J Mol Evol (2001) 52:2–5
DOI: 10.1007/s002390010128

JOURNAL OF MOLECULAR EVOLUTION

Intron Length and Codon Usage

Alexander E. Vinogradov

Institute of Cytology, Russian Academy of Sciences, Tikhoretsky Ave. 4, St. Petersburg 194064, Russia

Received: 7 December 1999 / Accepted: 10 May 2000

Abstract. The correlation was shown between the length of introns and the codon usage of the coding sequences of the corresponding genes, which in some cases can be related to the level of gene expression. The link is positive in the unicellular organisms, i.e., genes with the longer introns show the higher bias of codon usage. It is most pronounced in baker's yeast, where it is definitely related to the level of gene expression—genes with the higher level of expression have the longer introns. The correlation is inverted in multicellular organisms as compared to unicellular ones. Some organisms, however, do not show the link. The presence or absence of the link does not seem to be related to the GC percent of the coding sequences.

Key words: Genome evolution — Noncoding DNA — Junk DNA — Intervening sequence — Codon usage — Codon bias — Gene expression — Unicellular — Multicellular — Baker's yeast

The role of introns in eukaryotic genes remains enigmatic. In some cases they were reported to encode for small RNA molecules involved in splicing (Santoro et al. 1994; Caffarelli et al. 1998) or participate in the gene regulation at the pretranscriptional (Bhattacharyya and Banerjee 1999; Kawada et al. 1999) or post-transcriptional level (Santoro et al. 1994; Barta and Iggo 1995). They may also facilitate exon shuffling at genetic recombination (Hall et al. 1989; Long et al. 1995; de Souza et al. 1998). No general relation has so far been demonstrated between any intron parameters and the informational content of a given gene. Moreover, in the mouse–rat

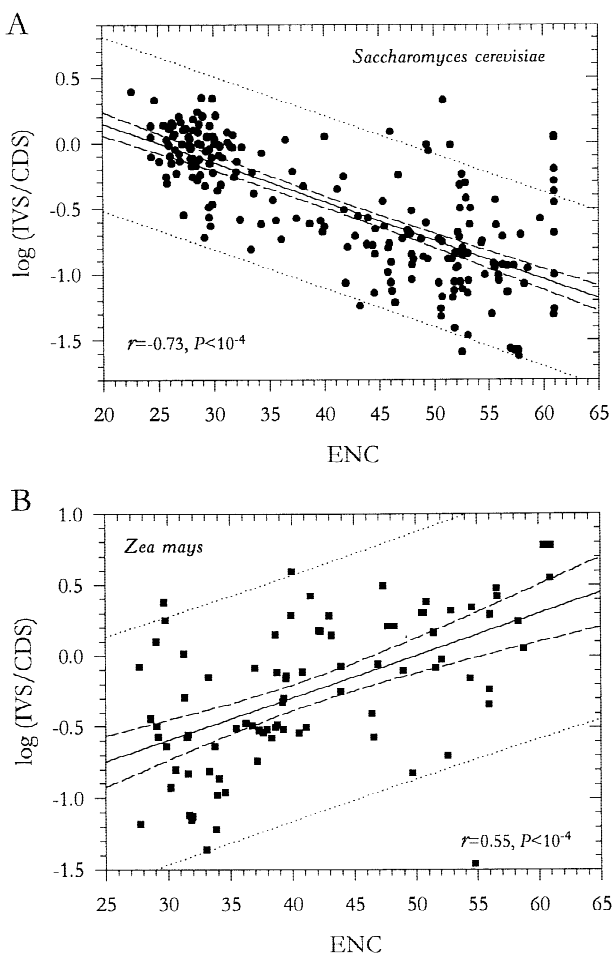

Fig. 1. The relationship between the ratio of intervening to coding sequence lengths (IVS/CDS) and the effective number of codons (ENC) in a unicellar (**A**) and a multicellular (**B**) organisms. Correlation coefficients are Pearson product-moment.

