Proc. Natl. Acad. Sci. USA
Vol. 83, pp. 2133–2137, April 1986
Evolution

Origin of eukaryotic introns: A hypothesis, based on codon distribution statistics in genes, and its implications

(DNA sequence randomness/stop codons/computer simulation/statistical analysis/intron evolution)

PERIANNAN SENAPATHY

Division of Computer Research and Technology, National Institutes of Health, Bethesda, MD 20892

Communicated by Marshall Nirenberg, November 12, 1985

ABSTRACT **A hypothesis for the origin of introns in eukaryotic genes is developed. By computer simulation it was found that the reading-frame lengths in a random nucleotide sequence are distributed in a negative exponential manner and that there exists an upper limit of about 200 codons in the length of the reading frames (RFs). These characteristics suggest that, if primordial DNA contained a random nucleotide sequence, the most primitive cells would have been under selective pressure to eliminate interfering stop codons in order to increase the length of RFs. Further, they indicate that the only possible way that a coding sequence that is considerably longer than 600 nucleotides could be derived from the short coding sequences occurring in a random sequence would be to splice the short coding sequences and to eliminate the stretches of sequences containing clusters of in-frame stop codons. Thus, introns are suggested to be those stretches of sequences containing interfering stop codons that were originally earmarked in the first primitive cells to be eliminated in order to enable the coding for long polypeptides. Because the statistical characteristics of codon distributions in today's eukaryotic DNA sequences resemble closely those of a random sequence and because the upper limit in the length of RFs (200 codons) in a random sequence corresponds precisely to the observed maximum length of exons in today's eukaryotic genes (600 nucleotides), it is suggested that introns originated in the most primitive unicellular eukaryotes when they evolved from primordial sequences. The data from the prokaryotic gene sequences indicate that prokaryotic genes may have been derived originally from primitive unicellular eukaryotic genes by losing introns from them.**

Eukaryotic protein-coding sequences (exons) are interrupted by intervening sequences (introns), whereas the coding sequences of prokaryotic DNA are continuous (1). The length of introns varies between 10 and 10,000 nucleotides (nt), but the length of exons has an upper limit of about 600 nt in most of the eukaryotic genes (1, 2). Because exons code for protein sequences, they are very important for the cell, yet constitute only 10% of the cell's gene sequences. Introns, in contrast, constitute 90% of the cell's gene sequences but seem to have no crucial functions in genes, except for functions such as containing enhancer sequences and developmental regulators in rare instances (3, 4).

The origin and function of eukaryotic introns have been an enigma since their discovery and there have been contrasting discussions—whether the introns were introduced when eukaryotic genes evolved from more ancient prokaryotic intronless genes or whether the primitive single-celled eukaryotes were the most ancient to evolve along with introns (5–9). This question is still unsettled due to lack of corroborative evidence.

This paper presents a hypothesis for the origin of introns based on statistical analyses of codon distributions in DNA sequences from databanks. The analyses reveal novel features of DNA sequences and the hypothesis leads to important implications. It is shown here that (*i*) the lengths of sequences between the successive repetitions of given codons are distributed in a negative exponential manner in a random sequence and (*ii*) there exists an upper limit of ≈200 codons in the reading-frame lengths (RFLs) in a finite length of a random sequence. By these general criteria of the distribution of codons, the DNA sequences from databanks are shown to behave similarly to a random sequence. These properties were analyzed in eukaryotic and prokaryotic groups of DNA sequences from databanks and were compared with those of random sequences. The upper limit in the RFL (200 codons) in a random sequence exists precisely in eukaryotic DNA sequences, and it is exactly the same as the existing upper limit in the length of exons (600 nt) in today's eukaryotes. Based on these data, it is proposed that the DNA sequences of the most primitive unicellular eukaryotes may have originated from primordial DNA sequences. In the first primitive cells, the main selective pressure in evolving a coding gene must have been to generate long coding sequences from short coding sequences (which existed with an upper limit of 600 nt in the primordial sequences). This must have been achieved by splicing the short coding sequences and excising the sequences containing clusters of interfering stop codons (S codons). For evolutionary reasons, these sequences (introns) were conserved at the DNA level and were removed only at the mRNA level. The specific upper limit in RFLs is greatly exceeded in prokaryotic DNA sequences, in contrast to the upper limit that exists in sequence lengths between their nonstop codons (NS codons). These data suggest that prokaryotic genes could have evolved from the split genes of primitive single-celled eukaryotes by means of the "gene processing" mechanism.

