The Alu Family of Dispersed Repetitive Sequences

Carl W. Schmid and Warren R. Jelinek

As an example of the tremendous complexity of eukaryotic DNA's, human DNA consists of about 2.5 × 10^9 base pairs (bp) (*1*). Even with the recent advances in DNA sequencing technology, it is unlikely that the base sequence of more than a few percent of such a complex DNA will ever be determined, and it is equally unlikely that the overall sequence organization of such a complex DNA will be elucidated exclusively by base sequence analyses. Alternatively, much of our understanding of eukaryotic DNA sequence organization stems from renaturation rate studies of denatured DNA. Such studies revealed that the DNA of most eukaryotic organisms includes multiple copies of certain sequences and single copies of other sequences (*2*).

Summary. A family of related sequences that includes approximately 500,000 members is the most prominent short dispersed repeat family in primate and rodent DNA's. The primate sequence is approximately 300 base pairs in length and is composed of two imperfectly repeated monomer units, whereas the rodent repeat consists of only a single monomer. Properties of this repeat sequence, its flanking sequences in chromosomal DNA, and RNA's transcribed from it suggest that it may be a mobile DNA element inserted at hundreds of thousands of different chromosomal locations.

Approximately 20 to 30 percent of human DNA consists of repetitive sequences (*3, 4*). In almost all eukaryotic DNA's, including those of such diverse species as sea urchins, *Xenopus*, and humans, a portion of the repeated sequences are interspersed with single-copy sequences (*1, 3–6*). These dispersed repeats consist in part of short sequences approximately 100 to 300 bp in length, interspersed with longer single-copy sequences of approximately 1000 to 2000 bp (Fig. 1) (*3, 5, 7*). This sequence arrangement is called the short-period interspersion pattern. Its presence in an organism's DNA does not exclude the presence of other types of sequence organization within the same DNA. For example, in addition to the short-period interspersion pattern, human DNA also contains (i) clustered repeats that are not interspersed with single-copy sequences (*8*), (ii) repeated sequences that are significantly longer than 300 bp (*9*), and (iii) long single-copy sequences that are not interspersed with short repeated sequences (*10*). Most eukaryotic DNA's also contain inverted repeated sequences (*1, 7, 11–13*) that consist of complementary sequences covalently linked in the same DNA strand (Fig. 1). Many of these inverted repeats in primate and rodent DNA's consist of the most prominent family of interspersed repeats that account for a significant proportion of the total short-period interspersion pattern.

A Prominent Family of Dispersed Repeats in Primate and Rodent DNA

Although DNA's of the sea urchin, *Xenopus*, and probably all mammals share the short-period interspersion pattern, there is at least one major difference between the short period repeats in these organisms. In sea urchins they belong to many unrelated sequence families (*14*) with no single repeat family being the most abundant. The term "family" is used to denote those sequences that form stable base-paired duplexes after renaturation of denatured DNA. By definition, members of one sequence family do not hybridize with members of another sequence family, although sequences that do not hybridize at stringent conditions may do so at less stringent conditions (*15*). It then becomes necessary to define subfamilies (*15*), and possibly sub-subfamilies. In contrast to that of sea urchins, primate and rodent DNA's contain a prominent, dispersed sequence family—clearly more abundant than any other—which, for reasons discussed below, has been termed the Alu family of interspersed repeats. The evidence for this conclusion resulted from independent observations (*4, 12, 13, 16–19*). Studies of the renaturation rate kinetics of denatured human DNA, while not conclusive, suggested that the human interspersed repetitious sequences might contain one or only a few prominent sequence families that are significantly more abundant than all others (*19*). To test this possibility further, total human repetitive DNA was isolated and separated by electrophoresis in an agarose gel. A distinct band of DNA approximately 300 bp in length and accounting for approximately 3 percent of the mass of the human genome was observed in the gel superimposed on a background of essentially random lengths of duplex DNA (*4*). Of this crude preparation of 300-bp repetitious DNA 60 percent of the mass could be cleaved at a common site by the restriction endonuclease Alu I, suggesting that it might be composed predominantly of a single "Alu family" of sequences (*4*). Similar results were obtained when inverted repeated DNA sequences, 300 bp in length, were isolated and used as the substrate for Alu I restriction endonuclease digestions.

The identification of a prominent interspersed repetitious sequence family (the Alu family) in the human genome supported previous observations by Jelinek and co-workers, who demonstrated that HeLa cell heterogeneous nuclear RNA (hnRNA) contains transcripts of an abundant repetitious sequence family (*12, 13, 16, 20*). The repetitious RNA could be isolated either as intramolecular or intermolecular RNA:RNA duplexes (dshnRNA) or as RNA that hybridized to DNA under conditions where only repetitious sequences hybridize (*20*). Two dimensional nucleotide analyses of this RNA revealed a simple pattern containing six characteristic ribonuclease T1 oligonucleotides (*14*), suggesting that the

Dr. Schmid is an associate professor in the Department of Chemistry, University of California at Davis, Davis 95616. Dr. Jelinek is an associate professor in the Department of Biochemistry, New York University Medical Center, New York 10016.

