary (as opposed to the middle of a codon) significantly more often than expected. This observation is consistent with the exon shuffling hypothesis, because exons that end at codon boundaries can be concatenated without causing a frame shift and thus are evolutionarily advantageous. On the other hand, when introns interrupt at the middles of codons, they exist in between the first and second bases much more frequently than between the second and third bases, despite the fact that boundaries between the first and second bases of codons are generally far more important than those between the second and third bases. The reason for this is not clear and yet to be explained. We also show that the length of an exon is a multiple of 3 more frequently than expected. Furthermore, the total length of two consecutive exons is also more frequently a multiple of 3. All the observations above are consistent with results recently published by Long, Rosenberg, and Gilbert (1995).

Introduction

RNA splicing and protein synthesis are known to occur sequentially as two independent processes; the latter takes place only when the former has completed. In this regard, codon reading frames, which can be determined by the protein synthesis process, cannot be used as a factor of RNA splicing machinery in determining its splicing sites. In other words, locations of splicing sites and their codon positions should be independent of each other as far as the molecular mechanism of RNA splicing is concerned. Contrary to the argument above, however, we shall present in this paper evidence that there exists correlation between splicing sites and their codon positions in a reading frame. In particular, we show that:

- Introns are more likely to begin at codon boundaries; i.e., exons are more likely to end at codon boundaries.
- If not at codon boundaries, introns begin after first codon positions more often than second codon positions.
- The lengths of exons, as well as those of pairs of adjacent exons, are multiples of 3 more frequently than expected.

Those tendencies, presumably, have emerged by evolution. They may give us some hints about the question of the origin of introns, which is still under heated debate (Blake 1978; Gilbert 1978; Cavalier-Smith 1985; Senapathy 1986; Kersanach et al. 1993; Roger and Doolittle 1993; Logsdon et al. 1994; Mattick 1994; Berget 1995). Statistical data may also be taken into account by computer algorithms for exon/intron finding programs to improve their recognition accuracy.

Key words: sequence analysis, intron, exon, splicing.

Address for correspondence and reprints: Marasu Tomita, Department of Environmental Information and Department of Molecular Biology, Keio University, 5322 Endo, Fujisawa, 252, Japan. E-mail: mt@sfc.keio.ac.jp.

Mol. Biol. Evol. 13(9):1219–1223. 1996

Methods

Introns of primates, rodents, mammals, vertebrates, invertebrates, and plants were extracted from the GenBank database (NCBI-GenBank flat file release 90.0, 15 August 1995). The following entries were excluded from the analysis:

- Incomplete sequences.
- Pseudogenes.
- Introns which do not start with "GT" and introns which do not end with "AG." While a few introns are known to have different consensus sequences at their ends, most nonconsensus introns in the database are due to errors.

Furthermore, the following simple method was used to remove duplicate/homologous sequences: If three consecutive exons have the same length pattern in two different contexts, they are considered homologous. For example, if entry A consists of exons of lengths 24, 120, 40, 20, 65, and 61 and entry B consists of exons of lengths 33, 54, 81, 24, 120, and 40, then the parts with the three exons of lengths 24, 120, and 40 are considered homologous. The middle exon (120) and the two introns around it are excluded from our analysis. (A more conservative procedure, excluding all three exons and the four introns, was also tried, but essentially the same results were obtained.) Cases of two nonhomologous parts having same-length exons three times in a row by chance are quite rare and therefore ignorable. Note that this screening method can exclude not only homologous entries, but also intragenetic homologous exons. For instance, some genes, such as collagen genes, have many homologous exons lined up within the same gene; only one of those exons will be counted in our analysis, as they all have the same length.

All the computer programs are written in C programming language and run on Sparc stations under the UNIX operating system. The software is available from the authors on request.

