

On biased distribution of introns in various eukaryotes

A. Sakurai^{a,b}, S. Fujimori^{a,b}, H. Kochiwa^{a,b}, S. Kitamura-Abe^{a,c}, T. Washio^a, R. Saito^{a,d},
The RIKEN Genome Exploration Research Group Phase II Team

P. Carninci^d, Y. Hayashizaki^d, M. Tomita^{a,b,e,*}

^aInstitute for Advanced Biosciences, Keio University, Tsuruoka, 997-0035, Japan

^bBioinformatics Program, Graduate School of Media and Governance, Keio University, Fujisawa, 252-8520, Japan

^cDepartment of Applied Physics, Graduate School of Engineering, Hokkaido University, Kita 13 Nishi 8, Kita-ku, Sapporo, 060-8628, Japan

^dLaboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama-city, Kanagawa, 230-0045, Japan

^eDepartment of Environmental Information, Keio University, Fujisawa, 252-8520, Japan

Received 14 December 2001; received in revised form 28 July 2002; accepted 18 September 2002

Abstract

We conducted comprehensive analyses on intron positions in the *Mus musculus* genome by comparing genomic sequences in the GenBank database and cDNA sequences in the mouse cDNA library recently developed by Riken Genomic Sciences Center. Our results confirm that introns have a tendency to be located toward the 5' end of the gene. The same type of analysis was conducted in the coding region of seven eukaryotes (*Saccharomyces cerevisiae*, *Plasmodium falciparum*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus*, *Homo sapiens*, *Arabidopsis thaliana*). Introns in genes with a single intron have a locational bias toward the 5' end in all species except *A. thaliana*. We also measured the distance from the start codon to the position of the intron, and found that single introns prefer the location immediately after the start codon in *S. cerevisiae* and *P. falciparum*. We discuss three possible explanations for these findings: (1) they are the consequence of intron loss by reverse-transcriptase; (2) they are necessary to accommodate the function; and (3) they are concerned with the mechanism of pre-mRNA splicing. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Position of intron; Single intron; Intron loss; Regulatory sequences

1. Introduction

The length and position characteristics of introns have been investigated from various aspects. The positions of introns were reported to show a correlation with protein structure (Gilbert and Glynias, 1993; Go, 1981; Go and Nosaka, 1987), though Weber and Kabsch (1994) reported that intron positions in actin genes are unrelated to the secondary structure of the protein. Sahrawy et al. (1996) suggested that the intron position in thioredoxin genes could be a useful marker of evolution. In addition, it has been proposed that introns are hot spots for genetic recombination and that exon shuffling has been a major factor in protein evolution (Kolkman and Stemmer, 2001). On the

other hand, Long et al. (1995) and Tomita et al. (1996) independently showed that intron positions have a correlation with codon frames. Vinogradov (1999) recently found that the size of introns is correlated with their genome size.

The work presented in this paper was inspired by the report that introns in *Saccharomyces cerevisiae* have a tendency to be located towards the 5' end of the gene (Fink, 1987). As yet, no clear explanation for this phenomenon has been discovered, although Kriventseva and Gelfand (1999) suggested that this characteristic might be related to the fact that first introns are anomalously long.

In this study, we conducted a comprehensive analysis of intron positions by comparing genomic sequences in the GenBank database and cDNA sequences in the RIKEN mouse cDNA library. In addition, we also conducted similar analyses in the coding region of seven eukaryotes (*S. cerevisiae*, *Plasmodium falciparum*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus*, *Homo*

* Corresponding author. Institute for Advanced Biosciences, Keio University, 5322 Endo, Fujisawa-city, Kanagawa, 252-8520, Japan. Tel./fax: +81-466-47-5099.

E-mail address: mt@sfc.keio.ac.jp (M. Tomita).

Table 1
Data sets used

Species	Genes with introns (A)	Genes with one intron (B)	B/A (%)
<i>S. cerevisiae</i>	276	268	97.1
<i>P. falciparum</i>	191	119	62.3
<i>C. elegans</i>	14,946	1432	9.6
<i>D. melanogaster</i>	12,096	3240	26.8
<i>M. musculus</i>	1288	317	24.6
<i>H. sapiens</i>	3551	652	18.4
<i>A. thaliana</i>	8664	1571	18.1

sapiens, *Arabidopsis thaliana*) and especially in the genes with only one intron.