0036-8075/82/0604-1065$01.00/0 Copyright © 1982 AAAS

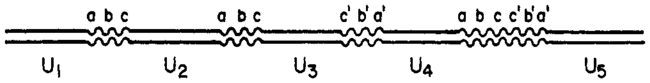

Fig. 1. Representation of the short-period interspersion pattern. The two lines represent double-stranded DNA. The short-period interspersion pattern includes 100 to 300 bp repeated sequences (abc) that are interspersed with variable length single-copy sequences (U) having an average length of approximately 2000 bp. The sequence abc can be present either repeated directly (. . . abc . . . abc . . .) or in inverted orientation, so that both it and its complementary sequence are present in the same DNA strand (. . . abc . . . c'b'a' . . .). In human DNA approximately one-third of the 300-bp repeated sequences form inverted repeats (*7*) and approximately one-third of the 300-bp inverted repeats have no detectable sequences between the two complements (. . . abcc'b'a' . . .) (*7, 11*).

dshnRNA fraction might contain transcripts from a predominant repetitious sequence family. Purified inverted repeated human DNA was vastly enriched in sequences found in the dshnRNA fraction (*12*). When either rodent or human dshnRNA was used to screen cloned libraries of either rodent DNA—the cloned DNA fragment size was about 5 kilobase pairs (kbp)—or human DNA (the cloned DNA fragment size was about 15 kbp), between 25 and 94 percent, respectively, of all clones were found to contain dshnRNA complementary sequences (*13, 18, 21*). Therefore, as was suggested by the two-dimensional nucleotide analyses, these sequences, subsequently demonstrated to be Alu family transcripts, are both highly repeated and widely distributed throughout rodent and human DNA's.

The Alu family accounts for approximately 3 to 6 percent of the mass of human DNA. The mass yield of this sequence family after S1 nuclease digestion of renatured DNA was 3 percent, indicating approximately 300,000 copies per haploid DNA complement (*4*). This is probably an underestimate. Renaturation rate studies indicated approximately 500,000 or more Alu family members per haploid human genome (*4, 22*). The abundance of this family is sufficient to account for a large proportion of all the interspersed 300-bp repeats in human DNA. A random distribution of at least 500,000 copies of this sequence throughout the entire human genome would give an average spacing of 5000 nucleotides between Alu family members (2.5×10^9 bp per 5×10^5 Alu copies). The distance between all short-period interspersed repeats in the human genome has been estimated at approximately 2200 bp (*3, 7*). Thus, one of approximately every two-and-a-half interspersed repeats could be an Alu family member. However, because of the uncertainty in the methods used to estimate these frequencies, each could be inaccurate although probably not by more than a factor of 2 (*3, 7, 20*). Regardless of these uncertainties, the Alu family accounts for a major proportion of all the 300 nucleotide interspersed repeats in human DNA.

The distance between Alu family members in human DNA is variable. The 56,000 bp of DNA containing the epsilon, A gamma, G gamma, delta, and beta globin genes also contain seven Alu family members (*23*), and thus the average spacing of the Alu sequences in this region of the human genome is approximately 8000 bp. However, some of these Alu family members are as close as 700 to 800 bp apart, but in inverted orientation. A 12-kbp human onc gene (c-sis) has three Alu family members located in two intervening sequences (*24*), while a 19-kbp human genomic DNA fragment containing the insulin gene has only one Alu sequence, and apparently no other dispersed repeats (*25*). To confirm the high frequency and wide distribution of Alu family members in human DNA, 100 randomly selected clones bearing DNA fragments approximately 15,000 to 20,000 bp long were screened by hybridization with purified, labeled Alu DNA sequences. Ninety-four of these clones hybridized with the labeled Alu sequence probe (*21*). In an analogous experiment 75 percent of unselected clones bearing African green monkey DNA fragments were found to hybridize with a purified human Alu sequence (*26*). The Alu family is therefore both abundant and widely dispersed throughout human and monkey DNA's.

Alu Family Members Have a Conserved Sequence in Mammals

Members of a family of repeated DNA sequences are related in sequence, but because of evolutionary divergence they are not necessarily identical (*2*). It is therefore necessary to represent the overall base sequence of such a family as the consensus or most frequent sequence of its members. The consensus sequence of the human Alu family has been estimated by two methods. The Alu family is so abundant and its members are so highly conserved in sequence that it was possible to determine a partial consensus sequence of the entire 300-bp DNA fraction liberated by S1 nuclease from renatured human DNA (*27*). Alternatively, this 300-bp DNA was cloned, and the nucleotide sequences of ten individual clones were determined (*27, 28*). The consensus sequence derived from these ten clones is shown in Fig. 2. Included within the Alu family sequence are the six characteristic oligonucleotides reported for HeLa cell dshnRNA (*16, 27, 29*), thus confirming that Alu family members serve as the template for these oligonucleotides in hnRNA.