Table 1
Intron-Insertion Patterns of 3,116 Primate Introns for Each of the 64 Codons[a]

!AAA = 26	A!AA = 5	AA!A = 17	AAA! = 13	!AAC = 31	A!AC = 4	AA!C = 8	AAC! = 4
!AAG = 13	A!AG = 9	AA!G = 43	AAG! = 348	!AAT = 24	A!AT = 6	AA!T = 5	AAT! = 26
!ACA = 17	A!CA = 4	AC!A = 1	ACA! = 5	!ACC = 15	A!CC = 2	AC!C = 2	ACC! = 1
!ACG = 6	A!CG = 0	AC!G = 6	ACG! = 23	!ACT = 18	A!CT = 6	AC!T = 3	ACT! = 5
!AGA = 13	A!GA = 7	AG!A = 47	AGA! = 4	!AGC = 14	A!GC = 20	AG!C = 57	AGC! = 3
!AGG = 14	A!GG = 3	AG!G = 183	AGG! = 92	!AGT = 12	A!GT = 8	AG!T = 32	AGT! = 3
!ATA = 12	A!TA = 1	AT!A = 10	ATA! = 0	!ATC = 51	A!TC = 4	AT!C = 5	ATC! = 0
!ATG = 11	A!TG = 7	AT!G = 5	ATG! = 48	!ATT = 25	A!TT = 4	AT!T = 4	ATT! = 8
!CAA = 5	C!AA = 0	CA!A = 2	CAA! = 8	!CAC = 10	C!AC = 4	CA!C = 5	CAC! = 3
!CAG = 4	C!AG = 5	CA!G = 6	CAG! = 245	!CAT = 6	C!AT = 0	CA!T = 1	CAT! = 4
!CCA = 4	C!CA = 0	CC!A = 1	CCA! = 8	!CCC = 16	C!CC = 0	CC!C = 0	CCC! = 0
!CCG = 4	C!CG = 4	CC!G = 2	CCG! = 33	!CCT = 5	C!CT = 2	CC!T = 1	CCT! = 7
!CGA = 2	C!GA = 4	CG!A = 6	CGA! = 0	!CGC = 8	C!GC = 2	CG!C = 3	CGC! = 1
!CGG = 4	C!GG = 3	CG!G = 31	CGG! = 29	!CGT = 1	C!GT = 2	CG!T = 0	CGT! = 1
!CTA = 7	C!TA = 0	CT!A = 3	CTA! = 0	!CTC = 109	C!TC = 0	CT!C = 0	CTC! = 0
!CTG = 48	C!TG = 1	CT!G = 14	CTG! = 44	!CTT = 14	C!TT = 5	CT!T = 1	CTT! = 1
!GAA = 53	G!AA = 45	GA!A = 5	GAA! = 10	!GAC = 43	G!AC = 60	GA!C = 3	GAC! = 6
!GAG = 67	G!AG = 81	GA!G = 8	GAG! = 163	!GAT = 34	G!AT = 79	GA!T = 2	GAT! = 30
!GCA = 27	G!CA = 12	GC!A = 1	GCA! = 3	!GCC = 35	G!CC = 31	GC!C = 0	GCC! = 3
!GCG = 5	G!CG = 12	GC!G = 7	GCG! = 20	!GCT = 37	G!CT = 50	GC!T = 0	GCT! = 10
!GGA = 38	G!GA = 91	GG!A = 9	GGA! = 3	!GGC = 52	G!GC = 127	GG!C = 3	GGC! = 1
!GGG = 41	G!GG = 129	GG!G = 18	GGG! = 35	!GGT = 86	G!GT = 139	GG!T = 5	GGT! = 1
!GTA = 27	G!TA = 5	GT!A = 4	GTA! = 0	!GTC = 42	G!TC = 12	GT!C = 2	GTC! = 1
!GTG = 104	G!TG = 51	GT!G = 0	GTG! = 19	!GTT = 29	G!TT = 18	GT!T = 0	GTT! = 0
!TAA = 0	T!AA = 0	TA!A = 0	TAA! = 0	!TAC = 12	T!AC = 11	TA!C = 3	TAC! = 3
!TAG = 0	T!AG = 0	TA!G = 0	TAG! = 0	!TAT = 7	T!AT = 6	TA!T = 0	TAT! = 8
!TCA = 2	T!CA = 1	TC!A = 1	TCA! = 4	!TCC = 4	T!CC = 12	TC!C = 1	TCC! = 7
!TCG = 0	T!CG = 1	TC!G = 5	TCG! = 16	!TCT = 4	T!CT = 6	TC!T = 0	TCT! = 9
!TGA = 0	T!GA = 0	TG!A = 0	TGA! = 0	!TGC = 11	T!GC = 4	TG!C = 4	TGC! = 1
!TGG = 4	T!GG = 3	TG!G = 29	TGG! = 13	!TGT = 9	T!GT = 8	TG!T = 3	TGT! = 2
!TTA = 3	T!TA = 2	TT!A = 0	TTA! = 0	!TTC = 22	T!TC = 5	TT!C = 0	TTC! = 4
!TTG = 10	T!TG = 4	TT!G = 1	TTG! = 7	!TTT = 11	T!TT = 1	TT!T = 2	TTT! = 22

Totals[b]
!*** = 1,368 *!** = 1,128 **!* = 628 ***! = 1368

[a] The symbol "!" indicates a breaking point; an intron is present at this location.

