MOLECULAR AND CELLULAR BIOLOGY, Nov. 1982, p. 1362–1371 Vol. 2, No. 11
0270-7306/82/111362-10$02.00/0

DNA Sequence Analysis of a Mouse Proα1(I) Procollagen Gene: Evidence for a Mouse B1 Element Within the Gene

JANET M. MONSON,†* JAMES FRIEDMAN, AND BRIAN J. McCARTHY‡

Department of Biochemistry and Biophysics, University of California, San Francisco, California 94143

Received 8 February 1982/Accepted 17 August 1982

In a 3.8-kilobase mouse DNA sequence encoding amino acid sequences for the proα1(I) chain of type I procollagen, 14 coding sequences were identified which specify a sequence 95% homologous to amino acid residues 568 to 963 of the bovine α1(I) chain. All of these coding sequences were flanked by appropriate splice junctions following the GT/AG rule. These observations suggest, but do not prove, that this proα1(I) gene is transcriptionally active. Of the 14 coding sequences, 7 were 54 base pairs in length, whereas the remainder were higher multiples of 54 base pairs. Nonrandom utilization of codons pertained throughout all of the coding sequences showing a preference (56%) for U in the wobble position. Two of the intervening sequences encoded imperfect vestiges of coding sequences which exhibited a codon preference different from that of the proα1(I) gene proper and were not flanked by splice junctions. One intervening sequence encoded a member of the mouse B1 family of middle repetitive sequences. It was flanked by 8-base-pair direct repeats and had a truncated A-rich region, suggesting that it may be a mobile element. Within this element were sequences which could function as a RNA polymerase III split promoter.

Much of the abiding interest in the collagen gene family stems from the key role that these extracellular structural proteins play in development, where they are primarily responsible for establishing and maintaining tissue architecture. As a prerequisite to studies of developmental gene regulation, considerable attention is being focused on the structures of the genes encoding the constituent polypeptide (proα) chains of type I procollagen. In particular, these are the proα2(I) genes from chickens (54, 55) and sheep (45) and the proα1(I) gene from mice (39). The proα2(I) genes are the largest, most highly interrupted genes yet identified in eucaryotes. For example, the chicken proα2(I) gene may have more than 50 intervening sequences distributed over 38 kilobases (kb) of genomic DNA. The preponderance of 54-base-pair (bp) coding sequences observed in this gene led to the hypothesis that procollagen genes arose by the amplification of a primordial 54-bp unit (55). However, a more compact genomic organization is observed in a proα1(I) procollagen gene (39). Therefore, a more complex evolutionary history, involving successive unequal cross overs within coding sequences (CSs) or precise deletions or insertions of intervening sequences or both, has been postulated for procollagen genes (39).

The preponderance of intervening sequences in procollagen genes inevitably raises the issue of their possible function. One possibility is that they serve to stabilize the gene by reducing recombination within the homologous CSs. A more intriguing possibility is that they might encode other gene products. Candidates for such products might be regulators or proteins belonging to an extended family of genes, including those for the post-transcriptional and post-translational processing required for collagen maturation. Therefore, to inquire into these possibilities and to extend the knowledge of collagen CSs, we chose to establish the sequence of a 3.8-kb segment of a proα1(I) procollagen gene. Here we analyze that nucleotide sequence.

MATERIALS AND METHODS

DNA sequencing. The isolation of the mouse proα1(I) gene segment analyzed in this paper has been described previously (39). The DNA sequence of the inserts from two subclones, pMPC1C and pMPC1A, was established essentially by the protocol of Maxam and Gilbert (38). These two inserts span the first 3.8 kb of the MPC1 clone.

Intragenic genomic repetitive sequence. Genomic DNAs other than that from BALB/c mice were gifts from Lou Kunkel. The genomic DNA samples were cleaved with the appropriate restriction endonuclease,

† Present address: Zymos Corp., Seattle, WA 98103.

‡ Present address: Department of Molecular Biology and Biochemistry, University of California, Irvine, CA 92717.

fractionated by agarose gel electrophoresis, and transferred to nitrocellulose by the Southern procedure (47) after partial depurination (51). The *Kpn*I-*Bam*HI subfragment of MPC1A was ^{32}P labeled with radioactive dCTP in a T4 DNA polymerase reaction (39). Hybridizations were performed in 50% formamide–5×SSC–50 mM Tris-hydrochloride (pH 7.4)–1 mM EDTA–0.1% sodium dodecyl sulfate–10× Denhardt solution (13) for 18 h at 37°C. Filters were washed three times for 20 min each at 50°C in 0.1× SSC–0.1% sodium dodecyl sulfate. For the localization of the genomic repetitive sequence within the MPC1A clone, mouse (BALB/c) genomic DNA was ^{32}P labeled by a modified nick-translation reaction (44), and the hybridization to a Southern blot of MPC1A restriction fragments was performed as described above except that 10% dextran sulfate (51) was included in the reaction.

RESULTS

DNA sequence. The nucleotide sequence of two subclones representing a segment of a mouse proα1(I) procollagen gene was completed. The strategy employed to establish this 3.8-kb DNA sequence is sumamrized in Fig. 1. The sequence of each DNA strand was established from a set of overlapping determinations. Data from a previous report (39) are included in Fig. 1 by dashed arrows to illustrate all of the overlaps. The entire 3.8-kb DNA sequence is shown in Fig. 2. Fourteen CSs were identified within this DNA segment in comparison with the known amino acid sequence of the calf proα1(I) chain (17, 18, 33, 52). A schematic of the genomic organization of these CSs is depicted in Fig. 3A. In addition to the eight reported CSs (39), six new CSs (7–9, 12–14) were identified which encode 198 amino acids. Taken together, the 14 CSs specify residues 568 to 963 of a mouse α1(I) chain. In comparison to the corresponding calf sequence, there is 95% homology at the amino acid level. All of the 18 amino acid substitutions can be accounted for by single base changes in the mouse codons. The location of each substitution is indicated in Fig. 2 above the specified mouse sequence.

The 14 CSs exhibit a unique degree of size regularity. All are 54 bp in length or higher multiples of 54. The 162-nucleotide CS7 is the largest CS yet reported for the α-chain domain. Of the other 13 CSs, 6 are 108 bp, and 7 are 54 bp.