In addition to the nucleotide sequences of ten Alu family members used to determine the consensus sequence given in Fig. 2, the sequences of four other human Alu family members have been determined from genomic clones containing (i) an insulin gene (*25*), (ii) the epsilon globin gene (*30*), (iii) the G gamma globin gene (*31*), and (iv) the delta globin gene (*31*). These sequences demonstrated that one end of the Alu repeat is conserved among different Alu family members and can be precisely defined (left end of the sequence in Fig. 2), while the other end, which is rich in deoxyadenylate (dAMP) residues, is less precise and varies among different Alu family members (see below).

Like human DNA, rodent DNA has a short-period interspersion pattern containing a prominent family of interspersed repetitious sequences (*13, 18*). However, unlike the human Alu sequence, which is approximately 300 bp in length, the rodent equivalent sequence is only approximately 130 bp in length (*18, 32*). The human Alu sequence is an imprecise dimer formed of two directly repeated, approximately 130-bp monomer sequences with a 31-bp insertion in the second monomer (*25, 28, 32, 33*). The plus (+) overlining in Fig. 2 indicates a dAMP-rich sequence at the end of the first monomer unit. The consensus sequences derived by Krayev *et al.* (*18*) for the so-called mouse B1 repeat and that determined by Haynes *et al.* (*32*) for the equivalent Chinese hamster repeat are compared to the human Alu consensus sequence in Fig. 2. The nucleotide sequence of part of an African green monkey Alu family member found as an insert in an SV40 viral genome (*34*) is also compared. The mouse sequence is 129 bp long, and the Chinese hamster sequence is 134 bp long. Each of these sequences shows considerable sequence homology with the human Alu sequence.

The total length of the second human monomer agrees more closely than the first with the length of the rodent sequence. The rodent repeat contains a sequence (indicated by the dashed line above the mouse sequence in Fig. 2) of DNA corresponding approximately in length but not in sequence to the insert in the second human monomer (indicated by straight line above in Fig. 2). The second monomer of the human Alu sequence, the monkey sequence, the mouse B1 consensus sequence, and the Chinese hamster Alu-equivalent consensus sequence all agree well up to the position of the insert in the human sequence. The human and monkey sequences beyond the insert agree with the rodent sequences for nine bases that form a perfect inverted repeat with a single C (cytosine) residue at the position of symmetry (indicated by the asterisks above the human sequence). The homology between the rodent and human sequences again breaks down at the position indicated by the dashed line over the mouse sequence, but resumes again beyond this region. Presumably these rodent and primate repeats are descendants of a common ancestral sequence that has been well preserved during recent evolution. The possibility cannot be excluded that 300 bp Alu dimers are present in rodent DNA or that monomeric Alu family members are present in primate DNA's. However, they have not yet been identified.

As is indicated in Fig. 2, Alu family members are flanked on one end by dAMP-rich sequences that are not conserved among different Alu family members, but that do maintain the structure $[N(A)_n]_m$, where N represents any nucleotide, or in some instances more than one nucleotide, and n is usually less than 20 and m is usually less than 10. Figure 3 gives the sequences of the dAMP-rich regions abutting nine Alu and Alu-equivalent family members in three different mammalian species. As is discussed more completely below, these dAMP-rich regions may play a role in the dispersal of Alu family members to new chromosomal locations.

It was previously noted that Alu family members contain a G-rich sequence extensively homologous to a sequence at or near the origin of papovavirus DNA replication (*29*). This sequence is present in the viral DNA at a region believed to be involved in a number of molecular events required for the orderly progress of the viral replication cycle. Possibly, Alu family members provide analogous function in cellular DNA. Recently evidence consistent with the suggestion that Alu family members may play a role in DNA replication has been presented (*35*), although considerably more experimental evidence is required to confirm this idea.

Alu Family Members Flanked by Short Direct Repeats

The ubiquity of interspersed repeats in eukaryotic DNA's, and in particular of the Alu family sequence in mammalian DNA's, raises the issue concerning the mechanisms of dispersal of these repeats throughout eukaryotic genomes. Figure 3 gives the structure surrounding nine Alu or Alu-equivalent sequences. Each is flanked on either side by direct repeats ranging in size from 7 to 20 bp. The sequences of these flanking repeats are

```
Human         GGCTGGGCGTGGTGGCTCACACCTGTAATCCCAGCACTTTGGGAGGCCGAGGTGGGTGGATCACCTGAGGTCAGGAGTTCAAGACCAGCCTGGCCAACAT


                                 ++++++++++++++++
Human         GGTGAAACCCCGTCTCTACTAAAAATACAAAAATTAGCCGGGCGT GGTGGCGCGCCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATCGC
Monkey                   A.............G...........T...T.. .....TAT..A...T..G....T.T..G.......................T.
Mouse                                              ......A. ........AT....T..G....... ...........A........C.G..TT.
CHO                                                ..A...A.T.....CA.A.A...T. G....... ....A......A........C.G..TT.
CHO (alt)                                                        T

                                                                 *********
                                  -------------------------------
Human         TTGAACCCAGGAGGTGGAGGTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGG                                     GCA
Monkey        .......AC..........T...T.....A..A.....A.A...T..T............                                     ..G

                                                                          ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Mouse         . ..GTT.GA. .                                      .........TCTTC AGAGT    GAGTTCCAGGACACCAGGGCTA
CHO           ....GTT.AA...                                      .........TCTACCAGAGTTCCTGAGTT CAAGACA    GGCTA .
CHO (alt)             C                                                            T
                      T                                                            A


Human         ACAGAGCGAGACTCCATCTC A-rich sequence
Monkey        ......A
Mouse          .....  .A..C.TG... A-rich sequence
CHO           ......  .A..C.TG... A-rich sequence
```