RATIONALE

The origin and evolution of introns could be very closely related to that of exons. Since exons are nothing but protein-coding sequences and since coding sequences are common to all organisms, the evolution of exons could be explained by analyzing the earliest evolution of coding sequences in general. A fundamental question here is how protein-coding sequences could have evolved from primordial DNA sequences at the initial development of the very first primitive cells. To answer this, two basic assumptions were made: (*i*) before a self-replicating cell could come into existence, DNA molecules were synthesized in the primordial soup by random addition of the 4 nt without the help of templates and (*ii*) the nucleotide sequences that code for proteins were selected from the preexisting DNA sequences in the primordial soup by natural selection, not by construction from shorter coding sequences. If primordial DNA did contain random nucleotide

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.

Abbreviations: nt, nucleotide(s); S codon, stop codon; NS codon, nonstop codon; RF, reading frame; RFL, RF length.

sequences, the next question is: Was there an upper limit in the coding-sequence lengths, and, if so, did this limit play a crucial role in the formation of the structural features of genes at the very beginning of evolution?

First, an analysis with computer-simulated random sequences revealed that there actually existed an upper limit of about 200 codons in RFLs. The shortest RFL (zero) was the most frequent. At increasing lengths, the frequency of reading frames (RFs) decreased logarithmically, reaching zero at lengths greater than 600. This suggested that if a typical coding gene had to evolve from a random sequence, the only possible way was with a split structure of exons and introns.

Next, a striking difference in the properties of the RFLs was found between the eukaryotic and prokaryotic groups of DNA sequences. The present-day eukaryotic DNA sequences exhibited both of the characteristics of a random sequence—namely, (*i*) a negative exponential distribution of RFLs and (*ii*) an upper limit of 600 nt in RFLs. Based on this, it is suggested that the primitive unicellular eukaryotic cells evolved from the primordial DNA sequences, thus originating introns and exons in their genes. The data from the prokaryotic gene sequences (showing RFLs significantly exceeding the upper limit, with a negative exponential distribution of NS codon distances matching those in a random sequence) indicated that they could not have evolved directly from primordial DNA sequences but could have been derived from the primitive unicellular eukaryotic genes by losing introns. Thus, the first primitive living cells to come on earth seem to have been the primitive eukaryotic single-celled organisms, and, from them, the prokaryotic cells evolved by a retrograde evolution.

METHODS

All DNA sequences were taken from the nucleic acids sequence databank at the National Biomedical Research Foundation (10), currently listing 2.1 million bases in 1320 sequences. Because the present analyses concerned the evolution of protein-coding genes, most DNA sequences used consisted of protein-coding sequences. Further, since the analyses involved primarily the comparison of eukaryotic and prokaryotic groups of DNA sequences, the DNA sequences from bacteriophages, eukaryotic viruses, yeasts, and processed genes were not included in this study.

A typical random sequence was generated by independent selections of a nucleotide from 1 of the 4 nt, with each nucleotide occurring with a probability of 1/4. The distribution of distances between the successive repetitions of a given codon (or any subsequence) in the simulated random sequence follows a negative exponential distribution. In a random sequence, either a S codon occurs with a probability of 3/64 or any of the other 61 NS codons occurs (with a probability of 61/64). This paper deals mainly with finding the number of codons bounded between the successive repetitions of S codons in DNA sequences from databanks and comparing these lengths with those of a random sequence simulated as described above.

The upper limit in the RFLs in a finite length of a random sequence was computed using the equation $L = [21.3/(0.95)^{(F-1)}] - 20.33$, where F = the upper limit in the RFL and L = the finite length of the random sequence. In a random sequence of length 10^{10} nt, the length of a typical eukaryotic genome, the upper limit in RFL was found to be about 600 nt. It was observed, using the above equation, that each order of magnitude increase in the length of the finite sequence results in an increment in the upper limit of only about 40 nt. This was also verified by simulating random sequences using the computer and analyzing the RFLs.

