Volume 14 Number 11 1986 Nucleic Acids Research

Discrimination among multiple AATAAA sequences correlates with interspecies conservation of select 3' untranslated nucleotides

Jeanne C.Myers[1,2,3]*, Jane M.Brinker[1], Nicholas A.Kefalides[1,2], Joel Rosenbloom[4], Sho-Ya Wang[5] and Lorraine J.Gudas[5]

[1]Connective Tissue Research Institute, Departments of [2]Medicine and [3]Human Genetics, [4]Center for Oral Health Research, University of Pennsylvania, Philadelphia, PA 19104 and [5]Dana Farber Cancer Institute, Department of Pharmacology, Harvard Medical School, Boston, MA 02115, USA

Received 4 March 1986; Revised and Accepted 1 May 1986

ABSTRACT

The DNA sequence corresponding to the 1.3 kb 3' untranslated region of the 6.5 kb human procollagen α1(IV) mRNA was determined and compared with the mouse sequence obtained from 3' cDNA and genomic clones overlapping the reported 5' half (Oberbaumer et al., 1985, Eur. J. Biochem. 147:217). Although four AAUAAA hexanucleotides are found in the human and seven in the mouse RNAs, Northern blot hybridization showed almost exclusive utilization of the most 3' sequence, in contrast to the pattern seen when using α1(I), α2(I), α1(III) and α2(V) procollagen probes. Moreover, the ninety nucleotides 5' to the poly A tail in the major α1(IV) mRNAs exhibit a much greater degree of interspecies homology than those encompassing the other three shared AAUAAA recognition signals. Further examination of this highly conserved area revealed the presence of two "consensus sequences" found in the 3' noncoding region of a number of RNA polymerase II transcribed genes (Mattaj and Zeller, 1983, Embo J. 2:1883) and, unexpectedly, some similarity with the nucleotides 5' to the poly A attachment signals in other procollagen mRNAs.

INTRODUCTION

In the last several years, DNA sequencing of cDNA and genomic clones has allowed determination of a number of 3' untranslated regions (UTRs). These areas are receiving increasing attention in attempts to understand why a strong selective pressure is exerted upon the noncoding nucleotides (1). Very often striking interspecies homology is observed, especially in the vicinity of those sequences involved in polyadenylation and correct 3' end formation (2-16). Recently, Yaffe et al. presented an extensive description of the cardiac, skeletal and β-actin UTRs and, in addition, discussed the interspecies relationship of the 3' noncoding region of 10 other mammalian and avian genes (17 and refs. therein). These authors noted that while the length of the UTRs is quite similar in isotypic genes, both the lengths and the distribution of the conserved sequences vary greatly in nonhomologous genes. In this report we have concentrated on the human, mouse and avian procollagen UTRs which have not received the critical examination necessary

to assess the extent of nucleotide conservation.

Collagens are the primary structural components of the extracellular matrix and consist of 10 formerly classified homo- or heterotrimeric types. The closely related polypeptide chains are encoded by at least 20 genes exhibiting coordinate and differential developmental and tissue specific expression (for recent reviews see refs. 18-21). Of the major collagens, type I ($\alpha1(I)_2$ $\alpha2(I)$, $\alpha1(I)_3$) is found in bone and is the predominant collagen in most connective tissue except for cartilage and basement membranes. Type III ($\alpha1(III)_3$) and type V ($\alpha1(V)_3$, $\alpha1(V)_2$ $\alpha2(V)$, $\alpha1(V)$ $\alpha2(V)$ $\alpha3(V)$) have a distribution similar to type I but are usually present in lesser amounts (I> III> V). Type II ($\alpha1(II)_3$) is almost exclusively located in cartilage and only type IV ($\alpha1(IV)_2$ $\alpha2(IV)$, $\alpha1(IV)_3$) is present in basement membranes.

