Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 192–196, January 1989
Evolution

Evolution of the genome and the genetic code: Selection at the dinucleotide level by methylation and polyribonucleotide cleavage

(nuclease/stop codon)

ERNEST BEUTLER†, TERRI GELBART†, JIAHUAI HAN‡, JAMES A. KOZIOL†, AND BRUCE BEUTLER‡

†Scripps Clinic and Research Foundation, Research Institute of Scripps Clinic, 10666 North Torrey Pines Road, La Jolla, CA 92037; and ‡Howard Hughes Medical Institute, 5323 Harry Hines Boulevard, Dallas, TX 75235-9050

Contributed by Ernest Beutler, October 6, 1988

ABSTRACT **Noting the scarcity of CpG dinucleotide in total genomic DNA derived from higher organisms and the scarcity of TpA dinucleotide in total genomic DNA derived from most life forms, we examined the distribution of these dinucleotides in sequences derived from functionally distinct types of human DNA, including mitochondrial DNA, intergenic DNA, intron DNA, and DNA destined to be represented in the cytoplasm as mRNA, tRNA, or rRNA. While CpG frequency has fallen to its lowest levels in DNA that is transcriptionally silent, TpA is most stringently excluded in DNA destined to be expressed as mRNA in the cytosol. This observation suggests that the selective pressures leading to the removal of CpG and TpA operate at different levels. With respect to TpA, dinucleotide scarcity may reflect a requirement for mRNA stability and may indicate the action of UpA-selective ribonucleases. We propose that, by reason of its instability, UpA must have been very rare in primordial RNA. Therefore, tRNA with the anticodon for this dinucleotide may have failed to evolve, making UpA the primordial doublet "stop" codon. The modern triplet code has faithfully conserved this arrangement in the two universal stop codons, UAA and UAG.**

In 1961, at a time when the determination of DNA sequences was but a dream, Kornberg and his associates (1, 2) developed the ingenious technique of nearest-neighbor analysis and found that representation of some dinucleotides in DNA deviated significantly from that expected were the mononucleotides associated randomly with each other. Specifically, there seemed to be a marked deficiency in the number of CpG dinucleotides and a slight but significant underrepresentation of TpA dinucleotides in animal and plant cells. With the development of techniques for sequencing large stretches of DNA and with the establishment of computer-accessible data bases, sequence analyses also revealed a scarcity of CpG and TpA dinucleotides (3–5).

Underrepresentation of these dinucleotides could have evolved because of constraints operating at the level of DNA, primary messenger, processed messenger, or protein. Although a number of studies comparing the frequency of dinucleotides in different species have been performed (5–7), the distribution of dinucleotides in functionally different types of DNA has received scant attention.

We have now studied the occurrence of CpG and TpA dinucleotides in different functional domains of DNA. While CpG frequency is lowest in transcriptionally silent DNA, TpA is most stringently excluded from DNA destined to be expressed as cytosolic mRNA. We show that UpA dinucleotides in RNA are unstable to nucleosidic cleavage and suggest that instability of this dinucleotide may account for the fact that the conserved stop codons start with UpA.

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.

MATERIALS AND METHODS

Data were from the GenBank and the EMBL data libraries supplemented by sequences determined in our own laboratory and made available through the generosity of E. W. Davie (University of Washington, Seattle). Only human genes were examined and were selected for no other reason than that both complete cDNA and genomic sequences were available. Immunoglobulin genes and other genes subject to rearrangement were excluded from analysis. The sequences that were used as a basis for this analysis are summarized in Table 1.

Mononucleotide and dinucleotide compositions were determined by computer analysis (11). The results were usually expressed as the percentage of dinucleotides, uncorrected for the relative amounts of thymine, adenine, cytosine, and guanine in the sequence, because our interest was in the relative prevalence of the dinucleotides in different DNA domains, regardless of the mechanisms by which the levels that had been achieved had evolved. Because in most of the genes studied the proportion of A:T:C:G approaches 1:1:1:1, correction for prevalence of the mononucleotides does not change any of our conclusions. The genes for tRNA and rRNA, however, are unusually rich in guanine and cytosine, presumably because of the constraints imposed by the required secondary structure of the RNA. In these cases the analysis presented is based on the expected frequency of dinucleotides as derived from the mononucleotide frequency.