2. Materials and methods

We used the 21,076 RIKEN full-length cDNA sequences of *M. musculus* (<http://genome.gsc.riken.go.jp>) (The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium, 2001). We also downloaded genomic sequences from the National Center for Biotechnology Information (NCBI) (<ftp://ncbi.nlm.nih.gov/>). For comparative studies, we used complete genome sequences of *S. cerevisiae*, *P. falciparum* (chromosomes 2 and 3), *C. elegans*, and *A. thaliana* (chromosomes 1, 2 and 4) and GenBank release 122.0 entries for *D. melanogaster*, *M. musculus*, and *H. sapiens*. CLEANUP version 1.8.3 (Grillo et al., 1996) was run to eliminate redundant sequences/entries (Table 1).

The gene pairs of the cDNA sequences and the genomic

sequences of *M. musculus* were first screened by BLAST (Altschul et al., 1990) (<ftp://ncbi.nlm.nih.gov/blast/executables>) using the following criteria.

- There are at least two matching regions, each of which has a sequence similarity of 95% or higher.

Second, those selected gene pairs were further aligned by CLUSTAL-W (Thompson et al., 1994; Higgins et al., 1996) (<http://www.no.embnet.org/>) to detect introns using the following criteria.

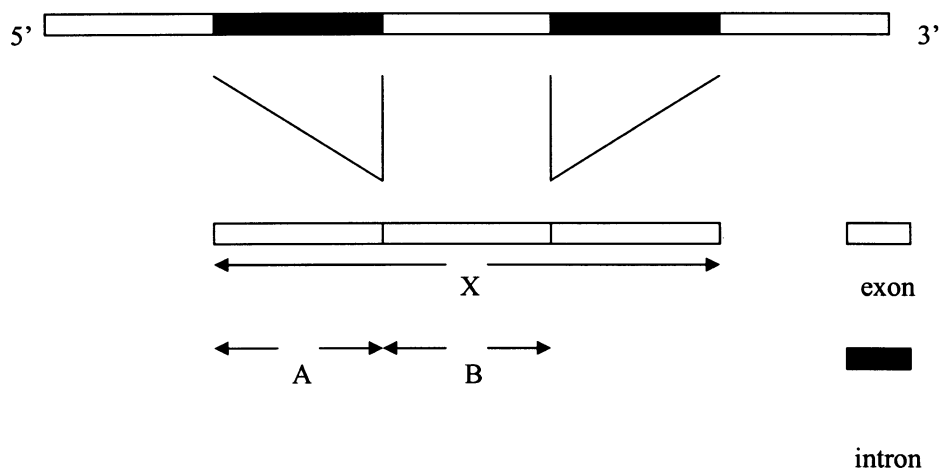
- All matching regions must have more than 95% sequence similarity.

Because introns can be presented as points in cDNA sequences, we define the ‘relative position’ of an intron as follows: the distance (bp) from the 5′ end of cDNA divided by the length (bp) of the cDNA (Fig. 1).

3. Results

3.1. The relative position of introns

First, we measured the relative position of introns in *M. musculus* mRNA (Fig. 2). Our result presented in Fig. 2 confirms that introns have a tendency to be located toward the 5′ end. This tendency may be due to the length of the 3′ UTR, because it is known that the length of the 3′ UTR is comparatively long and there are few introns in the 3′ UTR (Hawkins, 1988).



$$\text{The relative position of the first intron} = \frac{A}{X}$$

$$\text{The relative position of the second intron} = \frac{A+B}{X}$$

Fig. 1. The relative position of the intron.

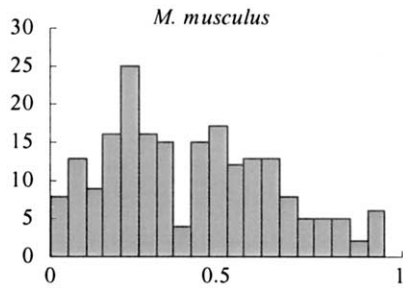


Fig. 2. Locational distribution of introns in mRNA. Horizontal axis: relative position in mRNA. 0 indicates the position of the 5' end, and 1.0 the position of the 3' end. Vertical axis: the number of introns located at the position.

