Inserted Sequences in Bovine Satellite DNA's

Abstract. *The nucleotide sequence of the 1413-base-pair repeat unit of bovine 1.711a satellite DNA (density in cesium chloride, 1.711 grams per cubic centimeter) has been determined. The repeat unit contains two segments consisting of variants of a basic 23-base-pair sequence that is closely related to sequences of bovine 1.706 satellite DNA. A third segment of the repeat unit contains an unrelated 611-base-pair sequence that is not internally repetitive. This segment is flanked by inverted repeats of 8 base pairs and, on one side, by a direct repeat of the terminal sequence. A related segment is present in bovine 1.711b satellite DNA and is inserted into sequences derived from the 1.715 satellite. These nucleotide sequences suggest the timing of some of the stages in the evolution of these complex, closely related satellite DNA's and indicate the mechanisms inherent in their divergence from a common ancestor.*

In the genome of all higher eukaryotes analyzed, repetitive sequences constitute a major fraction of the DNA (*1*). In addition, evidence is accumulating that transposable DNA elements may similarly be widespread (*2*). Whether repetitive and transposable sequences serve any present function is controversial (*3*), but they probably played a role in the arrangement and rearrangement of the DNA during the evolution of eukaryotic genomes.

Of the various classes of repetitive sequences present in eukaryotes, the highly repetitive components (satellite DNA's) have been studied in greatest detail (*1*). They have been found to consist of thousands to millions of copies of a tandemly repeated sequence that differ mainly in single base changes. Mutations tend to be nonrandomly distributed and to be clustered in adjacent copies of the repeated sequence. The repeat units are often composed of shorter repeat units, an indication of a stepwise amplification process during the evolution of these DNA's. In this report, I describe a new type of satellite DNA structure that suggests additional mechanisms for the generation of highly repetitive sequences.

The bovine genome contains eight satellite DNA's, two of which have a density of 1.711 g/cm^3 in CsCl and have been termed 1.711a and 1.711b (*4*). The 1.711a satellite represents 1.7 percent of the DNA (*4*). It has a repeat unit of 1413 base pairs (bp), which have been sequenced. Two segments of the repeat unit exhibit high homology to the 1.706 satellite DNA previously analyzed (*5*). A third segment (termed INS) is completely unrelated to the 1.706 satellite, and it is likely that this segment is the result of an insertion into a 1.706-like satellite DNA (Fig. 1a).

The 1.711b satellite DNA represents 7.1 percent of the bovine genome (*4*). Part of the repeat unit of this satellite DNA can be described as an insertion of at least 1200 bp into a 1400-bp repeat unit derived from the 1.715 satellite (the well-known satellite I of calf thymus DNA). As assessed by restriction mapping (Fig. 1c), hybridization experiments, and sequence analysis, part of the inserted sequence (INS) is homologous to the insert in the 1.711a satellite, and the rest (INS′) is of unknown origin.

For a more detailed comparison of the related sequences present in the different satellite DNA's, I have determined the nucleotide sequence of the repeat unit of the 1.711a satellite DNA. The nucleotide sequence of the 1.711b satellite is still unknown. The 1.711a repeat unit can be divided into three segments (Fig. 1a). The 550-bp segment (termed Sau) consists of 24.3 tandem repeats of 21 to 23 bp that are variants of a prototype sequence identical to the prototype sequence of the Sau segments (*5*) in the 1.706 satellite (Fig. 2a). These repeats contain both invariant and highly variable sections. From the regular alteration of single nucleotides, short sequences, and deletions (underlined in Fig. 2a), two different larger periodicities, locally superimposed on the 23-bp periodicity, are discernible. Such a structure is best explained by unequal crossing-over.

Smith (*6*) has proposed that unequal crossing-over is the basic mechanism by which repeated sequences are generated and maintained. According to computer simulations, a crucial structure that may distinguish unequal crossing-over from other possible mechanisms of generating repetitive DNA sequences is a tandem array of repeat units exhibiting heterogeneous lengths of superimposed larger repeats in different regions of the tandem array. Such a structure has not previously been found in any satellite DNA, but this may be because long enough stretches of adjacent tandem repeats have not been sequenced in these DNA's. The organization of the 24 tandem repeats in the Sau segment of bovine 1.711a satellite DNA closely matches the predicted structure and indicates that unequal crossing-over has indeed been an important mechanism in the evolution of this satellite DNA.

