Note

Natural Selection Affects Frequencies of AG and GT Dinucleotides at the 5′ and 3′ Ends of Exons

S. T. Eskesen, F. N. Eskesen[1] and A. Ruvinsky[2]

Institute for Genetics and Bioinformatics, University of New England, Armidale, New South Wales 2351, Australia

Manuscript received November 13, 2003
Accepted for publication January 16, 2004

ABSTRACT

GT and AG, located at the 5′ and 3′ ends of introns, are important for correct splicing. It is anticipated that natural selection decreases frequency of AG and GT near the 5′ and 3′ ends of exons, preventing appearance of cryptic splicing sites. The data presented in this article support the expectation.

IT is common knowledge that GT and AG dinucleotides, located at the 5′ and 3′ ends of introns, respectively, constitute an important part of the donor and acceptor splice sites. These sites are highly conserved and essential for correct splicing (Burset *et al.* 2000). Mutations in these sites inevitably lead to severe disruptions of normal splicing (Nissim-Rafinia and Kerem 2002). Still a question can be asked whether the presence of AG at the 5′ ends of nonfirst exons may confuse identification of the last 3′ AG of preceding introns, which are known to be involved in splicing. A similar question may apply to GT located at the 3′ ends of nonlast exons, which in turn may confuse recognition of the first 5′ intronic GT also essential in splicing. If the answer to both questions is positive, one may expect that natural selection would affect 5′ and 3′ ends of exons by reducing frequencies of AG and GT in the vicinity of the intron-exon boundaries. To test this hypothesis we collected and compared data concerning observed and expected frequencies of AG in the 5′ ends and GT in the 3′ ends of exons in three model species as well as conducted several other independent studies. It is, however, clear from numerous publications that the role of AG and GT in determining the splicing point is not exclusive and that there are other factors affecting the process (Fairbrother *et al.* 2002; Maniatis and Tasic 2002).

[1]*Present address:* T. J. Watson Research Center, 19 Skyline Dr., Hawthorne, NY 10532.

[2]*Corresponding author:* Institute for Genetics and Bioinformatics, University of New England, Armidale, NSW 2351, Australia.
E-mail: aruvinsk@metz.une.edu.au

DISTRIBUTION OF AG PAIRS AT 5′ ENDS OF EXONS

Information relevant to *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens* was extracted from the exon-intron database (EID), which was compiled in the W. Gilbert laboratory, Department of Molecular and Cellular Biology, Harvard University (Saxonov *et al.* 2000). The database contains protein-coding intron-containing genes and is available on the Internet at http://www.mcb.harvard.edu/gilbert/eid/. From the version of the database that we used, the following data were extracted: *C. elegans*, 14,836 genes and 98,581 exons; *D. melanogaster*, 13,361 genes and 58,801 exons; *H. sapiens*, 7150 genes and 47,908 exons. A program was written to align exons at the 5′ or 3′ ends and to search sequences within exons. The program gives an option to ignore the first or the last exon in a gene. The program is also capable of ignoring all exons except the first or the last. Expected frequencies of nucleotide pairs were calculated using the formula $fr_{AG} = fr_{A/n} \times fr_{G/n+1}$, where $fr_{A/n}$ is a frequency of A in position n from the beginning of the exon and $fr_{G/n+1}$ is a frequency of G in position $n + 1$ from the beginning of the exon. A similar procedure was used for GT pairs. As we show, 5′ and 3′ alignments of different and numerous exons negate possible codon usage biases (our unpublished data) and we believe that this factor does not affect results discussed in this article.

Pictograms representing frequencies of different nucleotides at the first 10 5′ positions of exons are shown in Figure 1. It is to be expected (Lewin 1994) that while the first position at the 5′ exonic end is predominantly occupied by A or G and the second position by mainly T or A (or G in humans) other positions starting from the third or fourth positions do not have strong bias