The codon usage for the additional 198 amino acids confirms and extends the nonrandom distribution of codons noted previously (39). For the 396 codon total, there is a marked preference (56%) for U in the wobble position, whereas C, A, and G occur in this position only 24, 14, and 6% of the time, respectively. The most striking bias in codon utilization occurs for alanine, where 35 of 44 codons are GCU. The dominant glycine codon is GGU (80/132), but GGC (31/132), GGA (19/132), and GGG (3/132) are also

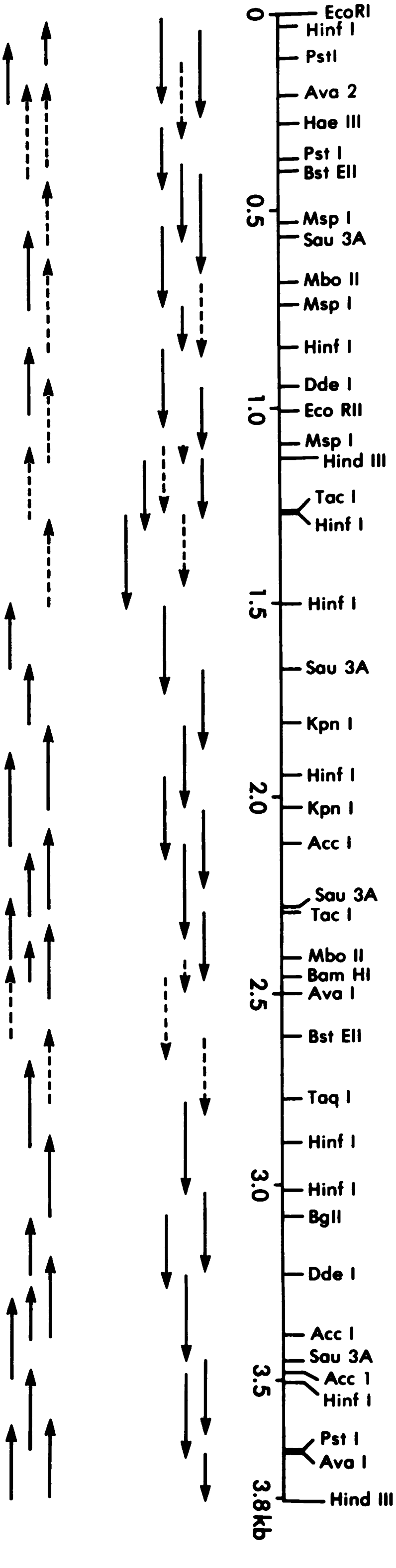


FIG. 1. Strategy employed to determine the sequence of the first 3.8 kb of MCP1. Dashed arrows from a previous report (39) are included to illustrate all of the overlaps. Only relevant restriction enzyme recognition sites are shown.

```
GAATTCAGGGGCCTCTTAACCCAGGTTCCACCTGAATCCCCAAGTAGGCCCCTTTAACCCCTGAAAGAACT

TTACATCCTTGACAGAGCTAAGGAGGGCCTCTTAGATATAGCTGCAGAATGCTGTAGCTCAACTCTTGGCC
                                                          568
                                                          Gly Asp Ala Gly
CCAATCTGGAAGTCTCAGGGGTTGTTTTGGTTTTTTTTTTTTTTCCTGTTTTTAG   GGT GAT GCT GGT
                        Ala                     Val
Pro Lys Gly Ala Asp Gly Ser Pro Gly Lys Asp Gly Ala Arg Gly Leu Thr Gly
CCC AAA GGT GCT GAT GGT TCT CCT GGT AAA GAT GGT GCC CGT GGT CTG ACT GGT
                                                    603
Pro Ile Gly Pro Pro Gly Pro Ala Gly Ala Pro Gly Asp Lys
CCC ATT GGT CCT CCT GGC CCT GCT GGT GCC CCT GGT GAC AAG  GTTAGTGGCTACCT

CTTCACCTTCTTTCTTAACTCAAAACCTCCTTCGCAAGCTCAGGTGGGGCTCTGCAGGGAAGGCAGGTCCT

GCCATATGAAGCTGGTGACCCAGGAAGTTCAAGGGACCAGGAGGGAGGGAAGGTGTCATTGGTTCATCTCC
                                    604                             Ala
                                    Gly Glu Ala Gly Pro Ser Gly Pro Pro
CAGGAGAGTTGAGAGTTCCTGTCTCTCCCTCATAG GGT GAA GCT GGT CCC AGT GGT CCT CCC
                                621
Gly Pro Thr Gly Ala Arg Gly Ala Pro
GGT CCC ACC GGA GCC CGT GGT GCT CCC  GTAAGTACAGAAGACCCTGATCTCTGTTCATCCC

TTCTCCCCTACCTGCTTTCTGCCCCCCACCGCAAACCCCACCCCTTTACTCTATCCGTTCCTCTCCTTCCC
                                            622
                                            Gly Asp Arg Gly Glu Ala Gly
TAATGTTGAGACATCTCTCCAAAGTCGTCTCCTTCTTCTAG   GGA GAC CGT GGT GAG GCT GGT
                                        639
Pro Pro Gly Pro Ala Gly Phe Ala Gly Pro Pro
CCC CCT GGT CCT GCT GGC TTT GCC GGC CCC CCT  GTGAGTATCAAGACCCTCCTCATTTT

CTGTCCCTAGCTGAGACACGAGGCATGGGACCTTGGGTGGCTGAATGAGGACAGAAGTGTTACCCTGAGTC

AGAGGAGAAGGGTGGGGAGGTACTGGTGTCTCCAAGTGTCTCTACATCTCCAAGTCCCTATCTGTGGCCCT

TCCTCTAGCCCAGAGGCCCTCTGCTCTCAGGCTGCCTCCTCCACTCCTCCACTCTCCATTCTCCCTCCTGC
        640                                                 Ala
        Gly Ala Asp Gly Gln Pro Gly Ala Lys Gly Glu Pro Gly Asp Thr Gly
CTAG    GGT GCT GAT GGC CAA CCT GGT GCG AAA GGT GAA CCT GGT GAT ACT GGT
Ala
Val Lys Gly Asp Ala Gly Pro Pro Gly Pro Ala Gly Pro Ala Gly Pro Pro Gly
GTT AAA GGT GAT GCT GGT CCT CCT GGC CCT GCT GGT CCT GCT GGA CCC CCC GGC
    675
Pro Ile
CCC ATT  GTAAGTATCTTGTCTTCTGCACCATAAGCTTTGGATAGCCTTGGACTTGGGGCTAGCCTGGA
                            676
                            Gly Asn Val Gly Ala Pro Gly Pro Lys Gly Pro
TCTCATACCTTGACACTGTCTTACAG  GGT AAC GTT GGT GCT CCT GGA CCC AAA GGT CCT
        Ser             693
Arg Gly Ala Ala Gly Pro Pro
CGT GGT GCT GCT GGT CCC CCT  GTGAGTATCATATGCATCTCTGTCGCGACTCCCCAAAGGCAG

AGACTGGAGATGAGGCCAGGTGACAGGTGACTGTTCACTTCTGACCACCCAATGTTCTCTCCTACCAG
694
                                                                    711
Gly Ala Thr Gly Phe Pro Gly Ala Ala Gly Arg Val Gly Pro Pro Gly Pro Ser
GGT GCT ACT GGC TTC CCT GGT GCT GCT GGC CGT GTC GGT CCC CCT GGT CCC TCT

GTGAGTATCTGTGGTTCTGGAATGAGGATGGGGTGAGACATGTATTGTCAGGACAGCAGGCCTGGCTGGG
                                                                712
                                                                Gly Asn
GCTTGCCACTATGATGCTTTGGAAGCCTGGACTCTGACAGTCCTTCTTGTGCCCATCTAG    GGA AAT
                                    Ala                 Ser
Ala Gly Pro Pro Gly Pro Pro Gly Pro Val Gly Lys Glu Gly Gly Lys Gly Pro
GCT GGA CCC CCT GGC CCT CCC GGT CCC GTT GGC AAA GAA GGG GGC AAA GGT CCC

Arg Gly Glu Thr Gly Pro Ala Gly Arg Pro Gly Glu Val Gly Pro Pro Gly Pro
CGT GGT GAG ACT GGC CCT GCT GGA CGT CCT GGT GAA GTT GGT CCC CCA GGT CCC
                                Ala                         765
Pro Gly Pro Ala Gly Glu Lys Gly Ser Pro Gly Ala Asp Gly Pro Ala
CCC GGT CCT GCT GGT GAG AAA GGA TCT CCT GGT GCT GAT GGA CCT GCT  GTAAGT

GCTAACTCACATCTCTGTGATTGTGGAGAGTTCCAGAGTTGTGTATGTGTTTCCTGTGCTACTGTGAGCCC
                            766 Ala
                            Gly Ser Pro Gly Thr Pro Gly Pro Gln Gly Ile
TTCTCACCCCTGTCTGCCTCCCACAG  GGC TCT CCT GGT ACC CCT GGA CCT CAG GGT ATT

Ala Gly Gln Arg Gly Val Val Gly Leu Pro Gly Gln Arg Gly Glu Arg Gly Phe
GCT GGA CAA CGT GGT GTG GTC GGT CTT CCC GGT CAG AGA GGA GAA AGA GGC TTC
                        801
Pro Gly Leu Pro Gly Pro Ser
CCT GGT CTT CCT GGC CCC TCT  GTGAGTGTTCTTTCCTCTTGGGGTGTCCAAGAAGAATCATCT

TAGGACTTGAGTACTAGAAGGGGCAGGGTAGCAGCAGTGGAGACAAGGAGAGCAAATGTGATAGAAATGCT

CTCATGGTACCCAGGTGGTGGTGGTGAACACCTTTAATCCCAGCACTAAGGAAGCAGAGGCAGGTAAATTC
```

