Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 3391–3395, June 1983
Evolution

The primitive code and repeats of base oligomers as the primordial protein-encoding sequence

(origin of life/periodical polypeptides)

SUSUMU OHNO AND JÖRG T. EPPLEN*

City of Hope Research Institute, Duarte, California 91010

Contributed by Susumu Ohno, February 18, 1983

ABSTRACT **Even if the prebiotic self-replication of nucleic acids and the subsequent emergence of primitive, enzyme-independent tRNAs are accepted as plausible, the origin of life by spontaneous generation still appears improbable. This is because the just-emerged primitive translational machinery had to cope with base sequences that were not preselected for their coding potentials. Particularly if the primitive mitochondria-like code with four chain-terminating base triplets preceded the universal code, the translation of long, randomly generated, base sequences at this critical stage would have merely resulted in the production of short oligopeptides instead of long polypeptide chains. We present the base sequence of a mouse transcript containing tetranucleotide repeats conserved during evolution. Even if translated in accordance with the primitive mitochondria-like code, this transcript in its three reading frames can yield 245-, 246-, and 251-residue-long tetrapeptidic periodical polypeptides that are already acquiring longer periodicities. We contend that the first set of base sequences translated at the beginning of life were such oligonucleotide repeats. By quickly acquiring longer periodicities, their products must have soon gained characteristic secondary structures—α-helical or β-sheet or both.**

Implicit in the immortal dictum *omnis cellula a cellula* of Rudolph Virchow was the realization that all the diverse living organisms of this earth are branches of the gigantic monophyletic tree that germinated in the primeval environs eons ago. The ultimate challenge to biologists then is to understand the process of spontaneous generation that gave rise to the very first cell, for it is this cell that served as the primordium of the gigantic monophyletic tree.

Regardless of its ultimate origin, terrestrial or extraterrestrial (1–3), the *conditio sine qua non* of this spontaneous generation was prebiotic nucleic acid self-replication (4) followed by the emergence of primitive tRNAs that began to translate the base sequence of nucleic acids to the amino acid sequence of functionally far more versatile polypeptide chains. It is at this stage, however, that the proposition of spontaneous generation suddenly appears to be a lost cause (2, 3). The reason is that the newly emerged prebiotic translation machinery had to cope with base sequences that were not preselected to be coding sequences. Consequently, the probability of these sequences being translated to polypeptide chains of meaningful lengths should have been practically nil.