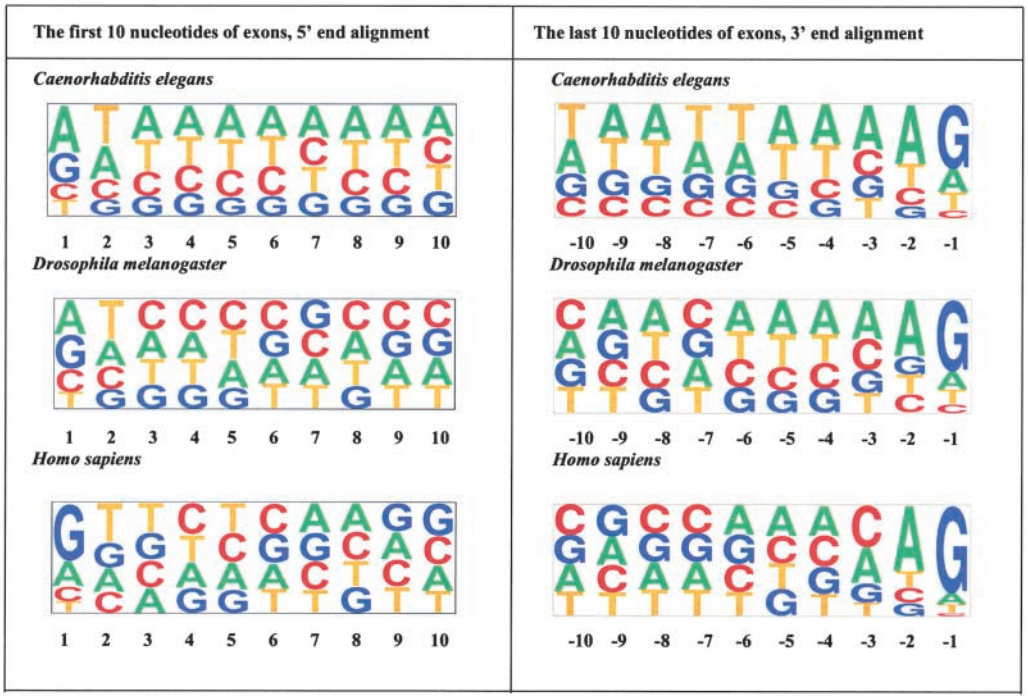

FIGURE 1.—Pictograms of the first 10 nucleotides at the 5′ ends and the last 10 nucleotides at the 3′ ends of exons in *C. elegans*, *D. melanogaster*, and *H. sapiens*. The first and last exons of each gene were excluded from this analysis. Size of a letter on a pictogram is proportional to frequency of relevant nucleotide. Pictograms were built by a program developed by C. Burge, which is available from http://genes.mit.edu/pictogram.html.

toward certain nucleotides in the studied species. Indeed no obvious similarities between *C. elegans*, *D. melanogaster*, and *H. sapiens* can be found beyond the first two positions.

To study the problem posed in the title of this article, we calculated observed and expected frequencies of AG pairs in the 5′ part of exons in the compared species. Figure 2 represents the distributions of AG frequencies starting from the 5′ end along nonfirst exons. The first exons were excluded from this count as they do not have the preceding intron-exon boundary and thus differ from all the rest. Periodic variations of AG frequencies, which can be seen in all distributions, are discussed in a separate article. As can be seen in Figure 2 the first five positions in the three compared species differ from the following positions. Taking this fact into account we compared observed and expected frequencies of AG pairs on the intervals 6–30 and 101–125 (not shown in Figure 2) positions. In the nonfirst exons of *C. elegans* expected frequencies are significantly higher within both compared intervals (Table 1), while the intensity of the differences slightly declines in the 3′ direction. Correlation ($r = -0.65$; $P < 0.0001$) between AG position in the exons and the difference between expected and observed values was found. In *D. melanogaster* significant differences are found in the first interval (positions 6–30) but not in the second (positions 101–125; Table 1). Correlation ($r = -0.40$; $P < 0.0001$) between AG position in the *D. melanogaster* exons and the difference between expected and observed values was also observed. Thus, the nonfirst exons of *C. elegans* and *D. melanogaster* demonstrate significantly lower than expected AG frequencies in the positions adjacent to the 5′ end of the exon and the difference diminishes downstream. Interestingly the observed frequencies of GA pairs (data not shown) are significantly higher than expected (for *C. elegans*, $t = 14.751$, $P = 0$ and for *D. melanogaster*, $t = 6.155$, $P = 0$, in the first 50 positions) and much higher than observed frequencies of AG pairs. One may presume that selective pressure against AG pairs could contribute to the observed phenomenon.

On the contrary the first exons do not have a preceding intron-exon junction and were not expected to show specific selection pressure against AG. Table 2 presents comparisons between observed and expected frequencies of AG on the same intervals as is shown above for nonfirst exons. In *C. elegans* and *D. melanogaster* no differences were observed in either interval in the first exons. This observation supports the idea that the preceding intron-exon junction likely contributes to selection against exonic AG in the vicinity of splicing site in the nonfirst exons.

In *H. sapiens* the situation seems to be different. Observed and expected frequencies of AG practically do not differ from expected in the first four positions, after which the pattern changes and observed frequencies of AG are higher than expected. Obviously there is a disparity between *H. sapiens* and the two other studied species. What could be a rational interpretation of this difference? One cannot rule out that splicing mechanisms in *H. sapiens* slightly differ and specific selection against AG typical for *C. elegans* and *D. melanogaster* might be much lower or even disappear in *H. sapiens* exons. It is also possible that in *H. sapiens* additional factors might be involved, which could mask possible selection pressure against AG. For instance, we found