FIG. 2. DNA sequence of the first 3.8 kb of MCP1 with a translation of the 14 identified coding sequences. Amino acid residues are numbered from the first glycine of the Gly-X-Y repeating pattern of the corresponding bovine α1(I) amino acid sequence. At the 18 position, where amino acid substitutions occur in the mouse sequence, the corresponding bovine α1(I) residue is written above the specified mouse amino acid.

```
CTGAGTTCAAGACCAGCCTGGTCTACAGAGCGAGTGCCAGGACAGCCAGAGCTACACAGAGAAACCCTGTC

TTGAAAACCAAACTAAACAAACACACAAAAGAAGTCTCATGGCTTGAGCCACCACATCTGACCTCCAGCCT
                802                                     Ala
                Gly Glu Pro Gly Lys Gln Gly Pro Ser Gly Ser Ser Gly Glu
TACTCTGTTCTTTAG GGT GAA CCT GGC AAA CAA GGT CCT TCT GGA TCA AGT GGT GAA

Arg Gly Pro Pro Gly Pro Met Gly Pro Pro Gly Leu Ala Gly Pro Pro Gly Glu
CGC GGT CCC CCT GGC CCC ATG GGG CCC CCT GGA TTG GCT GGT CCC CCT GGT GAA
            837
Ser Gly Arg Glu
TCT GGA CGT GAG  GTCAGCAGCCCCCACCTCCTAGCCAGTCCCCGTGCAGGACCACTTGTCCTACGC
                                                    838 Ala
                                                    Gly Ser Pro Gly Ala
CTGACCTCTTCCCTGGAGTGGACACTCATCTCTCCCCCTCCCTTTACAG   GGA TCC CCT GGT GCT
                                Ser             855
Glu Gly Ser Pro Gly Arg Asp Gly Ala Pro Gly Ala Lys
GAA GGC TCC CCT GGA AGG GAT GGT GCT CCC GGG GCC AAG  GTAAGAGATCATGCCACA

TATCAGGCTGGACTCTGGTGAGCCCTGGCCCCTCCCCAGACTGCACATTTCATCTGAGGCTGACCCCATGA
                        856
                        Gly Asp Arg Gly Glu Thr Gly Pro Ala Gly Pro Pro
CCTCTACCCTCTGTTCCCAG    GGT GAC CGT GGT GAG ACT GGC CCC GCT GGC CCC CCT

Gly Ala Pro Gly Ala Pro Gly Ala Pro Gly Pro Val Gly Pro Ala Gly Lys Asn
GGT GCC CCT GGT GCT CCC GGT GCT CCC GGC CCT GTT GGT CCC GCT GGC AAG AAT
                    891
Glv Asp Arg Gly Glu Thr
GGC GAT CGT GGT GAG ACT  GTAAGTTGCTGAGCTCAGATCAGACCTCTTCTTCAGACATGCTCAA

AGGCCTCGAAATGGATGAATTACCTCACTCAGGCTGGAGAAGAAGAGGGTTTTGGGATATGTCTGGGTCCT

TAGCTTCCAGGGAAAAACAGTGATTACTTAGCCTTCATCTAGGCAGGACTCCATCTTCCCCAAGGCATGGG

GCTTGCCCCAGCTTATCCCATGAAGCCTGGCTCTGGGGAGGTTTAATCGGGAATCAGGAGAGGTGGCCACA

GCAGAAGAGATGTGGGTCAGGAGGAGTCTAGCCAAGGGGAAGGCTCACCCTGAAGCCCTCAGCCTTGGCTT
                                        892                 Ile
                                        Gly Pro Ala Gly Pro Ala Gly Pro
TTCCAAGTCAAGATCTAACACACTGTTTTCTCTCCAG   GGT CCT GCT GGT CCT GCT GGT CCC
Val                                 909
Ile Gly Pro Ala Gly Ala Arg Gly Pro Ala
ATT GGC CCT GCT GGT GCC CGT GGC CCT GCT  GTAAGTGTCCCATGCCCACCTCATGCCCTC

GGAACACTGACCCAGAGGTGAATACCATCCCACTCAATACTCAGTGTGAACTTGCACCTGACAGCCTGTCT
                        910
                        Gly Pro Gln Gly Pro Arg Gly Asp Lys Gly Glu Thr
TGTCCCCTCCTCTCTTTTAG    GGA CCC CAA GGC CCC CGT GGT GAC AAG GGT GAG ACA

Gly Glu Gln Gly Asp Arg Gly Ile Lys Gly His Arg Gly Phe Ser Gly Leu Gln
GGC GAA CAA GGT GAC AGA GGC ATA AAG GGT CAT CGT GGC TTC TCT GGT CTC CAG
                Pro 945
Gly Pro Pro Gly Ser Pro
GGT CCT CCT GGT TCT CCT  GTGAGTATACTCAACTTCCCAGGCCCTGGCTGCCATCAGGCCTTTT

CCACCATACCCTGATGCAGATCACATGGTGGGAGGGGGACACTTGCCTCAGTGTCTACTCTAGATAGACAT

TCTGATTCCTTCCTAGTGAGGGTGAGGGGCAGGACACAGACAGACAGACAGAGGGCTCAAAGAAGGATTGT

GGGAGTGGTCTAAAATGAGGGGATATGTTACCCCCATTGCCCCTGAACTGCTTTTCTCTGTTCTTCAG
946                                                                 963
Gly Ser Pro Glv Glu Gln Gly Pro Ser Gly Ala Ser Gly Pro Ala Gly Pro Arg
GGT TCT CCT GGT GAA CAA GGC CCC TCT GGA GCT TCA GGT CCT GCA GGC CCC CGG

 GTAAGTTGCATCTTCCATTACTTCTTCCTGGGTTCCTCCCTGTCTTGGTTACTTTCAGATATCCCTCTGT

TCTCCAGAGAAGAGTTCAGAGACCAGACAGGGGGTGGGGAGAAAAAGAAGCTTT
```

FIG. 2. *Continued*

used. Proline codons are nearly evenly distributed between CCU (51/91) and CCC (39/91).