The 3' UTRs of this gene family have generated particular interest due to the presence of multiple poly A attachment signals generally resulting in the transcription of two or three different sized RNAs with an as yet unknown significance (22-33). Some AAUAAA or related hexanucleotides are utilized to a similar extent, whereas others are infrequently recognized. Since these sites are often located 0.5-1 kb apart, the UTRs of the larger RNAs are exceptionally long and their length surpasses almost all 3' noncoding regions examined to date (17). Here we have focused primarily on the $\alpha1$ type IV procollagen transcripts, but we have also obtained relevant information by comparing the types I, II, III and V human and/or avian 3' UTRs. Specific "subsegments" (34) of these regions display homology greatly exceeding the 50% value expected between human and mouse and the <30% value expected between human and avian in the absence of evolutionary constraints (17). Furthermore, most of these sequences are situated 5' and 3' to poly A attachment signals and show considerable inter- and/or intraspecies homology.

MATERIALS AND METHODS

Mouse and Human Procollagen DNA Clones

We previously reported the isolation and characterization of the following cDNA clones used in these studies: mouse $\alpha1(IV)$ pcI5 (35), human $\alpha1(IV)$ KK4, RR6 and NB3 (31), human $\alpha1(I)$ $\alpha12$ (27), human $\alpha2(I)$ Hf32 (25), human $\alpha1(III)$ E6 (27) and human $\alpha2(V)$ NH20 (33). The 2 kb mouse $\alpha1(IV)$ EcoRI genomic subclone I17 extends 5' from the EcoRI site in the cDNA clone pcI5 (Fig. 1) (Wang and Gudas, unpublished results).

DNA Sequencing Strategy

HindIII/BamHI, HindIII/PstI, EcoRI/HindIII, HindIII and PstI restriction fragments of the four α1(IV) cDNA clones shown in Fig. 1 were ligated to appropriately cleaved M13mp18,19 vectors. Sanger dideoxy-sequencing reactions were performed essentially as described by Messing (36). The universal primer of 17 nucleotides (Collaborative Research) was used for the sequencing in addition to four 15 or 17 nucleotide primers (noted in Fig. 1 by triangles) which were synthesized using sequences from the clones KK4, NB3 and pcI5 (University of Pennsylvania, Department of Chemistry). Mouse α1(IV) sequences were also determined from pcI5 (3'→ 5' of both PstI fragments) and from I17 (5'→3' of a PvuII/PstI fragment) by Maxam and Gilbert procedures (37). One hundred nucleotides obtained from I17 provided the 65 nucleotides linking the 3' part of the UTR to the 5' half reported by Oberbaumer et al. (29). DNA sequences of the 3' EcoRI/PstI fragment of the human α1(III) cDNA clone E6 (27) were obtained using the universal primer and M13mp18,19 vectors.

Alignment and Sources of DNA Sequences

The 3' untranslated regions of α1(IV), α2(I), α1(II) and α1(III) were aligned manually to show maximum homology with the least number of insertions and deletions. Unless otherwise stated, homology (%) was calculated as the ratio of identical bases or amino acids to the total number in which each gap was considered a mismatch. Sequences, included here but not determined in this report, were obtained from the following manuscripts: Myers et al. (33), Oberbaumer et al. (29), Yamada et al. (38), Stoker et al. (39), Ninomija et al. (40), Sandell et al. (41), Myers et al. (25), Fuller and Boedtker (42) and Aho et al. (23).

Poly A^+ RNA Isolation and Northern Blot Hybridization

Poly A+ RNA was isolated from normal mouse (43) and human fibroblast cell lines (GM3348 from the Human Genetic Mutant Cell Repository) by lysis and incubation of the cells in an SDS/proteinase K buffer followed by passage through an oligo dT-cellulose column (44). The RNA (0.2-1.5 ug per lane) was electrophoresed in a 1% agarose gel containing 2.5M formaldehyde for 22 hours at 30V. Transfer to nitrocellulose filters was carried out at 4^0 for 16-18 hrs in 10xSSC (1xSSC = 0.15M NaCl, 0.015M NaCitrate, pH 6.8). Hybridization to the ^{32}P nick-translated probes ($5\text{-}8x10^8$cpm/ug) was in a solution containing 50% formamide, 5xSSC for 18-24 hours at $38\text{-}40^0$ and final washing of the filters was at 65^0, 0.2xSSC or 60^0, 0.5xSSC for the human and mouse RNAs, respectively.