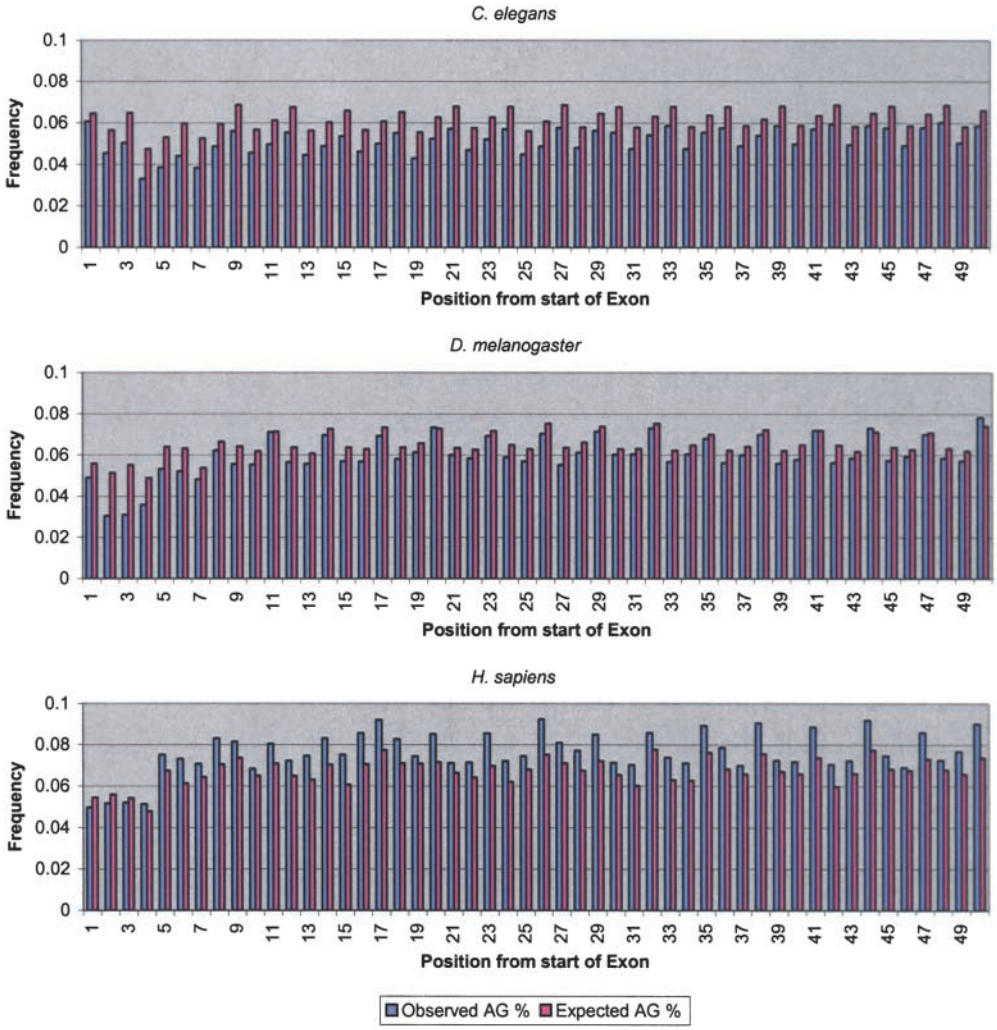

FIGURE 2.—Expected and observed frequencies of AG pairs at the 5′ ends of nonfirst exons in *C. elegans*, *D. melanogaster*, and *H. sapiens*. Blue bars represent observed and red bars represent expected frequencies of AG in certain positions of exons aligned at the 5′ ends. Numbers on the *x*-axis indicate position of nucleotide A in AG pairs from the start of an exon. Observed frequencies were calculated as proportion of AG among all possible 16 duplets in a certain position. Expected frequencies were calculated using the formula in the text. Exon-intron database was obtained from http://www.mcb.harvard.edu/gilbert/eid. Graphs were created using MS Excel's chart wizard tools.

that in *H. sapiens* exons observed frequency of CG is dramatically lower (~200% less, data not shown) than expected, while in the other compared species this was not the case. As frequencies of dinucleotides are interrelated, one may guess that an increased level of observed AG in *H. sapiens* could be a compensation for low frequency of CG and some other dinucleotides in exons. In any case the difference of *H. sapiens* from the two other species is apparent.

Distributions of AG frequencies shown in Figure 2 represent a mixture of three phases (0, 1, and 2) in each species. Phase separation of any of these distribu-

TABLE 1

Comparisons of differences between observed and expected frequencies of AG at the 5′ ends of the nonfirst exons and GT at the 3′ ends of the nonlast exons

	AG pairs at the 5′ ends of exons, difference between observed and expected frequencies			GT pairs at the 3′ ends of exons, difference between observed and expected frequencies[a]		
Species	Positions	*t* statistic	*P* value	Positions	*t* statistic	*P* value
C. elegans	6–30	7.97930	0.00000	−6 to −30	13.70754	0.00000
	101–125	4.89402	0.00001	−101 to −125	7.55123	0.00000
D. melanogaster	6–30	2.89362	0.00588	−6 to −30	11.30163	0.00000
	101–125	0.00461	0.99634	−101 to −125	10.26395	0.00000
H. sapiens	6–30	−6.2786	0.00000	−6 to −30	12.38340	0.00000
	101–125	−7.06285	0.00000	−101 to −125	10.10876	0.00000

[a] Positions were counted from the 3′ ends of exons upstream.

