J Mol Evol (1994) 39:439–447

Journal
of Molecular
Evolution

Biased Distribution of Adenine and Thymine in Gene Nucleotide Sequences

Jan Mrázek, Jaroslav Kypr

Institute of Biophysics, Academy of Sciences of the Czech Republic, CZ-61265 Brno, Czech Republic

Received: 2 March 1994 / Accepted: 15 April 1994

Abstract: We analyzed occurrences of bases in 20,352 introns, exons of 25,574 protein-coding genes, and among the three codon positions in the protein-coding sequences. The nucleotide sequences originated from the whole spectrum of organisms from bacteria to primates. The analysis revealed the following: (1) In most exons, adenine dominates over thymine. In other words, adenine and thymine are distributed in an asymmetric way between the exon and the complementary strand, and the coding sequence is mostly located in the adenine-rich strand. (2) Thymine dominates over adenine not only in the strand complementary to the exon but also in introns. (3) A general bias is further revealed in the distribution of adenine and thymine among the three codon positions in the exons, where adenine dominates over thymine in the second and mainly the first codon position while the reverse holds in the third codon position. The product $(A_1/T_1) \times (A_2/T_2) \times (T_3/A_3)$ is smaller than one in only a few analyzed genes.

Key words: Biased distribution — Adenine-to-thymine ratio — Exons — Complementary strand — Introns — Codon positions

Introduction

Nucleotide sequences of protein coding and noncoding regions substantially differ in several respects. Most of the differences relate to the triplet nature of the genetic code and to the fact that codon usage in genes is biased (Grantham et al. 1986). A simple consequence of unequal occurrence of codons in genes is a strong signal with a period of three nucleotides, appearing in nucleotide correlation analysis of protein coding regions (Fickett and Tung 1992).

The periodicity of three nucleotides arises as a result of several contributions. The first is gene or organism specific and connected with the fact that if a codon is preferred to code for an amino acid in a part of a gene, or in several genes, then it is usually preferred in the whole gene or all genes of the organism (Sharp and Li 1987), the latter statement being only valid with unicellular organisms (Ikemura 1985). For example, human genes prefer codons ending with C (Wada et al. 1991), and that is why the three-nucleotide periodicity in human protein coding sequences receives a contribution from the higher-than-average usage of C in the third codon position.

The contribution described in the previous paragraph is gene or organism specific and mainly concerns the third, i.e., the most degenerate, codon position. However, codon usage is also biased regarding the first two codon positions, which is rather surprising because the occurrences of nucleotides in the first and especially the second codon position are restricted by the fact that they bear most of the information coding for the protein structure—primary, secondary, and tertiary (Kypr and Mrázek 1987a). Several laboratories including ours have noted that the strongest and general bias concerns guanine, which is overrepresented in the first codon position but underrepresented in the second codon position (Kypr 1986; Trifonov 1987). The notion that this pattern is

Correspondence to: J. Kypr

Table 1. Adenine and thymine contents in exons

	Number of CDS-s[a] analyzed	Total length (bp)	A	T	A − T
Primates	3,933	4,954,665	25.2%	21.3%	+3.92%
Rodents	3,954	4,692,291	25.4%	21.8%	+3.57%
Other mammals	1,191	1,429,449	24.4%	21.1%	+3.31%
Other vertebrates	1,315	1,424,286	27.0%	21.6%	+5.38%
Invertebrates	2,307	3,140,769	28.3%	22.3%	+6.04%
Plants	2,136	2,072,325	26.3%	23.8%	+2.42%
Fungi	2,273	3,341,805	30.2%	26.9%	+3.22%
Prokaryotes	8,465	9,109,602	25.7%	23.0%	+2.72%

[a] CDS means "protein-coding sequence" and is equivalent to the CDS keyword in the database feature table. One CDS usually comprises all exons of a gene; the introns are excluded

required to maintain the reading frame during translation has recently been elaborated (Lagunez-Otero and Trifonov 1992) but experiments do not support this hypothesis (Curran and Gross 1994).

In previous work (Kypr and Mrázek 1987a), we have also noticed further nucleotide (especially adenine and thymine) preferences for the particular codon positions in genes, but the amount of known sequences was too small at that time to arrive at convincing conclusions in this direction. Here we make use of the dramatically enlarged database of the known gene nucleotide sequences to show that the occurrence of adenine and thymine is indeed biased in the three codon positions in protein coding regions. The bias is remarkably strong but is only a part of the uneven distribution of adenine and thymine in genes because adenine is generally more abundant in exons than thymine while thymine dominates over adenine in introns and, naturally, in the strands complementary to the exons.

