Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 9630–9634, December 1988
Evolution

Universal rule for coding sequence construction: TA/CG deficiency–TG/CT excess

SUSUMU OHNO

Beckman Research Institute of the City of Hope, Duarte, CA 91010

Contributed by Susumu Ohno, September 13, 1988

ABSTRACT **Each coding sequence is a finite resource as to the number and composition of four bases. Accordingly, the excessive recurrence of one base oligomer entails the noticeable underrepresentation by the other, so that if the former is the same in most, if not all, of the coding sequences, the latter too must necessarily be the same in all. Indeed, a previous series of studies on 20-odd divergent coding sequences established CTG as one of the most frequently recurring base trimers (if not the most frequent), and this excess was compensated by the underrepresentation by CG and TA dimer-containing base trimers. In this study, I have analyzed three additional coding sequences and reanalyzed one previously studied coding sequence. These four, derived from man, a plant, and a fish, were of variously lopsided base compositions that were not at all conducive to high recurrences of either CT dimer or CT and TG. Yet, the excess of CT and TG dimers accompanied by complementary deficiency of CG and TA dimers emerged as the common rule. Thus, I propose the above as the universal rule of coding sequence construction. The underrepresentation by CG and TA dimers within coding sequences explains why regulatory signals in intergenic spacers are of two kinds: one, TA dimer rich; and the other, CG dimer rich.**

During the past several years, I have identified primordial repeating units in more than 20 coding sequences of divergent functions and origins (see refs. 1 and 2). In all but one (3), the base trimer CTG was included in these primordial repeating units. This was no surprise, for these coding sequences (with the single exception noted above) encoded proteins rich in secondary structures: α-helices and/or β-sheets. Not only is leucine the major ingredient of proteins rich in secondary structures but also leucine-encoding CTG is the universally preponderant leucine codon from *Escherichia coli* to man (4).

Each coding sequence being a finite resource, the excessive recurrence of one base trimer entails the scarcity of another. Indeed, the low frequencies of recurrence of base trimers ending in CG and TA have been noted (2). Reexamination of the 20 previously analyzed coding sequences encoding proteins made mainly of either β-sheets or α-helices indeed confirmed that the excessive recurrence of two base dimers, CT and TG, was compensated by the low recurrence of CG and TA and that CT and TG invariably combined to make CTG one of the most frequently recurring base trimers, if not the most frequent.

The single exception was the H1 histone coding sequence of the rainbow trout (5). The peculiarity of the 210-residue-long protein containing more than 60 residues each of alanine and lysine but only 7 of leucine was reflected in the pentameric repeating unit CCAAG of its coding sequence that excluded CTG (3). Furthermore, this 630-base-long coding sequence contained only 43 thymidine bases (6.8%). Thus, it occurred to me that the universality of the rule—the excessive recurrence of CT/TG dimers being compensated by a deficiency of CG/TA dimers—can best be tested on those coding sequences whose cytidine or thymidine content decreases to 20% or less. Accordingly, four coding sequences that satisfied the above noted condition, including the rainbow trout H1 histone coding sequence, were chosen and subjected to the dimer analysis to be described below.