[b] Total numbers of insertion patterns, wherein !*** stands for intron insertion at the codon boundary, *!** for insertion after the first base position, and **!* for insertion after the second base position.

Results and Discussion

Intron-Insertion Patterns

Table 1 shows insertion patterns of 3,116 primate introns for each of the 64 codons. The symbol "!" indicates an insertion point (an intron is present at this location). At the bottom of the table, total numbers of insertion patterns are shown, where !*** stands for intron insertion at the codon boundary, *!** for insertion after the first base position, and **!* for insertion after the second base position.

It is not unexpected that each individual codon has a completely different pattern of splicing site preference, because splicing sites and their surrounding positions must accommodate sequence constraints posed by spliceosomes. Some codons such as CTC and AAG, however, have a particularly strong preference. We shall get back to this point later in this section.

The totals at the bottom of the table tell us that: (1) codon boundaries (!***) are the most preferred insertion points and (2) insertion points after the first base positions (*!**) are more preferred than those after the second positions (**!*). This observation holds for primates, rodents, mammals, vertebrates, invertebrates, and plants, as shown in table 2. It is thus natural to believe that some selectional force must have existed in the course of evolution.

The first point (frequent insertion at codon boundaries) can make some sense if we accept the exon shuffling hypothesis (Gilbert 1978), which states that introns play an important role in efficient evolution by allowing the shuffling of exons and, thus, rearrangements of genes far more effectively than without introns. In this model, exons that end at codon boundaries would be evolutionarily advantageous, since two such exons can be smoothly concatenated without causing a frame shift. On the other hand, if we accept the selfish DNA hypothesis (Cavalier-Smith 1985), which states that most

Table 2
Percentage of Each Intron-Insertion Position for Each Taxonomical Group

	!***	*!**	**!*
Primates	43.9	36.2	19.9
Rodents	44.5	38.9	16.6
Mammals	43.4	38.4	18.2
Vertebrates	50.7	32.8	16.5
Invertebrates	48.6	28.8	22.6
Plants	55.2	24.6	20.2

NOTE.—!*** stands for insertion at the codon boundary; *!** for insertion after the first base position, and **!* for insertion after the second base position.

Table 3
Profile of 3,116 Primate Splicing Sites

	← Exon →			← Intron →					← Intron →					← Exon →		
A	261	253	346	589	82	0	0	473	715	53	145	248	189	81	78	63	82	244	32	1,000	0	232	214	221	219	244
C	241	298	367	127	32	0	0	27	78	50	176	225	303	412	431	465	408	320	754	0	0	155	208	273	314	296
G	254	281	182	148	804	1,000	0	473	129	847	206	352	254	120	85	66	63	218	1	0	1,000	522	245	249	255	205
T	243	167	103	134	80	0	1,000	25	76	48	471	173	252	385	403	405	445	215	211	0	0	89	330	255	210	253

NOTE.—Introns that do not follow the "GT-AG rule" were excluded from the analysis. 1,000 = 100%. Vertical line between arrows indicates intron–exon boundary.

Table 4
Intron-Insertion Distribution for AAG and CTC

	!..AAG	!.AAG	!AAG	A!AG	AA!G	AAG!	AAG.!	AAG..!	!..CTC	!.CTC	!CTC	C!TC	CT!C	CTC!	CTC.!	CTC..!
OBS	46	14	13	9	43	348	27	22	18	57	109	0	0	0	8	20
EXP	31	38	39	15	79	128	40	37	53	52	44	2	5	5	12	46
CHI**2	7	15	17	2	16	378	4	6	22	0	93	2	5	5	1	15

NOTE.—The symbol "!" in the table indicates a splicing site, and the symbol "." stands for any nucleotide. Thus "!..AAG" indicates the number of splicing sites located two bases upstream of an AAG codon. OBS = observed number; EXP = expected number; CHI**2 = χ^2 value. Some entries, such as AAG! and !CTC, have unusually high numbers.