In the 611-bp region following the Sau

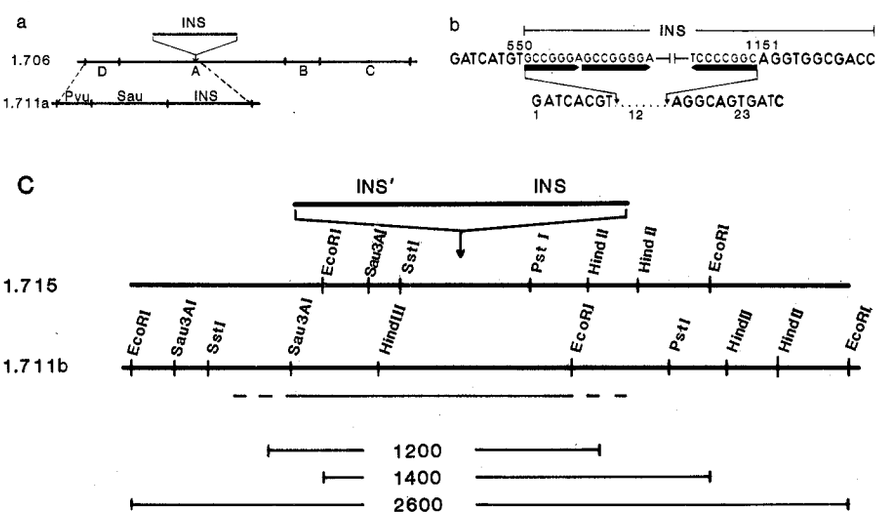

Fig. 1. Insertion model for the generation of the 1.711a and 1.711b satellite DNA's. (a) Comparison of the long-range repeat units of the 1.706 (*5*) and 1.711a satellite DNA's. The possible site of integration is indicated. (b) Terminal and flanking sequences of the INS segment. The terminal direct and inverted repeats are indicated by arrows. The flanking sequences (543 to 550 and 1151 to 1161) are compared to a sequence from the Sau segment (295 to 320) (see Fig. 2a). A deletion of 8 bp in the 23-bp repeat unit is indicated by dots. (c) Comparison of the long-range repeat units of the 1.715 and 1.711b satellites. Sizes are given in base pairs. The location of some restriction sites in the 1400-bp repeat unit of the 1.715 satellite (*5*, *13*) and in a 2600-bp fragment derived from the repeat unit of the 1.711b satellite is indicated. A region of nonhomology is underlined. The sequence denoted INS is homologous to the inserted sequence in the 1.711a satellite, and INS′ is a sequence of unknown origin. The possible region of insertion of the INS′INS sequence is indicated.

0036-8075/81/0724-0443$00.50/0

segment, no regular arrangement of a repeated sequence can be detected (Fig. 2b). There are regions open for reading of more than 150 bp in all reading frames, and ATG (A, adenine; T, thymine; G, guanine) codons are found in the beginning of two of them. The segment carries sequences complementary to the consensus sequence at the 3′ end of 18*S* ribosomal RNA's (*7*). Other remarkable sequences are a perfect Hogness box, TATAAATA, and a stretch of seven C's (C, cytosine) followed by seven A's. The presence of these sequences invites the speculation that the segment could be transcribed.

Several arguments support the notion that the INS region may have been inserted into sequences of the 1.706 or a 1.706-like satellite, as indicated in Fig. 1a. A first argument is based on the structure of the terminal sequences. There is an inverted repeat of 8 bp near the ends of the segment, including seven G · C pairs (Figs. 1b and 2b). On one side, the 8-bp sequence is also directly repeated (with a deletion of 1 bp). This structure is reminiscent of the sequences found at the ends of insertion sequences and other transposable elements in both prokaryotes and eukaryotes (*2*, *8*).