For each sequence studied, appropriate-order Markov chains were fit to the nucleotide frequencies in the various DNA domains. Frequencies of TpA and CpG dinucleotides were then taken to be approximately normally distributed, with means devolving from the limiting stationary probabilities for the appropriate-order transition matrix and covariances devolving from the corresponding asymptotic formulas (12, 13). In this way nucleotide frequencies in different genomic regions could be compared by means of test statistics that are expressible as linear combinations of these approximately normal random variables.

Analysis of preferred cleavage sites by macrophage cytoplasmic RNase was carried out with partially purified enzyme (J.H. and B.B., unpublished data), dissolved at a concentration of 200 units/ml in 10 mM Tris, pH 8.0, containing 0.1% octyl glucoside. Terminally labeled RNA oligonucleotides, bearing the cachectin-derived UUAUUUAU sequence (14) were prepared by *in vitro* transcription in the presence of [γ-^{32}P]GTP. The sequence (5′ to 3′) of this oligomer was *GAAUACAAUUAUUUAUUAUUUAUUUAUUAUUUAUUUAUUUC (asterisk denotes the label). An oligonucleotide bearing all 16 possible dinucleotides was prepared by chemical synthesis by use of an Applied Biosystems synthesizer and was then phosphorylated with [γ-^{32}P]ATP using polynucleotide kinase. The sequence (5′ to 3′) of this

Table 1. Source and data bank designation of human genes studied

Gene product (abbreviation)	Designation	Source or ref.
Interleukin 1 (IL-1)^G	HUMIL1AG.PR	GenBank
Interleukin 1^m	HUMIL1.PR	GenBank
Apolipoprotein AII (LIP)^G	HUMAPOA2I.PR	GenBank
Apolipoprotein AII^m	HSAPOA2A.EMBL	EMBL
Thymidine kinase (TK)^G	HUMTKRA.PR	GenBank
Thymidine kinase^m	HUMTK.PR	GenBank
Factor IX (FIX)^G	HUMFIXG.PR	GenBank
Factor IX^m	HUMFIX.PR	GenBank
Albumin (ALB)^G	HUMALBGC.PR	GenBank
Albumin^m	HUMALBAF1.PR	GenBank
Adenosine deaminase (ADA)^G	HUMADAG.PR	GenBank
Adenosine deaminase^m	HUMADA.PR	GenBank
Tubulin (TUB)^G	HUMTUBAG.PR	GenBank
Tubulin^m	HUMTUBAK.PR	GenBank
Protein C (PRC)^G	HUMPRCA.PR	GenBank
Protein C^m	HUMPRCM.PR	GenBank
GM-CSF (CSF)^G	HSGCSFG.EMBL	EMBL
GM-CSF^m	HSGCSFR.EMBL	EMBL
Erythropoietin (EPO)^G	HSERPG.EMBL	EMBL
Erythropoietin^m	HSERP.EMBL	EMBL
Haptoglobin (HPT)^G	HUMHPARS1.PR	GenBank
Haptoglobin^m	HUMHPAB.PR	GenBank
Cytochrome P-450 (CYP)^G	HUMCYP450.PR	GenBank
Cytochrome P-450^m	HUMCYP145.PR	GenBank
$\beta\gamma\delta\varepsilon$-Globin (Hb)^G	HUMHBB.PR	GenBank
β-Globin^m	HSBGL1.EMBL	EMBL
$\gamma\delta\varepsilon$-Globin^m		8
Glucocerebrosidase (GC)^G		9
Glucocerebrosidase (GC)^m	HUMGCB.PR	GenBank
α_1-Antitrypsin (AT)^G	HSA1ATP.EMBL	EMBL
α_1-Antitrypsin (AT)^m	HUMA1ATM.PR	GenBank
Prothrombin (PT)^G		*
Prothrombin (PT)^m	HUMTHBNA.PR	GenBank
Factor VII (F7)		
Factor VII (F7)^m	HUMFVII.PR	GenBank
Fibrinogen α peptide (FBA)^G		*
Fibrinogen α peptide (FBA)^m	HUMFBRA.PR	GenBank
Fibrinogen β peptide (FBB)^G		*
Fibrinogen β peptide (FBB)^m	HUMFBRB.PR	GenBank
Fibrinogen γ peptide (FBC)^G	HUMFBRG.PR	GenBank
Fibrinogen γ peptide (FBC)^m	HSFBRG.EMBL	EMBL
Met tRNA	HUMTGMI1.PR	GenBank
Leu tRNA	HUMTGPL1.PR	GenBank
Pro tRNA	HUMTGPL2.PR	GenBank
Thr tRNA	HUMTGPL2.PR	GenBank
Leu tRNA	HUMTRNLG.PR	GenBank
Phe tRNA	HUMTRF.ST	GenBank
Gly tRNA	HUMTRGCCC.ST	GenBank
Gly tRNA	HUMTRGGCC.ST	GenBank
His tRNA	HUMTRHJO1.ST	GenBank
Init Met tRNA	HUMTRMI.ST	GenBank
Asn tRNA	HUMTRN.ST	GenBank
Val tRNA	HUMTRV1B.ST	GenBank
5.85 RNA	HUMRRB.ST	GenBank
18.5 RNA	HUMRRN18S.ST	GenBank
28.5 RNA	HUMRGM.PR	GenBank
Mitochondrion (complete)	HUMNT.OR	GenBank
Human Y chromosome *Eco*RI fragment from pDP 1035		10