Table 5
Occurrences of Nine Different Types of Exons for Each of the Six Taxonomic Groups of Organisms

	Primates	
00 = 470	01 = 211	02 = 129
10 = 231	11 = 336	12 = 152
20 = 133	21 = 127	22 = 70
	Rodents	
00 = 323	01 = 167	02 = 108
10 = 180	11 = 234	12 = 112
20 = 108	21 = 97	22 = 37
	Mammals	
00 = 64	01 = 28	02 = 16
10 = 20	11 = 74	12 = 19
20 = 20	21 = 15	22 = 8
	Vertebrates	
00 = 182	01 = 72	02 = 63
10 = 73	11 = 90	12 = 38
20 = 63	21 = 32	22 = 18
	Invertebrates	
00 = 589	01 = 322	02 = 256
10 = 267	11 = 244	12 = 161
20 = 292	21 = 151	22 = 143
	Plants	
00 = 936	01 = 302	02 = 315
10 = 410	11 = 207	12 = 171
20 = 286	21 = 148	22 = 121

NOTE.—Both 5′ and 3′ ends of exons are classified into three categories (0, 1, 2), based on their codon phase. For instance, exon type 00 represents an exon whose 5′ and 3′ ends are at a codon boundary, whereas exon type 11 represents an exon whose 5′ end has two extra bases beyond a codon boundary (upstream) and whose 3′ end has one extra base downstream.

introns were transposed and inserted into exons, then it would be hard to explain why introns prefer certain codon positions to jump into.

The second point, the cases of introns breaking apart a codon (*!** and **!*), makes much less sense. Since third positions of codons, in general, carry less information, it would be logical to think that an exon would prefer to be broken apart (by an intron) after second base positions (**!*) rather than after first base positions (*!**). However, this is apparently not the case, as is shown in table 1. Indeed, introns prefer location

Table 6
Exon Length Classes: Modulus of 3

	3*N*	3*N* + 1	3*N* + 2
Primates	876 (47.1%)	496 (26.7%)	487 (26.2%)
Rodents	594 (43.5%)	387 (28.3%)	385 (28.2%)
Mammals	146 (55.3%)	67 (25.4%)	51 (19.3%)
Vertebrates	290 (46.0%)	173 (27.4%)	168 (26.6%)
Invertebrates	976 (40.2%)	775 (32.0%)	674 (27.8%)
Plants	1,264 (43.6%)	759 (26.2%)	873 (30.1%)

NOTE.—The first column in the table represents the total numbers of exons whose lengths are 3*N*, that is, divisible by 3. This can be readily obtained by adding the numbers of type 00, type 11, and type 22 exons. The second and third columns similarly represent the numbers of exons whose lengths are 3*N* + 1 (type 01 + type 12 + type 20) and 3*N* + 2 (type 02 + type 10 + type 21), respectively.

Table 7
All of the 27 Possible Pair Types of Two Consecutive Exons

00#00 = 1,439 (1,339)*	00#01 = 442 (508)	00#02 = 383 (416)
01#10 = 431 (369)*	01#11 = 292 (358)	01#12 = 193 (187)
02#20 = 379 (368)*	02#21 = 222 (238)	02#22 = 181 (174)
10#00 = 525 (586)	10#01 = 287 (222)*	10#02 = 179 (182)
11#10 = 316 (404)	11#11 = 497 (392)*	11#12 = 189 (205)
12#20 = 229 (254)	12#21 = 198 (164)*	12#22 = 112 (120)
20#00 = 401 (439)	20#01 = 168 (166)	20#02 = 173 (136)*
21#10 = 224 (196)	21#11 = 152 (190)	21#22 = 110 (99)*
22#20 = 180 (165)	22#21 = 91 (107)	22#22 = 80 (78)*

NOTE.—For example, 20#01 indicates a pair of an exon of type 20 and an exon of type 01. Numbers in parentheses following observed numbers are expected numbers based on frequencies of left exon types and right exon types, treated independently. The table shows that pairs of exons whose total lengths are multiples of 3 (entries indicated by an asterisk) are observed more frequently than expected.

after the first base positions of a codon (*!**). The reason is not clear and yet to be explained.