A more direct argument is based on the flanking sequences. The 8 bp preceding the direct repeat and the 11 bp following the inverted repeat are very similar to sequences from the Sau segment (for example, nucleotides 295 to 320) (Fig. 1b). This makes it likely that the segment has been integrated into a 23-bp repeat unit; 8 bp of this sequence are missing and have been replaced by the INS segment. The deleted sequence has

Fig. 2. Nucleotide sequence of the repeat unit of bovine 1.711a satellite DNA. Bovine 1.711a satellite DNA was purified by repeated isopycnic centrifugation (*4*) analogous to that described for bovine 1.706 satellite DNA (*5*). There are about 35,000 copies of the 1.711a repeat unit in the genome. The nucleotide sequence determined from uncloned restriction fragments (*14*) corresponds to the most abundant nucleotide at each position of the repeat unit. Most of the sequence (95 percent) was determined from at least two different overlapping restriction fragments, 55 percent being determined from both strands. (a) The Sau segment. The sequence is arranged for maximum homology between the short repeats. Single nucleotides and short sequences indicating a superimposed higher-order periodicity are underlined. A prototype sequence of the repeat unit, given below the line at the bottom, is compared to the prototype sequence of the Sau segments in bovine 1.706 satellite DNA (*5*). (b) The INS segment. Terminal direct and inverted repeats are indicated by arrows. Some sequences are in boxes: 591 to 600, opposite strand; 991 to 997, complementary to the 3′ end of 18*S* ribosomal RNA (*7*); 664 to 677, C_7A_7; 1086 to 1093, a Hogness box. (c) The Pvu segment. The sequence is arranged for maximum homology between 23-bp repeat units. Alu I and Pvu II sites are underlined. The prototype sequence of the Pvu segment is indicated below the line at the bottom, and the prototype sequence of segment D of 1.706 DNA is given for comparison.

the same length as the inverted repeat and corresponds to the highly variable section of the Sau repeat units.

A third argument is based on the observation that sequences similar to the INS segment occur elsewhere in the bovine genome. This was shown by hybridization of a cloned fragment comprising the INS segment to total calf thymus DNA and various fractions obtained by isopycnic centrifugation. A major region of homology was found in a 1600-bp Eco RI fragment from the 1.711b satellite DNA. Additional bands seen in the autoradiograms were partly a result of larger fragments from the 1.711b satellite that were resistant to Eco RI cleavage because of sequence divergence, and partly a result of minor repetitive components that have not been identified.

The third segment in the 1.711a repeat unit is remarkably similar to Pvu segments B and D, respectively, of the 1.706 satellite DNA (*5*). The 23-bp prototype sequence of this segment differs only by 2 bp from the prototype sequence of Pvu segment D (Fig. 1c) and is closely related to the prototype sequence of the 1.720 satellite (*9*). When the sequences of pairs of individual Pvu segments are compared, the following numbers of base changes are found: between 1.711a and 1.706-B, 52; between 1.711a and 1.706-D, 35; and between 1.706-B and 1.706-D, 20. This suggests that the evolutionary pathways leading to the 1.706 and 1.711a satellite DNA's separated before segment B was generated. Therefore, the 611-bp INS segment should have been incorporated into an ancestor of the 1.706 satellite that probably had only one Pvu and one Sau segment. Such an intermediate structure has been postulated (*5*).

The same sequence of events can be reconstructed from the deletions present in the Pvu segment. There is a deletion of 1 bp in all three segments, an additional deletion of 1 bp in segments B and D, and a further deletion of 4 bp only in segment B. This indicates again that the lines of descent leading to the 1.706 and 1.711a satellite DNA's separated before segment B in the 1.706 satellite was generated by duplication.

From the DNA sequences of cloned preproinsulin and globin genes, Perler *et al.* (*10*) have estimated a minimum value of 7×10^{-9} substitutions per nucleotide site per year for the mutation rate at silent positions and within introns. If this figure holds also for satellite DNA's, then the duplication of the Pvu segment in the 1.706 satellite should have occurred about 5 million years ago, and the separation of the 1.706 and 1.711a satellites, 10 million years ago. With additional sequence data of Pvu-like segments in satellite DNA's from species related to the cow it would be possible to decide whether the rate of mutation in satellite DNA is the same as in introns and silent positions of coding regions.

The structures of the 1.711 satellite DNA's suggest a novel mechanism for the generation of satellite DNA's: the insertion of DNA into repetitive sequences already present in the genome and the subsequent amplification of a new repeat unit containing the inserted sequence and a part of the repetitive sequences. This way of generating satellite DNA's may have been widespread. Related structures have been found for the highly repetitive sequences in the telomeric heterochromation of rye (*11*). The complex satellite DNA's of the human genome (*12*) may be similarly composed. The size variation of the repeating unit found in a satellite DNA from *Drosophila* has been explained by a related mechanism (*1*).