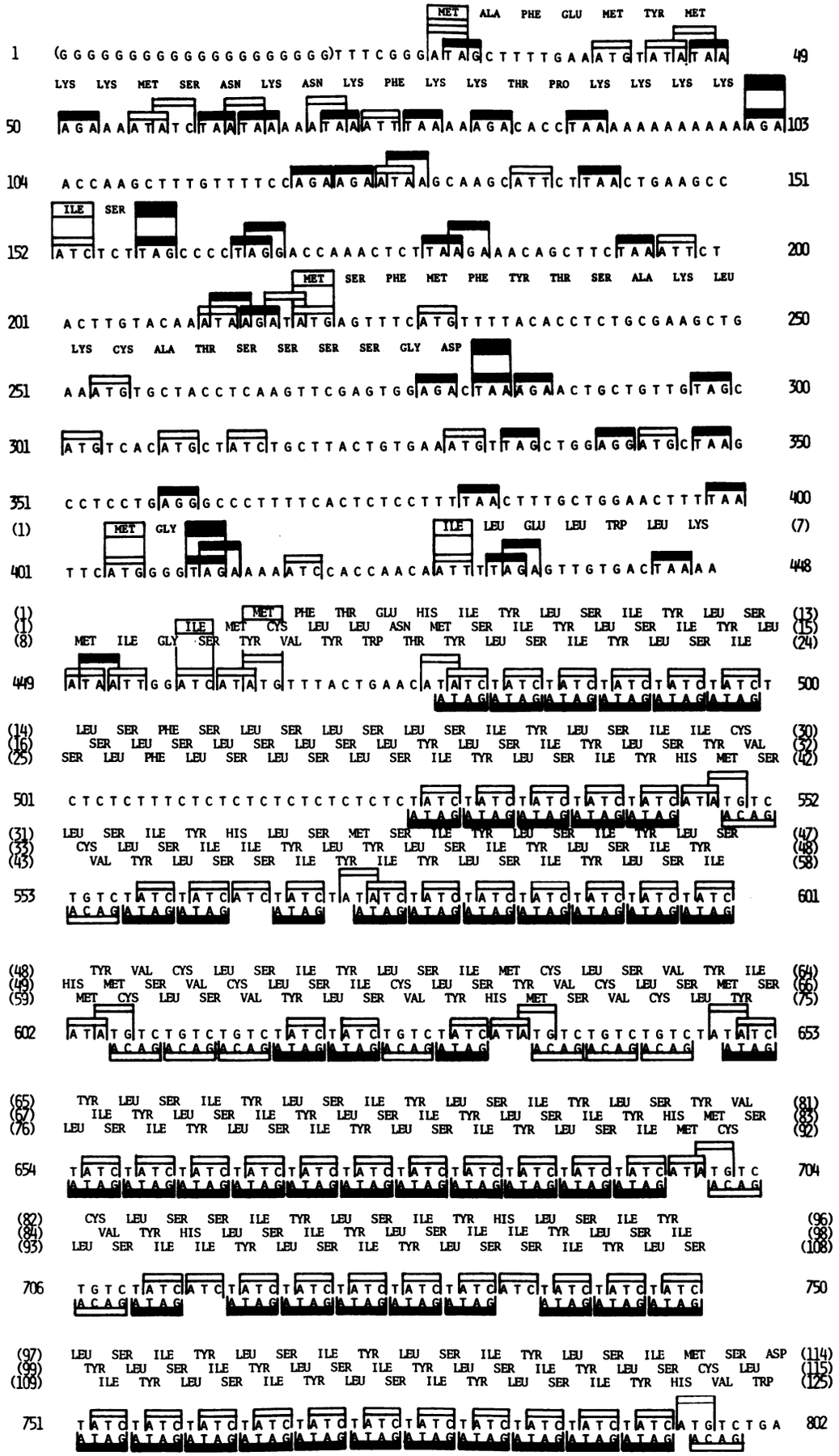
Recent findings on the simpler translation machinery of the mammalian mitochondrial genome (5) appears to indicate that the universal code as we understand it (6) might not have existed at the beginning of the life. Rather, life started with the simpler mitochondria-like code involving fewer species of tRNAs and, therefore, fewer anticodons of greater infidelity with respect to their codon recognition (7). There are two essential differences between the universal code and the simpler mitochondria-like code. (*i*) Whereas one base triplet—AUG for methionine—serves as the chain-initiator in the former, four base triplets—AUU, AUC, AUA, and AUG—fulfill this function in the latter (the first two triplets specify isoleucine and the last two specify methionine). (*ii*) Compared to three chain-terminators (UAA, UAG, and UGA) in the former, the latter has four (UAA, UAG, AGA, and AGG); UGA in the latter becomes the second tryptophan codon (5).

Under the primitive code nearly all the long randomly generated base sequences contain many short, open reading frames instead of one long one

In order to specify a polypeptide chain only 100 amino acid residues long, a 300-base-long coding sequence is required. However, because 4 of the 4^3 base triplets are chain terminators in the primitive mitochondria-like code, a vast majority of the 4^{300} randomly generated 300-base-long sequences should contain somewhere between 14 and 23 chain terminators in each; the average is 18.75. Because each should also contain nearly as many chain initiators as chain terminators, most of these randomly generated 300-base-long sequences can only specify a number of oligopeptides. This point is well illustrated by a 5′ noncoding segment of a mouse transcript shown at the top of Fig. 1 *Left*. Because the first 20 Gs of the vector origin should be discounted, this segment occupying the first $8\frac{1}{2}$ rows of Fig. 1 *Left* is 422 bases long. Included in it are 25 chain initiators (3 ATT, 4 ATC, 9 ATA, and 9 ATG) and 35 chain terminators (15 TAA, 8 TAG, 3 AGG, and 9 AGA). Of those, four pairs nullify each other by forming initiator-terminator hexamers (ATA/TAA of row 1, ATC/TAA and ATT/TAA of row 2, and ATA/AGA of row 5). Furthermore, in-frame initiators located within longer reading frames should be neglected. The above leaves us with 13 short, open reading frames within the 427-base-long nonrepetitious base sequence at the top of Fig. 1 *Left* that has not been vigorously surveyed by natural selection. In descending order of their lengths, these 14 oligopeptides are made of 24, 21, 14 (two), 10, 8, 6 (two), 5 (two), 4, 3, and 2 (two) amino acid residues. In Fig. 1 *Left*, the longest two open reading frames straddling rows 1 and 2 as well as rows 5 and 6, and two shortest dipeptide reading frames (rows 4 and 9) are accompanied by corresponding amino acid residues shown in smaller letters. For spontaneously generating the first cell, the just-emerged primitive translation machinery had to produce polypeptides of respectable lengths; random-sequence oligopeptides no doubt were already in existence in prebiotic environs.