Materials and Methods

The analysis was done on the sequences deposited with the EMBL Nucleotide Sequence Database, CD-ROM release 32 (Rice et al. 1993). The sequences of primates, rodents, other mammals, other vertebrates, invertebrates, plants, fungi, and prokaryotes were analyzed separately. Sequences of protein coding regions and introns were extracted using the data in the "feature" tables of the database entries. Some protein-coding sequences contain in-frame termination codons, which indicates a probable error in the exon location. Such sequences were excluded from the present analysis. Genes coding for polypeptides shorter than 100 amino acids were excluded as well. Introns containing 100 or more consecutive nontermination codons were also excluded because it is possible that these sequences code for a protein.

Results

Adenine Generally Dominates over Thymine in Exons

Genes are known to have different (A + T) contents but there is no study published in the literature concerning the ratio of A to T in genes. The database of genomic nucleotide sequences is now large enough to obtain statistically meaningful results regarding the comparison of the A and T contents in the protein coding regions. We did the study and the results are summarized in Table 1 to demonstrate several interesting facts. First, adenine dominates over thymine in all of the eight organism groups, starting with prokaryotes and ending with primates. The dominance is highest for the invertebrate and nonmammal vertebrate sequences. The average difference between adenine and thymine contents is close to 4% over the 30 Mb of the analyzed protein coding sequences, which certainly is not a negligible number. Adenine and thymine are complementary bases and thus the dominance of adenine over thymine immediately implies that thymine dominates over adenine in the exon complementary strands. In other words, adenine and thymine are distributed in an asymmetric way between the protein coding and the complementary strands, and the protein coding sequences prefer to be located in the adenine-rich strands. We analyzed exons of more than 25,000 genes here and 76% of them were located in the adenine-rich strand.

Thymine Dominates over Adenine in the Introns

Remarkably, introns provide just the opposite picture regarding the A-to-T ratio, compared to exons. Table 2 documents that introns contain more T than A for all analyzed groups of eukaryotic species. (In prokaryotes, the introns are quite rare and mostly very short, offering only a nonrepresentative data set.) The ends of introns contain splicing signals which are known to be rich in thymine (Stephens and Schneider 1992). That is why we repeated the calculations with truncated introns. Excluded were the 7-base 5′-intron termini and 26-base 3′-intron termini which show a degree of conservation indicating a participation in splicing. As expected, the bias in favor of thymine is rather diminished by the truncation, but it is still evident that thymine dominates over adenine even in the central parts of introns which do not bear any splicing signal.

Table 2. Adenine and thymine contents in introns and truncated[a] introns (values in parentheses)

	Number of introns analyzed	Total length (bp)	A	T	A-T
Primates	7,276	1,468,907	25.0%	27.8%	–2.73%
			(25.9%)	(27.2%)	(–1.37%)
Rodents	4,769	1,119,095	25.0%	27.9%	–2.94%
			(25.6%)	(27.2%)	(–1.62%)
Other mammals	1,132	267,024	26.2%	28.5%	–2.21%
			(27.2%)	(28.1%)	(–0.87%)
Other vertebrates	1,690	454,421	26.9%	31.1%	–4.21%
			(27.6%)	(30.5%)	(–2.86%)
Invertebrates	2,472	520,456	31.8%	33.4%	–1.63%
			(32.4%)	(32.5%)	(–0.15%)
Plants	1,978	475,981	28.7%	36.6%	–7.82%
			(29.4%)	(36.4%)	(–6.98%)
Fungi	1,008	97,495	26.9%	31.5%	–4.58%
			(27.0%)	(31.4%)	(–4.40%)

[a] Conserved termini involved in splicing (Stephens and Schneider 1992) were removed

Biased Distribution of Adenine and Thymine among the Three Codon Positions in the Exons

Figure 1 shows distributions of differences between the occurrence of adenine and thymine in the first, second, and third codon positions of almost 26,000 genes, originating from various groups of species and covering the whole range from prokaryotes to primates. It follows from Fig. 1 that adenine strongly dominates over thymine in the first codon position while the reverse is true in the third codon position. In the second position, adenine slightly dominates over thymine. This conclusion is strikingly independent of the type of organism and quite strong, so it should reflect a fundamental and general property of both prokaryotic and eukaryotic protein coding sequences. The preferences are not obeyed by only 4.5% of the analyzed genes as to the first codon position (Table 3), 28% regarding the second codon position, and 21% as to the third codon position. The product of ratios of A to T at codon positions 1 and 2 and of T to A at position 3, i.e., $(A_1/T_1) \times (A_2/T_2) \times (T_3/A_3)$, which reflects the reported bias, is smaller than unity in 4.9% of the analyzed genes, and only 3.5% of the eukaryotic genes.