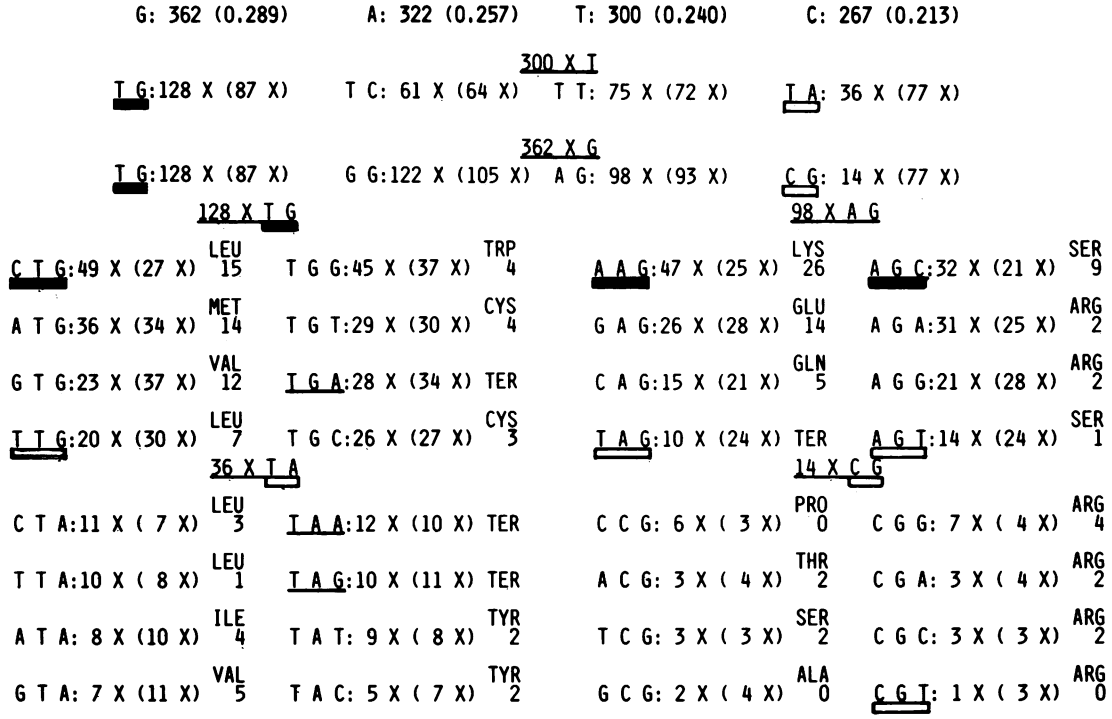
Human Phosphoglycerate Kinase Coding Sequence

As with other sugar-metabolizing enzymes, this kinase is comprised of alternating short α-helical and β-sheet segments (6). The base composition of the 1251-base-long human coding sequence for this enzyme is shown in the top line of Fig. 1 (7). From this base composition, one can calculate the expected occurrences of each dimer [e.g., AG dimer is expected to occur 93 times, which is 322 (the number of adenosine bases) times 0.289 (the fraction represented by guanosine)]. As the observed and expected appearances of four dimers in each set sharing the identical base either at the first or second position were compared, the excessive occurrence of TG compensated by the markedly less-than-expected occurrence of TA and CG became evident, as shown at the top of Fig. 1. Although not shown, CG is also a member of the four dimers starting with cytidine, and in this set, the occurrence of only 14 CG dimers was compensated by 100 CT dimers whereas only 64 were expected. Thus, the low cytidine content (21%) did not hinder the high recurrence of CT merely because of the gross underrepresentation by CG. Accordingly, with 49 appearances, CTG was the most frequently recurring base trimer; by contrast, GCG and CGT made only 2 and 1 appearances. The primordial repeating unit of this coding sequence was GCTG, with 21 appearances. This not only assured the abundance of alanine (40 residues) and leucine (37 residues) in this enzyme but also established codons ending in CT as the preponderant codons for alanine, threonine, and proline. The scarcity of CG and TA dimers, on the other hand, decreed that codons ending in CG and TA were the least utilized of alanine, threonine, serine, and proline codons as well as of leucine, valine, and isoleucine codons. Reflecting the scarcity of TA and CG dimers, there were only 4 tyrosine and 12 arginine residues in human phosphoglycerate kinase.