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.

* Present address: Max Planck Institute for Immunology, D7800 Freiburg, Federal Republic of Germany.



(Fig. 1. continues on next page.)

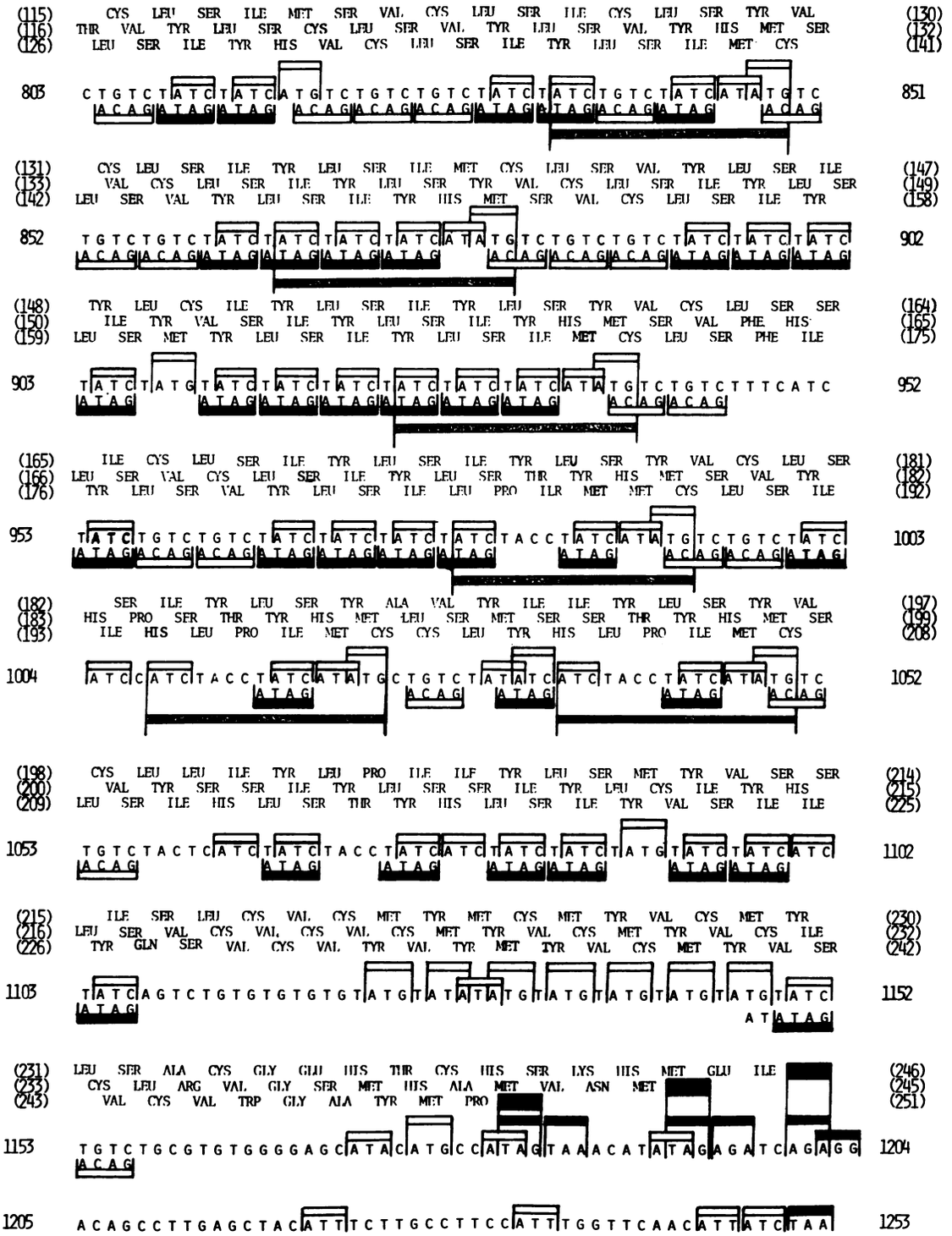


FIG. 1. The sense-strand nucleotide sequence of a cloned cDNA copy of one mouse transcript containing tetrameric repeats in open reading frames, shown in two parts: *Left*, up to base 802; *Right*, bases 803–1,253. After screening of approximately 100,000 clones of one male mouse cDNA library (8) by using a subfragment of pErs5 cloned snake satellite DNA (originally identified as Ch4A Ers6 in ref. 9) as a hybridization probe, this 2,500-base-long (total length) clone was identified. Its sequence was determined by the methods of both Maxam and Gilbert (10) and Sanger *et al.* (11). Detailed descriptions of the experimental procedures will be published elsewhere (12, 13). All three long reading frames in the segment stretching from row 9 of *Left* to the last row of *Right* are accompanied by corresponding amino acid residues in larger capital letters. The two longest (24 and 21 amino acid residues long) and two shortest (dipeptides) reading frames encountered in the 5′ apparently random base sequence (first $8^1/_2$ rows of *Left*) are also accompanied by amino acid residues shown in smaller letters. All chain initiators (ATT, ATC, ATA, and ATG) and chain terminators (TAA, TAG, AGG, and AGA) are individually identified by open (initiators) or solid (terminators) bars placed above these base triplets. Each cardinal base tetramer is shown in the double-stranded $\frac{TATC}{ATAG}$ form underlined by a solid bar, and so is its most numerous single-base-deviant $\frac{TGTC}{ACAG}$ also underlined by an open bar. Three exact two-single-base-deviant and one double-base-deviant copies of the nucleotide sequence A-T-C-T-A-C-C-T-A-T-C-A-T-A-T-G in *Right* are also underlined by shaded bars. Bases shown in rows of 50 each as a rule are identified by numbers at both sides of each row. Three numbers in parenthesis shown at each side of row 9 of *Left* to row 8 of *Right* are with regard to amino acid residues specifiable by three long open reading frames.