^{G}, Gene; ^{m}, mRNA; GM-CSF, granulocyte/macrophage colony-stimulating factor; init, initiator.
*E. W. Davie, personal communication.

oligomer was *AAAAAAAAAAAAUAGACUUGUCG-GCCA.

RESULTS

Frequency of TpA and CpG Dinucleotides in Different Functional Domains. The relative frequencies of CpG and TpA in different functional domains of the DNA genes were normalized with respect to the total frequency in each gene and averaged. The relative frequencies of these two dinucleotides in different domains of DNA are illustrated in Fig. 1.

Intron DNA vs. cDNA. A marked difference was found to exist between TpA and CpG frequencies in intron DNA as compared with cDNA. The cDNA had a strikingly low TpA content when compared with the remainder of the gene (χ^2_{40} = 469.1, $P < 0.00001$). The reverse was true of CpG dinucleotides (χ^2_{40} = 242.3, $P < 0.0001$); the deficiency of these dinucleotides was much less severe in cDNA than in the remainder of the gene. The dinucleotide frequencies of these regions are compared in Fig. 2.

As shown in this figure some cDNA sequences contained particularly few TpA dinucleotides, and in these genes dinucleotide frequency in the remainder of the gene was very low as well. For example, the cDNA for erythropoietin contained only 1.19% TpA dinucleotides, and the frequency in the introns was only 2.3%—both among the lowest in the entire series. In contrast, the cDNA for the C peptide of fibrinogen contained 5.74% TpA dinucleotides, whereas the frequency in the introns was 8.12%. The TpA frequencies in cDNA were correlated with those in introns ($r = 0.80$; $P < 0.001$). Similarly, the prevalence of CpG in the cDNA portion of the gene and the remainder were strongly correlated ($r = 0.82$; $P < 0.001$).

Coding vs. Noncoding Portions of the cDNA. There was a tendency for TpA dinucleotide frequencies to be lower in the coding than in the noncoding portion ($\chi^2_{40} = 99.0, P < 0.0001$), but the effect was smaller than that seen when intron DNA was compared with cDNA (Fig. 1). Conversely, the CpG frequency was lower in the noncoding portion than in the coding portion of the genes (χ^2_{40} = 142.1, $P < 0.0001$).

It is interesting to note that one of the highest relative TpA dinucleotide frequencies was found in the granulocyte/macrophage colony-stimulating factor cDNA, which is known to have a long TTATTTAT sequence in the 3'-untranslated portion (14).