Human Estrogen Receptor Coding Sequence

The three-dimensional configuration of steroid receptor proteins is not known. At any rate, the 595-codon-long coding sequence for human estrogen receptor (8) was noticeably deficient in thymidine (19%). Thus, it was to be contrasted with the human phosphoglycerate kinase coding sequence, which was deficient in cytidine. Yet, the story was the same, as shown in Fig. 2. The sequence was markedly deficient in two base dimers, TA and CG, and their lower-than-expected occurrences were compensated by 146 TG dimers. Although not shown, CT dimer also occurred 130 times. Accordingly, with 55 occurrences, CTG was again the most frequently recurring base trimer. The primordial repeating unit, how-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.



```
         G: 362 (0.289)       A: 322 (0.257)     T: 300 (0.240)       C: 267 (0.213)

                                              300 X T
        T G:128 X (87 X)     T C: 61 X (64 X)   T T: 75 X (72 X)     T A: 36 X (77 X)

                                              362 X G
        T G:128 X (87 X)     G G:122 X (105 X)  A G: 98 X (93 X)     C G: 14 X (77 X)
              128 X T G                                                98 X A G

                      LEU                    TRP                       LYS                    SER
C T G:49 X (27 X)  15    T G G:45 X (37 X)   4     A A G:47 X (25 X)  26    A G C:32 X (21 X)  9

                      MET                    CYS                       GLU                    ARG
A T G:36 X (34 X)  14    T G T:29 X (30 X)   4     G A G:26 X (28 X)  14    A G A:31 X (25 X)  2

                      VAL                                              GLN                    ARG
G T G:23 X (37 X)  12    T G A:28 X (34 X) TER     C A G:15 X (21 X)   5    A G G:21 X (28 X)  2

                      LEU                    CYS                                              SER
T T G:20 X (30 X)   7    T G C:26 X (27 X)   3     T A G:10 X (24 X) TER    A G T:14 X (24 X)  1
              36 X T A                                                 14 X C G
                      LEU                                              PRO                    ARG
C T A:11 X ( 7 X)   3    T A A:12 X (10 X) TER     C C G: 6 X ( 3 X)   0    C G G: 7 X ( 4 X)  4

                      LEU                                              THR                    ARG
T T A:10 X ( 8 X)   1    T A G:10 X (11 X) TER     A C G: 3 X ( 4 X)   2    C G A: 3 X ( 4 X)  2

                      ILE                    TYR                       SER                    ARG
A T A: 8 X (10 X)   4    T A T: 9 X ( 8 X)   2     T C G: 3 X ( 3 X)   2    C G C: 3 X ( 3 X)  2

                      VAL                    TYR                       ALA                    ARG
G T A: 7 X (11 X)   5    T A C: 5 X ( 7 X)   2     G C G: 2 X ( 4 X)   0    C G T: 1 X ( 3 X)  0
```

FIG. 1. The 417-codon-long human phosphoglycerate kinase coding sequence (7). Base composition in numbers and fractions of four bases are given in the top row. In the next two rows are the observed and expected (in parentheses) occurrences of four dimers in each of the two sets: the first set of four dimers has thymidine for the first base, whereas the second set of four dimers shares guanosine as the second base. TG dimer, which occurred far more often than expected (128 times versus 87 times) is underlined by the solid bar, while TA (36 occurrences observed versus 77 expected) and CG (14 versus 77) that appeared far less than expected are underlined by open bars. Shown next in four columns are four sets, each containing eight base trimers, with members of each set sharing the same base dimer either as the second and third bases or as the first and second bases. Aside from the excessively recurring TG (upper half of columns 1 and 2) and grossly underrepresented TA and CG (bottom half of columns 1 and 2 and columns 3 and 4, respectively), AG dimer was chosen as the control because its frequency of occurrence was more or less as expected. In calculating the expected occurrence frequency of each trimer, the occurrence of the reference dimer (e.g., TG 128 times) was taken as *fait accompli*. The frequency with which each trimer is utilized as a codon is also shown. Within each set, a trimer that is grossly overabundant is underlined by the solid bar and one that is not noticeably underrepresented is underlined by an open bar. Since the expected occurrence frequency of each base trimer is calculated on the basis of the observed occurrence of the reference dimer, the expected frequencies for the same trimer included in different sets may be different. Base trimers whose excessive or deficient recurrences are of borderline significance are merely underlined, and so are three base triplets that are potential chain terminators.

ever, was CCTG tetramer instead of GCTG for human phosphoglycerate kinase. Because of this, while CTG remained the preponderant codon for leucine, encoding 34 of the 72 leucine residues, codons ending in CC rather than CT emerged as preponderant codons for alanine, threonine, and proline. Reflecting the scarcity of TA and CG dimers, those codons ending in CG and TA remained the least utilized of alanine, threonine, serine, valine, isoleucine, and leucine codons. A single exception was CCG, which occurred far more often than expected (37 times observed versus 24 times expected); thus, CCG encoded 10 of 37 proline residues, ranking second to CCC among four proline codons (Fig. 2, column 3).