TABLE 2

Comparisons of differences between observed and expected frequencies of AG at the 5′ ends of the first exons and GT at the 3′ ends of the last exons

	AG pairs at the 5′ ends of exons, difference between observed and expected frequencies			GT pairs at the 3′ ends of exons, difference between observed and expected frequencies[a]		
Species	Positions	*t* statistic	*P* value	Positions	*t* statistic	*P* value
C. elegans	6–30	1.38884	0.17121	−6 to −30	2.83826	0.00685
	101–125	1.10813	0.27310	−101 to −125	2.15865	0.03713
D. melanogaster	6–30	0.06501	0.94844	−6 to −30	2.85932	0.00687
	101–125	−0.25218	0.80203	−101 to −125	1.76302	0.08578
H. sapiens	6–30	−2.51296	0.01526	−6 to −30	4.05915	0.00018
	101–125	−1.57308	0.12268	−101 to −125	2.14499	0.03746

The *t*-test statistics were calculated using the two-sample calculator from http://calculators.stat.ucla.edu/twosamp/.

[a] Positions were counted from the 3′ ends of exons upstream.

tions into three distributions (phase 0, phase 1, and phase 2) reveal the same pattern, which is more sharply expressed (data not shown).

In an attempt to further test the selection hypothesis we compared frequencies of pairs of synonymous codons located at the 5′ end of exons (phase 0) in the three studied species, which differ in the third position. The compared codons are AAA and AAG (lysine), CAA and CAG (glutamine), and GAA and GAG (glutamic acid). If natural selection really operates against AG pairs located near the splice junction (5′ end of exons), one can also expect selection against AG-carrying co-