Correlations of the Biased Distribution of Adenine and Thymine among the Three Codon Positions in the Exons

We analyzed correlations of all possible pairs of (A + T) and (A – T) in all three codon positions. None is very strong, i.e., close to 1, but some are statistically significant, i.e., higher than approximately 0.3 (Table 4). The strongest correlation exists between the (A + T) contents in the three codon positions, suggesting that the global exon (A + T) content is usually reflected in all three codon positions. In other words, if an exon is (A + T)-rich/poor, then all of the three codon positions have a relatively high/low (A + T) content. Another relatively strong correlation was found between the (A – T) content at the first codon position and the (A + T) content at the second codon position. This correlation is strongest with the mammalian and prokaryotic genes. Low but still significant correlation also exists among the A – T differences at all of the three codon positions. The correlations indicate that the A/T bias has two components, one being gene-specific and the other being codon position-specific.

We have also analyzed dinucleotide occurrences at the three codon positions in the exons and compared them with expectations based on the nucleotide occurrences in the three codon positions, including the biased distribution of A and T. If the biased distribution of A and T were entirely a consequence of a biased dinucleotide distributions, then the numbers given in Table 5 should all be zero, which evidently is not the case. We took those dinucleotides where the difference between the true occurrence and expectation was the highest and determined an expected consequence of the biased dinucleotide distribution on the A/T ratio in the particular codon positions. Table 6 demonstrates that the biased dinucleotide distribution among the three codon positions in genes parallels the observed A/T bias as frequently as it goes in the opposite direction to suggest that the biased distribution of A and T among the three codon positions hardly results from the nonrandom distribution of dinucleotides in the three codon positions in genes.

Genomes Having Extreme (A + T) Contents and RNA Genomes

Further information about the origin and biological meaning of the biased A/T distribution might follow from an analysis of genomes or their parts which are extremely (A + T)-rich or poor. Table 7 shows that some

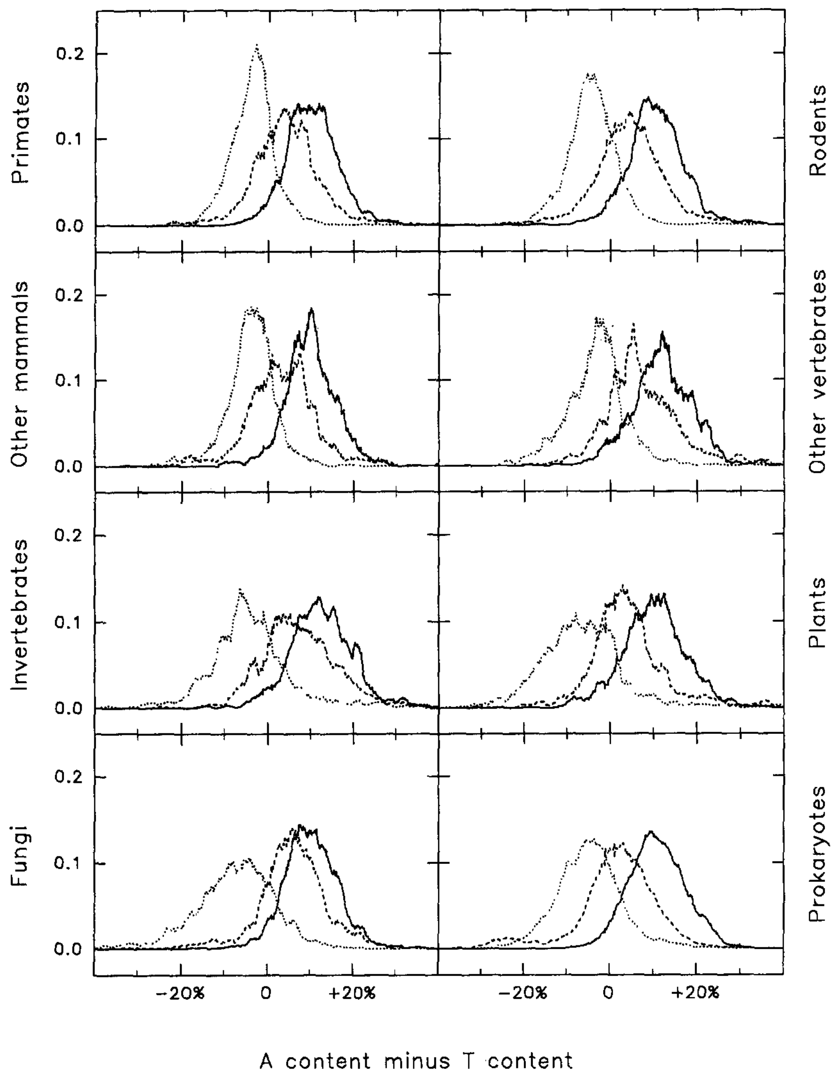

Fig. 1. Distribution of the difference between adenine and thymine contents (normalized to the number of nucleotides in the gene) at codon positions 1 (*solid line*), 2 (*dashed*), and 3 (*dotted*). The ordinate shows the fraction of analyzed genes having the A − T difference within the interval of ±1% around the value given at the abscissa.