Pollen–Stigma Self-Incompatibility Protein of the Mustard Plant

Next, we come to the 405-codon-long coding sequence for the pollen–stigma self-incompatibility glycoprotein of the mustard plant (9). This coding sequence was relatively deficient in cytidine (21%) as was the case with the coding sequence for human phosphoglycerate kinase. However, the sequence for human phosphoglycerate kinase was purine-rich (55% A + G), whereas that for pollen–stigma self-incompatibility protein of the mustard plant was A+T rich (53% A + T). It should be recalled that human estrogen receptor coding sequence was G+C rich (57% G + C) as already shown in Fig. 2. Fig. 3 shows that this mustard plant protein showed again the less-than-expected occurrences of TA and CG, and that these underrepresentations were compensated as before by the 104 occurrences of TG. Although not shown, among four dimers starting with cytidine, the deficiency of CG was once again compensated by 79 occurrences of CT, whereas 68 were expected. In this case, however, 77 occurrences of CA also compensated for the paucity of CG. Not surprisingly, CT and TG did not combine to make CTG the most frequently recurring base trimer of this coding sequence. Instead, the honor went to TGG, which occurred 39 times (Fig. 3, column 2). Nevertheless, CTT, which occurred 28 times as the base trimer, was the preponderant leucine codon, encoding 15 of the 36 leucine residues. TTG, which occurred 32 times as the base trimer, was second, encoding 7 leucine residues. In spite of the low recurrence of CG in its coding sequence, this mustard plant pollen–stigma self-incompatibility protein was relatively rich in arginine (30 residues). This was because 16 of 30 arginine residues were encoded by AGA and AGG (Fig. 3, column 4).

Rainbow Trout H1 Histone

At last we come to the 210-codon-long rainbow trout H1 histone coding sequence, which has the most lopsided base composition, with thymidine accounting for only 7% of the total (5). The previous analysis identified the CCAAG pentamer with 25 occurrences as the primordial repeating unit of this relatively short coding sequence (3). As the periodicity decay in the coding sequence follows the golden mean (2), the above pentamer often became a part of octameric repeating units

C: 515 (0.289)	G: 506 (0.283)	A: 417 (0.234)	T: 347 (0.194)

347 X T			
T G:146 X (98 X)	T C: 92 X (100 X)	T T: 57 X (67 X)	T A: 52 X (81 X)

506 X G			
T G:146 X (98 X)	A G:138 X (116 X)	G G:134 X (143 X)	C G: 86 X (143 X)

146 X T G				138 X A G			
C T G:55 X (41 X)	LEU 34	T G G:49 X (41 X)	TRP 5	C A G:52 X (40 X)	GLN 16	A G G:41 X (39 X)	ARG 9
A T G:41 X (34 X)	MET 23	T G C:38 X (41 X)	CYS 7	G A G:46 X (39 X)	GLU 28	A G A:38 X (32 X)	ARG 6
G T G:32 X (41 X)	VAL 10	T G A:37 X (34 X)	TER	A A G:37 X (32 X)	LYS 16	A G C:37 X (40 X)	SER 10
T T G:18 X (29 X)	LEU 12	T G T:22 X (29 X)	CYS 4	T A G: 3 X (27 X)	TER	A G T:23 X (27 X)	SER 6

52 X T A				86 X C G			
C T A:29 X (15 X)	LEU 6	T A C:28 X (15 X)	TYR 15	C C G:37 X (24 X)	PRO 10	C G C:36 X (24 X)	ARG 9
G T A:11 X (15 X)	VAL 1	T A T:14 X (10 X)	TYR 6	A C G:23 X (20 X)	THR 3	C G G:23 X (24 X)	ARG 3
A T A: 7 X (12 X)	ILE 2	T A A: 7 X (12 X)	TER	G C G:19 X (24 X)	ALA 9	C G A:19 X (20 X)	ARG 4
T T A: 5 X (10 X)	LEU 1	T A G: 3 X (15 X)	TER	T C G: 7 X (17 X)	SER 3	C G T: 8 X (17 X)	ARG 2

FIG. 2. The 595-codon-long human estrogen receptor coding sequence (8).

such as GCCAAGAA and CCGCCAAG, thus encoding tripeptides Ala-Lys-Lys and Ala-Ala-Lys or Pro-Ala-Lys (3). In fact, 122 of the 210 amino acid residues were alanine and lysine.