For the sake of comparison, we also conducted similar analyses for six other eukaryotes without considering 5' UTR and 3' UTR. Given that the sequences used for this analysis are all from the GenBank database, the locational distribution of introns was analyzed only for the coding

region. We found that introns in the coding region have a tendency to be located toward the 5' end in *S. cerevisiae* and *P. falciparum* (Fig. 3). Given that most introns in *S. cerevisiae* (97.1%) and *P. falciparum* (62.3%) are single introns (the gene contains only one intron) (Table 1), the locational bias may be due solely to those single introns. In order to clarify this, the same type of analysis was conducted using only genes with a single intron, and the results for the seven species are shown in Fig. 4. Our results confirmed that single introns have a general tendency to be located toward the start codon.

3.2. Distance from the start codon

We also measured distances from the start codon to the intron in genes whose coding sequences are more than 500 bps long, containing only a single intron (Fig. 5). The results confirmed that the single introns have a tendency to be

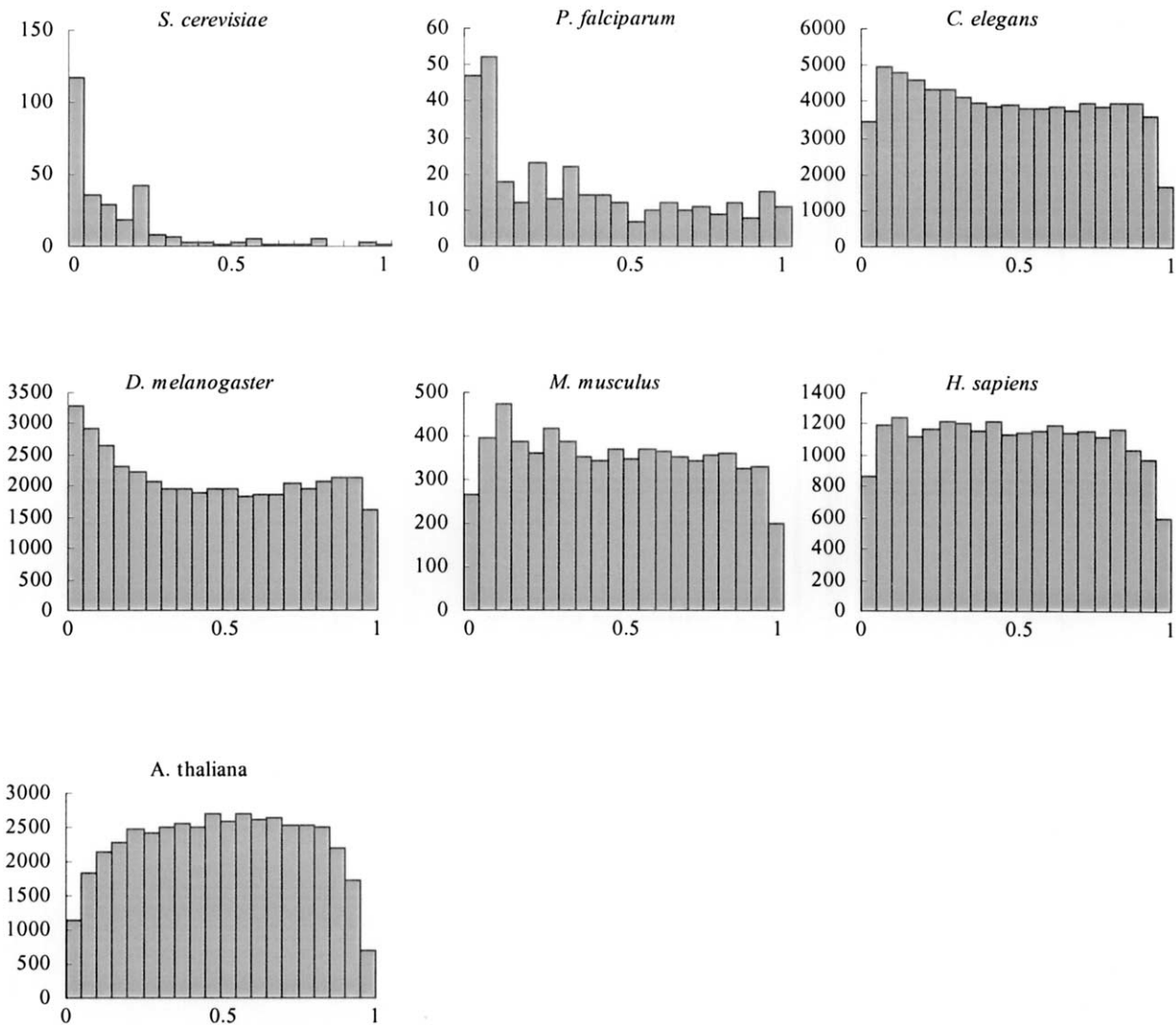


Fig. 3. Locational distribution of introns in coding sequence (all genes). Horizontal axis: relative position within the coding region. 0 indicates the position of the start codon, and 1.0 the position of the stop codon. Vertical axis: the number of introns located at the position.