Fig. 2. Comparison of the human Alu consensus sequences, an African green monkey Alu-equivalent sequence, the mouse B1 consensus sequence, and the Chinese hamster Alu-equivalent sequence. The top line is the consensus of the human Alu repeat extending from the first nucleotide of the repeat to the first nucleotide before the 3′ dAMP-rich region (text and Fig. 3). Position 0 is the G residue at the Alu I (restriction endonuclease) cleavage site. All other sequences are compared to the human sequence. The second line represents the sequence of part of an African green monkey Alu family sequence found as an insert in an SV40 viral genome (*34*). The third line represents the mouse B1 consensus sequence determined by Krayev *et al.* (*18*), and the fourth line represents the Chinese hamster Alu-equivalent sequence determined by Hayes *et al.* (*32*). A dot at any position indicates the same nucleotide as in the human Alu consensus sequence at the corresponding position. Occasionally a blank space has been inserted to facilitate the alignment of the sequences. The plus (+) overlining indicates the end of the first monomer unit of the human Alu consensus sequences. The asterisk (*) overlining indicates a nine-base sequence, perfectly conserved in the three consensus sequences and in the single African green monkey sequence. The second monomer of the human Alu sequence has a 31-bp insert not present in the first monomer unit that is located immediately to the left of the nine-nucleotide conserved sequence and is indicated by straight overlining (−). The mouse and Chinese hamster sequences each have a 32-bp insert not represented in the human or monkey sequences. This 32-bp sequence is located immediately to the right of the nine-nucleotide conserved sequence and is indicated by tilde overlining (~).

not conserved among different Alu family members, but are unique to each different Alu sequence. By analogy with the direct repeats flanking known bacterial transposons or insertion sequences and mobile DNA elements in eukaryotic DNA's (*36*), the direct repeats flanking Alu family members may have resulted from the duplication of the DNA sequence at the target site of Alu insertion into chromosomal DNA. If so, Alu family members may be mobile DNA elements.

The 3′ dAMP-rich sequence and the flanking direct repeats that are characteristic of Alu family members are not structures that occur only in full-length Alu sequences. Haynes *et al.* (*32, 37*) and Page *et al.* (*38*) described the sequence of a rodent dispersed repeat [termed the type 2 rodent Alu-equivalent sequence by Haynes and Jelinek (*37*)] having the middle 62 residues of the rodent Alu-equivalent sequence followed by 96 residues of a second repetitive sequence, followed in turn by a dAMP-rich sequence. The entire structure is flanked on either side by approximately 20-bp direct repeats reminiscent of the structures that flank standard Alu family members. Likewise, Van Arsdell *et al.* (*39*) found similar structures surrounding pseudogenes for the U1, U2, and U3 species of small nuclear RNA's (snRNA's) and a U6 snRNA gene described by Hyashi (*40*) is also flanked by them. Whether members of still other families of dispersed repeats are also flanked by these types of sequences awaits further investigation. If so, then they might also have been dispersed throughout eukaryotic DNA's by the same mechanism that disperses Alu family members.

Alu Family Members Within and Near hnRNA Transcription Units

Heterogeneous nuclear RNA molecules, like their DNA templates, contain interspersed repetitive sequences (*41*). In some molecules these repeats occur more than once in inverted orientation so that they can be isolated as "fold back" or inverted repeated double-stranded hnRNA segments (dshnRNA) (*12, 13, 16, 20, 42*). Comparisons of the sequences of major ribonuclease T1 oligonucleotides from the dshnRNA of cultured human cells with the human Alu family DNA sequence confirmed that Alu family members are heavily represented in hnRNA molecules (*12, 16, 27, 29*). As much as 25 percent of the mass of HeLa cell hnRNA might be composed of repetitious sequence now known to be transcripts of Alu family members (*20*). Likewise, HeLa and Chinese hamster ovary cell cytoplasmic, polyribosomal associated, poly(A) (polyadenylate)-terminated RNA molecules also contain these sequences but at a considerably lower frequency per molecule than in hnRNA (*20, 21, 43*). Some Alu family members are therefore located within RNA polymerase II transcription units. Apparently they are more frequently represented in regions of hnRNA that are removed during processing or turnover than in regions that are conserved and transported to the cytoplasm. Therefore, they probably do not serve a function in mature messenger RNA (mRNA) molecules.