Table 1. Coefficients of Spearman rank correlation between the frequency of optimal codons (FOP) and effective number of codons (ENC) and the within-gene average internal intron length and ratio of intervening to coding sequence lengths (IVS/CDS)

Organism	Number of genes analyzed	GC3s% of CDS mean CV, %	Frequency of optimal codons (FOP)		Effective number of codons (ENC)	
			Avg. intron length	IVS/ CDS	Avg. intron length	IVS/ CDS
Saccharomyces cerevisiae	224	39.0	0.61	0.69	−0.59	−0.72
(baker's yeast)		17.1	($< 10^{-4}$)	($< 10^{-4}$)	($< 10^{-4}$)	($< 10^{-4}$)
Schizosaccharomyces pombe (fission yeast)	833	30.9	—	—	−0.14	−0.16
		15.4			($< 10^{-4}$)	($< 10^{-4}$)
Emericella nidulans (mold)	138	60.7	0.05	0.38	−0.07	−0.32
		12.8	(< 0.6)	($< 10^{-4}$)	(< 0.4)	($< 10^{-3}$)
Neurospora crassa (fungus)	133	70.6	0.15	0.36	−0.14	−0.35
		12.0	(< 0.1)	($< 10^{-4}$)	(< 0.2)	($< 10^{-3}$)
Candida albicans	17	25.4	—	—	−0.82	−0.54
(pathogenic yeast)		26.6			($< 10^{-3}$)	(< 0.03)
Dictyostelium discoideum	96	16.1	0.36	0.18	0.04	−0.23
(cellular slime mold)		42.6	($< 10^{-3}$)	(< 0.1)	(< 0.7)	(< 0.03)
Tetrahymena thermophila	18	34.2	—	—	−0.54	−0.40
(ciliate)		36.3			(< 0.03)	(< 0.1)
Chlamydomonas reinhardtii	45	88.0			−0.25	0.36
(unicellular plant)		5.7			(< 0.1)	(< 0.02)
Caenorhabditis elegans	8,447	37.7	−0.04	−0.08	0.14	0.17
(nematode)		23.6	($< 10^{-3}$)	($< 10^{-4}$)	($< 10^{-4}$)	($< 10^{-4}$)
Xenopus laevis	53	50.1			0.37	0.29
(clawed frog)		16.6			(< 0.01)	(< 0.05)
Gallus gallus	97	68.1	—	—	0.34	0.35
(chicken)		24.3			($< 10^{-3}$)	($< 10^{-3}$)
Mus musculus	455	63.1	—	—	−0.04	0.08
(mouse)		16.3			(> 0.4)	(< 0.08)
Rattus norvegicus	250	62.3			0.19	0.30
(rat)		17.0			(< 0.01)	($< 10^{-4}$)
Homo sapiens	785	65.7	—	—	0.26	0.28
(human)		22.2			($< 10^{-4}$)	($< 10^{-4}$)
Arabidopsis thaliana	1,730	40.2	—	—	0.12	0.04
(thale cress)		15.1			($< 10^{-4}$)	(< 0.12)
Zea mays	91	76.6	—	—	0.26	0.55
(corn)		23.4			(< 0.02)	($< 10^{-4}$)

Significance levels are shown in parentheses; the empty cells are because optimal codons are not known for the corresponding species.
Sequences were extracted from GenBank (only the complete ones). Genes were checked for duplicates on the ground of gene names. The intron/exon boundaries were taken from annotations.

comparison, no correlation was found between the rates of nucleotide substitutions in the intervening and the coding DNA (Hughes and Yeager 1997).

The bias in codon usage (an unequal use of different codons for the same amino acid) is supposed to relate to the abundance of the corresponding isoacceptor tRNAs and the efficiency of translation (Sharp et al. 1995; Xia 1996), as well as to some other factors (Karlin and Mrazek 1996). In several organisms it was shown that the optimal codons occur most frequently in the highly expressed genes (Sharp et al. 1995; Chiapello et al. 1998). A number of parameters are used to estimate the codon bias: frequency of optimal codons (FOP), codon bias index (CBI), and codon adaptation index (CAI) (Ikemura 1981; Bennetzen and Hall 1982; Sharp and Li 1987). In those organisms where the optimal codons are not known, the effective number of codons (ENC) can be used as estimate of codon bias (Wright 1990). The lower ENC, the higher codon bias is.