Table 3. Numbers of genes that do no obey the general bias of the adenine and thymine distribution among the three codon positions

Genes of	Total amount of genes	Numbers of genes having[a]			
		$A_1 < T_1$	$A_2 < T_2$	$T_3 < A_3$	$A_1A_2T_3 < T_1T_2A_3$
Primates	3,933	166	1,015	707	119 (3.0%)
Rodents	3,954	174	1,024	719	118 (3.0%)
Other mammals	1,191	52	364	227	50 (4.2%)
Other vertebrates	1,315	60	195	337	39 (3.0%)
Invertebrates	2,307	101	394	593	50 (2.2%)
Plants	2,136	171	613	403	160 (7.5%)
Fungi	2,273	94	383	437	68 (3.0%)
Prokaryotes	8,465	324	3,073	1,864	661 (7.8%)

[a] The subscript indicates codon position

(A + T)-rich genomes prefer to have adenine even in the third codon position of their genes. This tendency is most obvious in lentiviruses. However, there are other (A + T)-rich organisms where T dominates over A in the third codon position in genes, like in most other organisms where the (A + T) genomic content is not extreme. Another exception is the second codon position in chloroplast genes, where T dominates over A, in contrast to the prevailing dominance of A over T in the nuclear genomes. *Streptomyces* genes, which are (G + C)-rich, also have more T than A in the second codon position. Probably, a certain amount of hydrophobic amino acids, which mostly have T in the second codon position (Woese et al. 1966; Volkenstein 1966; Weber and Lacey

Table 4. Correlation coefficients among adenine and thymine composition characteristics of genes

	Correlation coefficient							
Characteristics[a]	Pri[b]	Rod	Mam	Vrt	Inv	Pln	Fun	Pro
$(A + T)_1$ and $(A + T)_2$	0.29	0.16	0.20	0.01	0.44	0.33	0.39	0.55
$(A + T)_1$ and $(A + T)_3$	0.47	0.37	0.41	0.28	0.41	0.40	0.59	0.74
$(A + T)_1$ and $(A - T)_1$	0.20	0.04	0.33	–0.04	0.27	0.20	0.16	0.29
$(A + T)_1$ and $(A + T)_2$	–0.10	–0.14	0.02	–0.09	0.00	–0.16	0.06	0.15
$(A + T)_1$ and $(A - T)_3$	–0.10	–0.15	–0.07	–0.02	–0.02	–0.31	0.31	0.01
$(A + T)_2$ and $(A + T)_3$	0.21	0.06	0.20	0.17	0.12	0.26	0.42	0.55
$(A + T)_2$ and $(A - T)_1$	0.44	0.37	0.42	0.25	0.24	0.21	0.24	0.32
$(A + T)_2$ and $(A - T)_2$	0.18	0.06	0.21	0.15	0.14	0.11	0.29	0.27
$(A + T)_2$ and $(A - T)_3$	0.01	0.07	–0.05	0.02	–0.03	–0.01	0.37	0.09
$(A + T)_3$ and $(A - T)_1$	0.22	0.09	0.33	–0.01	0.01	–0.07	0.08	0.29
$(A + T)_3$ and $(A - T)_2$	0.16	0.09	0.18	0.09	0.23	0.05	0.14	0.27
$(A + T)_3$ and $(A - T)_3$	–0.02	–0.02	–0.06	0.01	0.21	–0.14	0.16	–0.01
$(A - T)_1$ and $(A - T)_2$	0.29	0.29	0.26	0.33	0.24	0.21	0.46	0.35
$(A - T)_1$ and $(A - T)_3$	0.18	0.30	0.11	0.21	0.20	–0.14	0.35	0.22
$(A - T)_2$ and $(A - T)_3$	0.30	0.23	0.22	0.32	0.26	0.28	0.32	0.21

[a] The subscript indicates codon position

[b] Abbreviations indicate groups of species: Pri = primates, Rod = rodents, Mam = other mammals, Vrt = other vertebrates, Inv = invertebrates, Pln = plants, Fun = fungi, Pro = prokaryotes

1978; Lacey et al. 1985), should be contained in any globular protein to build the hydrophobic core, and this requirement is stronger than the mechanism increasing the amount of adenines in genes.

With the exception of lentiviruses, RNA genomes, both single-stranded and double-stranded, have equal amounts of T and A in the third codon position, but otherwise they obey the general rules reported in the present article.