One region of the Alu sequence resembles a splice junction of an intervening sequence in the human beta globin gene (*29*). This resemblance is restricted to a relatively short sequence unusually guanylate (G)-rich in composition so that the resemblance might be fortuitous. Likewise, regions of the mouse Alu equivalent sequence have also been compared with sequences at splice junctions (*18*). The presence of Alu sequences in hnRNA molecules, their depletion in mRNA molecules, and possible involvment in secondary structure suggest a role in hnRNA packaging or maturation (or both). However, it is not clear whether their removal during hnRNA processing is a passive event or whether they help direct this process. The abundance of Alu sequences in hnRNA could simply reflect their ubiquity in DNA.

Alu family members do not appear to be located exclusively within RNA transcription units. The seven Alu family members located in the human beta and beta-like gene cluster are between pairs of genes that are closely related and that are coordinately active in transcription during human development (*23*). Thus, these Alu sequences do not appear to lie within the beta and beta-like globin transcription units. However, final judgment on this issue should be reserved until the 3′ end points of transcription are determined, since transcription of the mouse beta major gene has been demonstrated to extend well beyond the poly(A) addition site for the mRNA (*44*) and may continue into an Alu-type repeat sequence. The Alu family members 5′ to the delta globin and 3′ to the beta globin genes are located in DNA regions conserved during recent evolution (*45*) and thought to be involved in the cis-acting suppression of fetal globin expression in adults (*23*). Whether the Alu sequences themselves participate in this gene "regulation" remains an unresolved but exciting possibility.

1. AAACAAGCAGGAGAGGCT...human alu...$\mathrm{A_7CA_5CA_7TCA_4CA_2TCA_3}$AAAACAAGCAGGAGGGGCT
2. AAGATTCACTTGTTTAG.....human alu...$\mathrm{A_{12}GAGAGATTGATTGA_2}$AAGATTCACTTGTTTAG
3. AAAGAAATGG............human alu...$\mathrm{A_{14}GA_3GA_3GA_4GA_5GA_6GA_3}$AAATAAATGG
4. GTTTAGATAAG...........human alu...$\mathrm{A_{25}}$GTTTAGATAAA
5. AAAAGAAACTTGGAAAGAG.....cho alu...$\mathrm{A_2TA_3TA_3TA_4TCTTA_7}$AAAAGGAAACTTGGAAAGGA
6. AACATACTAATTTTG.........cho alu...$\mathrm{A_4CA_2}$AACTATAATTTTTG
7. GTCAGCC.................cho alu...$\mathrm{TGA_5CCA_5GA_7GA_5GA_5GA_3}$GTTCCAGGCCAGTCAGCC
8. AGCTCATGAATGAAG.........cho alu...$\mathrm{CCA_5CA_3TCA_4}$CCAGACAGGCACAGCCCCAGCCCAT
9. GAGACAACAAATCAGAG.....mouse alu...$\mathrm{A_7CCA_3CCA_3CCA_3CCA_6}$CCGAGACAACAAATCAAAT

Fig. 3. Comparison of the sequences flanking Alu family members in three different mammalian species. The sequences flanking nine Alu family members are compared. The dotted line represents the Alu sequences. The first human sequence is from an Alu family member located downstream from an insulin gene (*25*). The second, third, and fourth human Alu sequences are located in the beta and beta-like gene cluster (*30, 31*). The four Chinese hamster sequences and the one mouse sequence were observed in cloned genomic DNA fragments selected for their ability to hybridize with dshnRNA isolated from cultured cells from these rodents (*18, 32*). In each example, the Alu sequence is immediately flanked on the 3′ side by a dAMP-rich sequence of the general structure $[N(A)_n]_m$, where N represents any nucleotide and in some examples more than one. Each of the nine Alu examples including their associated dAMP-rich regions are flanked on either side by direct repeats (indicated by underscoring). The generalized structure for these Alu flanking sequences is represented in the bottom line of the figure.

Discrete Low Molecular Weight RNA's Contain Alu Sequences

Alu family members are transcribed as discrete short RNA molecules, presumably by RNA polymerase III. Originally Jelinek and Leinwand (*46*) identified a discrete low molecular weight RNA, known as 4.5*S* RNA, that could be purified from Chinese hamster ovary cells because it formed base paired duplexes with poly(A)-terminated hnRNA and mRNA molecules. The nucleotide sequence of this RNA isolated from mouse, Syrian hamster, and Chinese hamster cells has been determined (*32, 47*). The molecule is 96 residues long and is terminated by pppGp at its 5′ end and by a short oligo(U) (uridylate) sequence of variable length at its 3′ end, both of which are characteristic of RNA polymerase III transcription products. It contains a 26-residue region at its 5′ end having 65 percent homology with the rodent Alu-equivalent consensus sequence and a 50-residue sequence at its 3′ end having 88 percent homology with the rodent Alu-equivalent consensus sequence. The 4.5*S* RNA can thus be considered as a member of the rodent Alu-equivalent family with a sequence somewhat divergent from the consensus or most frequent rodent Alu-equivalent sequence. Presumably this sequence homology allows it to undergo base pairing with hnRNA molecules that contain transcripts of both strands of the Alu sequence in high abundance (*13, 32*). Whether this association between the 4.5*S* RNA and the poly(A)-terminated hnRNA and cytoplasmic RNA molecules occurs within living cells is unknown.