Nevertheless, Fig. 4 shows that two dimers, CG and TA, as usual recurred at the considerably less-than-expected rates. Among four dimers having guanosine as the second base, however, the deficiency of CG was no longer compensated by TG but by AG, which occurred 87 times, simply because there were not enough thymidine bases. Similarly, among four base dimers having thymidine as the first base, the occurrence of only six TA dimers was compensated not by 12 TG dimers but by 23 TC dimers. Yet, the fact remained that TG held its own ground (12 observed versus 13 expected), while CT as usual occurred in excess (17 observed versus 14 expected). It thus appears that even under the most trying of circumstances, this rule of excessive CT/TG and deficient CG/TA was bent but not broken.

A: 326 (0.268)	T: 320 (0.263)	G: 311 (0.256)	C: 258 (0.212)

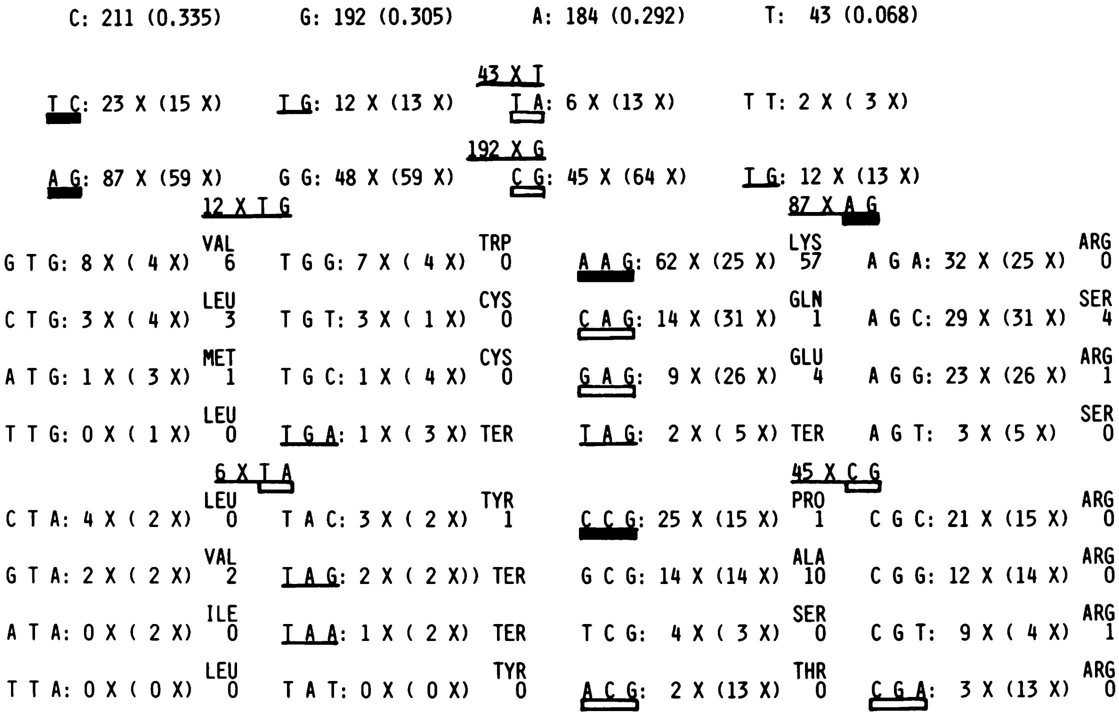
320 X T			
T G:104 X (82 X)	T C: 82 X (68 X)	T T: 73 X (84 X)	T A: 60 X (86 X)

311 X 6			
T G:104 X (82 X)	6 G: 83 X (80 X)	A G: 72 X (86 X)	C G: 52 X (66 X)

104 X T G				72 X A G			
T T G:32 X (27 X)	LEU 7	T G G: 39 X (27 X)	TRP 14	C A G:27 X (15 X)	GLN 7	A G A:28 X (19 X)	ARG 10
A T G:28 X (28 X)	MET 10	T G T: 27 X (27 X)	CYS 7	A A G:21 X (19 X)	LYS 6	A G T:17 X (19 X)	SER 9
G T G:26 X (27 X)	VAL 8	T G A: 21 X (28 X)	TER	G A G:17 X (18 X)	GLU 8	A G G:15 X (18 X)	ARG 6
C T G:18 X (22 X)	LEU 3	T G C: 17 X (22 X)	CYS 4	T A G: 7 X (19 X)	TER	A G C:11 X (15 X)	SER 4