Two codons, AAG and CTC, in table 1 have a particularly interesting distribution, and deserve special attention. We first constructed a profile of splicing sites of all the primate introns used in our analysis (table 3). We then computed, based on the profile, expected intron-insertion patterns for those two codons, as shown in table 4. The symbol "!" in the table indicates a splicing site, and the symbol "." stands for any nucleotide. Thus, "!..AAG" indicates the number of splicing sites located two bases upstream of an AAG codon. In the row labeled "OBS" are observed numbers and in that labeled "EXP" are expected numbers. "CHI**2" stands for χ^2 value. Some entries, such as AAG! and !CTC, have unusually high numbers. It remains to be seen whether those are merely due to GenBank data biases or there are other unknown causes.

Exon Patterns

Both 5′ and 3′ ends of exons can be classified into three categories (0, 1, 2), based on the intron phases with respect to reading frames (phase 0 for codon boundaries, phase 1 for the first codon position, and phase 2 for the second codon position). For instance, exon type 00 represents an exon whose 5′ and 3′ ends are at a codon boundary, whereas exon type 11 represents an exon whose 5′ end has two extra bases beyond a codon boundary (upstream) and whose 3′ end has one extra base downstream. Occurrences of nine different types of exons are counted for each of the six taxonomic groups of organisms and summarized in table 5. Based on these data, table 6 further classifies exons by their lengths divided by 3. The first column in table 5 represents the total number of exons whose length is $3N$, that is, divisible by 3. This can be readily obtained by adding the numbers of type 00, type 11, and type 22 exons. The second and third columns similarly represent the number of exons whose length is $3N + 1$ (type 01 + type 12 + type 20) and $3N + 2$ (type 02 + type 10 + type 21), respectively.

The results show that exons of length $3N$ are preferred by all primates, rodents, mammals, vertebrates, invertebrates, and plants. This observation is, again, consistent with the exon shuffling hypothesis because those exons, when inserted in a gene, would not cause a frame shift and thus would be evolutionarily advantageous. It is, therefore, easier to imagine that exons, not introns, have been moving around the genome in the course of evolution.

This view is further supported by another result shown in table 7. All of the 27 possible pair types of two consecutive exons are shown in the table. For example, 20#01 indicates a pair of an exon of type 20 and an exon of type 01. (There is a certain type constraint when two exons are concatenated. You cannot have 12#02, for instance. That is why there are not 81 possible pair types.) Numbers in parentheses are expected numbers based on frequencies of left exon types and right exon types, treated independently. A total of 8,073 pairs of exons from primates, rodents, mammals, vertebrates, invertebrates, and plants together were used for the analysis. The table shows that pairs of exons whose total length is a multiple of 3 (entries indicated by an asterisk) are observed more frequently than expected. All the observations above are consistent with the recently published results of Long, Rosenberg, and Gilbert (1995).

Acknowledgments

The concept of selection of exons based on properties of their ends with respect to codon boundaries was derived from earlier work by Lee Altenberg (personal communication).

LITERATURE CITED

BERGET, S. M. 1995. Exon recognition in vertebrate splicing. J. Biol. Chem. **270**(6):2411–2414.

BLAKE, C. C. F. 1978. Do genes-in-pieces imply proteins-in-pieces. Nature **273**:267.

CAVALIER-SMITH, T. 1985. Selfish DNA and the origin of introns. Nature **315**:283–284.

GILBERT, W. 1978. Why genes in pieces? Nature **271**:501.

KERSANACH, R., H. BRINKMANN, M. F. LIAUD, D. X. ZHANG, W. MARTIN, and R. CERFF. 1993. Five identical intron positions in ancient duplicated genes of eubacterial origin. Nature **367**:387–389.

LOGSDON, J. M. JR., J. D. PALMER, A. STOLTZFUS, R. CERFF, W. MARTIN, and H. BRINKMANN. 1994. Origin of introns—early or late? Nature **369**:526–528.

LONG, M., C. ROSENBERG, and W. GILBERT. 1995. Intron phase correlations and the evolution of the intron/exon structure of genes. Proc. Natl. Acad. Sci. USA **92**:12495–12499.

MATTICK, J. S. 1994. Introns: evolution and function. Curr. Opin. Genet. Dev. **4**:823–831.

ROGER, A. J., and F. DOOLITTLE. 1993. Why introns-in-pieces. Nature **364**:289–290.

SENAPATHY, P. 1986. Origin of eukaryotic introns: a hypothesis, based on codon distribution statistics, and its implications. Proc. Natl. Acad. Sci. USA **83**:2133–2137.

STANLEY SAWYER, reviewing editor

Accepted July 30, 1996