Primordial protein-encoding sequences must have been repeats of base oligomers that specified periodical polypeptides

What if those prebiotic base sequences subjected to the initial translation attempt were repeats of base oligomers? Oligomeric repeats may generate a chain terminator not only within that oligomer itself but also at the junction—e.g., A-T-A-G-T-A contains not only a TAG chain terminator within but also its repeats generate TAA chain terminator at each junction. Accordingly, a fraction free of chain terminators in the M-X repeats of n-base-long randomly generated oligomers is represented by $4^n - (4^{n-3} \times 4n)$, regardless of whether n is 2 or 10,000. It follows then that 68.75% of the randomly generated pentameric repeats and 60.01% of the hexameric repeats shall forever remain free of chain terminators, no matter how long they become. Were the just-emerged prebiotic translation machinery to have encountered not a collection of long randomly generated base sequences but repeats of randomly generated base oligomers, it would have produced various long periodical polypeptides instead of a collection of random sequence oligopeptides.

Because the transition from the ordered state to the disordered state is always a far more likely event than the converse, it may be assumed with reasonable certainty that all of the modern polypeptide chains of diverse functions have descended from those that assumed either of the two characteristic secondary structures—α-helical and β-sheet. The point is that such secondary structures are far more readily formed by descendants of periodical polypeptides specified by oligomeric repeats than those of so-called unique-sequence polypeptides. For example, the modern coding sequence for vertebrate α-helical collagen is repeats of the 54-base-long unit sequence (14), thus giving the 18-amino-acid-residue-long periodicity to modern collagen polypeptide chains. The ultimate ancestral coding sequence for this class of α-helical polypeptides, nevertheless, appears to have been nonomeric repeats that specified Gly-X-Pro or X-Gly-Pro (14).