The splice junctions flanking the 14 CSs all follow the AG/GT rule, as shown in Fig. 4. They all exhibit substantial self-homology, as well as homology to similar sequences reported for a variety of other genes (8). In addition, the consensus sequence for these splicing sites shows considerable sequence complementarity to rat U1a RNA, suggesting that a small nuclear RNA resembling U1a may be involved in the processing of the primary transcript as postulated for other genes (37, 43). The degree to which both the encoded sequence and the splice junctions are conserved suggests, but does not prove, that this proα1(I) gene is transcriptionally active.

Intervening sequences containing CS vestiges. Two vestiges of CSs were identified within intervening sequences. Vestige 1 is located in the second intervening sequence as shown in Fig. 3A. Two apparent insertions of 10 and 2 nucleotides interrupt a pattern of nine Gly-X-Y triplets

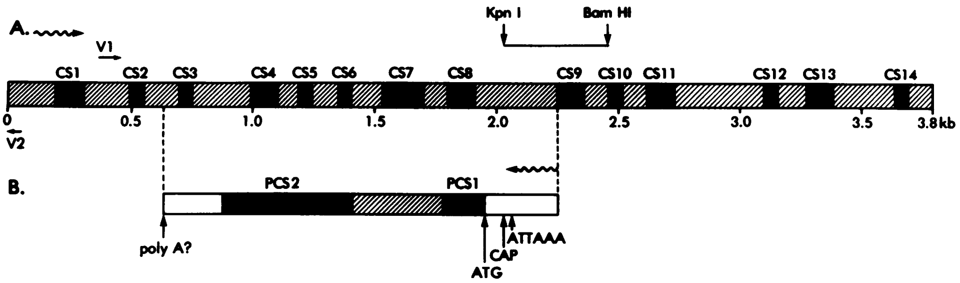


FIG. 3. (A) Genomic organization of an internal segment of a mouse proα1(I) gene. CSs are depicted as blackened boxes, and the intervening sequences are designated by cross-hatching. The locations of two vestiges of CSs are indicated by the arrows designated V1 and V2. The location of the *Kpn*I-*Bam*HI fragment which hybridizes as a genomic repetitive sequence is indicated by the bar above the gene. (B) Position and genomic organization of a putative gene identified on the complementary strand of the proα1(I) gene. The locations of the imperfect Goldberg-Hogness sequence (ATTAAA), a potential capping site (CAP), and initiating methionyl residue (ATG) are indicated by designated arrows. A possible poly(A) addition signal site is also illustrated. The direction of transcription is denoted by wavy arrows in (A) and (B).

(Fig. 5). Vestige 2 appears as an inversion within the first intervening sequence as depicted in Fig. 3A. In this DNA segment, an apparent single-base insertion interrupts six Gly-X-Y repeats (Fig. 5). Vestige 2 is also imperfect by virtue of the fact that it specifies both ocher and stop codons. Neither vestige is flanked by splice junction sequences, and the codon usage for each is unlike that of the proα1(I) gene proper.

Long, open reading frame on the complementary strand. Based on an analysis of the DNA sequence, a long, open reading frame of 594 nucleotides was identified complementary to nucleotides 1,483 to 890 of the proα1(I) sequence shown in Fig. 2. In searching the sequences upstream from this open reading frame for other gene characteristics, an imperfect (ATTAAA) Goldberg-Hogness box could be identified complementary to nucleotides 2,059 to 2,054 of Fig. 2. A potential capping site is located 31 bases downstream from this, and an ATG, followed by an open reading frame of 167 nucleotides, is located 112 nucleotides downstream from the imperfect Goldberg-Hogness box. By selecting an appropriate splicing regime, a putative gene could be constructed which contains two putative CSs which together yield a 699-nucleotide sequence, encoding 233 amino acids. The genomic organization of such a putative gene relative to the proα1(I) gene is illustrated in Fig. 3B. In such a construction, the putative CSs are complementary to both intervening and CSs within the mouse proα1(I) gene. A relevant point is that not all of the intervening sequences in the proα1(I) gene are multiples of three nucleotides. Consequently, when the complementary strand is read in a single frame, the amino acid sequence does not reflect the Gly-X-Y repeat of the proα1(I) gene CSs as might otherwise be expected. A search of the Dayhoff computer atlas of protein sequences failed to identify any protein with greater than 22% homology to the 233-amino acid sequence encoded by the two putative CSs described above; no clear polyadenylic acid addition signal corresponding to the AAUAAA observed for several mRNAs (42) could be identified for this putative gene; and there is as yet no evidence that it is transcriptionally active.

CS

1.	TTTTCCTGTTTTTAG	GTTAGT
2.	GTCTCTCCCTCATAG	GTAAGT
3.	GTCTCCTTCTTCTAG	GTGAGT
4.	CTCCCTCCTGCCTAG	GTAAGT
5.	GACACTGTCTTACAG	GTGAGT
6.	GTTCTCTCCTACCAG	GTGAGT
7.	CTTGTGCCCATCTAG	GTAAGT
8.	GTCTGCCTCCCACAG	GTGAGT
9.	TACTCTGTTCTTTAG	GTCAGC
10.	CCCCTCCCTTTACAG	GTAAGA
11.	ACCCTCTGTTCCCAG	GTAAGT
12.	CTGTTTTCTCTCCAG	GTAAGT
13.	CCTCCTCTCTTTTAG	GTGAGT
14.	TTCTCTGTTCTTCAG	GTAAGT

FIG. 4. Splice junction sequences.

By a similar analysis, a "virtual" gene encoding 246 amino acids has been reported on the

Vestige 1

```
GGCTCTGCAGGGAAGGCAGGTCCTGCCATATGAAGCTGGTGACCCAGGAAGTTCAAGGGACCAGGAGGGAGGGAAGGTGTCATTGGTTCATCT

GGCTCTGCAGGGAAGGCAGGTCCTGCC          GGTGACCCAGGAAGTTCA  GGACCAGGAGGGAGGGAAGGTGTCATTGGTTCATCT

GlySerAlaGlyLysAlaGlyProAla          GlyAspProGlySerSer  GlyProGlyGlyArgGluGlyValIleGlySerSer
```

Vestige 2

```
AGTCCCCGGAGAATTGGGTCCAAGGTGGACTTAGGGGTTCATCCGGGGAAATTGG

AGTCCCCGGAGAATTGGGTCCAAG TGGACTTAGGGGTTCATCCGGGGAAATTGG

OP ProGlyArgLeuGlyProGlu GlySerAspGlyLeuLeuGlyArgOC Gly
```

FIG. 5. Sequences for two vestiges of CSs. The positions of apparent nucleotide insertions are underscored and in bold type. The translation of the nucleotide sequence minus the apparent insertions is also indicated.

strand complementary to the human ε-globin gene (11), but the longest open reading frame in this case is 322 nucleotides as compared with the 594-nucleotide open reading frame described above.