Discussion

Studies of rules of nucleotide distribution in genes have so far been mostly focused on the third codon position (Ikemura 1985; Aota and Ikemura 1986; Bernardi et al. 1993). However, the nucleotide composition is also biased in the first and second, i.e., mostly nondegenerate, codon positions. The biased distributions of nucleotides between the first two codon positions are general and not gene or organism specific as with the degenerate position. The strongest bias which is translated into the structure of proteins concerns the abundance of G in the first codon position and its avoidance in the second position (Kypr 1986; Trifonov 1987; Kypr and Mrázek 1987a; Lagunez-Otero and Trifonov 1992). We demonstrate here that the biased distribution of A and T is almost as strong and that not only the (G + C) or (A + T) contents but each base has its own specific role in coding for the biological information.

The biased distribution of adenine and thymine in genes means that the amino acids encoded by the codons ANN, NAN, and NNT are more frequent in proteins than the amino acids encoded by the codons TNN, NTN, and NNA. Thus it is possible that the amino acid composition of proteins is the origin of the observed A/T bias. Yet we search for other possible explanations because, e.g., the protein secondary structure is also important for protein function (Dufton 1985; Soto et al. 1985) and thus can contribute to the A/T bias. Computer simulations show that random nucleotide sequences code for much lower amounts of protein secondary structure than the gene nucleotide sequences while the introduction of the observed A/T bias into the random sequences increases the amount of encoded beta sheets and especially alpha helices (Kypr and Mrázek 1987a; Mrázek and Kypr unpublished data). Therefore it remains open which factor stands behind the biased A/T distribution in the nondegenerate codon positions in genes.

Another aspect that might contribute to the A/T bias in genes is the uneven distribution of A and T in the termination codons. In their first position, T is always present while A is absent, so 26% of the 61 sense codons begin with A and only 21% with T. Therefore, a dominance of A over T by 5% in the first codon position is a maximum possible effect of the elimination of the termination codons from protein coding sequences on the results given here. However, this difference is far below the observed effect (Fig. 1). On the other hand, there are 16 sense codons having T but only 14 having A in the second and third codon positions. As far as the second codon position is concerned, this is the opposite of the present observations, so the elimination of termination codons either has nothing in common with the biased

Table 5. Dinucleotide preferences given in the form of observed minus expected occurrence expressed in percent of the expected occurrence

	Primates (codon position)			Rodents (codon position)			Other mammals (codon position)			Other vertebrates (codon position)		
	1 – 2	2 – 3	3 – 1	1 – 2	2 – 3	3 – 1	1 – 2	2 – 3	3 – 1	1 – 2	2 – 3	3 – 1
AA	+13	+10	–5	+13	+7	–13	+13	+16	–5	+13	+11	–7
AC	–10	–18	–21	–10	–15	–14	–9	–17	–22	–14	–15	–14
AG	–1	+20	+28	–2	+23	+33	–4	+18	+29	–2	+21	+21
AT	–5	–9	–16	–5	–13	–22	–5	–11	–18	–2	–15	–13
CA	–7	+38	+30	–8	+41	+34	–11	+36	+27	–10	+33	+33
CC	+6	+20	+17	+6	+15	+13	+4	+20	+13	+8	+12	+10
CG	–30	–63	–56	–30	–66	–59	–28	–5	–49	–26	–58	–51
CT	+25	+28	+36	+25	+34	+37	+28	+30	+33	+24	+27	+32
GA	+16	+1	–2	+17	+6	+1	+18	+7	–3	+17	+4	+2
GC	–6	+11	+3	–7	+8	0	–3	+9	+3	–5	+14	+4
GG	+16	–6	+11	+17	–8	+9	+15	–8	+12	+14	–15	+4
GT	–26	–9	–22	–26	–5	–17	–28	–8	–21	–27	–4	–16
TA	–42	–46	–41	–41	–47	–43	–41	–51	–42	–43	–47	–40
TC	+19	–4	–13	+20	0	–8	+15	–3	–11	+25	–2	–6
TG	+15	+36	+49	+12	+35	+51	+17	+32	+49	+10	+37	+45
TT	+22	–6	–9	+22	–10	–14	+23	–6	–12	+26	–2	–11

distribution of A and T in genes, or it has, and then the other effect causing the uneven distribution of A and T in the second codon position is even stronger than indicated by the data in Fig. 1, which do not take the contingent effects of the termination codons into account. In contrast, the termination codons would entirely explain the observed dominance of T over A in the third codon position in vertebrates, while invertebrates, plants, fungi, and prokaryotes still have more T than A in the third codon position even after the correction for termination codons.

The first codon position is much looser than the second codon position regarding the restraints imposed by the encoded protein structure, but it is related to the encoded amino acid biosynthetic pathway (Taylor and Coates 1989). Underrepresentation of T compensates for the strong dominance of G in the first codon position, and because A is also overrepresented, though less than G, in the first codon position (Kypr and Mrázek 1987a), the A/T ratio is higher than 1. It has previously been shown that by preferring G and avoiding T in the first codon position, genes minimize the detrimental effects of random single point mutations on the stability of the encoded proteins (Kypr 1986).