60 X T A				52 X C G			
A T A:22 X (16 X)	ILE 7	T A C: 25 X (13 X)	TYR 12	T C G:15 X (14 X)	SER 4	C G A:23 X (14 X)	ARG 6
T T A:15 X (16 X)	LEU 0	T A T: 17 X (16 X)	TYR 6	A C G:14 X (14 X)	THR 6	C G G:17 X (13 X)	ARG 6
G T A:14 X (16 X)	VAL 1	T A A: 11 X (16 X)	TER	G C G:12 X (13 X)	ALA 1	C G T: 8 X (14 X)	ARG 1
C T A: 9 X (13 X)	LEU 2	T A G: 7 X (16 X)	TER	C C G:11 X (11 X)	PRO 2	C G C: 4 X (11 X)	ARG 1

FIG. 3. The 405-codon-long pollen–stigma self-incompatibility protein of the mustard plant, *Brassica oleraceae* (9).


43 X T
192 X G
12 X T G
87 X A G
6 X T A
45 X C G

C: 211 (0.335)	G: 192 (0.305)	A: 184 (0.292)	T: 43 (0.068)
T C: 23 X (15 X)	T G: 12 X (13 X)	T A: 6 X (13 X)	T T: 2 X (3 X)
A G: 87 X (59 X)	G G: 48 X (59 X)	C G: 45 X (64 X)	T G: 12 X (13 X)

G T G:	8 X (4 X)	VAL 6	T G G:	7 X (4 X)	TRP 0	A A G:	62 X (25 X)	LYS 57	A G A:	32 X (25 X)	ARG 0
C T G:	3 X (4 X)	LEU 3	T G T:	3 X (1 X)	CYS 0	C A G:	14 X (31 X)	GLN 1	A G C:	29 X (31 X)	SER 4
A T G:	1 X (3 X)	MET 1	T G C:	1 X (4 X)	CYS 0	G A G:	9 X (26 X)	GLU 4	A G G:	23 X (26 X)	ARG 1
T T G:	0 X (1 X)	LEU 0	T G A:	1 X (3 X)	TER	T A G:	2 X (5 X)	TER	A G T:	3 X (5 X)	SER 0
C T A:	4 X (2 X)	LEU 0	T A C:	3 X (2 X)	TYR 1	C C G:	25 X (15 X)	PRO 1	C G C:	21 X (15 X)	ARG 0
G T A:	2 X (2 X)	VAL 2	T A G:	2 X (2 X))	TER	G C G:	14 X (14 X)	ALA 10	C G G:	12 X (14 X)	ARG 0
A T A:	0 X (2 X)	ILE 0	T A A:	1 X (2 X)	TER	T C G:	4 X (3 X)	SER 0	C G T:	9 X (4 X)	ARG 1
T T A:	0 X (0 X)	LEU 0	T A T:	0 X (0 X)	TYR 0	A C G:	2 X (13 X)	THR 0	C G A:	3 X (13 X)	ARG 0

FIG. 4. The 210-codon-long rainbow trout H1 histone coding sequence (5).

Various Consequences of TA/CG Deficiency in Coding Sequences

The observed deficiency of TA dimer in all coding sequences insures that two base trimers TAA and TAG shall not recur too often, even in A+T-rich coding sequences (Fig. 3). This appears to be the very reason that, at the beginning of life on this earth eons ago, these two were assigned the role of being chain terminators. On the contrary, the third chain terminator, TGA as the base trimer, recurs rather often within coding sequences (Figs. 1–3). The view of Jukes (10) is that TGA was originally a tryptophan codon not a chain terminator, as it still is in the mitochondrial coding system. Transcription of coding sequences by RNA polymerase II has to start in front of each coding sequence. The scarcity of TA in coding sequences must have been a very useful marker in distinguishing coding sequences from their adjacent intergenic spacers. Both CAT box and TATA box as transcription initiation signals are deliberately rich in TA. Accordingly, these signal sequences are not likely to be present within coding sequences. The scarcity of TA in all coding sequences entails that tyrosine encodable only by TAT and TAC shall be scarce in all proteins, including those encoded by A+T-rich sequences (Fig. 3). Fig. 3 shows that the tyrosine content of the pollen–stigma self-incompatibility protein of the mustard plant remained at 4.4% in spite of the A+T richness of its coding sequence.