The 4.5*S* RNA has not been detected in human cells. However, Weiner (*48*) demonstrated that another discrete low molecular weight RNA in HeLa cells, the 7*S* RNA, could undergo base pairing with human Alu family members. After hybridization of this RNA to Alu sequences a specific region (or regions) comprising at most one-half of the RNA was protected from mild digestion by ribonuclease T1. Busch and his colleagues (*49*) have determined the sequence of the rat 7*S* RNA. As predicted from Weiner's observations, the molecule is extensively homologous to the Alu sequence and, like the 4.5*S* RNA, the 7*S* RNA could be considered as an Alu family member with a sequence somewhat divergent from the consensus or most frequent Alu sequence.

The 7*S* RNA has been conserved during recent evolution. Two-dimensional analyses of ribonuclease T1 oligonucleotides indicate that avian and murine 7*S* RNA's are similar in sequence (*50*). If the sequence homology between Alu family members and the 7*S* RNA is biologically significant, then avian species might be expected to have Alu related sequences in their DNA's. In this respect, it is germane that Czernilofsky *et al.* (*51*) have identified Alu-like sequences in the vicinity of three candidates for splice junctions of the avian sarcoma virus *src* mRNA, and Breathnach and Chambon refer to an Alu-like sequence in intron C of the chicken conalbumin gene (*52*). Thus, Alu sequences may also be present in the DNA's of some nonmammalian vertebrate species as well as in various mammalian species.

Like the Alu sequences described above, the type 2 rodent Alu-equivalent sequence (*37*) is also extensively homologous to an abundant low molecular weight RNA. Busch and his colleagues (*53*) have determined the sequence of a rat RNA known as the 4.5S_I RNA that shares most of its sequence with type 2 Alu family members, although the RNA and DNA sequences are not completely colinear, but scrambled with respect to one another. The biological function (or functions) of neither the 4.5*S*, the 4.5S_I, nor the 7*S* RNA's are known. Since each appears to be a single molecular sequence rather than a group of closely related sequences as are Alu family members, each must be encoded by a single gene or group of invariant genes, unlike other low molecular weight RNA's that are transcribed from Alu family members themselves.

Alu Sequences Transcribed by RNA Polymerase III

Duncan *et al.* (*17*) demonstrated that two DNA sequences in the human beta and beta-like globin gene region are transcribed in vitro by RNA polymerase III. Further studies demonstrated that the transcription templates for these RNA's are Alu family members (*32, 43*). Detailed mapping of these RNA transcripts revealed that the start site for transcription was close to or coincident with the first nucleotide of the Alu sequence. Transcription proceeded through the Alu sequence and terminated beyond the dAMP-rich region at the 3′ side of the Alu sequence in a DNA sequence having two regions of four adenylic acid residues, resulting in an oligo(U) sequence at the 3′ end of the RNA.

Although the human Alu sequence is composed of two monomer units similar to one another in nucleotide sequence, transcription in the cell-free system has only been observed to begin within the first monomer. Elder *et al.* (*43*) identified the sequence GAGTTCPuAGACC in the first monomer unit as the most likely internal "control" region for RNA polymerase III transcription (Pu is purine). A modified second copy of this sequence is present in the second monomer unit of the human Alu sequence, but is interrupted by the 31-bp insert that is absent from the first monomer (see above and Fig. 2). By analogy to the internal "control" region of *Xenopus* 5*S* DNA (*54*), this internal control region in the Alu sequence lies downstream from the transcription initiation site. Each Alu family member seems to produce one or only a few transcription products of defined length. However, the RNA transcripts from different Alu family members vary in length because transcription termination occurs at different positions external to the Alu sequence. The entire human Alu family (approximately 500,000 copies per haploid genome) would thus produce a heterogeneous set of RNA's having the shared Alu sequence at their 5′ ends, but individual sequences at their 3′ ends.

Like human Alu family members, rodent type 2 Alu-equivalent family members also serve as templates for RNA polymerase III in vitro, and analogous RNA's have been isolated from the nuclei of growing Chinese hamster cells (*37*). Standard Chinese hamster Alu family members whose consensus sequence is represented in Fig. 2 could not be demonstrated to serve as transcription templates in vitro. In this respect, they act like the second monomer unit of the dimeric human Alu sequence while the type 2 sequences act like the first monomer unit.