Independence of Dinucleotide Frequency and Codon Preference. The low levels of TpA observed in coding regions was not simply a function of codon choice. Although codon usage fixes the frequency with which a dinucleotide is present in

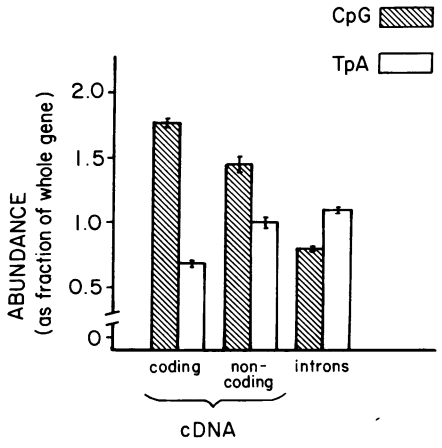

FIG. 1. Frequency of CpG dinucleotides in different functional portions of DNA expressed as a fraction of their occurrence in the whole gene. CpG dinucleotides occur least frequently in introns and most frequently in the coding region. TpA dinucleotides occur most frequently in the introns and least frequently in the coding region. The deviation bars represent one SE of the estimate.

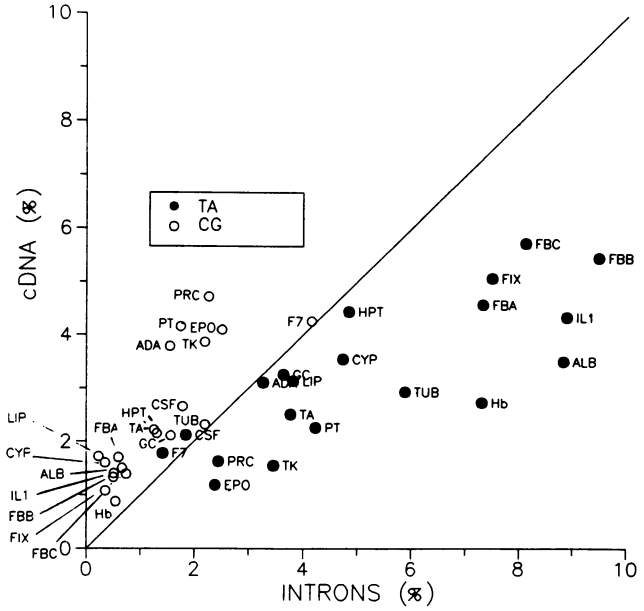

FIG. 2. Frequency of occurrence of TpA (●) and CpG (○) dinucleotides in the portion of DNA destined to become cDNA and in the introns. Abbreviations are given in Table 1. With only two exceptions TpA frequencies are higher in introns than in cDNA. In every case CpG frequency is higher in cDNA than in introns.

positions 1 and 2 and in positions 2 and 3 of a codon, it has no effect on the incidence of junctional TpA dinucleotides, in which the thymine occupies the third position of a codon and adenine the first position of the following codon. We calculated the observed frequencies of TpA dinucleotides in the 3/1 position for the 20 coding sequences studied; we also determined the first three moments of the permutation distributions of these frequencies—the reference set for each sequence being the particular observed codon triplet frequencies. In all cases the observed frequencies of TpA in the 3/1 junctional position was less than would be expected under a random permutation of the triplets.

We also obtained approximate P values (15) for assessing the hypothesis that the incidence of TpA in the 3/1 junctional position was congruent with the permutation distribution. Combining as before, we found Fisher's statistic $\chi^2_{40} = 185.5$, $P < 0.0001$; the incidence of TpA in the junctional position is significantly less than would be expected by chance. Thirty-eight and one-half percent of the TpA dinucleotides tallied occupied junctional positions; 35.1% occupied the 1/2 position, and 26.4% occupied the 2/3 position.

Mitochondrial DNA. Mitochondrial DNA contained 8.29% TpA and 2.63% CpG dinucleotides. Even taking into account the fact that mitochondrial DNA is somewhat poor in G+C, the TpA frequency was higher and the CpG lower than that expected on the basis of the frequency of the individual nucleotides.