The third codon positions in genes are much more frequently occupied by pyrimidines than purines (Shepherd 1981) and specifically by T than A (Kypr and Mrázek 1987a). The avoidance of A in the third codon position has recently been attributed (Yomo et al. 1992) to the fact that it creates termination codons on the complementary DNA strand (the termination codons TAA, TAG, and TGA arise from the triplets of TTA, CTA, and TCA, respectively, on the complementary strand, all of which have A in the third codon position) while it has been proposed that a potential of the complementary strand to code for proteins is essential in the evolution (Yomo et al. 1992). However, the avoidance can also be a trivial consequence of an unusual codon usage, for which reasons can be found easier than for the vague evolutionary potential (Ikehara and Okazawa 1993).

Many codons ending with A occur (Table 7) in the genes of HIV (Kypr and Mrázek 1987b), other lentiviruses (Kypr et al. 1989), the human malaria parasite *Plasmodium falciparum* (Weber 1987), and the bacterium *Mycoplasma capricolum* (Ohkubo et al. 1987). Furthermore, the A-ending codons frequently appear in chloroplasts and yeast whose genomes are also (A + T)-rich (Wada et al. 1991). On the other hand, certain codons ending with A do not appear at all or are unassigned in the coding regions of the (G + C)-rich bacterium *Micrococcus luteus* (Kano et al. 1993). Thus an extremely strong (G + C) pressure completely eliminates A-ending codons with a concomitant disappearance of the corresponding tRNAs, and converts them to other synonymous codons. A similar phenomenon occurs with a G-ending codon in an extremely (A + T)-rich *Mycoplasma capricolum* genome (Ohkubo et al. 1987).

The leucine codon TTA, whose sequence is quite opposite to the preferences in the distribution of T and A among the three codon positions in genes, is not used at all in genes required for vegetative growth of *Streptomyces* spp. However, it occurs in a few genes used during differentiation, so the ability to translate the TTA codon is confined to only late stages of colony development (Leskiw et al. 1991). This is an example of a developmental regulatory mechanism operating on the translational level which uses a codon whose bases are opposite in all three positions to the preferences reported here.

We have shown previously that randomized gene nucleotide sequences code for hypothetical proteins which

Table 5. Extended

Invertebrates (codon position)			Plants (codon position)			Fungi (codon position)			Prokaryotes (codon position)		
1 – 2	2 – 3	3 – 1	1 – 2	2 – 3	3 – 1	1 – 2	2 – 3	3 – 1	1 – 2	2 – 3	3 – 1
+19	+10	+4	+18	+12	–1	+16	+20	+6	+17	+45	+16
–16	–18	–3	–19	–8	–2	–1	–5	–3	–8	–18	–15
–20	+15	–8	–9	+23	+1	+1	+5	–7	–33	–17	–5
+2	–4	+10	+1	–21	+2	–5	–15	+4	+8	+4	+3
–3	+22	+23	+4	+28	+26	+6	+5	+15	–9	+5	0
+3	+11	–16	+13	–2	–1	+17	+8	–1	–14	+2	–10
–5	–23	–14	–32	–41	–25	–24	–36	–12	+20	+10	+6
+5	–9	+8	+5	+18	+7	–9	+12	–3	+9	–17	0
+8	+1	–6	+4	+6	+7	+16	–11	+2	+8	–3	–7
0	+21	+19	–1	+5	+7	–8	–6	+11	+12	+40	+31
+23	–42	+1	+31	–8	–5	+29	–10	–1	+21	–25	–6
–27	+18	–16	–25	–2	–7	–29	+19	–9	–31	+10	–18
–46	–34	–25	–42	–43	–32	–53	–22	–17	–34	–29	–6
+24	–1	+2	+15	+8	–6	+27	+3	–3	+6	–7	–10
–5	+30	+22	–9	+16	+30	–25	+30	+15	–20	+26	+4
+43	+1	–2	+41	+9	–2	+56	–2	+5	+45	+4	+16