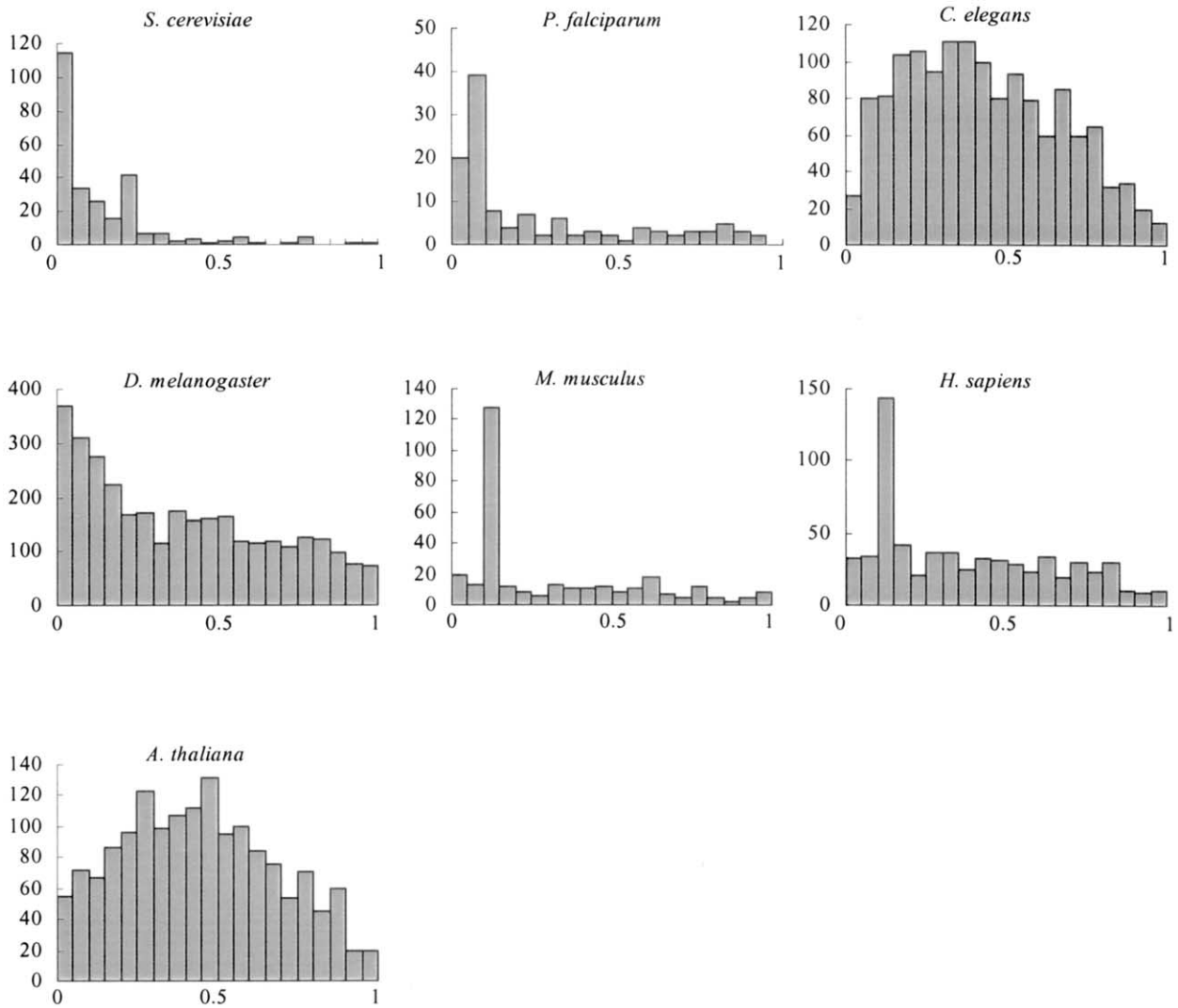


Fig. 4. Locational distribution of introns with genes containing only one intron. Horizontal axis: relative position within the coding region. 0 indicates the position of the start codon, and 1.0 the position of the stop codon. Vertical axis: the number of introns located at the position.

located near the start codon in all of the species investigated except *A. thaliana*. In *S. cerevisiae* and *P. falciparum*, those single introns in particular prefer to locate immediately after the start codon.

4. Discussion

We have confirmed that the introns in the gene with a single intron have a tendency to reside near the 5' end in the following six species: *S. cerevisiae*, *P. falciparum*, *C. elegans*, *D. melanogaster*, *M. musculus*, and *H. sapiens*. We consider three possible explanations for this phenomenon.

First, it might result from the mechanism of intron loss. It has been suggested that introns in the genomic DNA can be deleted through homologous recombination with a cDNA produced from an mRNA of the gene by reverse transcriptase (Derr and Strathern, 1993). If homologous recombination with cDNA is the major force of intron loss,

then introns located near the 3' end are likely to be lost more frequently than those near the 5' end; all reverse transcription begins at the 3' poly (A) tract, and many of the cDNAs would not extend to the 5' end. This scenario was previously suggested by Fink (1987) as a possible cause of the unusual distribution of single introns in the genes of *S. cerevisiae*. Our results indicate that the same mechanism of intron loss may exist in many other species, and could cause the asymmetric distribution of their single introns, given that most of the species are known to possess reverse transcriptase in their transposons (e.g. copia in *D. melanogaster* (Mount and Rubin, 1985) and Ty in *S. cerevisiae* (Garfinkel et al., 1985)).

Second, the single introns that locate near the start codon may be biologically more important, given that they sometimes contain the function related to the transcription. Lai et al. (1995, 1997) showed that the single intron in Zfp-36, the gene encoding the putative zinc finger protein tristetraprolin, contains the transcriptional regulatory

