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

Masaru Tomita
Department of Environmental Information
and Department of Molecular Biology
Keio University, 5322 Endo, Fujisawa, 252, JAPAN
Phone: +81-3-3440-4083, Fax: +81-3-3440-7281, mt@sfc.keio.ac.jp
Nobuyoshi Shimizu
Department of Molecular Biology, Keio University
and
Doug Brutlag
Department of Biochemistry, Stanford University

Corresponding Author: Masaru Tomita

Keywords: Sequence Analysis, Intron, Exon, Splicing

Running Head: Introns and Reading Frames

Abstract

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 taxonomical groups: primates, rodents, mammals, vertebrates, invertebrates and plants.
We found that reading frames are interrupted by an intron at a codon boundary (as oppose 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 middle of codons, they exist in between the first and second bases much more frequently than 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 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 the recently published results 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 are to be determined by the protein synthesis process, cannot be 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 that there exists correlation between splicing sites and their codon positions in reading frame. In particular, we show:

- 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 pairs of adjacent exons, are a multiple of 3 more frequently then 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 (Gilbert 1978; Blake 1978; Senapathy 1986; Kersanach et al. 1993; Logsdon 1994; Roger and Doolittle 1993; Cavalier-Smith 1985; 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.

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 of 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 length (24, 120, 40, 20, 65, 61) and entry B consists of exons of length (33, 54, 81, 24, 120, 40), then the parts with three exons (24, 120, 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 the three exons and the four introns, was also tried, but essentially the same results were obtained.) The cases of two non-homologous parts having same-length exons three times in a row by chance is quite rare and therefore ignorable.
Note that this screening method can exclude not only homologous entries, but also intra-genetic homologous exons. For instance, some genes, such as collagen genes, have many homologous exons lined up within the same gene; and only one of those exons will be counted in our analysis, as they all have the same length.
All the computer programs have been written in C programming language, and running on Sparc stations under the UNIX operating system. The software is available from the authors by request.

Results and Discussion

Intron Insertion Patterns

Table 1 shows insertion patterns of 3116 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 nearby 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 total numbers at the bottom of the table tell us that: (1) codon boundaries `!***` are the most preferred insertion points, and (2) after the first base positions `*!**` are more preferred than 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 have been playing 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 of 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 introns were transposed and inserted into exons, then it would be hard to explain why introns prefer a 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 more logical to think that an exon prefers to be broken apart (by an intron) at the location after second base positions (**!*) rather than after first base positions (*!**). However, it is apparently not the case, as shown in Table 1. Introns indeed prefer location 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 at 2 bases upstream of an *aag* codon. Rows labeled "OBS" are observed numbers and "EXP" expected numbers. "CHI**2" stands for χ^2 value.
Some entries, such as `aag!` and `!ctc` have unusually high numbers. It is yet to see 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 3 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 exons whose 5' and 3' ends are right at a codon boundary; whereas exon type 11 represents exons whose 5' end has two extra bases beyond a codon boundary (up stream) and whose 3' end has one extra base down stream. Occurrences of 9 different types of exons are counted for each of the six taxonomical groups of organisms and summarized in Table 5.
Based on these data, Table 6 further classifies exons by their length 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 of 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 does not exit 81 possible pair types.) Numbers in blankets are expected numbers based on frequencies of left exon types and right exon types, treated independently. A total of 8073 pairs of exons from primates, rodents, mammals, vertebrates, invertebrates and plants altogether 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 by Long, Rosenberg and Gilbert (1995).

Acknowledgements

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**. The Journal of Biological Chemistry 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. Current Opinion in Genetics and Development 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.
Reviewing Editor: STANLEY SAWYER

Table 1
Intron-insertion patterns of 3116 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
TOTAL
!***=1368 *!**=1128 **!*=628 ***!=1368  (b)
------------------------------------------------------------------------
```

[a]The symbol "!" indicates a breaking point; an intron is present at this location.
[b]Total numbers of insertion patterns, where `!***` stands for intron insertion at the codon boundary, `*!**` for insertion after the first base position, and `**!*` for insertion after the second base position.

Table 2
Intron-insertion positions for each taxonomical groups

	!***	*!**	**!* (a)
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%

[a] `!***` 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 3116 Primate Splicing Sites[a]

TOTAL=3116; 1000 = 100%

	-------exon---------->						<----------intron---------					
A	261	253	346	589	82	0	0	473	715	53	145	248	189
C	241	298	367	127	32	0	0	27	78	50	176	225	303
G	254	281	182	148	804	1000	0	473	129	847	206	352	254
T	243	167	103	134	80	0	1000	25	76	48	471	173	252

	----------intron---------->						<---------exon----------					
A	81	78	63	82	244	32	1000	0	232	214	221	219	244
C	412	431	465	408	320	754	0	0	155	208	273	314	296
G	120	85	66	63	218	1	0	1000	522	245	249	255	205
T	385	403	405	445	215	211	0	0	89	330	255	210	253

[a]Introns that do not follow the "GT-AG rule" were excluded from the analysis.

Table 4
Intron-insertion Distribution[a] for *aag* and *ctc*

Sites	!..aag	!.aag	!aag	a!ag	aa!g	aag!	aag.!	aag..!
OBS	46	14	13	9	43	348	27	22
EXP	31	38	39	15	79	128	40	37
CHI**2	7	15	17	2	16	378	4	6

Sites	!..ctc	!.ctc	!ctc	c!tc	ct!c	ctc!	ctc.!	ctc..!
OBS	18	57	109	0	0	0	8	20
EXP	53	52	44	2	5	5	12	46
CHI**2	22	0	93	2	5	5	1	15

[a]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 at 2 bases upstream of an `aag` codon. Rows labeled "OBS" are observed numbers and "EXP" expected numbers. "CHI**2" stands for χ^2 value. Some entries, such as `aag!` and `!ctc` have unusually high numbers.

Table 5
Occurrences of 9 different types of exons for each of the six taxonomical groups of organisms[a]

```
[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
```

[a]Both 5' and 3' ends of exons are classified into 3 categories (0, 1, 2), based on their codon phase. For instance, exon type 00 represents exons whose 5' and 3' ends are right at a codon boundary; whereas exon type 11 represents exons whose 5' end has two extra bases beyond a codon boundary (up stream) and whose 3' end has one extra bases down stream.

Table 6
Exon Length Classes[a]: Modulus of 3

	3N	3N+1	3N+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	1264 (43.6%)	759 (26.2%)	873 (30.1%)

[a]The first column in the table 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.

Table 7
All of the 27 possible pair types of two consecutive exons[a]

	OBS	EXP		OBS	EXP		OBS	EXP	
00#00=	1439	(1339)	*	00#01=	442	(508)	00#02=	383	(416)
01#20=	431	(369)	*	01#21=	292	(358)	01#22=	193	(187)
02#10=	379	(368)	*	02#11=	222	(238)	02#12=	181	(174)
10#00=	401	(439)		10#01=	168	(166)	10#02=	173	(136) *
11#20=	224	(196)		11#21=	152	(190)	11#22=	110	(99) *
12#10=	180	(165)		12#11=	91	(107)	12#12=	80	(78) *
20#00=	525	(586)		20#01=	287	(222) *	20#02=	179	(182)
21#20=	316	(404)		21#21=	497	(392) *	21#22=	189	(205)
22#10=	229	(254)		22#11=	198	(164) *	22#12=	112	(120)

For example, 20#01 indicates a pair of an exon of type 20 and an exon of type 01. Numbers in blankets 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 length is a multiple of 3 (entries indicated by an asterisk) are observed more frequently than expected.