DNA That Is Transcribed But Not Translated. We studied DNA that encodes RNA used by the cell for specialized purposes other than that of serving as a template for protein synthesis (namely, rRNA and tRNA). Data from DNA encoding all three ribosomal subunits and all transfer RNAs were pooled. The TpA dinucleotide frequency found in DNA of these types was extremely low (2.83%), and that of the CpG dinucleotide was high (10.8%). However, DNA specifying rRNA and tRNA is extremely rich in G+C, presumably because of the requirements for a stable secondary structure. The expected values of dinucleotides were therefore obtained adjusting for nucleotide frequencies. Using omnibus χ^2 statistics (16) the hypothesis that the observed frequencies are congruent with those expected from the permutation distribution may readily be assessed: the relevant statistic is suggestive with the tRNA data, $\chi^2_9 = 16.1$, $P = 0.065$ and is overwhelmingly significant with the rRNA data, $\chi^2_9 = 210.7$, $P < 0.0001$. The TpA frequency of the pooled data from the tRNA and rRNA was significantly decreased ($P = 0.042$), and that of CpG was modestly increased ($P = 0.11$).

DNA That Is Not Transcribed. Y chromosomal DNA (given to us by D. Page, Whitehead Institute, Cambridge, MA) representing DNA that had presumably been transcriptionally silent for a long evolutionary period was sequenced. The CpG frequency of 0.46% was among the lowest encountered in any of the DNA examined, and the TpA frequency at 6.4% did not differ significantly from binomial expectation.

UpA Specificity of Cytoplasmic Ribonucleases. A macrophage ribonuclease purified by a procedure to be described elsewhere (J.H. and B.B., unpublished work) was allowed to partially degrade terminally labeled RNA oligomers so as to determine base specificity. As shown in Fig. 3, preferential hydrolysis at UpA dinucleotides was seen.