RESULTS AND DISCUSSION

Upper Limit in the Length of Coding Sequences in DNA Sequences from Databanks. If we take a particular codon, say AAA, in a random sequence, the probability that the next codon is also AAA is 1/64, and the probability that it is a non-AAA is 63/64. Thus, the probability that the codon AAA is repeated after *n* non-AAA codons (i.e., the probability of $AAAX_nAAA$, where X = any of 63 non-AAA codons) is $(63/64)^n$ (1/64). As *n* increases, the probability decreases exponentially. Thus, the lengths of sequences occurring between successive repetitions of the codon AAA has a negative exponential distribution. Further, in a random sequence of a finite length found in biological systems, statistically there should exist an upper limit in the distances between successive repetitions of a given codon. This limit also applies to the distances that separate S codons (RFLs).

The parameters of these properties were analyzed in the databank DNA sequences and were then compared with those in the simulated random sequence. First, the databank DNA sequences were analyzed to see if there existed any definite distribution characteristics of RFLs in the DNA sequences of today's living organisms. Taking sequences that were longer than 2000 nt (totaling 600,000 nt) from the National Biomedical Research Foundation databank, the distances between the successive repetitions of S codons were analyzed for their distribution characteristics. The control in this analysis was to measure the distances between 3 NS codons randomly chosen from the 61 NS codons. To compare the patterns with those in a random sequence, similar measurements were done in a random sequence of length 600,000 nt generated using the computer.

All of the RFs in the random sequence (plotted serially) were shorter than 600 nt (Fig. 1*c*). The pattern for the three NS codons in the databank DNA sequences (Fig. 1*b*) was roughly similar to that for the random sequence. The pattern for the S codons in the databank sequences (Fig. 1*a*) was strikingly different from that for the NS codons, because a number of RFs significantly longer than 600 nt occurred.

The frequencies of these lengths reveal a negative exponential distribution (Fig. 2). The frequencies of the lengths between the successive repetitions of NS codons in the databank sequences (Fig. 2*b*) resembled the frequencies of the lengths in the simulated random sequence (Fig. 2*c*). In both DNA groups, the lengths clearly had an upper limit of about 600 nt. A χ^2 test of the frequency data indicated that the NS codons were close to being randomly distributed in the DNA sequences (data not shown).

The frequencies of the RFLs in the databank DNA sequences (Fig. 2*a*) showed that abnormally long RFs existed in the DNA sequences. It was not surprising, because, in bacteria at least, coding sequences as long as 6000 nt occur

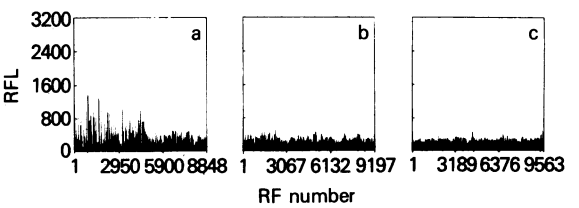

FIG. 1. Upper limit in RFLs in databank DNA sequences. Starting with the first nucleotide in an individual DNA sequence, the first and the subsequent occurrences of a S codon were searched in steps of 3 nt until the DNA sequence was completed. The distances in number of nucleotides between each pair of consecutive occurrences starting with the first were serially computed. Then the search was again carried out from the second and third nucleotides of this DNA successively. This process was repeated for all of the DNA sequences. The consecutive lengths were plotted serially. The patterns shown are lengths of sequences between the successive repetitions of S codons (TGA, TAA, TAG) in databank DNA sequences (totaling 600,000 nt) (*a*) and a simulated random sequence of length 600,000 nt (*c*) and between NS codons (ACT, TTA, GGC) in databank DNA sequences (*b*).

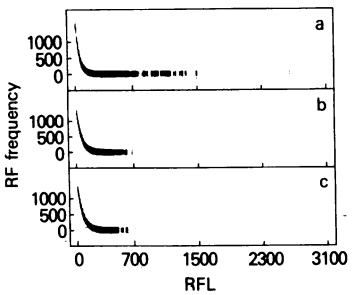

FIG. 2. Frequency distributions of RFLs in databank DNA sequences. The RFLs from Fig. 1 were grouped as frequencies according to increasing lengths. Only the lengths that have a non-zero frequency were plotted. The patterns shown are for the frequencies of the lengths of sequences between the successive repetitions of S codons (*a*) and NS codons (*b*) in databank DNA sequences and between S codons in a computer-simulated random sequence (600,000 nt in length) (*c*).

in a continuous fashion and code for proteins of length 2000 amino acids (10). However, it was interesting to note that only the distances between S codons were distinctly so long, against the background of the distances between all other codons (which distances were similar to those in the random sequence having an upper limit of 600 nt).