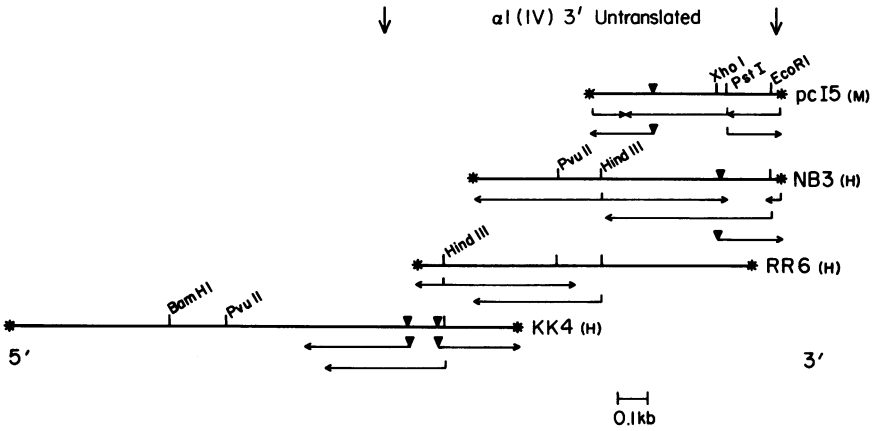

FIGURE 1. Mouse and Human α1(IV) cDNA Clones and Sequencing Strategy. Asterisks indicate insertion into the PstI site of the pBR322 vector (31,35). Lines below the restriction maps show the sequencing strategy described in Methods. Triangles denote the location of specific α1(IV) oligonucleotide primers also used for the DNA sequencing.

RESULTS

Hybridization of Procollagen Clones to Human and Mouse Fibroblast RNAs

In our previous studies aimed at determining the structure of the human α1(IV) COOH-terminal noncollagenous peptide, we sequenced most of a 1.7 kb cDNA clone KK4 which contains the 5' part of the 3' untranslated region (31). Hybridization of KK4 (Fig. 1) to normal human fibroblast poly A^+ RNA identified a major 6.5 kb transcript and two 5.9 and 5.3 kb minor species (Fig. 2). The 6.5 and 5.9 kb α1(IV) species were also observed using the clone RR6, but only the 6.5 kb transcript was seen with NB3 (Figs. 1 and 2). In parallel experiments using mouse RNA and probing with homologous α1(IV) DNA clones I17 (a 2 kb genomic subclone extending 5' from the EcoRI site in pcI5) and pcI5 (Fig. 1), the same basic profile was found except for the absence of the intermediate-sized RNA (Fig. 2). Polymorphic RNAs are characteristic of all human procollagen genes examined so far with the possible exclusion of α1(II) (45,46). The differences in length of the α1(I), α2(I), α1(III) and α2(V) transcripts are primarily, if not entirely, attributable to utilization of various polyadenylation sites in the 3' UTRs (25, 28, 30, 33).

To investigate the evolutionary persistence of other multiple procollagen mRNAs, we hybridized human α1(I), α2(I), α1(III) and α2(V) cDNA clones to filter-bound mouse and human fibroblast poly A^+ RNAs (Fig.