Intragenic genomic repetitive sequence. A mouse genomic repetitive sequence was localized within the proα1(I) gene. Initially, when the entire cloned insert MPC1 was ^{32}P labeled and hybridized to a Southern blot of *Eco*RI-cleaved mouse DNA, a hybridization signal was observed throughout the lane of genomic DNA rather than at the expected 5.5-kb band. When ^{32}P-labeled inserts from each of three subclones of MPC1 were hybridized to identical blots, only the MPC1A insert exhibited this anomalous hybridization pattern. To further localize this repetitive sequence, a Southern blot of a series of digests of MPC1A was hybridized with BALB/c mouse DNA which had been ^{32}P labeled by nick translation. The *Kpn*I-*Bam*HI fragment shown in Fig. 3A was identified as the primary site of hybridization. Confirmatory evidence was provided by hybridizing this ^{32}P-labeled *Kpn*I-*Bam*HI fragment to a Southern blot (Fig. 6) of genomic DNAs isolated from several different species. It hybridized as a genomic repetitive sequence when tested against mouse liver (Fig. 6, lanes A to C) or A9-cell (Fig. 6, lanes D to F) DNAs or rat (Fig. 6, lane K) and hamster (Fig. 6, lane M) DNAs, but failed to cross-react with genomic repetitive sequences in human (Fig. 6, lanes G to I) or chicken (Fig. 6, lane J) DNAs. Therefore, the *Kpn*I-*Bam*HI fragment contains an apparently rodent-specific genomic repetitive sequence.

An analysis of the DNA sequence within this *Kpn*I-*Bam*HI fragment identified a region homologous to the mouse B1 genomic repetitive sequence (34). This 168-bp homolog is flanked by 8-bp direct repeats and is located between nucleotides 121 and 289 of IVS9, as illustrated in Fig. 7. The sequence extending from base 125 to 256 closely resembles the core consensus sequence for the mouse B1 and type I-CHO, Alu-equivalent families of repetitive sequences (27, 34). The 3′ flanking A-rich region is very similar

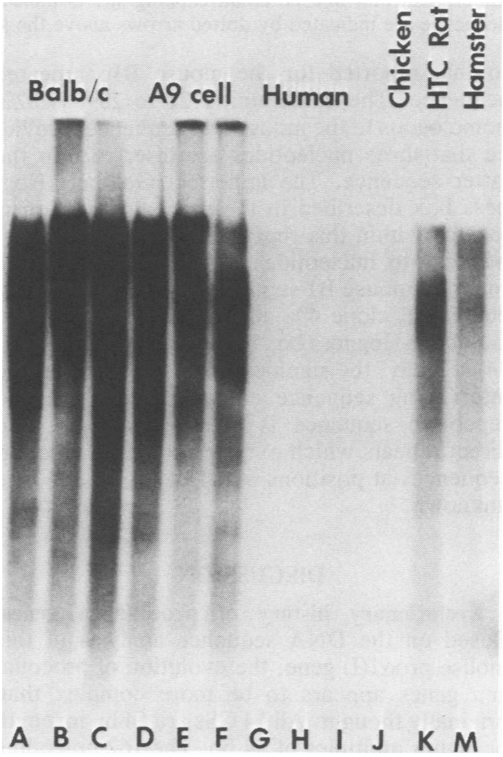
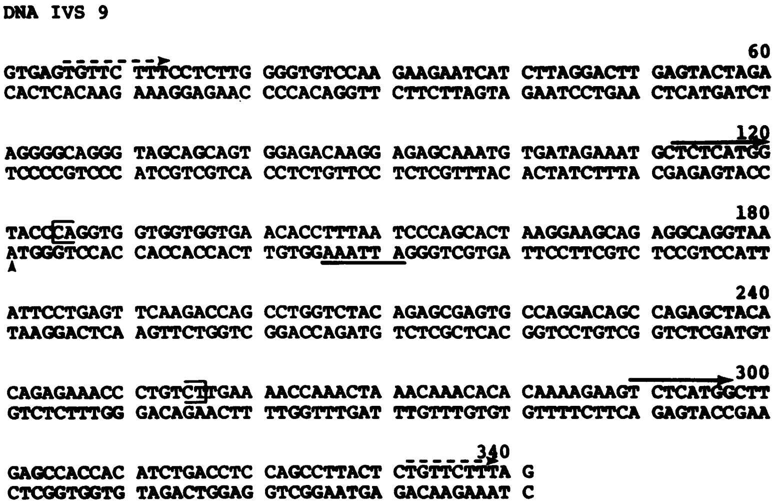


FIG. 6. Southern blot of cleaved genomic DNAs from several species after hybridization with the ^{32}P-labeled *Kpn*I-*Bam*HI fragment. BALB/c (lanes A to C), A9 cell (lanes D to F), and human (lanes G to I) DNAs were each cleaved with *Eco*RI, *Hin*dIII, and *Bam*HI, respectively. Chicken (lane J), rat (lane K), and hamster (lane M) DNAs were cleaved with *Eco*RI before electrophoresis through a 0.8% agarose gel. A 5-μg amount of DNA was loaded in each lane. Hybridization conditions were described in the text. Exposure time was 6 h.



```
DNA IVS 9

                                                                  60
GTGAGTGTTC TTTCCTCTTG GGGTGTCCAA GAAGAATCAT CTTAGGACTT GAGTACTAGA
CACTCACAAG AAAGGAGAAC CCCACAGGTT CTTCTTAGTA GAATCCTGAA CTCATGATCT

                                                                 120
AGGGGCAGGG TAGCAGCAGT GGAGACAAGG AGAGCAAATG TGATAGAAAT GCTCTCATGG
TCCCCGTCCC ATCGTCGTCA CCTCTGTTCC TCTCGTTTAC ACTATCTTTA CGAGAGTACC

                                                                 180
TACCCAGGTG GTGGTGGTGA ACACCTTTAA TCCCAGCACT AAGGAAGCAG AGGCAGGTAA
ATGGGTCCAC CACCACCACT TGTGGAAATT AGGGTCGTGA TTCCTTCGTC TCCGTCCATT

                                                                 240
ATTCCTGAGT TCAAGACCAG CCTGGTCTAC AGAGCGAGTG CCAGGACAGC CAGAGCTACA
TAAGGACTCA AGTTCTGGTC GGACCAGATG TCTCGCTCAC GGTCCTGTCG GTCTCGATGT

                                                                 300
CAGAGAAACC CTGTCTTGAA AACCAAACTA AACAAACACA CAAAAGAAGT CTCATGGCTT
GTCTCTTTGG GACAGAACTT TTGGTTTGAT TTGTTTGTGT GTTTTCTTCA GAGTACCGAA

                                           340
GAGCCACCAC ATCTGACCTC CAGCCTTACT CTGTTCTTTA G
CTCGGTGGTG TAGACTGGAG GTCGGAATGA GACAAGAAAT C
```

FIG. 7. Ninth intervening sequence (Fig. 2 and 3A), illustrating the encoded mouse B1 genomic repetitive sequence. The direct repeats flanking this repetitive sequence are illustrated by arrows above the sequence. The core sequence is indicated in brackets. The putative Goldberg-Hogness box on the complementary strand is underlined, and one potential capping site is indicated by an arrow. Direct repeats flanking the intervening sequence are indicated by dotted arrows above the sequence.

to that reported for the mouse B1c repeated sequence. The entire unit (121 to 289) is 82% homologous to the mouse B1c sequence, provided that three nucleotides are inserted into the latter sequence. The imperfect Goldberg-Hogness box described in the preceding section is located within this repetitive sequence complementary to nucleotides 146 to 151. All of the reported mouse B1 sequences and the CHO-Alu equivalent clone 49a also contain an imperfect Goldberg-Hogness box at this same position (27, 34). Finally, the significance of the fact that the intervening sequence containing this genomic repetitive sequence is itself flanked by 8-bp direct repeats which overlap the splice junction sequences at positions 6 to 13 and 332 to 339 is unknown.