These data indicated two possibilities concerning the initial evolution of DNA sequences. (*i*) Since all of the 61 NS codons were distributed in today's DNA sequences in a manner similar to that seen in the random sequence, it is possible that the primordial DNA sequences from which the DNA of primitive living cells probably originated contained random nucleotide sequences. (*ii*) Since the lengths only between S codons were up to 6000 nt in today's DNA sequences, alterations to specifically increase the lengths between the S codons of the primordial DNA probably occurred at some later stage in evolution.

Differences in Codon Distribution Between Eukaryotic and Prokaryotic DNA. Because relating these data to the difference in the split and unsplit structures of the genes of eukaryotes and prokaryotes suggested that probably there existed basic differences in the mechanisms of their coding-sequence evolution and that these differences may be reflected in the present-day eukaryotic and prokaryotic DNA RFLs, the prokaryotic and eukaryotic DNA sequences were grouped separately, and the same analyses of RFLs were performed with these new groups.

The resulting data revealed a striking difference in the RFLs in the prokaryotic and eukaryotic DNA sequences (Fig. 3). A serial plot of the sequence lengths between a sampling of three NS codons in eukaryotic DNA sequences (totaling 250,000 nt in length, Fig. 3*b*) and in prokaryotic DNA sequences (totaling 170,000 nt, Fig. 3*d*) indicated that the patterns in both sequences were similar to those in the simulated random sequence (250,000 nt, Fig. 3*e*). The RFLs in the eukaryotic DNA (Fig. 3*a*) seemed to behave similarly to those in the random sequence (Fig. 3*e*) and showed an upper limit of 600 nt. On the other hand, very long RFs occurred in prokaryotic DNA sequences (Fig. 3*c*). The RFs longer than 600 nt were now found to be associated only with the prokaryotic DNA sequences.

As would be expected from the results in Fig. 3, the frequency of the above lengths (Fig. 4) indicated again a negative exponential distribution. The patterns of sequence lengths for NS codons (Fig. 4 *Left, b*) and for S codons in eukaryotic DNA sequences (Fig. 4 *Left, a*) were similar to those in the random sequence (Fig. 4 *Left, c*). The frequency patterns of sequence lengths between NS codons in prokaryotic DNA (Fig. 4 *Right, b*) were similar to those in the random sequence (Fig. 4 *Right, c*). However, a number of RFs in

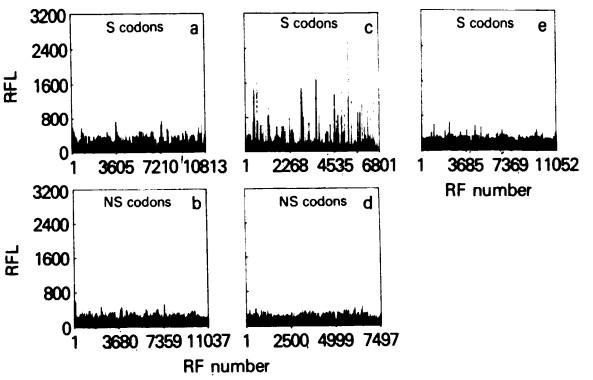

FIG. 3. Differences in RFLs in eukaryotic and prokaryotic DNA sequences. Eukaryotic DNA sequences (250,000 nt) and prokaryotic DNA sequences (170,000 nt) were grouped separately and the distances between the successive occurrences of the S codons and NS codons in them were measured as described for Fig. 1. The patterns shown are lengths of sequences between the successive repetitions of S codons (*a* and *c*) and between NS codons (*b* and *d*) in eukaryotic (*a* and *b*) and prokaryotic (*c* and *d*) DNA sequences and for S codons in a computer-simulated random sequence (250,000 nt) (*e*).

prokaryotic DNA sequences were significantly longer than the 600-nt upper limit (Fig. 4 *Right, a*). Again, one observes that only the distances between S codons in today's prokaryotic DNA were longer than 3000 nt in contrast to all other codons, which displayed a random-sequence behavior clearly showing an upper limit of about 600 nt.