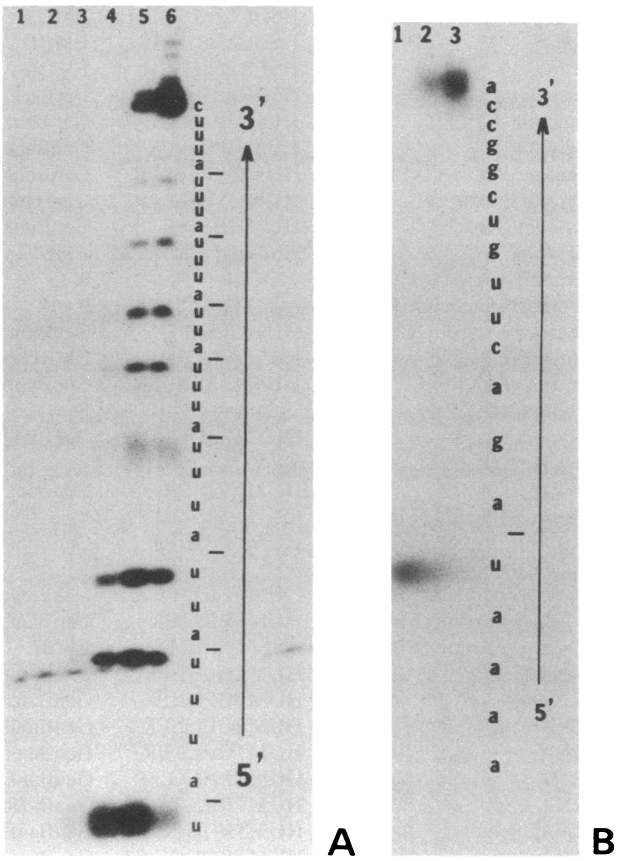

FIG. 3. Degradation of oligonucleotides by cytoplasmic macrophage RNase. The RNase was serially diluted in 3-fold increments with the same buffer as a diluent. One microliter of enzyme solution was then added to 4 μl of RNA oligonucleotide (≈500 dpm per system) in 10 mM Tris, pH 8.0/0.04% octyl glucoside. Five-minute digestion at 37°C was stopped by adding 5 μl of DNA sequencing buffer containing 50% (wt/vol) formamide. After the mixture was heated to 65°C for 5 min, electrophoresis was performed in a 15% polyacrylamide gel under denaturing conditions. (*A*) Degradation of the cachectin-derived UUAUUUAU oligomer by serially diluted enzyme (lanes 1–6). (*B*) Degradation of an oligomer containing all 16 dinucleotides by serially diluted enzyme (lanes 1–3). Hydrolysis of each oligomer occurs with high specificity at UpA linkages. Even when great care was taken to inactivate any contaminating ribonucleases, faint bands were present at the UpA sites, indicating a low level of spontaneous selective cleavage of the polymer at that dinucleotide (data not shown).

DISCUSSION

Examination of the CpG and TpA dinucleotide patterns in DNA reveal striking differences in the degree of depletion of these dinucleotides in different functional domains of DNA. Combined with other characteristics of coding DNA sequences, such as the existence of long open reading frames, the relatively low TpA and high CpG frequencies of coding portions may prove to aid in identification of coding regions when the genome is sequenced. Much more interesting to us, however, is the question of how these differences in dinucleotide frequencies arose in the evolution of all species in the case of TpA and in vertebrates in the case of CpG.

Some explanations have previously been proffered for the underrepresentation of CpG and TpA dinucleotides. The avoidance of CpG is striking in vertebrates and has received the most attention. Salser (17) proposed that the depletion of this dinucleotide might be related to the fact that CpG is a mutational hot spot. This view has been supported by Bird (18) but has been rejected by Lennon and Fraser (19) based on an analysis of globin genes. The latter authors and Nussinov (3) have preferred the view that the low CpG frequency points to a structural constraint operating at the DNA level.

Recent analyses of human mutations (20, 21) have emphasized how frequently mutations of CpG dinucleotides are responsible for the development of disease, and our data are entirely compatible with the possibility that vertebrate DNA has gradually undergone depletion of CpG dinucleotides because of the methylation and subsequent oxidation of cytidine to thymidine. It appears reasonable that a strong selection against this conversion would exist in functional portions of genes, whereas in other portions of the genome this change would be relatively unopposed. The degree of depletion predicted by this scenario is precisely what we found. Y chromosomal DNA contains very few CpG dinucleotides, whereas the highest level of this dinucleotide is observed in cDNA and genes coding for tRNA and rRNA—sequences known to be highly conserved.

The low level of TpA dinucleotides in DNA has received relatively little attention. Unlike CpG, TpA is relatively scarce even in prokaryotes, and its deficiency probably has quite a different basis than the deficiency of CpG (22). The present investigations show that the frequency of TpA dinucleotides is not reduced in DNA coding for RNA that is not expressed in the cytoplasm. In this context we note that the occurrence of TpA dinucleotides found in human DNA by Swartz *et al.* (1) was nearly as expected, whereas Nussinov (5) found the level of TpA dinucleotides to be very low. The explanation for this discrepancy is now apparent; the sequences studied by Nussinov came from a data base heavily biased by the inclusion of cDNA and transcribed DNA sequences, whereas Swartz *et al.* investigated sheared whole genomic DNA.

It seemed to us possible that the avoidance of TpA might have nothing to do with the properties of DNA. Rather TpA scarcity might be related to properties of RNA, particularly because the UUAUUUAU sequence can destabilize mRNA (23). If the constraint limiting the frequency of TpA dinucleotides functioned at the primary messenger level, then there should be differences in TpA prevalence in transcribed when compared with untranscribed regions of DNA. If only processed mRNA were unable to tolerate these sequences, then the TpA content of this portion of the DNA should differ from others. Our findings show that the TpA deficiency is, indeed, most marked in DNA destined to become processed mRNA. DNA destined to be transcribed into introns is much less affected.

The avoidance of UpA in mRNA might, of course, merely reflect constraints produced by the genetic code for reasons not yet understood. Two of the three stop codons contain the TpA dinucleotide, and inspection of codon utilization (24, 25) indicates that TpA dinucleotides are found in the least frequently utilized codons. Tyrosine is an exception; all tyrosine codons contain this dinucleotide. Then, too, tyrosine is a relatively infrequent constituent of protein. Did the genetic code itself, as adopted by prokaryotes and their successors, dictate the scarcity of TpA? Several considerations lead us to believe that it did not.