DISCUSSION

Evolutionary history of procollagen genes. Based on the DNA sequence analysis of this mouse proα1(I) gene, the evolution of procollagen genes appears to be more complex than originally thought. All 14 CSs are 54 bp in length or higher multiples of 54 bp. The 162-bp coding unit is the largest yet found encoding the Gly-X-Y repeat in any type 1 collagen gene. Two pathways for the generation of 108-bp CSs starting from 54-bp units have been proposed (39). One pathway involves precise deletions of intervening sequences, and the other involves successive, unequal crossovers within homologous CS. The extension of either proposition could account for the 162-bp CS. However, since four successive, unequal crossovers would be required to generate a 162-bp CS from 54-bp units, this possibility seems unlikely. Thus, the presence of a 162-bp CS, as well as numerous 108-bp CSs, suggests that proα chain genes may be examples of partially processed genes. The fact that the junctions between the α chain, telopeptide, and propeptide domains at each end of the α chain are each fused into a single CS in the chicken proα2(I) gene (54) is also consistent with this view.

The processing mechanism by which intervening sequences might be precisely deleted remains somewhat obscure. However, one possibility is that a proα gene genomic fragment may have been incorporated into a retrovirus or cellular equivalent, as suggested for the α-ψ3 pseudogene (36, 40, 50). The transcription of these proviral sequences into RNA followed by partial RNA processing before reintegration could generate CSs that are higher multiples of 54 bp. Other examples may be the rat insulin I gene which has cleanly lost one intervening sequence (3) and certain retrovirus oncogenes which lack the intervening sequences found in their cellular analogs (4, 20, 24). More convincing evidence for this mechanism of dispersal and processing has recently been provided by the discovery of an immunoglobulin pseudogene having both a spliced J and C region and a poly(A)-rich tail (29) and a β-tubulin pseudogene lacking intervening sequences but also containing a poly(A) tail (53). Another possible mechanism might involve processing at the DNA level, but there is as yet no precedent for this in eucaryotes.

The discovery of vestiges of CSs within two of the intervening sequences (Fig. 5) points to an even more complex evolutionary history for procollagen genes. Whether this is a unique feature of procollagen genes remains to be seen; in most other genes, the amino acid sequence lacks the regularity that would allow evolutionarily related remnants of CSs to be recognized. In this respect, the Gly-X-Y repeat is an unambiguous indication that these two portions of intervening sequences are derived from some collagen-like CSs. Although the origin of these vestiges is unclear, it is noteworthy that both exhibit a codon utilization which is unlike that observed for the proα1(I) gene proper. Most noticeable is the fact that the glycine codon GGG is used twice in each vestige, whereas it is rarely used in collagen genes in general (21, 39, 54, 55). Thus, it would seem unlikely that these vestiges recently arose from the surrounding, bonafide CSs. The imperfections in these vestiges, including the fact that one is present as an inversion, and the absence of flanking splice junction sequences make it highly improbable that they are expressed. The extent to which the Gly-X-Y pattern is maintained suggests that some selective pressure may be operating on these vestiges.

Function of intervening sequences. The preponderance of intervening sequences within type I procollagen genes raises a question as to their function. In general, intervening sequences have been postulated to separate gene sequences encoding different conformational or functional domains, thereby allowing their independent evolution (23). At first glance, type I procollagen genes appear to be an exception to this; the bulk of the interruptions occur within the region specifying the large α-chain domain and the junctions between the α-chain, telopeptide, and propeptide domains are each fused into a single CS (54). Still, the intervening sequences might divide the α-chain into functional subdomains. One example of this could be CS8, which encodes amino acids 766 to 801. It specifies both the collagenase cleavage site at the gly-ileu bond (aa 775 to 776) and the fibronectin binding site (aa 766 to 788) (32, 33). Collagen types I, II, and III are all cleaved at the identical site by the same enzyme, but types IV and V are not (7). Moreover, different attachment proteins appear to exist for individual collagen types (32), although the collagen-binding sites for these other proteins have yet to be established. Nonetheless, CS8 could be identified as a subdomain exhibiting significant functional variability between collagen isotypes. Although more subtle examples of such subdomains may be recognized in the future, it does seem unlikely that every CS can be accounted for in this fashion.

A more probably explanation is that the preponderance of intervening sequences stabilizes type I procollagen genes by reducing the frequency of recombination within the repetitious CSs as first suggested for immunoglobulin genes (46). Consistent with this view is the fact that the intervening sequences tend to be less guanine-cytosine rich (52%) than the CSs (65%), and they exhibit little self-homology compared with that observed between CSs.

Mouse B1 sequence within an intervening sequence. One of the motivations for establishing the sequence of intervening sequences, as well as CSs, was to inquire into the possibility that other gene products might be encoded within the procollagen gene. The only published example of such a phenomenon in nonviral genes has been the discovery that an intervening sequence within the cytochrome *b* gene encodes an RNA maturase which is responsible for the splicing of the cytochrome *b* mRNA (35). It is conceivable that this mechanism for establishing an autoregulatory gene may extend to other split genes as well. Given the complexity of the mouse genome, a selective advantage for arrangements which economically utilize the same DNA segment in multiple ways seems unlikely. However, configurations which lead to overlapping genes, genes within genes, or symmetrical transcription of genes may have significant advantages for encoding regulators or gene products which require coexpression or alternating expression. A priori such gene products could be RNA or protein.

Based on hybridization data (Fig. 6) and an analysis of the DNA sequence (Fig. 2), the ninth intervening sequence encodes a member of the mouse B1 family of repetitive sequences (34). The fact that the three previously sequenced members of this family were cloned from the most abundant class of mouse foldback RNA suggests that the present example may also be transcribed in vivo. This view is supported by the presence of sequences characteristic of genes transcribed by RNA polymerase III. For example, the sequence TCCTGAGTTCAAGACC (nucleotides 187 to 198; Fig. 7) is a strong candidate for the 3′ element of a RNA polymerase III split promoter (22, 28). This sequence is a perfect match for the putative Alu-family consensus promoter GAGTTCPuAGACC (16). It is also nearly identical to a sequence TCCTGAGTTCAATTCC present in transcriptionally active, type 2 CHO Alu-equivalent genes (clones 49 and 250), but altered in the transcriptionally inactive type I CHO Alu-equivalent clones examined (26, 27). Furthermore, a sequence identical to this type 2 CHO sequence is found in the Alu-equivalents located within two rat growth hormone genes (2, 41). A very similar sequence, GAGTTCGAGGCC, is present in the mouse B1 consensus sequence and in mouse and hamster 4.5S RNAs (25, 34). Moreover, this sequence motif is highly conserved in the consensus sequence (GGGTTCGANACC) for Ad VaI and II genes and tRNA genes (19, 22, 28), but less so for 5S RNA (6). A kinship also exists between the 5′ promoter element identified in eucaryotic tRNAs and mouse 4.5S RNA and the sequence located between positions 128 and 140 in Fig. 7 (19, 22, 28). Transcription of Alu or Alu-equivalent sequences is known to proceed beyond the 3′ flanking direct repeat and terminate within the single-copy genomic sequences downstream (14, 16, 26). The efficient termination of transcription by RNA polymerase III is observed when a T cluster is surrounded by guanine-cytosine-rich sequences (5). Therefore, it is conceivable that the postulated transcription could terminate at position 298 or 326, or even more likely, at the end of the ninth intervening sequence (nucleotides 332 to 339). Based on this comparative sequence analysis, it does seem likely that this example of a mouse B1 family of genomic repetitive sequences would be transcribed in vivo by RNA polymerase III. However, this eventuality remains to be demonstrated. Genomic repetitive sequences have been postulated to function in numerous ways (9, 12, 30), but a clear demonstration of function is still lacking.