Conclusion

The function (or functions) of the short interspersed repeats in eukaryotic DNA's has been the subject of speculation since their discovery (*5, 29, 55, 56*). It has not yet been possible to perform directed experiments to elucidate these functions; and in the absence of such experiments we can only speculate on their biological role (or roles) on the basis of the properties they exhibit. The Alu family is the single most abundant family of interspersed repeats in the mammalian genomes that have been examined and as such have been the most completely described. Whether they function in the control of gene activity as has been suggested, but not demonstrated for short-period interspersed repeats in general (*5*), remains to be determined. From sequence analyses, they appear to be mobile DNA elements, possibly the most successful such elements in primate and rodent genomes. It has been suggested that the low molecular weight RNA's transcribed from them may function in their dispersal to new chromosomal locations, possibly through reverse transcript intermediates primed by the 3′ oligouridylate sequences basepaired to the A-rich sequences (*39, 57*). If Alu family members are nothing more than mobile DNA elements, then like other mobile DNA elements (*58*) they may have profound influences on gene expression at some of their sites of insertion into chromosomal DNA. The properties of Alu family members may also be attributes of other families of short-period dispersed repeats in both mammalian and nonmammalian species. An understanding of Alu family members may provide insights into their properties.

References and Notes

1. B. Lewin, *Gene Expression*, vol. 2, *Eucaryotic Chromosomes* (Wiley, London, 1974).
2. R. J. Britten and D. E. Kohne, *Science* **161**, 529 (1968).
3. C. W. Schmid and P. L. Deininger, *Cell* **6**, 345 (1975).
4. C. M. Houck, F. P. Rinehart, C. W. Schmid, *J. Mol. Biol.* **132**, 289 (1979).
5. E. H. Davidson and R. J. Britten, *Q. Rev. Biol.* **48**, 565 (1973).
6. J. Bonner, W. Garrard, J. Gottesfeld, D. S. Holmes, J. S. Sevall, M. Wilke, *Cold Spring Harbor Symp. Quant. Biol.* **38**, 303 (1974); M. E. Chamberlain, R. J. Britten, E. H. Davidson, *J. Mol. Biol.* **96**, 317 (1975); R. B. Goldberg *et al.*, *Chromosoma* **251**, 225 (1975); E. H. Davidson, G. A. Galau, R. C. Angerer, R. J. Britten, *ibid.*, p. 253.
7. P. L. Deininger and C. W. Schmid, *J. Mol. Biol.* **106**, 773 (1976).
8. G. Corneo, E. Ginelli, E. Polli, *ibid.* **33**, 331 (1968).
9. J. W. Adams, R. E. Kaufman, P. J. Kretschner, M. Harrison, A. W. Nienhuis, *Nucleic Acids Res.* **8**, 6133 (1980).
10. A. R. Wyman and R. White, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 6754 (1980).
11. D. Wilson and C. Thomas, *J. Mol. Biol.* **84**, 115 (1974); P. J. Dott, C. R. Chuang, G. F. Saunders, *Biochemistry* **15**, 4120 (1976).
12. W. R. Jelinek, *J. Mol. Biol.* **115**, 591 (1977).
13. ———, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 2679 (1978).
14. F. D. Costantini, R. J. Britten, E. H. Davidson, *Nature (London)* **287**, 111 (1980); W. H. Klein *et al.*, *Cell* **14**, 889 (1978).
15. R. H. Scheller, D. M. Anderson, J. W. Posakony, L. B. McAllister, R. J. Britten, E. H. Davidson, *J. Mol. Biol.* **149**, 15 (1981).
16. H. D. Robertson, E. Dickson, W. R. Jelinek, *ibid.* **115**, 571 (1977).
17. C. Duncan *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 5095 (1979).
18. A. S. Krayev, D. A. Kramerov, K. G. Skryabin, A. P. Ryskov, A. A. Bayev, G. P. Georgiev, *Nucleic Acids Res.* **8**, 1201 (1980).
19. C. M. Houck, F. P. Rinehart, C. W. Schmid, *Biochim. Biophys. Acta* **518**, 37 (1978).
20. W. R. Jelinek, R. Evans, M. Wilson, M. Salditt-Georgieff, J. E. Darnell, *Biochemistry* **17**, 2276 (1978).
21. W. R. Jelinek, unpublished work.