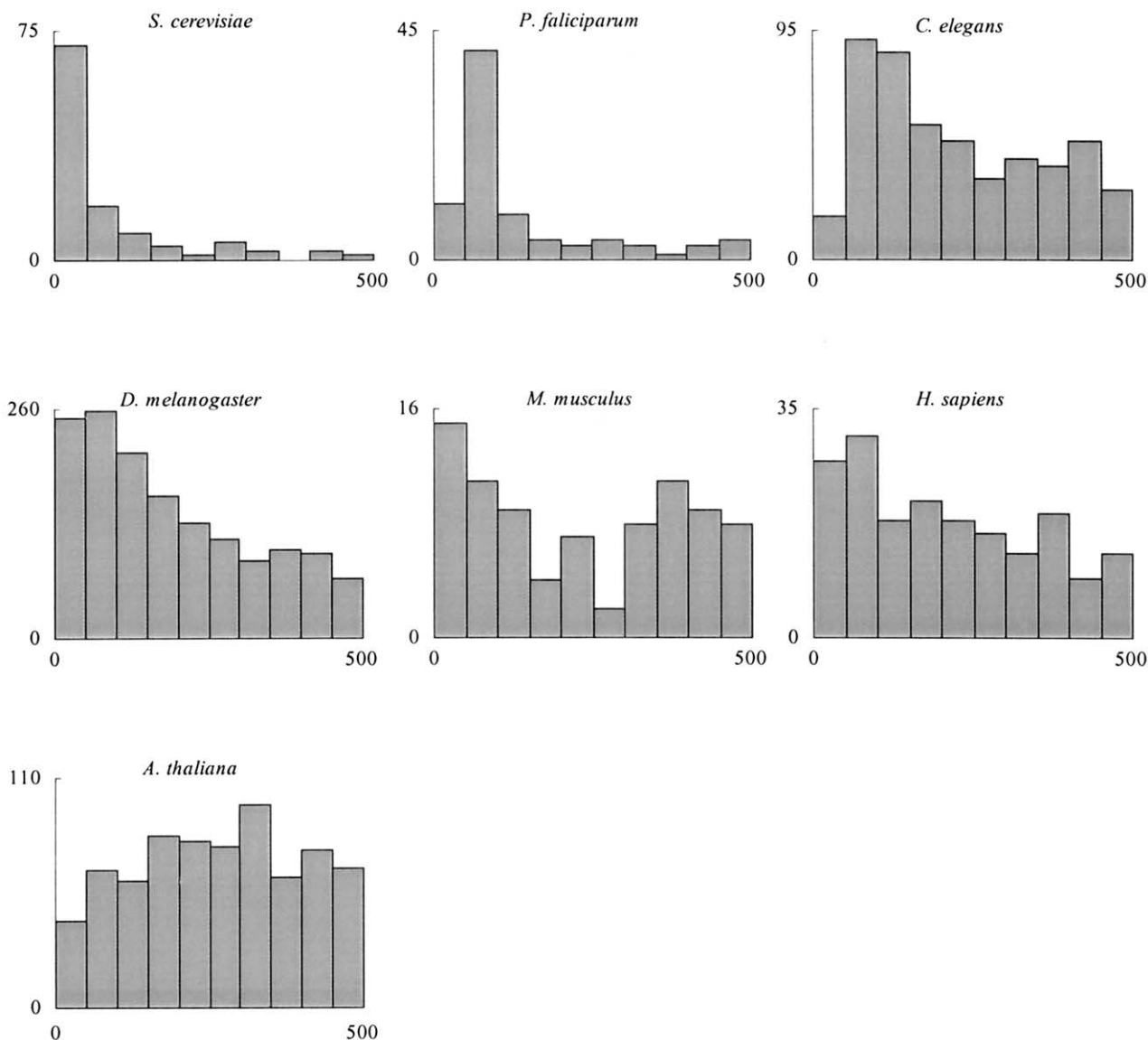


Fig. 5. Location of single introns in terms of distance from the start codon. Vertical axis: the number of introns located at the position. Horizontal axis: the distance (bps) from the start codon to the intron. The number of genes whose coding sequences are more than 500 bps long (the number of first introns that are located in the first 500 bp in these genes) are as follows: *S. cerevisiae*, 184 (118); *P. falciparum*, 106 (83); *C. elegans*, 608 (493); *D. melanogaster*, 2361 (1475); *M. musculus*, 136 (84); *H. sapiens*, 314 (185); *A. thaliana*, 1128 (709).

elements such as NFB-like binding element and Sp1 binding motifs. They also concluded that the intron location contributes significantly to the regulated expression of this gene. Similarly, the first introns in the genes carrying multiple introns sometimes contain the same function. [Tourmente et al. \(1993\)](#), for example, found that enhancer and silencer elements exist within the first intron of the $\beta 3$ tubulin gene in *Drosophila* Kc cells. In addition, this first intron has additional hormone-dependent negative and positive regulatory elements, which can act in both directions and in a position-independent manner ([Bruhat et al., 1990](#)). [Kolb et al. \(1998\)](#) found that the first intron of the human growth hormone (hGH-N) gene contains a novel eukaryotic promoter element. They also demonstrated that this element could activate a promoterless luciferase

reporter gene and thus functions as an independent eukaryotic promoter element. We confirmed that the single intron of the Zfp-36 gene and these first introns of the $\beta 3$ tubulin gene and the hGH-N gene were near the start codon ([Fig. 6](#)). Likewise, introns near the start codon may often contain the function and they may be more likely to be conserved. This could explain why most single introns tend to be located near the start codon.