Mobility of mouse B1 family sequences. The genomic repetitive sequence described above is closely related to the mouse B1c sequence (34), except that it is flanked by 8-bp direct repeats and the A-rich region is truncated. Duplications of target DNA also flank procaryotic and eucaryotic transposable elements (10) and retrovirus proviruses which have been postulated to have originated from cellular movable genetic elements (48). Recently, a model for the generation of truncated snRNA pseudogenes and Alu-family members flanked by direct repeats has been proposed (30, 49). According to this model, a self-primed reverse transcript of the RNA beginning within the 3′ A-rich region and extending to the 5′ end of the RNA would be inserted into chromosomal DNA at a staggered break, thereby generating a truncated gene flanked by direct repeats. Our findings of the first mouse B1 family member which is flanked by direct repeats and possesses a truncated A-rich region is consistent with the extension of this model for mouse B1 sequences and suggests that they may be mobile elements.

ACKNOWLEDGMENTS

J.M.M. thanks L. Kunkel for his gift of genomic DNAs, H. Martinez for providing access to the biocomputation laboratory, and J. Gampell for assistance in the preparation of the manuscript.

This work was supported by Public Health Service grant AM 25157A from the National Institute of Arthritis, Diabetes, and Digestive and Kidney Diseases.

LITERATURE CITED

1. **Baralle, F. E., C. C. Shoulders, and N. J. Proudfoot.** 1980. The primary structure of the human α-globin gene. Cell **21:**621–626.
2. **Barta, A., R. I. Richards, J. D. Baxter, and J. Shine.** 1981. Primary structure and evolution of rat growth hormone gene. Proc. Natl. Acad. Sci. U.S.A. **78:**4867–4871.
3. **Bell, G. I., R. L. Pictet, W. J. Rutter, B. Cordell, E. Tischer, and H. Goodman.** 1980. Sequence of the human insulin gene. Nature (London) **284:**26–32.
4. **Bishop, J. M.** 1981. Enemies within: the genesis of retrovirus oncogenes. Cell **23:**5–6.
5. **Bogenhagen, D. F., and D. D. Brown.** 1981. Nucleotide sequences in Xenopus 5S DNA required for transcriptional termination. Cell **24:**261–270.
6. **Bogenhagen, D. F., S. Sakonjv, and D. D. Brown.** 1980. A control region in the center of the 5S RNA gene directs specific initiation of transcription. II. The 3′ border of the region. Cell **19:**27–35.
7. **Bornstein, P., and H. Sage.** 1980. Structurally distinct collagen types. Annu. Rev. Biochem. **49:**957–1003.
8. **Breathnach, R., and P. Chambon.** 1981. Organization and expression of eucaryotic split genes coding for proteins. Annu. Rev. Biochem. **50:**349–383.
9. **Britten, R. J., and E. H. Davidson.** 1969. Gene regulation for higher cells: theory. Science **165:**349–357.
10. **Calos, M. P., and J. H. Miller.** 1980. Transposable elements. Cell **20:**579–595.
11. **Casino, A., M. Cipollaro, A. M. Guerrini, G. Mastrocinque, A. Spena, and V. Scarlato.** 1981. Coding capacity of complementary DNA strands. Nucleic Acids Res. **9:**1499–1518.
12. **Davidson, E. H., and R. J. Britten.** 1979. Regulation of gene expression: possible role of repetitive sequences. Science **204:**1052–1059.
13. **Denhardt, D. T.** 1966. A membrane filter technique for detection of cDNA. Biochem. Biophys. Res. Commun. **23:**641–646.
14. **Duncan, C. H., P. Jagadeeswaran, R. R. C. Wang, and S. M. Weissman.** 1981. Structural analysis of templates and RNA polymerase III transcripts of Alu family sequences interspersed among the human β-like globin genes. Gene **13:**185–196.
15. **Efstradiatis, A., J. W. Posakony, T. Maniatis, R. M. Lawn, C. O'Connell, R. A. Spritz, J. K. DeRiel, B. G. Forget, S. M. Weissman, J. L. Slightom, A. E. Blechl, O. Smithies, F. E. Baralle, C. C. Shoulders, and N. J. Proudfoot.** 1980. The structure and evolution of the human β-globin gene family. Cell **21:**653–668.
16. **Elder, J. T., J. Pan, C. H. Duncan, and S. M. Weissman.** 1981. Transcriptional analysis of interspersed repetitive polymerase III transcription units in human DNA. Nucleic Acids Res. **9:**1171–1189.
17. **Fietzek, P. P., F. W. Rexrodt, K. E. Hopper, and K. Kuhn.** 1973. The covalent structure of collagen 2. The amino-acid sequence of α1-CB7 from calf-skin collagen. Eur. J. Biochem. **38:**396–400.
18. **Fietzek, P. P., F. W. Rexrodt, P. Wendt, M. Stark, and K. Kuhn.** 1972. The covalent structure of collagen amino-acid sequence of peptide α1-CB6-C2. Eur. J. Biochem. **30:**163–168.
19. **Fowlkes, D. M., and T. Shenk.** 1980. Transcriptional control regions of the adeno VA1 RNA gene. Cell **22:**405–413.
20. **Franchini, G., J. Even, C. J. Sherr, and F. Wong-Staal.** 1981. onc sequences (v-fes) of Snyder-Theilen feline sarcoma virus are derived from noncontiguous regions of a cat cellular gene (c-fes). Nature (London) **290:**154–157.
21. **Fuller, F., and H. Boedtker.** 1981. Sequence determination

and analysis of the 3′ region of chicken proα1(I) and proα2(I) collagen mRNAs including the carboxy terminal propeptide sequences. Biochemistry 20:996–1006.