(*i*) TpA is strongly disfavored at codon junctions (3/1 positions), despite the fact that TpA has no defined coding function in this frame. Thus, codon choice alone cannot explain the observed scarcity of TpA.

(*ii*) Although the genetic code used by mitochondria is very similar to that used in the translation of nuclear genes, mitochondrial DNA (though largely comprised of coding sequence) is rich in TpA. This fact suggests that the intracellular milieu in which an RNA is expressed, rather than codon choice, dictates TpA frequency in the corresponding DNA sequence.

(*iii*) DNA directing the production of RNA that is not translated (i.e., genes encoding rRNA and tRNA) also contains significantly less than the expected levels of TpA.

As these considerations also indicate, a common attribute of TpA-poor DNA seems to be that it is transcribed into RNA that is expressed in the cell cytoplasm. Nontranscribed (Y-chromosomal) DNA, DNA encoding mRNA that is expressed only in mitochondria, and DNA encoding RNA that is degraded within the nucleus (e.g., intron DNA) are relatively rich in TpA.

Given that codon choice does not dictate the scarcity of TpA, we may well wonder whether the converse occurred. Did the selective pressure that led to scarcity of TpA in DNA encoding cytoplasmically expressed RNA shape the modern genetic code, so as to cause the relative dearth of codons employing this dinucleotide?

The recent discovery of a stereotypic, repeating TTATTTAT sequence in the 3′-untranslated region of certain genes (14) and the observation that the corresponding UUAUUUAU sequences are destabilizing to mRNA molecules that contain them (23) suggest that the selective pressure responsible for avoidance of TpA might be nucleolytic in character. The instability sequence is very rich in TpA dinucleotide—28.6% in the octamer. Our investigations of the cleavage of short strands of RNA by a cytoplasmic RNase strongly support the possibility that UpA sequences play a role in stability of the messenger. Digestion of various RNA constructs showed that highly selective cleavage takes place between 5′ uracil and 3′ adenine residues, corresponding to the TpA dinucleotide in DNA.

In most modern mRNAs UAA, UAG, and UGA serve as stop codons. The assignment of UGA to this function is the least robust of the three, because UGA is not a stop codon in mitochondria (26), coding for tryptophan instead and for selenocysteine in mammalian glutathione peroxidase (27) and in *Escherichia coli* formate dehydrogenase (28). Thus, only UAA and UAG may be considered to be universal stop codons, and both begin with UpA. UpA may well have served as the primordial "stop" codon in an ancestral doublet code and was then adopted by the modern triplet code.

The "choice" of UpA as the primordial stop codon was probably the result of its inherent instability. In primordial systems in which replication of RNA-like polymers may have been the basis of life (29), catalyzed cleavage of the polymer between uridine and adenine would eliminate all UpA dinucleotides. Because no UpA dinucleotides would have been present, no tRNA would have evolved for their translation when RNA began to serve as a template for the assembly of amino acids into protein. Thus, even when mechanisms for

the stabilization of UpA dinucleotides against cleavage evolved by the development of secondary structure and/or protective ligands, no mechanism for their translation would have existed: they would have functioned as "stop" codons. We propose that in this way the instability of UpA dinucleotides led to their selection as "stop" in the primordial genetic code and that this early mechanism is now reflected in the modern triplet code that has evolved.

Undoubtedly many factors regulate the stability of RNA within the cell. The occurrence of UpA dinucleotides in the RNA is one of these factors and may be the most ancient. That intron RNA is relatively rich in these dinucleotides suggests that selection against UpA occurs in RNA at the cytoplasmic level rather than operating in the nucleus, where the introns are removed. Moreover, it may be that messenger formed in mitochondria is shielded from UpA-specific nucleases that exist in the cytoplasm, because UpA dinucleotides are quite common in the mitochondria.