Now five of the eight major satellite DNA's, which together form almost one-fifth of the bovine genome, are known. The nucleotide sequences of the various related satellites have suggested mechanisms for their evolution from common ancestors. Moreover, sequence analyses have shown that complex satellite DNA's may contain sequences believed to be essential for the transcription of eukaryotic genes [see also (*5*)]. This encourages experiments to investigate whether heterochromatin is really genetically as inert as it is presently considered to be.

ROLF E. STREECK*
Institut für Physiologische Chemie, Physikalische Biochemie und Zellbiologie der Universität München, 8000 München 2, Federal Republic of Germany

References and Notes

1. D. L. Brutlag, *Annu. Rev. Genet.* **14**, 121 (1980).
2. J. R. Cameron, E. Y. Loh, R. W. Davis, *Cell* **16**, 739 (1979); P. J. Farabough and G. R. Fink, *Nature (London)* **286**, 352 (1980).
3. W. F. Doolittle and C. Sapienza, *Nature (London)* **284**, 601 (1980); L. E. Orgel and F. H. C. Crick, *ibid.*, p. 604.
4. G. Macaya, J. Cortadas, G. Bernardi, *Eur. J. Biochem.* **84**, 179 (1978).
5. R. E. Streeck and H. G. Zachau, *ibid.* **89**, 267 (1978); M. Pech, R. E. Streeck, H. G. Zachau, *Cell* **18**, 883 (1979).
6. G. P. Smith, *Science* **191**, 528 (1976).
7. O. Hagenbüchle *et al.*, *Cell.* **13**, 551 (1978).
8. P. Starlinger, *Plasmid* **3**, 241 (1980).
9. E. Pöschl and R. E. Streeck, *J. Mol. Biol.* **143**, 147 (1980).
10. F. Perler *et al.*, *Cell* **20**, 555 (1980).
11. J. R. Bedbrook, J. Jones, M. O'Dell, R. D. Thompson, R. B. Flavell, *ibid.* **19**, 545 (1980).
12. A. R. Mitchell, R. S. Beauchamp, C. J. Bostock, *J. Mol. Biol.* **135**, 127 (1979).
13. P. Philippsen, R. E. Streeck, H. G. Zachau, W. Müller, *Eur. J. Biochem.* **45**, 479 (1975); G. Roizès, *Nucleic Acids Res.* **3**, 2677 (1976).
14. A. M. Maxam and W. Gilbert, *Methods Enzymol.* **65**, 499 (1980).
15. I thank A. Pluciennicczak who participated in the sequencing work on part of the repeat unit during a short-term fellowship from the European Molecular Biology Organization, and D. Brutlag and H. Kössel for a computer analysis of the sequence. I also thank K. Beer and V. Heinemann for expert technical assistance and Deutsche Forschungsgemeinschaft for supporting this work.
* Present address: Department of Biology, Stanford University School of Medicine, Stanford, Calif. 94305.

12 December 1980; revised 26 March 1981

Size of the Chloroplast Genome in *Codium fragile*

Abstract. *Chloroplasts isolated from the siphonous green alga* Codium fragile *yield circular DNA molecules averaging 27.3 micrometers in length and* 56×10^6 *daltons in molecular size. This chloroplast genome is 25 to 30 percent smaller than any reported. The small size of the* Codium *chloroplast genome may represent a primitive evolutionary condition in green plants.*

In 1962 it was demonstrated that chloroplasts contain DNA (*1*). Over the past decade chloroplast DNA (ctDNA), extracted from a wide variety of plant groups, has exhibited marked uniformity in its physical and chemical properties, existing as covalently closed circular molecules of 37 to 62 μm in length and molecular size of 80×10^6 to 134×10^6 daltons (*2*). A single exception has been reported in *Acetabularia*, where the estimated genome size is 1500×10^6 daltons, as judged from kinetic complexity measurements (*3*).

We now report that the ctDNA from the green alga *Codium fragile* can be isolated as covalently closed circular molecules with an average contour length of 27.3 μm (measured by electron microscopy) (*4*) and a molecular size of 56×10^6 daltons, as determined by gel electrophoresis of restriction endonuclease fragments. This size is smaller than any yet described for ctDNA.

Until recently it was thought that ctDNA was entirely prokaryotic in character (*5, 6*). With the discovery of introns (intervening sequences) in *Chlamydomonas* ctDNA (*7*), a characteristic of eukaryotic DNA, the picture has become more complicated. Two hypotheses have been advanced to account for the evolu-

0036-8075/81/0724-0445$00.50/0