22. **Galli, G., H. Hofstetter, and M. L. Birnstiel.** 1981. Two conserved sequence blocks within eukaryotic tRNA genes are major promoter elements. Nature (London) 294:626–631.
23. **Gilbert, W.** 1979. Introns and exons: playgrounds of evolution, p. 1–12. *In* R. Axel, T. Maniatis, and C. F. Fox (ed.), Eucaryotic gene regulation, ICN-UCLA Symposium, vol. XIV. Academic Press, Inc., New York.
24. **Goff, S. P., E. Gilboa, O. N. Witte, and D. Baltimore.** 1980. Structure of the Abelson murine leukemia virus genome and the homologous cellular gene: studies with cloned viral DNA. Cell 22:777–785.
25. **Harada, F., and N. Kato.** 1980. Nucleotide sequences at 4.5S RNAs associated with poly(A)-containing RNAs of mouse and hamster cells. Nucleic Acids Res. 8:1273–1285.
26. **Haynes, S. R., and W. R. Jelinek.** 1981. Low molecular weight RNAs transcribed *in vitro* by RNA polymerase III from Alu-type dispersed repeats in Chinese hamster DNA are also found *in vivo*. Proc. Natl. Acad. Sci. U.S.A. 78:6130–6134.
27. **Haynes, S. R., T. P. Toomey, L. Leinwand, and W. R. Jelinek.** 1981. The Chinese hamster Alu-equivalent sequence: a conserved, highly repetitious, interspersed deoxyribonucleic acid sequence in mammalians has a structure suggestive of a transposable element. Mol. Cell. Biol. 1:473–583.
28. **Hofstetter, H., A. Kressmann, and M. L. Birnsteil.** 1981. A split promoter for a eucaryotic tRNA gene. Cell 24:573–585.
29. **Hollis, G. F., P. A. Hieter, O. W. McBride, D. Swan, and P. Leder.** 1982. Processed genes: a dispersed human immunoglobulin gene bearing evidence of RNA-type processing. Nature (London) 296:321–325.
30. **Jagadeeswaran, P., B. G. Forget, and S. M. Weissman.** 1981. Short interspersed repetitive DNA elements in eucaryotes: transposable DNA elements generated by reverse transcription of RNA Pol III transcripts? Cell 26:141–142.
31. **Jelinek, W. R., T. P. Toomey, L. Leinwand, C. H. Duncan, P. A. Biro, P. V. Choudary, S. M. Weissman, C. M. Rubin, C. M. Houck, P. L. Deininger, and C. W. Schmid.** 1980. Ubiquitous, interspersed repeated sequences in mammalian genomes. Proc. Natl. Acad. Sci. U.S.A. 77:1398–1402.
32. **Kleinman, H. K., R. J. Klebe, and G. R. Martin.** 1981. Role of collagenous matrices in the adhesion and growth of cells. J. Cell. Biol. 88:473–485.
33. **Kleinman, H. K., E. B. McGoodwin, G. R. Martin, R. J. Klebe, P. P. Fietzek, and D. E. Woolley.** 1978. Localization of the binding site for cell attachment in the α(I) chain of collagen. J. Biol. Chem. 253:5642–5646.
34. **Krayer, A. S., D. A. Karmerov, K. G. Skryabin, A. P. Ryskov, A. A. Bayer, and G. P. Georgiev.** 1980. The nucleotide sequence of the ubiquitous repetitive DNA sequence B1 complementary to the most abundant class of mouse fold-back RNA. Nucleic Acids Res. 8:1201–1215.
35. **Lazowska, J., C. Jacq, and P. P. Slonimski.** 1980. Sequence of introns and flanking exons in wild type and box 3 mutants of cytochrome b reveals an interlaced splicing protein coded by an intron. Cell 22:333–348.
36. **Leder, A., D. Swan, F. Ruddle, P. D'Eustachio, and P. Leder.** 1981. Dispersion of α-like globin genes of the mouse to three different chromosomes. Nature (London) 293:196–200.
37. **Lerner, M. R., J. A. Boyle, M. S. Mount, L. S. Wolin, and J. A. Steitz.** 1980. Are snRNPs involved in splicing? Nature (London) 283:220–224.
38. **Maxam, A., and W. Gilbert.** 1980. Sequencing end-labeled DNA base-specific chemical cleavages. Methods Enzymol. 65:499–560.
39. **Monson, J. M., and B. J. McCarthy.** 1981. Identification of a Balb/c mouse proα1(I) procollagen gene: evidence for insertions or deletions in gene coding sequences. DNA 1:59–69.
40. **Nishiokia, Y., A. Leder, and P. Leder.** 1980. Unusual α-globin-like gene that has cleanly lost both globin intervening sequences. Proc. Natl. Acad. Sci. U.S.A. 77:2806–2809.
41. **Page, G. S., S. Smith, and H. M. Goodman.** 1981. DNA sequence of the rat growth hormone gene: location of the 5′ terminus of the growth hormone mRNA and identification of an internal transposon-like element. Nucleic Acids Res. 9:2087–2104.
42. **Proudfoot, N. J., and G. G. Brownlee.** 1976. 3′ non-coding region sequences in eukaryotic messenger RNA. Nature (London) 263:211–214.
43. **Rogers, J., and R. Wall.** 1980. A mechanism for RNA splicing. Proc. Natl. Acad. Sci. U.S.A. 77:1877–1879.
44. **Sanchez, F., J. E. Natzle, D. W. Cleveland, M. W. Kirschner, and B. J. McCarthy.** 1980. A dispersed multigene family encoding tubulin in Drosophila melanogaster. Cell 22:845–854.
45. **Schafer, M., C. Boyd, P. Tolstoshev, and R. Crystal.** 1980. Structural organization of a 17-kb segment of the α2 collagen gene: evaluation of R loop mapping. Nucleic Acids Res. 8:2241–2253.
46. **Seidman, J. G., A. Leder, M. Nau, B. Norman, and P. Leder.** 1978. Antibody diversity: the structure of cloned immunoglobulin genes suggests a mechanism for generating new sequences. Science 202:11–17.
47. **Southern, E. M.** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. 98:503–517.
48. **Temin, H. M.** 1980. Origin of retroviruses from cellular movable genetic elements. Cell 21:599–600.
49. **Van Arsdell, S. W., R. A. Denison, L. B. Bernstein, A. M. Weiner, T. Manser, and R. F. Gesteland.** 1981. Direct repeats flank three small nuclear RNA pseudogenes in the human genome. Cell 26:11–17.
50. **Vanin, E. F., G. I. Goldberg, P. W. Tucker, and O. Smithies.** 1980. A mouse α-globin-related pseudogene lacking intervening sequences. Nature (London) 286:222–226.
51. **Wahl, G. M., M. Stern, and G. R. Stark.** 1979. Efficient transfer of large DNA fragments from agarose gels to diazobenzylomethyl paper and rapid hybridization by using dextran sulfate. Proc. Natl. Acad. Sci. U.S.A. 76:3683–3687.
52. **Wendt, R., K. von der Mark, F. Rexrodt, and K. Kuhn.** 1972. The covalent structure of collagen. The amino-acid sequence of 112-residues amino-terminal part of peptide α1-CB6 from calf-skin collagen. Eur. J. Biochem. 30:169–183.
53. **Wilde, C. D., C. E. Crowther, T. P. Cripe, M. Gwo-Shu Lee, and N. J. Cowan.** 1982. Evidence that a human β-tubulin pseudogene is derived from its corresponding mRNA. Nature (London) 297:83–84.
54. **Wozney, J., D. Hanahan, V. Tate, H. Boedtker, and P. Doty.** 1981. Structure of the proα2(I) collagen gene. Nature (London) 294:129–135.
55. **Yamada, Y., V. E. Avvedimento, M. Mudryi, H. Ohkubo, G. Vogel, M. Irani, I. Pastan, and B. de Crombrugghe.** 1980. The collagen gene: evidence for its evolutionary assembly by amplification of a DNA segment containing an exon of 54 bp. Cell 22:887–892.