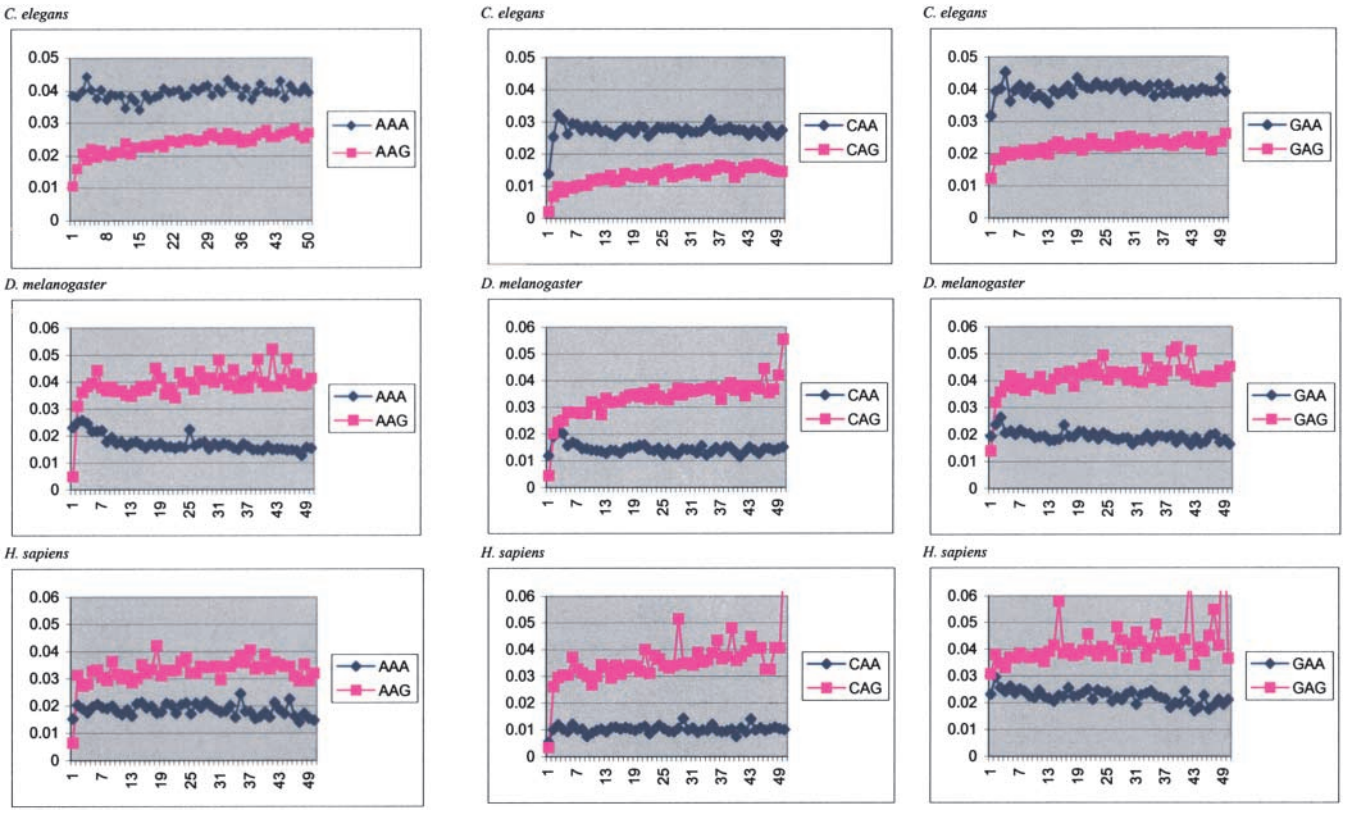

FIGURE 3.—Comparisons of frequencies of synonymous codons differing in the third position at the 5′ end of exons (phase 0) in the three studied species. The compared codons are AAA and AAG (lysine), CAA and CAG (glutamine), and GAA and GAG (glutamic acid). Numbers on the *x*-axis indicate position of codons starting from the 5′ end of exons. Exon-intron database was obtained from http://www.mcb.harvard.edu/gilbert/eid. Graphs were created using MS Excel's chart wizard tools.

TABLE 3

Comparisons of AG differences (expected − observed) between three codon positions

Comparisons between codon positions	*C. elegans*	*D. melanogaster*	*H. sapiens*
First "deleterious" − second "less deleterious"	$t = 8.31901$	$t = 11.456397$	$t = 3.67899$
	$P = 0.000001$[a]	$P = 0.000000$[a]	$P = 0.001500$[a]
Third "deleterious" − second "less deleterious"	$t = 3.528592$	$t = 11.04588$	$t = -0.234401$
	$P = 0.003452$[a]	$P = 0.000000$[a]	$P = 0.816946$

Exons starting from phase 0 (phase of the preceding intron) only were used in this comparison; 10 codons were compared starting from the second codon.

[a] Differences are statistically significant.

dons, while AA-carrying codons should not be affected in the same degree or at all. This type of comparison based on studying two observed values is very much different from hitherto used comparisons of observed and expected values. Figure 3 supports the hypothesis of specific selection against codons carrying AG. It is quite obvious that in all nine cases presented in Figure 3 the frequency of AG-containing codons in the first few 5′ positions and particularly the first position is lower than the frequency of AA-containing codons. There is also a distance effect: frequencies of AG-containing codons increase in the 3′ direction. Comparisons of the first 10 codons located at the very 5′ end of exons with the 10 codons located at positions 41–50 are highly significant ($P < 0.01$) except codon AAG in *H. sapiens*. However, even in this case the first codon is considerably below the average value. AA-containing codons behave differently and do not show strong distance-related selection effects, while some compensatory increases near the 5′ end of exons are possible. Such contrasting behavior of synonymous codons could hardly be explained by other causes and thus supports the tested selection hypothesis quite convincingly. Possibly this observation could add an extra factor affecting codon choices during evolution (Karlin and Mrazek 1996).

Next we investigated whether positions of AG within codons in the 5′ region of exons may have different selective values. Clearly there are three possible positions. When AG occupies position 1 within a codon (AG|*N*) and confuses splicing machinery, which may accept this particular AG as the last 3′ AG of the previous intron, it leads to abnormal splicing and causes a frameshift in the downstream part of the gene. The same is true when AG occupies position 3 within a codon (*NN*AG|); both of these positions (1 and 3) should be deleterious. However, position 2 (*N*AG|) seems to be less deleterious in this regard as it will not cause a frameshift but rather causes a loss of one or a few codons in the case of abnormal splicing.

To test this hypothesis we calculated differences between expected and observed frequencies of AG and compared these values among three codon positions using the approach explained above. The results presented in Table 3 show that the differences between expected and observed values were significantly smaller in the second position of codons ("less deleterious") than in the two other positions in all studied cases, except the third position in humans. The same conclusion is correct for exons in all three phases (data not shown). A smaller difference between expected and observed frequencies of AG in the second codon position indicates that this position is subjected relatively less to the specific selection pressure than are two other positions. These data provide additional support for the tested selection hypothesis.

Several independent types of evidence presented here create sufficient grounds to believe that AG dinucleotides located on the 5′ ends of exons experienced negative selection pressure in order to reduce the risk of their being mistakenly recognized as the last 3′ end intronic AG signal and thus to diminish the chance of deleterious splicing.

DISTRIBUTION OF GT PAIRS AT 3′ ENDS OF EXONS

A similar approach was applied for studying distribution of GT pairs at the 3′ end of exons. We investigated how distribution of the 3′ exonic GT could be affected by GT pairs nearly always located at the 5′ splice sites of introns. The last exons of genes were separated from the rest, as they do not have the following intron-exon boundary and probably exist under different selection pressures. Position −1 represents the last two 3′ positions of exons, which were aligned by their 3′ ends (Figure 4). As expected (Lewin 1994) in *C. elegans*, *D. melanogaster*, and *H. sapiens* the last two positions in Figure 4 (corresponding to the last 3′ positions in exons) have high frequencies of AG pairs (see also Figure 1), which obviously prevent appearance of other pairs, including GT, in considerable frequencies. This would explain a lack