As the representative of β-sheet formers, the prolific β_2-microglobulin family of polypeptide chains shall be considered. The cryptic repetitiousness inherent in this family of coding sequences was revealed by the abundant recurrence of base decamers to hexamers within each of them (15–17). The compilation of these recurring base oligomers led us to the view that the coding sequences of this β-sheet-forming family are variously truncated degenerate repeats of the 45- to 48-base-long primordial building block sequence, thus giving the 15- to 16-amino-acid-residues-long periodicity to the immediately ancestral polypeptide of this family. On the other hand, the internal repetitiousness evident within the primordial building sequence suggested that the ultimate ancestor of this and related families specified a polypeptide chain with a much shorter periodicity (17).

Of oligonucleotides of various lengths, those that are multiples of trimers appear to be least qualified to have served as the building block of primordial coding sequences. First, whereas repeats of a base pentamer already specify pentapeptidic periodical polypeptides, hexameric repeats can only give the dipeptidic periodicity to their polypeptide chains. Furthermore, whereas the periodicity of polypeptides specified by all other base oligomers remains impervious to frame shifts due to base deletions and insertions, the peptidic periodicity may be completely disrupted by frame shifts sustained by repeats of a multiple of 3-oligomer. For example, repeats of A-C-T-G-A specify pentapeptidic Thr-Glu-Leu-Asn-Trp repeats in all three reading frames, whereas repeats of A-T-G-C-C-A hexamer in three reading frames are translated to quite dissimilar dipeptidic repeats: Met-Pro, Cys-Gln, and Ala-Asn. Nevertheless, it should be noted here that a group of giant polypeptides secreted by larval salivary glands of a dipteran insect (*Chironomus tentans*) are specified by a family of genes that apparently arose as repeats of base nonomers (18). The same is apparently true of the ultimate ancestor of vertebrate collagen genes (14).

Under the mitochondria-like primitive code, the 770-base-long tetrameric repeat portion of a mouse transcript can specify three tetrapeptidic periodical polypeptides >240 residues long

In a series of studies, Singh and Jones have shown that repetitious elements contained in satellite IV or Bkm DNA of the two East Asiatic colubrid snake species (*Elaphe radiata* and *Bungarus fasciatus*) are also present in all the eukaryotic genomes studied, from bakers yeast to man (19) and that these repeats in higher vertebrates are concentrated in the W chromosomes of the female heterogamety as well as in the Y chromosome of the male heterogamety. Furthermore, in the case of laboratory mice, the concentration of these repeats in the functionally critical testis-determining portion of the Y chromosome was genetically proven (20). We have previously identified the cardinal base tetramer $\frac{\text{GATA}}{\text{CTAT}}$ and its single-base-deviant tetramer $\frac{\text{GACA}}{\text{CTGT}}$ as evolutionary conserved elements in these repeats (9). When G-A-T-A, G-A-C-A strand and its complementary T-A-T-C, T-G-T-C strand were used separately as hybridization probes on poly-A-containing putative mRNAs of the mouse, at least two, and possibly more, transcripts containing T-A-T-C and T-G-T-C repeats were identified (9, 12).

Their possible role in sex determination will be discussed elsewhere (13). We present here, in Fig. 1, the pertinent 1,253-base-long portion of one such mouse transcript (roughly 2,500 bases long) merely as an example of the oligomeric repeats that can yield three long periodical polypeptides in all three reading frames even under the primitive mitochondria-like code with four chain-terminating base triplets (5). It should be noted that the 777-base-long stretch starting from row 9 of Fig. 1 *Left* and ending in the second from the last row of Fig. 1 *Right* is totally free of chain-terminating base triplets, except for four at the very beginning (rows 9 and 10) and six at the very end (second from the last row) whereas the same segment is loaded with chain-initiating base triplets. This is because the cardinal tetramer T-A-T-C contains the chain-initiating ATC within it. Accordingly, in its three reading frames, this stretch is capable of specifying polypeptide chains that are 245, 246, and 251 amino acid residues long. These three amino acid sequences are similar in their tetrapeptidic Leu-Ser-Ile-Tyr periodicity specified by tandem repeats of the cardinal T-A-T-C tetramer. Nevertheless, these three are neither overly homologous with each other nor excessively monotonous. Sequence homology between the three does not exceed 60%, and the monotony in these sequences is broken by propagation of the three single-base-deviant tetramers of the cardinal T-A-T-C—i.e., 33 copies of T-G-T-C are scattered along the entire stretch, whereas 5 copies of T-C-T-C are concentrated in row 11 of Fig. 1 *Left*, and 7 copies of T-A-T-G are in rows 6 and 7 of Fig. 1 *Right*. The monotony is further broken by random base substitutions, deletions, and insertions. Accordingly, all 20 amino acid residues are represented in these three amino acid sequences.