Evolution of Splicing Machinery to Eliminate S Codons. The above results indicate that random distribution of S codons places an upper limit of 600 nt (200 amino acids) in the RFLs. Further, all of the NS codons are distributed in a manner similar to that in a random sequence in both eukaryotic and prokaryotic DNA sequences, with an upper limit of 600 nt. However, both classes of organisms, the eukaryotes and the prokaryotes, contain very long polypeptides with chain lengths up to 2000 amino acids (corresponding to a RFL of 6000 nt), with similar statistical distributions (10). These results strongly suggest that in the earliest evolution of polypeptide-coding genes from primordial DNA sequences, there could have been selective pressure for the primitive cells to specifically increase the RFLs.

The present study suggests that the best possible way to arrive at long protein-coding sequences from random sequences having only short RFs would be to splice the available short bits of coding sequences and to eliminate the regions of sequences containing interfering S codons. It is also reasonable to suggest that the most primitive cells

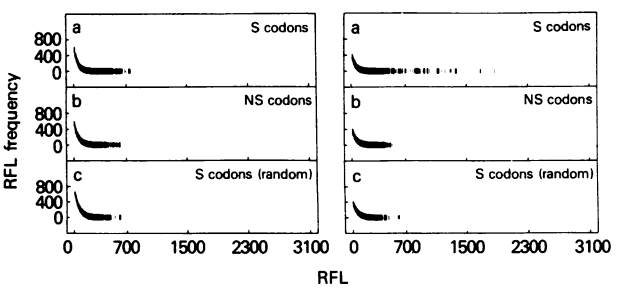

FIG. 4. Differences in the frequency distributions of RFLs in eukaryotic and prokaryotic DNA sequences. The RFLs from Fig. 3 were grouped as frequencies according to increasing lengths. The patterns shown are for eukaryotic (*Left*) and prokaryotic (*Right*) DNA sequences. Each pattern shows the frequencies of the lengths of sequences between the successive repetitions of S codons (*a*) and NS codons (*b*) and between S codons in a computer-simulated random sequence (250,000 nt for eukaryotes and 170,000 nt for prokaryotes) (*c*).

arrived at their long coding sequences from the primordial sequences using such a mechanism—assuming, as stated earlier, that the primordial sequences had random nucleotide sequences and that the primitive cells originally evolving from them required proteins significantly longer than 200 amino acids for their evolution and survival. Thus, introns are suggested to be those regions of sequences that were originally earmarked to be edited out to avoid S codons.

This study's data indicated that all of the 64 possible codons in today's eukaryotic DNA sequences appeared to be distributed in a manner similar to that in the random sequence. Further, the upper limit in the length of present-day exons (600 nt) coincides with the upper limit in the RFLs in the random sequence. In addition, only eukaryotes, not prokaryotes, have a split structure of genes. These data lead to the logical interpretation that the most ancient primitive cells, those that selected and evolved their coding sequences from primordial DNA with the split exon–intron structure of genes, must have been the primitive unicellular eukaryotic organisms. Thus, removing introns through splicing seems to be a mechanism developed by the primitive cell to deal with the cluster-distribution-of-S-codons characteristic in primordial sequences revealed in the present study. Introns probably did not have any functions at the beginning. The few functions observable (rarely) in today's introns (3, 4) may have been acquired later in evolution. Why introns were not totally eliminated from living systems is a complicated issue.

The present hypothesis suggests that originally in evolution the only function of splicing was to eliminate the stretches of sequences containing interfering S codons. If the basic properties of codon distributions of the primordial DNA were still conserved in today's higher eukaryotes, then splicing the exon sequences together in a gene of a present-day eukaryote should increase only the length of RFs without affecting the lengths between any NS codons. This was found to be true when the exon sequences in a gene were spliced together and the distributions of S codons and NS codons were analyzed in the spliced and the unspliced sequences for all of the eukaryotic-coding gene sequences (other than those of yeasts and viruses) existing in the databank. When the exon sequences of each of the three representative genes (Fig. 5 *a*, *c*, and *e*) were spliced together, a RF (containing the coding sequence) significantly longer than 600 nt was generated (Fig. 5 *b*, *d*, and *f*). Further, the lengths between NS codons in the unspliced sequences (Fig. 5 *g*, *i*, and *k*) were not altered in the spliced sequences (Fig. 5 *h*, *j*, and *l*). This clearly indicated that what had been altered because of splicing were only the lengths of sequences between S codons. Thus, splicing seems to edit out only the stretches of sequences containing interfering S codons and to increase the RFLs. This phenomenon was observed in each of the 50 eukaryotic genes.