Until now our examination of higher-order patterns in which UpA dinucleotides occur in mRNA has been unproductive, except, of course, for the specific UUAUUUAU pattern in the 3'-untranslated portion (14). Because such patterns might well involve the secondary structure of the RNA, the relationship between overall sequence and stability could be very complex.

This work was supported by National Institutes of Health Grants DK36639, R01-CA45525, and RR00833 and by a fund provided by the Sam Stein and Rose Stein Charitable Trust. This is publication 5373BCR from the Research Institute of Scripps Clinic.

1. Swartz, M. N., Trautner, T. A. & Kornberg, A. (1962) *J. Biol. Chem.* **237**, 1961–1967.
2. Josse, J., Kaiser, A. D. & Kornberg, A. (1961) *J. Biol. Chem.* **236**, 864–875.
3. Nussinov, R. (1984) *J. Mol. Evol.* **20**, 111–119.
4. Nussinov, R. (1981) *J. Biol. Chem.* **256**, 8458–8462.
5. Nussinov, R. (1984) *Nucleic Acids Res.* **12**, 1749–1763.
6. Elton, R. A. (1973) *J. Mol. Evol.* **2**, 293–302.
7. Setlow, P. (1976) in *Handbook of Biochemistry and Molecular Biology*, ed. Fasman, G. D. (CRC Press, Cleveland), pp. 312–318.
8. Collins, F. S. & Weissman, S. M. (1984) *Prog. Nucleic Acid Res. Mol. Biol.* **31**, 439–458.
9. Horowitz, M., Wilder, S., Horowitz, Z., Reiner, O., Gelbart, T. & Beutler, E. (1989) *Genomics*, in press.
10. Page, D. C., Mosher, R., Simpson, E. M., Fisher, E. M. C., Mardon, G., Pollack, J., McGillivray, B., de la Chapelle, A. & Brown, L. G. (1987) *Cell* **51**, 1091–1104.
11. Devereux, J., Haeberli, P. & Smithies, O. (1984) *Nucleic Acids Res.* **12**, 387–395.
12. Billingsley, P. (1961) *Ann. Math. Stat.* **32**, 12–40.
13. Good, I. J. (1963) *J. R. Stat. Soc. Ser. B* **25**, 383–391.
14. Caput, D., Beutler, B., Hartog, K., Brown-Shimer, S. & Cerami, A. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 1670–1674.
15. Mueller, P. H. & Vahl, H. (1976) *Biometrika* **63**, 191–194.
16. Anderson, T. W. & Goodman, L. A. (1957) *Ann. Math. Stat.* **28**, 89–110.
17. Salser, W. (1975) *Cold Spring Harbor Symp. Quant. Biol.* **40**, 985–1002.
18. Bird, A. P. (1980) *Nucleic Acids Res.* **8**, 1499–1504.
19. Lennon, G. G. & Fraser, N. W. (1983) *J. Mol. Evol.* **19**, 286–288.
20. Youssoufian, H., Kazazian, H. H., Jr., Phillips, D. G., Aronis, S., Tsiftis, G., Brown, V. A. & Antonarakis, S. E. (1986) *Nature (London)* **324**, 380–382.
21. Cooper, D. N. & Youssoufian, H. (1988) *Hum. Genet.* **78**, 151–155.
22. Santibanez-Koref, M. & Reich, J. G. (1986) *Biomed. Biochim. Acta* **45**, 1105–1109.
23. Shaw, G. & Kamen, R. (1986) *Cell* **46**, 659–667.
24. Lathe, R. (1985) *J. Mol. Biol.* **183**, 1–12.
25. Alff-Steinberger, C. (1987) *J. Theor. Biol.* **124**, 89–95.
26. Fox, T. D. (1987) *Annu. Rev. Genet.* **21**, 67–91.
27. Chambers, I. & Harrison, P. R. (1987) *Trends Biochem. Sci. Rev.* **12**, 255–256.
28. Zinoni, F., Birkmann, A., Leinfelder, W. & Böck, A. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 3156–3160.
29. Cech, T. R. & Bass, B. L. (1986) *Annu. Rev. Biochem.* **55**, 599–629.