Third, it might result from the mechanism of pre-mRNA splicing. The exon definition model ([Berget, 1995](#); [Robberston et al., 1990](#)) accounts for the pairing between the 3' splice site and its downstream 5' splice site in vertebrate splicing. Exon definition suggests that terminal exons, both first and last exons, will require special mechanisms for their recognition. First exons can be recognized via interactions

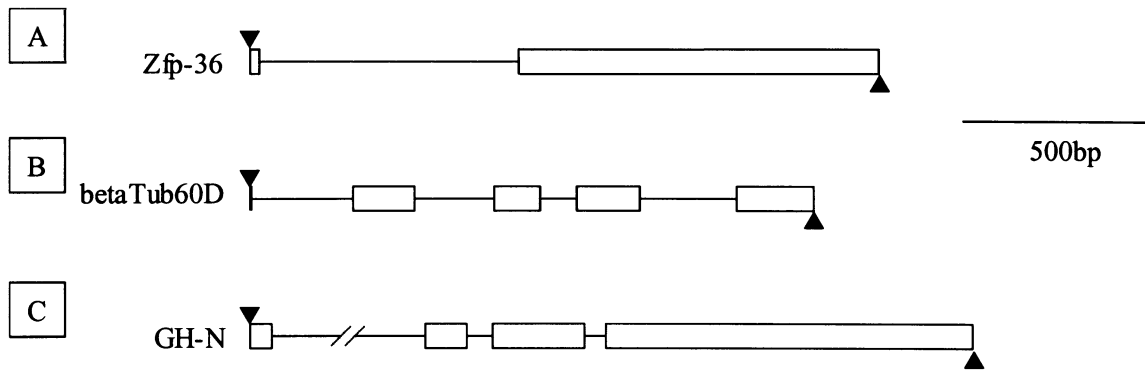


Fig. 6. The intron/exon structures. The boxes represent exons, and the lines represent introns. The positions of the start codon (▼) and the stop codon (▲) are shown.

between the factors that recognize caps and 5' splice sites. The cap and nuclear proteins that bind the cap are essential for splicing of the gene with a single intron (Izaurrealde et al., 1994). Factors recognizing 3' splice sites can interact with factors recognizing poly (A) sites to recognize last exons. Last exons are often the largest exon (Brunak et al., 1991; Hawkins, 1988). Based on the above, we suggested that the introns in the gene with a single intron consequently have a tendency to reside near the 5' end. Simpson and Filipowicz (1996) reported that the process of pre-mRNA splicing in plants exhibits significant differences from that in yeast or mammals, which might explain the different patterns between *A. thaliana* and the other species.

We believe that additional analyses of the genome sequences of many organisms will yield clearer answers to questions regarding the biological and evolutionary significance of introns.

Acknowledgements

We are grateful to Yoshihito Niimura for useful discussion. This study was supported by a research grant for the RIKEN Genome Exploration Research Project from the Science and Technology Agency of the Japanese Government, CREST (Core Research for Evolutional Science and Technology) and ACT-JST (Research and Development for Applying Advanced Computational Science and Technology) of the Japan Science and Technology Corporation (JST) to Y.H. This work was also supported in part by a grant from the Ministry of Agriculture, Forestry and Fisheries of Japan (Rice Genome Project SY-1104).