This result was also observed when the DNA sequences of 10 eukaryotic coding genes were pooled, the exons of each gene were spliced using the computer, and the RFLs were measured in the unspliced and spliced sequences. In the unspliced sequences (Fig. 6*a*) there were no RFs longer than 600 nt, whereas there were 10 RFs significantly longer in the spliced sequences (Fig. 6*b*). The pattern of distances between NS codons after splicing (Fig. 6*d*) was not different from that before splicing (Fig. 6*c*), showing that the distances between the NS codons were not at all affected by splicing.

The coincidence of the same upper limit existing in the length of exons (600 nt) as in the length of RFs in the random sequence offers a strong support to the present hypothesis. Further, this hypothesis suggests that, in the evolution of coding genes, a S codon search process—analogous to the reading of mRNA by ribosomes—aided in the identification of interfering S codons. The occurrence of all of the three S codons (TAG, TGA, TAA) at highest frequencies in splice signal sequences immediately inboard of both donor and acceptor splice junctions on the intron sides (1) indicates this may be true and the splice signals may have evolved, by natural selection, to include S codons among their integral parts.

Possible Evolution of Prokaryotes from Primitive Unicellular Eukaryotes. The data described above indicated that the genes of primitive unicellular eukaryotes evolved from primordial DNA sequences and not from prokaryotic genes. They also indicated that it was not possible by point mutations to arrive at RFs significantly longer than 600 nt and possibly as long as 6000 nt from random sequences due to the cluster-distribution-of-S-codons characteristic as shown in Figs. 2, 4, and 5. If a S codon in a random sequence was mutated to a NS codon, another S codon very close to it on either side would be encountered. It was found by using the above equation that even in a random sequence of length 10^{15} nt, the upper limit in the RFL was only about 800 nt. Thus, the very long RFs of the prokaryotic coding