Codon usage bias is positively correlated with the intron length or relative intron length (ratio of intervening to coding sequence lengths) in some unicellular organisms and negatively in some multicellular ones (Fig. 1, Table 1; correlation of CAI and CBI, where available, was qualitatively the same as of FOP). The most pronounced link was observed in baker's yeast (Fig. 1A). In this organism, intron length is definitely related to the FOP (Table 1) and the level of gene expression (Fig. 2). It should also be stressed that for this organism nearly all genes containing introns were analyzed, so there is no sampling error.) It is mostly ribosomal proteins that constitute a group of highly expressed genes with the longer introns. It can be suggested that their introns may play a role in the positive regulation of these genes as it was

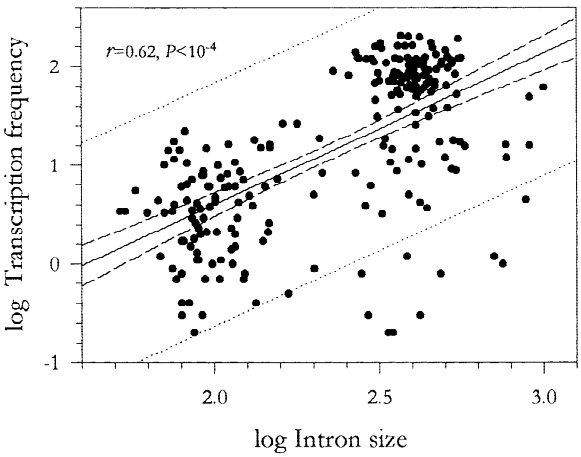

Fig. 2. The relationship between intron length and transcription frequency of the corresponding genes (i.e., the number of transcripts produced per unit of time) determined by the microarray experiments in the baker's yeast. The data on the transcription frequencies are taken from the work by Holstege et al. (1998) using the yeast intron database (Lopez and Seraphin 2000).

shown for some higher eukaryote genes (Bhattacharyya and Banerjee 1999; Kawada et al. 1999).

It is unclear why some organisms show a statistically significant link when others do not. The case of the mouse versus rat is noteworthy (Table 1). These species are close relatives, and the samples of their genes show similar means and coefficients of variation for GC percent of the third silent codon position (Table 1). It was shown that the rate of synomymous substitutions is related to the GC percent of silent positions (Alvarez-Valin et al. 1998) and of the third codon position (Alvarez-Valin et al. 1999). It is also unclear what parameter—average intron length or the ratio of intervening to coding sequence length—is a better expression of the character under study.

The occurrence of the positive link between intron length and the level of gene expression is especially intriguing (Table 1, Fig. 2). If introns are a useless "junk" just tolerated by selection, then the negative link between the intron length and the level of gene expression could be explained by a shift of balance between the selection and drift. However, the greater size of introns in the highly expressed genes of some unicellular organisms suggests a functional role for introns (as, for instance, participation in the positive feedback of gene regulation). Then the negative link between the intron length and the level of gene expression in the multicellular organisms can be explained by the hypothesis that introns are necessary for correct chromatin structure (Zuckerkandl 1981, 1997; Trifonov 1993). It was shown that in mammals and chicken the longer introns occur in the light isochores (Duret et al. 1995), which reside in the late-replicating and more condensed chromatin (Federico et al. 1998). The codon usage is known to be linked to the DNA curvature, the intergenic regions are more curved than the genic ones and random sequences (Jauregui et al. 1998). Therefore, introns may serve to increase the ability of genic DNA to bend. The reverse of the link at the evolutionary transition from the unicellular to the multicellular level may be connected to the appearance of a tighter chromatin condensation caused by the need to permanently turn off the tissue-specific genes. Thus, the maximum chromatin condensation is fivefold lower in yeast as compared to mammals (Russell and Nurse 1986). The intervening sequences are also significantly shorter in the unicellular organisms (Vinogradov 1999). It is possible that interplay of both effects—the necessity for longer introns in the more condensed chromatin and the direct participation of introns in gene regulation—leads to the apparent absence of the overall link between intron length and codon bias in some organisms.

Acknowledgment. This work was supported by a grant from the Russian Foundation for Basic Research (RFBR).