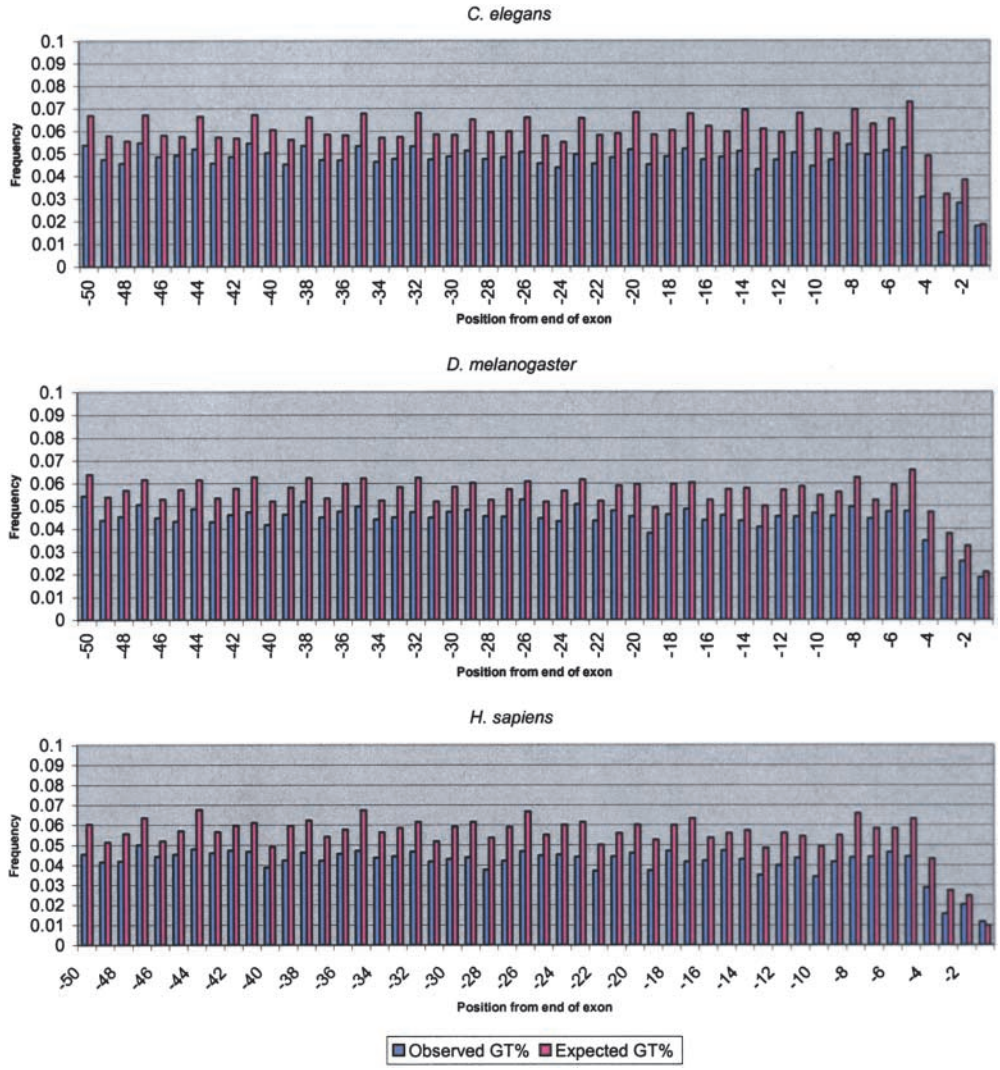

FIGURE 4.—Expected and observed frequencies of GT at the 3′ ends of the nonlast exons in *C. elegans*, *D. melanogaster*, and *H. sapiens*. Blue bars represent observed and red bars represent expected frequencies of GT in certain positions of exons aligned at the 3′ ends. Numbers on the *x*-axis indicate position of nucleotide G of GT pairs starting from the second-to-last nucleotide. Observed frequencies were calculated as proportion of GT among all possible 16 duplets in a certain position. Expected frequencies were calculated using the formula in the text. Exon-intron database was obtained from http://www.mcb.harvard.edu/gilbert/eid. Graphs were created using MS Excel's chart wizard tools.

of GT at the last 3′ positions of exons. Differences between expected and observed GT frequencies are similar in all studied species; observed frequencies are lower throughout.

Interestingly the observed frequencies of TG pairs (data not shown) do not differ from expected frequencies for *C. elegans* ($t = 0.7788$, $P = 0.4379$) and for *D. melanogaster* ($t = 1.1286$, $P = 0.2618$) in the first 50 positions. In *H. sapiens* observed frequencies of TG were even higher than expected ($t = 12.1069$, $P = 0$, in the first 50 positions). These data indicate that frequencies of GT and TG pairs behave differently at the 3′ ends of exons.