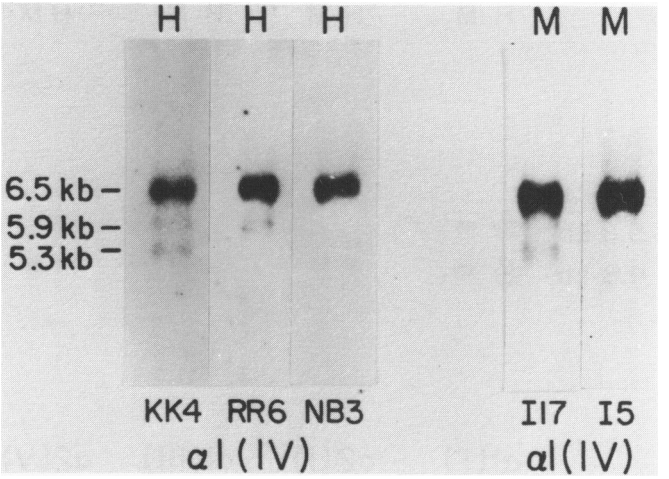

FIGURE 2. Hybridization of Human (H) and Mouse (M) Fibroblast RNA to α1(IV) Procollagen Clones. In the left lanes are the profiles obtained from hybridization of human RNA to the 5' clone KK4 and the 3' clones RR6 and NB3 (Fig. 1). The 5.9 kb transcript detected with RR6 was also seen using NB3 when the filters were washed at 56° (not shown). In the right two lanes of mouse RNA, the probes were I17, a 2 kb mouse α1(IV) genomic subclone encoding all of the 3' UTR except for the 3' 25 bases, and pcI5, the 3' mouse α1(IV) cDNA clone (Fig. 1). RNA sizes were determined from 35S poliovirus RNA and 28S and 18S ribosomal RNA markers (31).

3). The ratios of the type I (α1 and α2) mouse transcripts are very similar to the human except for the absence of the minor α2(I) RNA. In contrast, two prominent α1(III) species of 5.4 and 4.9 kb are present in the human population, whereas only the 4.9 kb RNA was detected in the mouse. A similar, but more pronounced, difference is found in the α2(V) profile. Although the larger of the two human transcripts predominates, only the smaller 5.0 kb mouse RNA is retained.

Comparison of the Complete Human and Mouse α1(IV) 3' Untranslated Regions

Several observations prompted investigation of the mouse and human α1(IV) 3' UTRs at the nucleotide level. Restriction analysis revealed this area to be unusually long and the Northern blot profiles implied the presence of multiple, infrequently used polyadenylation sites, as compared to those in other procollagen genes (Figs. 2 and 3). Furthermore, the human clone NB3 was identified by strong hybridization to the mouse pcI5 insert (31), and both clones were subsequently found to consist entirely of noncoding sequences. Therefore, we asked the following questions: Was the homology in the 3' α1(IV) UTRs localized or interdispersed? Were the major

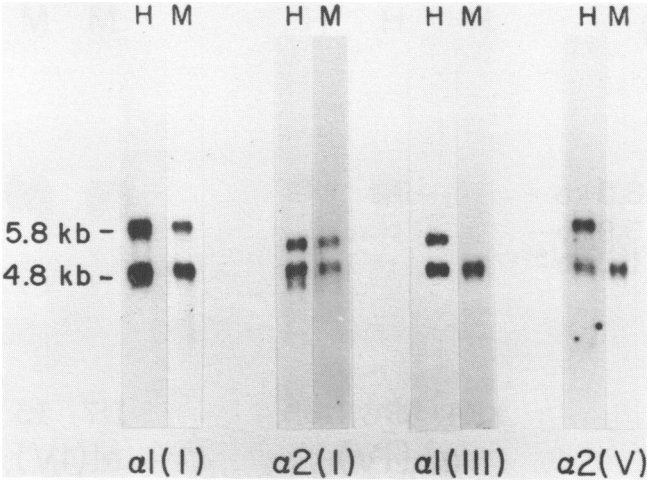

FIGURE 3. Hybridization of Human (H) and Mouse (M) Fibroblast RNA to Types I, III and V Human Procollagen cDNA Clones. Sizes of the α1(I) RNAs (5.8 and 4.8 kb) were determined by DNA sequencing of the human α1(I) procollagen gene (28). Using these values, the human α2(I) transcripts are estimated to be 5.2, 4.6 and 4.4 kb (28), and the human α1(III) and α2(V) RNAs, 5.4 and 4.9 kb, and 6.3 and 5.0 kb, respectively (33). The 2.4 kb insert of the α1(I) clone α12 codes for about 800 residues of the collagenous region and the 2.1 kb insert of the α2(I) clone Hf32 codes for almost half of the collagenous region and 80% of the COOH-propeptide (47,48). The 2.4 kb insert of the α1(III) clone E6 codes for 411 collagenous residues, the COOH-propeptide and the entire 3' UTR of the 4.9 kb RNA (27 and Fig. 7). The 1.35 kb insert of the α2(V) clone NH20 codes for 165 collagenous residues, the COOH-propeptide and 45 bases of the 3' UTR (33).

poly A attachment signals in the same positions and, if so, did the adjacent nucleotides display any features distinguishing them from those surrounding the polyadenylation sequences not generally utilized?