Table 6. Expected consequences of dinucleotide preferences on A/T bias

Dinucleotide	Codon position	Average preference[a]	Expected consequence	Correspondence with the observation
CG	2-3	–59%		No effect on A/T
CG	3-1	–54%		No effect on A/T
TG	3-1	+49%	$T_3 > A_3$	Yes
TA	2-3	–48%	$A_2 > T_2$	Yes
			$T_3 > A_3$	Yes
TA	1-2	–42%	$A_1 > T_1$	Yes
			$T_2 > A_2$	No
TA	3-1	–41%	$A_3 > T_3$	No
			$T_1 > A_1$	No
CA	2-3	+37%	$A_3 > T_3$	No
TG	2-3	+35%	$T_2 > A_2$	No
CT	3-1	+34%	$T_1 > A_1$	No
CA	3-1	+31%	$A_1 > T_1$	Yes
CT	2-3	+30%	$T_3 > A_3$	Yes
AG	3-1	+28%	$A_3 > T_3$	No
CG	1-2	–28%		No effect on A/T
GT	1-2	–27%	$A_2 > T_2$	Yes
CT	1-2	+25%	$T_2 > A_2$	No
TT	1-2	+23%	$T_1 > A_1$	No
			$T_2 > A_2$	No
AG	2-3	+20%	$A_2 > T_2$	Yes
TC	1-2	+20%	$T_1 > A_1$	No
GT	3-1	–19%	$A_1 > T_1$	Yes
AC	3-1	–18%	$T_3 > A_3$	Yes
AT	3-1	–17%	$T_3 > A_3$	Yes
			$A_1 > T_1$	Yes
CC	2-3	+17%		No effect on A/T
GA	1-2	+17%	$A_2 > T_2$	Yes
AC	2-3	–16%	$T_2 > A_2$	No

[a] The average was calculated from only vertebrate sequences since invertebrate, plant, fungi, and prokaryotic genes often show different dinucleotide preferences

Table 7. Adenine and thymine contents at the three codon positions in genomes having extreme A + T contents, and in RNA viruses

Genome	G + C content	Number of genes	Position 1		Position 2		Position 3	
			A	T	A	T	A	T
S. cerevisiae	40.4%	1,450	32.7%	21.6%	35.2%	27.9%	28.3%	33.1%
P. falciparum	30.3%	132	37.8%	19.8%	44.3%	23.7%	43.4%	40.1%
M. capricolum	30.8%	14	38.0%	16.5%	34.8%	29.9%	47.7%	40.8%
Lentiviruses	43.1%	578	33.4%	18.0%	33.1%	25.2%	36.6%	24.5%
Chloroplasts	40.5%	382	28.3%	21.0%	28.6%	30.6%	30.3%	39.6%
M. luteus	64.2%	9	23.8%	12.4%	31.2%	28.4%	4.6%	6.9%
Streptomyces	71.5%	304	17.5%	11.4%	23.9%	25.5%	3.7%	3.6%
ss-RNA viruses[a]	45.9%	2,133	31.3%	19.0%	30.3%	27.7%	27.1%	26.9%
ds-RNA viruses	41.4%	237	32.3%	20.7%	31.7%	29.8%	30.6%	30.7%

[a] Not including lentiviruses

are significantly less hydrophobic than the real proteins (Mrázek and Kypr 1992). However, hydrophobicity of proteins is strongly conserved in the evolution (Mrázek and Kypr 1992), and therefore it is the gene nucleotide composition which should change with time while an increasing adenine content in genes can explain the apparently decreased hydrophobicity of the encoded proteins. Existence of the mechanism increasing the amount of A in DNA is also indicated by the observation that the amount of A in *Drosophila* pseudogenes is higher than in functional genes (Morijama and Gojobori 1992). It has been proposed (Kypr 1990) that the mechanism is based on the A-rule, which says that adenine is preferentially inserted opposite no-base positions in DNA (Sagher and Strauss 1983; Randall et al. 1987) that frequently arise during repair of damaged bases.

Sensitivity to mutations is remarkably different in the two strands of DNA (Veaute and Fuchs 1993). In combination with the A-rule, this difference might have given rise to the asymmetry of adenine and thymine distribution between the protein coding and the complementary strand. On the other hand, A and T are distributed in a symmetric way in genomes on a large scale (Fickett et al. 1992), which is achieved by the opposite A/T biases in the protein coding and noncoding regions as we show here.

Acknowledgments. The database used in this work was created during a visit of J.M. to the EMBL in Heidelberg. The authors are grateful to Chris Sander for his kind hospitality and advice. This work was supported by grant No. 204/93/2370 to J.K. from the grant agency of the Czech Republic.