Again as in the previous section of this article, we compared observed and expected frequencies of GT pairs in the 3′ part of nonlast exons on two intervals, −6 to −30 and −101 to −125 (not shown in Figure 4) in the studied species. In *C. elegans* expected frequencies of GT are significantly higher within both compared intervals (Table 1), while the intensity of the differences declines in the 5′ direction. In *D. melanogaster* significant differences are also found on both intervals (Table 1). A similar pattern was found in *H. sapiens*. Thus the compared species demonstrate significant differences between expected and observed GT frequencies in the positions adjacent to the 3′ ends of the exons, which diminish in the 5′ direction. One may presume that selective pressure against GT pairs could contribute to the observed phenomenon.

The last exons do not have a following exon-intron junction and were not expected to show specific selection pressure against GT. Table 2 presents comparisons between observed and expected frequencies of GT on the same intervals calculated for the last exons. In all but one case there is a statistically detectable difference. This result is contradictory to our expectation that in the last exons there would not be specific selection pressure against GT. However, statistics for the last exons indicate a dramatic decline in the differences between observed and expected frequencies of GT, as compared to the nonlast exons. This may show a reduced selection pressure against GT; however, other

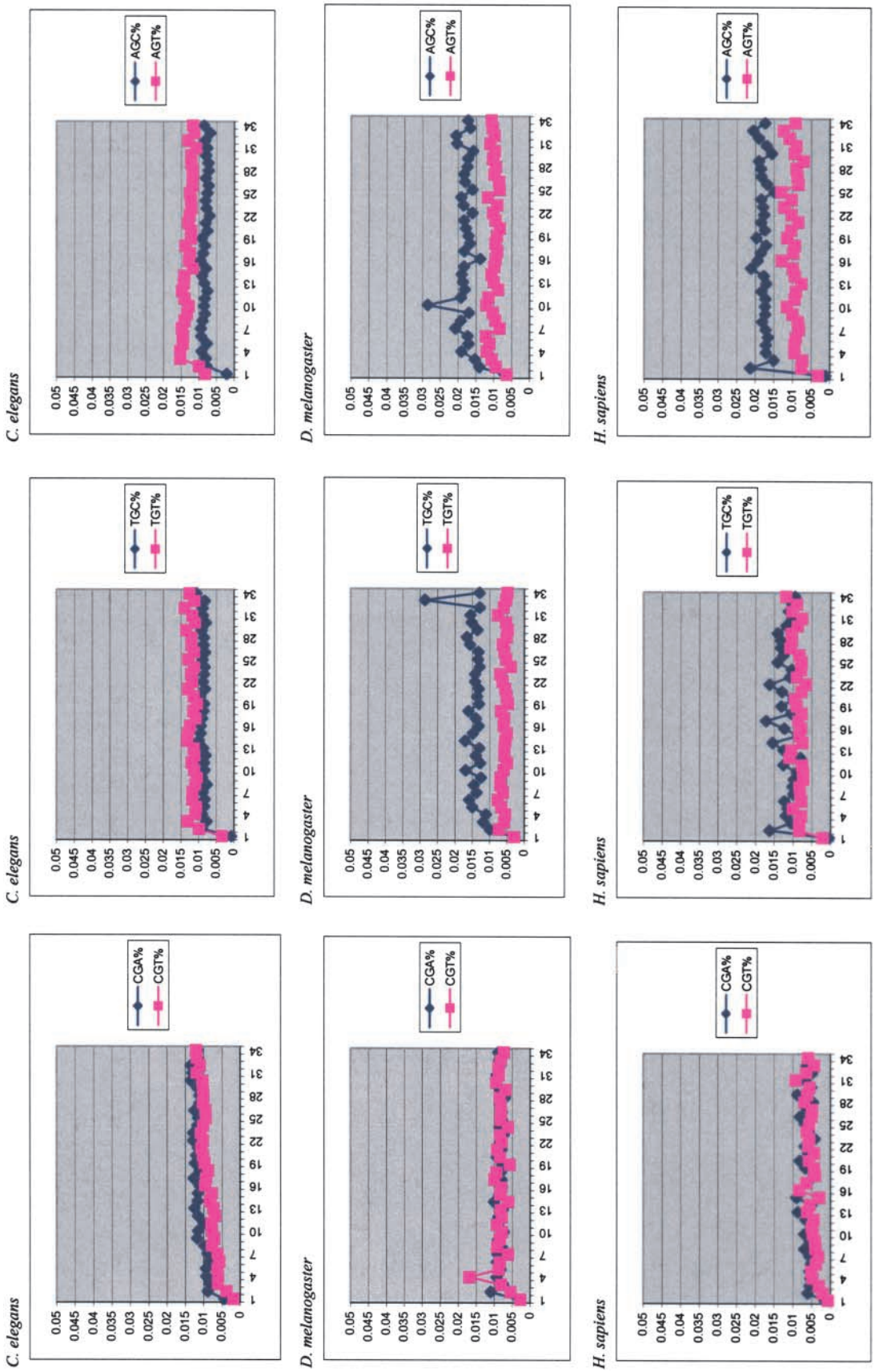

FIGURE 5.—Comparisons of frequencies of synonymous codons differing in the third position at the 3′ end of exons (phase 0) in the three studied species. The compared codons are CGA and CGT (arginine), TGC and TGT (cysteine), and AGC and AGT (serine). Numbers on the *x*-axis indicate the position of codons starting from the 3′ end of exons in the upstream direction. The first nucleotide of the last codon is represented by 1, the first nucleotide of the second-to-last codon is represented by 2, and so on. Exon-intron database was obtained from http://www.mcb.harvard.edu/gilbert/eid. Graphs were created using MS Excel's chart wizard tools.