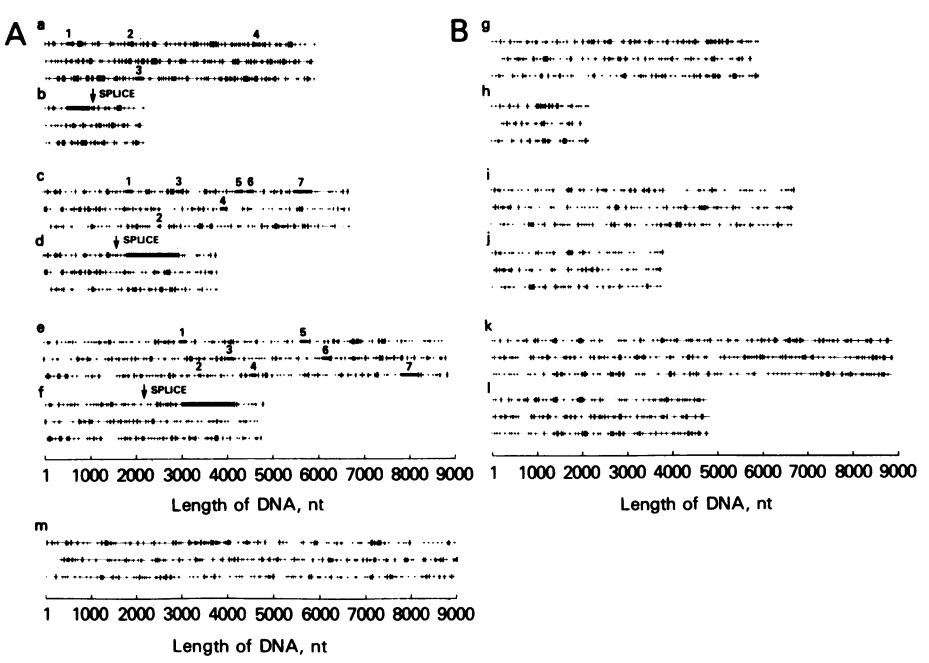

FIG. 5. S codon and exon distribution in three eukaryotic unspliced and spliced genes shown in a linear fashion. The three lines in each sequence represent the distribution of S codons (tick marks) in the three consecutive reading phases. The exons are indicated by numbered thickening of the horizontal lines. With the aid of the computer the genes were spliced and the S codon distribution was plotted (three short lines below the arrow mark). The thick line portion of the spliced sequences represents the reconstructed polypeptide-coding sequence of the corresponding gene. The patterns show the distribution of S codons (*A*) and NS codons (*B*) in three individual eukaryotic unspliced (*a*, *c*, *e*, *g*, *i*, and *k*) and spliced genes (*b*, *d*, *f*, *h*, *j*, and *l*). For comparison with a random sequence, the distribution of S codons in the three different RFs of a computer-generated random sequence was plotted (*m*). The three eukaryotic genes shown here are: human interferon γ precursor gene (*a*, *b*, *g*, and *h*); ovalbumin-related chicken gene Y (*c*, *d*, *i*, and *j*); chicken ovalbumin gene (*e*, *f*, *k*, and *l*).

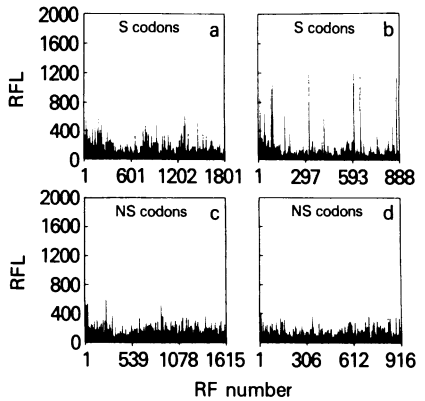

FIG. 6. Specific increase in the lengths of sequences between S codons in eukaryotic genes with splicing. Ten eukaryotic genes were spliced and all the RFLs were measured in spliced and unspliced sequences. The patterns shown are for sequence lengths between S codons in unspliced genes (*a*) and spliced genes (*b*) and between NS codons in unspliced genes (*c*) and spliced genes (*d*). The 10 longest eukaryotic genes described in ref. 10 were analyzed here.

genes could have neither occurred by chance in a random sequence nor been derived by point mutations. The question here then was: When the distances between all other codons behaved similarly to those in a random sequence and were shorter than 600 nt, at what stage in evolution and by what mechanism were only the distances between S codons increased up to 10 times longer? This study shows that, once splicing had evolved, the generation of very long RFs could be achieved by simply splicing the coding sequences, which splicing would lead to precisely the codon distribution characteristics found in today's prokaryotic genes. Thus, it is possible that the prokaryotic genomes could have evolved their very long RFs from the spliced genes of primitive unicellular eukaryotic genomes.

Comparison of the data from eukaryotic genes in Fig. 5 with those in prokaryotic genes (Fig. 7) showed that the codon distribution statistics in prokaryotic DNA were strikingly similar to those in spliced eukaryotic genes, supporting the above suggestion. All eukaryotic genes (50) and prokaryotic genes (30) existing in the databank were analyzed and gave similar results (data not shown). The same phenomenon was observed by comparing the data in Figs. 3 and 6.

Gene processing, a mechanism by which spliced eukaryotic mRNA containing the regulatory sequences can be reverse-transcribed and integrated into the chromsome, has been suggested for the evolution of some eukaryotic genes and has been shown to be operative in today's retroviral genes (11–13). If such a reverse-transcription existed in the primitive unicellular eukaryotes, then the selection of appropriate processed genes from the eukaryotic genome and their modifications to suit a prokaryotic environment could have led to the evolution of a prokaryotic cell. Thus, the prokaryotes could have lost the introns when primitive single-celled eukaryotes became prokaryotes.

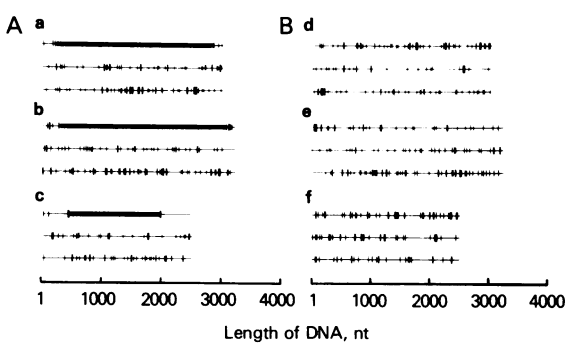

FIG. 7. Occurrence of S codons and NS codons in three prokaryotic genes. (*A*) The S codons in each gene are indicated by tick marks on three horizontal lines representing the three RFs. The coding sequences are shaded thick. The distributions of NS codons in the corresponding RFs are shown in *B*. The occurrence of S codons (*a–c*) and NS codons (*d–f*) is shown for the genes alanyl-tRNA synthetase gene (*a* and *d*), DNA polymerase I gene (*b* and *e*), and amidophosphoribosyl transferase gene (*c* and *f*) of *Escherichia coli*.