The sole function of thyroglobulin is to supply iodinated tyrosine for the synthesis of thyroid hormones. Yet, this giant precursor, 2,769-residue-long in the case of bovine, contains only 68 tyrosine residues (2.5%). Ironically, there are twice more (138) phenylalanine residues in thyroglobulin (11). Although tyrosine and phenylalanine are exceedingly similar, the latter is totally irrelevant to the assigned function of thyroglobulin. Such is the consequence of TA underrepresentation in coding sequences. Were tyrosine as common as phenylalanine, thyroid hormones would probably have been derived directly from a precursor protein containing Tyr-Tyr-Tyr-Tyr homotetrapeptide.

The scarcity of the CG dimer in mammalian coding sequences has usually been attributed to methylation of CpG, methylated CpG being converted to TG and CA (12). Indeed, the underrepresented CG is compensated by the overabundance of TG as we have seen, and CA dimer too tends to recur more often than expected in many coding sequences. Nevertheless, the fact remains that DNA methylation, including CpG methylation, is of very ancient origin, having been present already in bacteria in conjunction with restriction enzymes. If certain modern organisms such as *Drosophila* have lost the methylation mechanism, it must merely represent the secondary loss. Thus, the question arises, since the conversion of methylated CG to TG and CA has been going on in genomes of most organisms since time immemorial, why should there still be G+C-rich coding sequences today? Such coding sequences as seen in Fig. 2 contain a large number of CG dimers in the absolute sense but not in the observed-versus-expected sense. From the above considerations, I would rather think that the deficiency of CG was also the general characteristic of coding sequences from the very beginning. Perhaps, in multicellular organisms, CpG methylation is largely avoided in the germ line.

While tyrosine is a rare residue in all proteins, this is not the case with arginine as we have seen in Figs. 2 and 3. This is in part due to the assignment of four codons beginning with CG to arginine, in contrast to only two beginning with TA to tyrosine; but assignment of two additional codons AGA and AGG to arginine insured that, in spite of CG underrepresentation, there shall be no shortage of arginine. This assignment of two additional CG-free codons to arginine at the beginning of life on this earth indicates, it seems to me, that the deficiency of CG was also the general characteristic of all coding sequences from the very beginning. The scarcity of CG dimer in coding sequences must also have served as a useful marker in distinguishing them from their intergenic spacers. This must be the very reason that methylation of CG islands that are to be found only in intergenic spacers acquired a regulatory role in higher vertebrates (13).

1. Ohno, S. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 6486–6490.
2. Ohno, S. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 4378–4382.
3. Ohno, S. (1987) *J. Mol. Evol.* **25,** 325–329.
4. Grantham, R., Gautier, C., Gouy, M., Jacob, M. & Mercier, R. (1981) *Nucleic Acids Res.* **9,** 543–574.

5. Mezquita, J., Connor, W., Winkfein, R. J. & Dixon, G. H. (1985) *J. Mol. Evol.* **21**, 209–219.
6. Watson, H. C., Walker, N. P. C., Shaw, P. J., Bryant, T. N., Wendell, P. L., Fothergill, L. A., Perkins, R. E., Conroy, S. C., Dobson, M. J., Tuite, M. F., Kingsman, A. J. & Kingsman, S. M. (1982) *EMBO J.* **1**, 1635–1640.
7. Michelson, A. M., Markham, A. F. & Orkin, S. H. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 472–476.
8. Greene, G. L., Gilna, P., Waterfield, M., Baker, A., Hort, Y. & Shine, J. (1986) *Science* **231**, 1150–1154.
9. Nasrullah, J. B., Kao, T.-H., Chen, C.-H., Goldberg, M. L. & Nasrullah, M. E. (1987) *Nature (London)* **326**, 617–619.
10. Jukes, T. H. (1983) *Nature (London)* **301**, 19–20.
11. Mercken, L., Simons, M.-J., Swillens, S., Massaer, M. & Vassart, G. (1985) *Nature (London)* **316**, 647–651.
12. Liskay, R. M. & Evans, R. J. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 4895–4898.
13. Riggs, A. D. (1975) *Cytogenet. Cell Genet.* **14**, 9–25.