22. F. P. Rinehart, T. Rich, P. L. Deininger, C. W. Schmid, *Biochemistry* **20**, 3003 (1981).
23. E. F. Fritsch, R. M. Lawn, T. Maniatis, *Cell* **19**, 959 (1980); L. W. Coggins *et al.*, *Nucleic Acids Res.* **8**, 3319 (1980).
24. R. Della Fovira, E. P. Gelman, R. C. Gallo, F. Wong-Stall, *Nature (London)* **292**, 31 (1981).
25. G. I. Bell, R. Pictet, W. J. Rutter, *Nucleic Acids Res.* **8**, 4091 (1980).
26. G. Grimaldi, T. McCutchan, M. Singer, unpublished result.
27. C. M. Rubin, C. M. Houck, P. L. Deininger, T. Friedman, C. W. Schmid, *Nature (London)* **284**, 372 (1980).
28. P. L. Deininger, D. J. Jolly, C. M. Rubin, T. Friedmann, C. W. Schmid, *J. Mol. Biol.* **151**, 17 (1981).
29. W. R. Jelinek *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 1398 (1980).
30. F. E. Baralle, C. C. Shoulders, S. Goodbourn, A. Jeffreys, N. J. Proudfoot, *Nucleic Acids Res.* **8**, 4393 (1980).
31. C. H. Duncan, P. Jagadeeswaran, R. C. C. Wang, S. Weissman, *Gene* **13**, 185 (1981).
32. S. R. Haynes, T. P. Toomey, L. Leinwand, W. R. Jelinek, *Mol. Cell. Biol.* **1**, 573 (1981).
33. J. Pan, J. T. Elder, C. H. Duncan, S. M. Weissman, *Nucleic Acids Res.* **9**, 1151 (1981).
34. B. R. Dhruva, T. Shenk, K. N. Subramanian, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 4514 (1980).
35. J. H. Taylor and S. Watanabe, *ICN-UCLA Symp. Mol. Cell. Biol.* **22**, 597 (1981).
36. J. A. Shapiro, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 1933 (1979); A. Flavell, *Nature (London)* **289**, 10 (1981); M. P. Calos and J. H. Miller, *Cell* **20**, 579 (1980).
37. S. R. Haynes and W. Jelinek, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
38. G. S. Page, S. Smith, H. Goodman, *Nucleic Acids. Res.* **9**, 2987 (1981).
39. S. W. Van Arsdell, R. A. Denison, L. B. Bernstein, A. M. Weiner, T. Maser, R. F. Gestland, *Cell*, in press.
40. K. Hayashi, *Nucleic Acids Res.* **9**, 3379 (1981).
41. J. E. Darnell and R. Balint, *J. Cell. Physiol.* **76**, 349 (1970); G. R. Molloy, W. R. Jelinek, M. Salditt, J. E. Darnell, *Cell* **1**, 43 (1974).
42. W. R. Jelinek and J. E. Darnell, *Proc. Natl. Acad. Sci. U.S.A.* **69**, 2537 (1972); A. P. Ryskov, V. R. Farashyan, G. P. Georgiev, *Biochim. Biophys. Acta* **262**, 562 (1972); W. R. Jelinek, G. R. Molloy, R. Fernandez-Munoz, M. Salditt, J. E. Darnell, *J. Mol. Biol.* **82**, 361 (1974); N. V. Federoff, P. K. Wellauer, R. Wall, *Cell* **10**, 597 (1977).
43. J. T. Elder, J. Pan, C. H. Duncan, S. M. Weissman, *Nucleic Acids Res.* **9**, 1171 (1981).
44. E. Hofer and J. E. Darnell, *Cell* **23**, 585 (1981).
45. P. A. Barrie, A. J. Jeffreys, A. F. Scott, *J. Mol. Biol.* **149**, 319 (1981).
46. W. R. Jelinek and L. Leinwand, *Cell* **15**, 205 (1978).
47. F. Harada and N. Kato, *Nucleic Acids Res.* **8**, 1273 (1980).
48. A. Weiner, *Cell* **22**, 209 (1980).
49. W. Y. Li, R. Reddy, T. Hennig, P. Epstein, H. Busch, unpublished result.
50. E. Erickson, R. L. Erickson, B. Henry, N. R. Pace, *Virology* **53**, 40 (1973).
51. A. P. Czernilofsky, A. D. Levinson, H. E. Varmus, J. M. Bishop, E. Tischer, H. M. Goodman, *Nature (London)* **287**, 1978 (1980).
52. R. Breathnach and P. Chambon, *Annu. Rev. Biochem.* **50**, 349 (1981).
53. R. Reddy, P. C. Hennig, H. Busch, unpublished results.
54. S. Sakonju, D. F. Bogenhagen, D. D. Brown, *Cell* **19**, 13 (1980).
55. E. H. Davidson and R. J. Britten, *Science* **204**, 1052 (1979).
56. W. F. Doolittle and C. Sapienza, *Nature (London)* **284**, 601 (1980); L. Orgel and F. H. C. Crick, *ibid.*, p. 604.
57. P. Jagadeeswaran, B. G. Forget, S. M. Weissman, *Cell*, in press.
58. B. McClintock, *Cold Spring Harbor Symp. Quant. Biol.* **16**, 13 (1952); *ibid.* **21**, 197 (1957); *Proc. Stadler Genet. Symp.* **10**, 25 (1978).
59. Supported in part by PHS grants GM21346 (to C.W.S.) and GM26704 (to W.R.J.), and a Faculty Research Award from the American Cancer Society (to W.R.J.).