TABLE 4

Comparisons of GT differences (expected − observed) between three codon positions

Comparisons between codon positions	*C. elegans*	*D. melanogaster*	*H. sapiens*
Second "deleterious" − first "less deleterious"	$t = 29.103178$	$t = 7.986985$	$t = 9.233449$
	$P = 0.00000$[a]	$P = 0.000001$[a]	$P = 0.000002$[a]
Third "deleterious" − first "less deleterious"	$t = 5.919415$	$t = -9.953226$	$t = -4.148004$
	$P = 0.00011$[a]	$P = 0.00000$[a]	$P = 0.000788$[a]

Exons starting from phase 0 (phase of the preceding intron) only were used in this comparison; 10 codons were compared starting from the second codon.

[a] Differences are statistically significant.

factors may be the cause of the differences between observed and expected values in the last exons. It is quite possible that there are other independent selection pressures, which may reduce observed GT frequencies in the last exons, making the situation more complex.

Distributions of GT frequencies shown in Figure 4 represent a mixture of three phases (0, 1, and 2) in each species. Phase separation of any of these distributions into three distributions (phase 0, phase 1, and phase 2) reveal the same pattern, which is just more sharply expressed (data not shown).

Comparative analysis of pairs of synonymous codons located at the 3′ end of exons similar to those described earlier in the article is presented in Figure 5. It can be seen that in many cases the last few 3′ positions (1–5) and particularly the last position (1) have lower frequency of GT-containing codons compared with frequency of alternative synonymous codons. There are also distance effects in several cases: frequencies of GT-containing codons slightly increase in the 5′ direction. However, it was not common for all studied cases. The data provide some support for the hypothesis, while not as compelling as the data for AG pairs (Figure 3).

Finally we compared differences between expected and observed frequencies of GT using the same logic that was applied to AG analysis earlier. The only difference is that in this case the first position of codons is "less deleterious." The results presented in Table 4 show that the differences between expected and observed values were significantly smaller between the second and the first position of codons, while the difference between the third and the first positions was significant only for *C. elegans* (Table 4).

It is well known that nucleotides surrounding intron-exon boundaries are essential for correct splicing. The latest data also demonstrate the importance of spliceosome structures (Maniatis and Tasic 2002), exonic splicing enhancers (Fairbrother *et al.* 2002), and various aspects of mRNA metabolism (Mendell and Dietz 2001). These factors contribute to fidelity of splicing and could make identification of specific selection pressures on the ends of exons less simple.

Recent publications provide good reasons to exercise a cautious approach in assuming possible selection pressures on dinucleotide frequencies (Duret and Galtier 2000) or on codon usage (Urrutia and Hurst 2001). We hope the data presented here were considered cautiously enough to avoid possible traps of complex processes operating on the molecular level.

We thank A. Fedorov for advice concerning the exon-intron database (EID) and E. Koonin and I. Ruvinsky for useful comments. We are also grateful to an anonymous reviewer for very helpful suggestions.

LITERATURE CITED

Burset, M., I. A. Seledtsov and V. V. Solovyev, 2000 Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res. **28:** 4364–4375.

Duret, L., and N. Galtier, 2000 The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochors is due to a mathematical artifact. Mol. Biol. Evol. **17:** 1620–1625.

Fairbrother, W. G., R. F. Yeh, P. A. Sharp and C. B. Burge, 2002 Predictive identification of exonic splicing enhancers in human genes. Science **297:** 1007–1013.

Karlin, S., and J. Mrazek, 1996 What drives codon choices in human genes. J. Mol. Biol. **262:** 459–472.

Lewin, B., 1994 *Gene V*, p. 914. Oxford University Press, New York.

Maniatis, T., and B. Tasic, 2002 Alternative splicing pre-mRNA splicing and proteome expansion in metazoans. Nature **418:** 236–243.

Mendell, J. T., and H. C. Dietz, 2001 When the message goes awry: disease-producing mutations that influence mRNA content and performance. Cell **107:** 411–414.

Nissim-Rafinia, M., and B. Kerem, 2002 Splicing regulation as a potential genetic modifier. Trends Genet. **18:** 123–127.

Saxonov, S., I. Daizadeh, A. Fedorov and W. Gilbert, 2000 The exon intron database: an exhaustive database of protein-coding intron-containing genes. Nucleic Acids Res. **28:** 185–190.

Urrutia, A. S. O., and L. D. Hurst, 2001 Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. Genetics **159:** 1191–1199.

Communicating editor: M. A. F. Noor