References

Alvarez-Valin F, Jabbari K, Bernardi G (1998) Synonymous and nonsynonymous substitutions in mammalian genes: intragenic correlations. J Mol Evol 46:37–44

Alvarez-Valin F, Jabbari K, Carels N, Bernardi G (1999) Synonymous and nonsynonymous substitutions in genes from Gramineae: intragenic correlations. J Mol Evol 49:330–42

Barta I, Iggo R (1995) Autoregulation of expression of the yeast Dbp2p 'DEAD-box' protein is mediated by sequences in the conserved DBP2 intron. EMBO J 14:3800–3808

Bennetzen JL, Hall BD (1982) Codon selection in yeast. J Biol Chem 257:3026–3031

Bhattacharyya N, Banerjee D (1999) Transcriptional regulatory sequences within the first intron of the chicken apolipoproteinAI (apoAI) gene. Gene 234:371–380

Caffarelli E, Losito M, Giorgi C, Fatica A, Bozzoni I (1998) In vivo identification of nuclear factors interacting with the conserved elements of box C/D small nucleolar RNAs. Mol Cell Biol 18:1023–1028

Chiapello H, Lisacek F, Caboche M, Henaut A (1998) Codon usage and gene function are related in sequences of *Arabidopsis thaliana.* Gene 209:GC1–GC38

de Souza SJ, Long M, Klein RJ, Roy S, Lin S, Gilbert W (1998) Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. Proc Natl Acad Sci USA 95:5094–5099

Duret L, Mouchiroud D, Gautier C (1995) Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J Mol Evol 40:308–317

Federico C, Saccone S, Bernardi G (1998) The gene-richest bands of human chromosomes replicate at the onset of the S-phase. Cytogenet Cell Genet 80:83–88

Hall DH, Liu Y, Shub DA (1989) Exon shuffling by recombination between self-splicing introns of bacteriophage T4. Nature 340:575–576

Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA (1998) Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95:717–728

Hughes AL, Yeager M (1997) Comparative evolutionary rates of introns and exons in murine rodents. J Mol Evol 45:125–130

Jauregui R, O'Reilly F, Bolivar F, Merino E (1998) Relationship be-

tween codon usage and sequence-dependent curvature of genomes. Microb Comp Genomics 3:243–253

Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* system. J Mol Biol 151:389–409

Karlin S, Mrazek J (1996) What drives codon choices in human genes? J Mol Biol 262:459–472

Kawada N, Moriyama T, Ando A, Koyama T, Hori M, Miwa T, Imai E (1999) Role of intron 1 in smooth muscle alpha-actin transcriptional regulation in activated mesangial cells in vivo. Kidney Int 55:2338–2348

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 92:12495–12499

Lopez PJ, Seraphin B (2000) YIDB: the yeast intron database. Nucl Acids Res 28:85–86

Russell P, Nurse P (1986) *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae:* a look at yeast divided. Cell 45:781–782

Santoro B, De Gregorio E, Caffarelli E, Bozzoni I (1994) RNA-protein interactions in the nuclei of *Xenopus* oocytes: complex formation and processing activity on the regulatory intron of ribosomal protein gene L1. Mol Cell Biol 14:6975–6982

Sharp PM, Li WH (1987) The codon adaptation index a measure of directional synonymous codon usage bias, and its potential applications. Nucl Acids Res 15:1281–1295

Sharp PM, Averof M, Lloyd AT, Matassi G, Peden JF (1995) DNA sequence evolution: the sounds of silence. Phil Trans R Soc Lond B Biol Sci 349:241–247

Trifonov EM (1993) Spatial separation of overlapping messages. Comp Chem 117:27–31

Vinogradov AE (1999) Intron-genome size relationship on a large evolutionary scale. J Mol Evol 49:376–384

Wright F (1990) The effective number of codons used in a gene. Gene 87:23–29

Xia X (1996) Maximizing transcription efficiency causes codon usage bias. Genetics 144:1309–1320

Zuckerkandl E (1981) A general function of noncoding polynucleotide sequences. Mol Biol Rep 7:149–158

Zuckerkandl E (1997) Junk DNA and sectorial gene repression. Gene 205:323–343