Elimination of Nuclear Membranes in Prokaryotes in Evolution. In conclusion, this study further suggests a reason for the presence of a nucleus in the eukaryotic cell and its absence in the prokaryotic cell. If the protein-synthesizing machinery was not separated from the unspliced mRNA in a primitive unicellular eukaryote, then there would have been abortive and enormously wasteful translation of the unspliced RNA, due to the presence of the clusters of interfering S codons in the introns. Thus, it would have been evolutionarily advantageous to have separated the unspliced RNA from the protein-synthesizing machinery allowing only the spliced RNA to be translated. The nucleus probably evolved for this purpose in the first primitive unicellular eukaryotes. The fact that splicing always occurs within the nucleus before the RNA is transported to the cytoplasm in today's eukaryotic cell (1) corroborates this suggestion. Since in the prokaryotic cells, the genes were already in the spliced form, there was no need for separation of the RNA and the protein-synthesizing machinery by a nuclear membrane as in a eukaryotic cell. Thus, it seems that, during their retrograde evolution from the primitive single-celled eukaryotes when introns were lost, the prokaryotes lost the nuclear boundary and became nucleus-less cells.

The statistics of S codon distribution in yeast genomes and eukaryotic viral genomes, which also retain introns to some extent, are closer to those of prokaryotic genomes than to those of eukaryotic genomes (unpublished). It is possible that they also evolved by a partial gene-processing mechanism from the most ancient eukaryotic cells. The backdating of the fossils of primitive eukaryotic cells from 700 to 1500 million years (14) and the evidence that aerobic respiration preceded oxygen-releasing photosynthesis (15) show that geological evidence does not exclude the possibility of the precedence of primitive eukaryotic cells over prokaryotes in evolution.

I am grateful to Ms. M. Hodges and Dr. J. Buck for help in improving the manuscript, Dr. G. Weiss for deriving the probability equation, Dr. J. Mosimann and P. Munson for help with statistics, and Drs. D. Mischaud, G. Knott, and M. Shapiro for initial help in computer programming. This work was carried out when the author was in the visiting program at the National Institute of Arthritis, Diabetes, and Digestive and Kidney Diseases during 1980–1984.

1. Breathnach, R. & Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349–384.
2. Naora, H. & Deacon, N. J. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 6196–6200.
3. Gilles, S. D., Morrison, D., Or, V. T. & Tonegawa, S. (1983) *Cell* **33**, 717–728.
4. Mercola, M., Wang, X., Olsen, J. & Calame, K. (1983) *Science* **221**, 663–665.
5. Doolittle, W. F. (1978) *Nature (London)* **272**, 581–582.
6. Darnell, J. E., Jr. (1978) *Science* **202**, 1257–1261.
7. Tiemeier, D. C., Tilghman, S. M., Polsky, F. I., Seidman, G. J., Leder, A., Edgell, M. H. & Leder, P. (1978) *Cell* **14**, 237–245.
8. Crick, F. (1979) *Science* **204**, 264–271.
9. Gilbert, W. (1978) *Nature (London)* **271**, 501.
10. Dayhoff, M. O., Chen, H. R., Hunt, L. T., Barker, W. C., Yeh, L. S., George, D. G. & Orcutt B. C. (1983) *Nucleic Acids and Protein Sequence Database* (Natl. Biomed. Res. Found., Washington, DC).
11. Leder, A., Swan, D., Ruddle, F., D'Eustachio, P. & Leder, P. (1981) *Nature (London)* **293**, 196–200.
12. Goff, S. P., Gilboa, E., Witte, O. N. & Baltimore, D. (1980) *Cell* **22**, 777–785.
13. Bernstein, L. B., Mount, S. M. & Weiner, A. M. (1983) *Cell* **32**, 461–472.
14. Schopf, J. W. & Oehler, D. Z. (1976) *Science* **193**, 47–48.
15. Schwartz, R. M. & Dayhoff, M. O. (1978) *Science* **199**, 395–403.