References

Aota S, Ikemura T (1986) Diversity in G + C content at the third codon position in vertebrate genes and its cause. Nucleic Acids Res 14: 6345–6355

Bernardi G, Mouchiroud D, Gautier C (1993) Silent substitutions in mammalian genomes and their evolutionary implications. J Mol Evol 37:583–589

Curran JF, Gross BL (1994) Evidence that GHN phase bias does not constitute a framing code. J Mol Biol 235:389–395

Dufton MJ (1985) Genetic code redundancy and the evolutionary stability of protein secondary structure. J Theor Biol 116:343–348

Fickett JW, Torney CT, Wolf DR (1992) Base compositional structure of genomes. Genomics 13:1056–1064

Fickett JW, Tung CS (1992) Assessment of protein coding measures. Nucleic Acids Res 20:6441–6450

Grantham R, Perrin P, Mouchiroud D (1986). Patterns in codon usage of different kinds of species. Oxford Surv Evol Biol 3:48–81

Ikehara K, Okazawa E (1993) Unusually biased nucleotide sequences on sense strands of Flavobacterium sp. genes produce nonstop frames on the corresponding antisense strands. Nucleic Acids Res 21:2193–2199

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

Kano A, Ohama T, Abe R, Osawa S (1993) Unassigned or nonsense codons in Micrococcus luteus. J Mol Biol 230:51–56

Kypr J (1986) A part of codon bias in genes protects protein spatial structures from destabilization by random single point mutations. Biochem Biophys Res Commun 139:1094–1097

Kypr J (1990) Possible reason for the preferential insertion of adenine opposite abasic lesions in DNA. J Theor Biol 135:125–126

Kypr J, Mrázek J (1987a) Occurrence of nucleotide triplets in genes and secondary structure of the coded proteins. Int J Biol Macromol 9:49–53

Kypr J, Mrázek J (1987b) Unusual codon usage of HIV. Nature 327: 20

Kypr J, Mrázek J, Reich J (1989) Nucleotide composition bias and CpG dinucleotide content in the genomes of HIV and HTLV 1/2. Biochim Biophys Acta 1009:280–282

Lacey JC,Hall LM, Mullins DW (1985) Rationalization of some genetic anticodonic assignments. Orig Life 16:69–79

Lagunez-Otero J, Trifonov EN (1992) mRNA periodical infrastructure complementary to the proofreading site in the ribosome. J Biomol Struct Dyn 10:455–464

Leskiw BW, Bibb MJ, Chater KF (1991) The use of a rare codon specifically during development? Mol Microbiol 5:2861–2867

Mrázek J, Kypr J (1992) Nucleotide composition of genes and hydrophobicity of the encoded proteins. FEBS Lett 305:163–165

Morijama EN, Gojobori T (1992) Rates of synonymous substitution and base composition of nuclear genes in Drosophila. Genetics 130:855–864

Ohkubo S, Muto A, Kawauchi Y, Yamao F, Osawa S (1987) The ribosomal protein gene cluster of Mycoplasma capricolum. Mol Gen Genet 210:314–322

Randall SK, Eritja R, Kaplan BE, Petruska J, Goodman MF (1987)

Nucleotide insertion kinetics opposite abasic lesions in DNA. J Biol Chem 262:6864–6870

Rice CM, Fuchs R, Higgins DG, Stoehr PJ, Cameron GN (1993) The EMBL Data Library. Nucleic Acids Res 21:2967–2971

Sagher D, Strauss B (1983) Insertion of nucleotides opposite apurinic/apyrimidinic sites in deoxyribonucleic acid during in vitro synthesis: uniqueness of adenine nucleotides. Biochemistry 22:4518–4526

Sharp PM, Li W-H (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential application. Nucleic Acids Res 15:1281–1295

Shepherd JCW (1981) Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc Natl Acad Sci USA 78:1596–1600

Soto MA, Sepúlveda A, Tohá J (1985) Conservation of the secondary structure of protein during evolution and the role of the genetic code. Orig Life 16:157–164

Stephens RM, Schneider TD (1992) Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. J Mol Biol 228:1124–1136

Taylor FJR, Coates D (1989) The code within codons. Biosystems 22:177–187

Trifonov EN (1987) Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16S rRNA nucleotide sequences. J Mol Biol 194:643–652

Veaute X, Fuchs RPP (1993) Greater susceptibility to mutations in lagging strand of DNA replication in Escherichia coli than in leading strand. Science 261:598–600

Volkenstein MV (1966) The genetic coding of the protein structure. Biochim Biophys Acta 119:421–424

Wada K, Wada Y, Doi H, Ishibashi F, Gojobori T, Ikemura T (1991) Codon usage tabulated from the GenBank genetic data. Nucleic Acids Res 19:1981–1986

Weber AL, Lacey JC Jr (1978) Genetic code correlations: amino acids and their anticodon nucleotides. J Mol Evol 11:199–211

Weber JL (1987) Analysis of sequences from the extremely A + T-rich genome of Plasmodium falciparum. Gene 52:103–109

Woese CR, Dugre DH, Saxinger WC, Dugre SA (1966) The molecular basis for the genetic code. Proc Natl Acad Sci USA 55:966–974

Yomo T, Urabe I, Okada H (1992) No stop codons in the antisense strands of the genes for nylon oligomer degradation. Proc Natl Acad Sci USA 89:3780–3784