References

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. *J. Mol. Biol.* 215 (3), 403–410.
- Berget, S.M., 1995. Exon recognition in vertebrate splicing. *J. Biol. Chem.* 270 (6), 2411–2414.
- Bruhat, A., Tourmente, S., Chapel, S., Sobrier, M.L., Couderc, J.L., Dastugue, B., 1990. Regulatory elements in the first intron contribute to transcriptional regulation of the beta 3 tubulin gene by 20-hydroxyecdysone in *Drosophila* Kc cells. *Nucleic Acids Res.* 18 (10), 2861–2867.
- Brunak, S., Engelbrecht, J., Knudsen, S., 1991. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* 220 (1), 49–65.
- Derr, L.K., Strathern, J.N., 1993. A role for reverse transcripts in gene conversion. *Nature* 361 (6408), 170–173.
- Fink, G.R., 1987. Pseudogenes in yeast? *Cell* 49 (1), 5–6.
- Garfinkel, D.J., Boeke, J.D., Fink, G.R., 1985. Ty element transposition: reverse transcriptase and virus-like particles. *Cell* 42 (2), 507–517.
- Gilbert, W., Glynias, M., 1993. On the ancient nature of introns. *Gene* 135 (1–2), 137–144.
- Go, M., 1981. Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature* 291 (5810), 90–92.
- Go, M., Nosaka, M., 1987. Protein architecture and the origin of introns. *Cold Spring Harbor Symp. Quant. Biol.* 52, 915–924.
- Grillo, G., Attimonelli, M., Liuni, S., Pesole, G., 1996. CLEANUP: a fast computer program for removing redundancies from nucleotide sequence databases. *Comput. Appl. Biosci.* 12 (1), 1–8.
- Hawkins, J.D., 1988. A survey on intron and exon lengths. *Nucleic Acids Res.* 16 (21), 9893–9908.
- Higgins, D.G., Thompson, J.D., Gibson, T.J., 1996. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* 266, 383–402.
- Izaurrealde, E., Lewis, J., McGuigan, C., Jankowska, M., Darzynkiewicz, E., Mattaj, I.W., 1994. A nuclear cap binding protein complex involved in pre-mRNA splicing. *Cell* 78 (4), 657–668.
- Kolb, A.F., Gunzburg, W.H., Brem, G., Erfle, V., Salmons, B., 1998. A functional eukaryotic promoter is contained within the first intron of the hGH-N coding region. *Biochem. Biophys. Res. Commun.* 247 (2), 332–337.
- Kolkman, J.A., Stemmer, W.P., 2001. Directed evolution of proteins by exon shuffling. *Nat. Biotechnol.* 19 (5), 423–428.
- Kriventseva, E.V., Gelfand, M.S., 1999. Statistical analysis of the exon-intron structure of higher and lower eukaryote genes. *J. Biomol. Struct. Dyn.* 17 (2), 281–288.
- Lai, W.S., Thompson, M.J., Taylor, G.A., Liu, Y., Blackshear, P.J., 1995. Promoter analysis of Zfp-36, the mitogen-inducible gene encoding the zinc finger protein tristetraprolin. *J. Biol. Chem.* 270 (42), 25266–25272.
- Lai, W.S., Thompson, M.J., Blackshear, P.J., 1997. Characteristics of the intron involvement in the mitogen-induced expression of Zfp-36. *J. Biol. Chem.* 273 (1), 506–517.
- Long, M., Rosenberg, C., Gilbert, W., 1995. Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc. Natl. Acad. Sci. USA* 92 (26), 12495–12499.
- Mount, S.M., Rubin, G.M., 1985. Complete nucleotide sequence of the

- Drosophila* transposable element copia: homology between copia and retroviral proteins. *Mol. Cell. Biol.* 5 (7), 1630–1638.
- Robberson, B.L., Cote, G.J., Berget, S.M., 1990. Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol. Cell. Biol.* 10 (1), 84–94.
- Sahrawy, M., Hecht, V., Lopez-Jaramillo, J., Chueca, A., Chartier, Y., Meyer, Y., 1996. Intron position as an evolutionary marker of thioredoxins and thioredoxin domains. *J. Mol. Evol.* 42 (4), 422–431.
- Simpson, G.G., Filipowicz, W., 1996. Splicing of precursors to mRNA in higher plants: mechanism, regulation and sub-nuclear organisation of the spliceosomal machinery. *Plant Mol. Biol.* 32 (1–2), 1–41.
- The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium, 2001. Functional annotation of a full-length mouse cDNA collection. *Nature* 409 (6821), 685–690.
- Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22 (22), 4673–4680.
- Tomita, M., Shimizu, N., Brutlag, D.L., 1996. Introns and reading frames: correlation between splicing sites and their codon positions. *Mol. Biol. Evol.* 13 (9), 1219–1223.
- Tourmente, S., Chapel, S., Dreau, D., Drake, M.E., Bruhat, A., Couderc, J.L., Dastugue, B., 1993. Enhancer and silencer elements within the first intron mediate the transcriptional regulation of the beta 3 tubulin gene by 20-hydroxyecdysone in *Drosophila* Kc cells. *Insect Biochem. Mol. Biol.* 23 (1), 137–143.
- Vinogradov, A.E., 1999. Intron-genome size relationship on a large evolutionary scale. *J. Mol. Evol.* 49 (3), 376–384.
- Weber, K., Kabsch, W., 1994. Intron positions in actin genes seem unrelated to the secondary structure of the protein. *EMBO J.* 13 (6), 1280–1286.