The evolutionary fate of repetitious base sequences was clearly defined by Southern (21). Each original family of repeats starts as numerous exact copies of the characteristically short base oligomer. Soon, randomly sustained base substitutions, dele-

tions, and insertions diversify the sequence to yield a number of subfamilies of repeats, each subfamily now consisting of less-exact copies of a much longer unit sequence. With further accumulation of mutational degeneracy, the repetitiousness of these subfamilies becomes progressively cryptic. Indeed, in rows 1 to 5 of Fig. 1 *Right* we can already witness the birth of a subfamily with a 16-base-long unit sequence. Three exact copies of A-T-C-T-A-C-C-T-A-T-C-A-T-A-T-G are seen in rows 4 and 5. Also marked in rows 1–3 of Fig. 1 *Right* are two single-base-deviant copies and a double-base-deviant copy of the above-noted nucleotide sequence whereas its two-base-deviants occupy row 6.

All in all, the 770-base-long segment of Fig. 1 appears to serve as the model of the primordial coding sequence of eons ago. By quickly acquiring longer periodicities, derivatives of such a primordial coding sequence must have begun to specify α-helical- or β-sheet-forming polypeptide chains.

This work was supported in part by National Institutes of Health Grant AI 15620 and research grants from the Bixby Foundation and Wakunaga Pharmaceutical Company of America.

1. Orgel, L. E. (1968) *J. Mol. Biol.* **38,** 381–393.
2. Hoyle, F. (1979) *Ten Faces of the Universe* (Freeman, London).
3. Crick, F. H. C. (1982) *Life Itself: Origin and Nature* (McDonald/Simon & Schuster, New York).
4. Bridson, P. K. & Orgel, L. E. (1980) *J. Mol. Biol.* **144,** 567–577.
5. Bibb, M. J., Van Etten, R. A., Wright, C. T., Walberg, M. W. & Clayton, D. A. (1981) *Cell* **26,** 167–180.
6. Crick, F. H. C. (1968) *J. Mol. Biol.* **38,** 367–379.
7. Jukes, T. H. (1983) *Nature (London)* **301,** 19–20.
8. Reyes, A. A., Johnson, M. J., Schöld, M., Ito, H., Ike, Y., Morin, C., Itakura, K. & Wallace, R. B. (1981) *Immunogenetics* **14,** 383–392.
9. Epplen, J. T., McCarrey, J. R., Sutou, S. & Ohno, S. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 3793–3802.
10. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560.
11. Sanger, F., Coulson, A. R., Barell, B. G., Smith, A. J. H. & Roe, B. (1980) *J. Mol. Biol.* **143,** 161–178.
12. Epplen, J. T., Cellini, A., Shorte, M. M. & Ohno, S. (1983) *Differentiation* **24,** in press.
13. Epplen, J. T., Cellini, A., Romero, S. & Ohno, S. (1983) *J. Exptl. Zool.*, in press.
14. Yamada, Y., Avedimento, V. E., Murdry, J. M., Ohkubo, H., Vogeli, G., Irani, M., Pastan, I. & de Crombrugge, B. (1980) *Cell* **22,** 287–292.
15. Ohno, S., Matsunaga, T. & Wallace, R. B. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 1999–2002.
16. Ohno, S., Matsunaga, T., Epplen, J. T., Itakura, K. & Wallace, R. B. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 6342–6346.
17. Yazaki, A. & Ohno, S. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 2337–2340.
18. Sümegi, J., Wieslander, L. & Daneholt, B. (1982) *Cell* **30,** 579–587.
19. Singh, L., Purdom, I. F. & Jones, K. W. (1981) *Cold Spring Harbor Symp. Quant. Biol.* **45,** 805–813.
20. Singh, L. & Jones, K. W. (1982) *Cell* **28,** 205–216.
21. Southern, E. I. (1972) in *Modern Aspects of Cytogenetics: Constitutive Heterochromatin in Man*, ed. Pfeiffer, R. A. (Schattauer, Stuttgart, Federal Republic of Germany), pp. 19–28.