Initial sequencing of the 3' human (NB3) and mouse (pcI5) α1(IV) cDNA clones disclosed poly A tails of 25 and 17 bases, respectively, which were preceded by an AATAAA hexanucleotide. From the three overlapping human α1(IV) clones 1300 3' UTR nucleotides were obtained, while 619 nucleotides were derived from the mouse 3' clone. During the course of these experiments, the human α1(IV) UTR sequences were also derived by Pihlajaniemi et al. (32) and 594 nucleotides containing the 5' half of the mouse 3' UTR were reported by Oberbaumer et al. (29). Since the mouse 5' sequences did not overlap with our 3' sequences, 100 nucleotides beginining at the PvuII site, common to both the mouse and human clones (Fig. 1), were determined from a genomic subclone. These provided the linking 65 bases

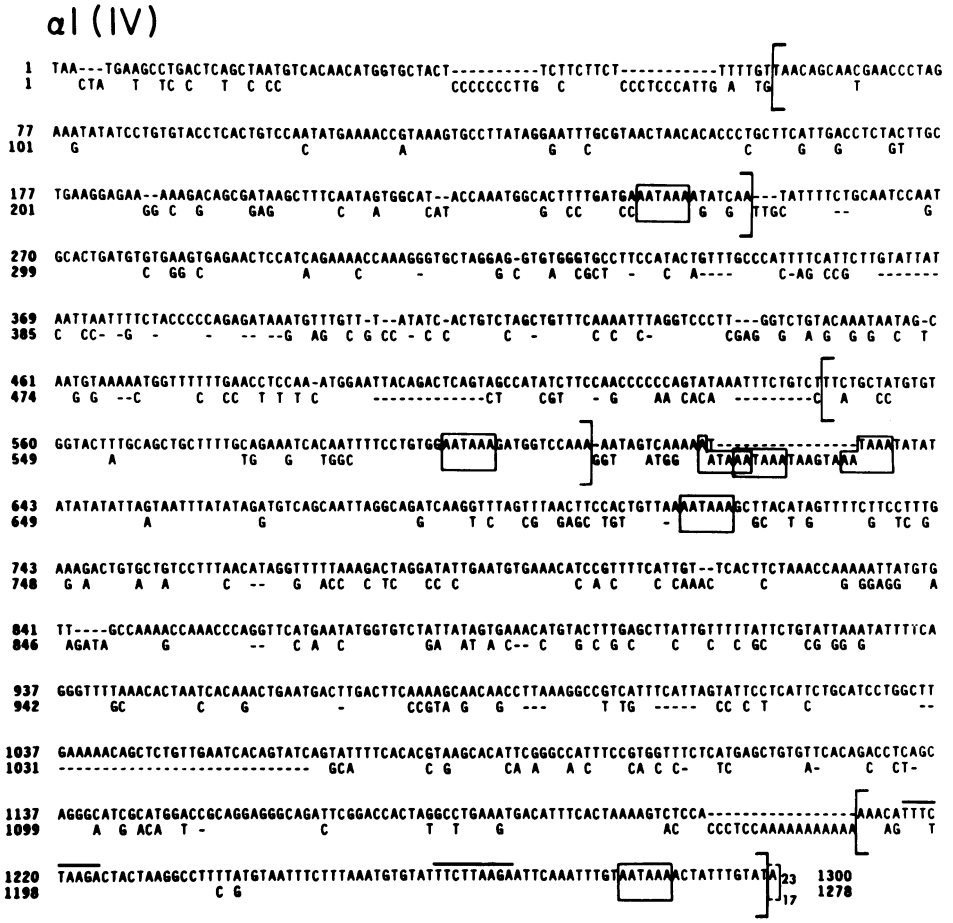

FIGURE 4. Complete DNA Sequence of the Human and Mouse α1(IV) 3' UTRs. Human sequences are shown on the top line and only differences in the mouse UTR are on the bottom line. Dashes indicate absence of a corresponding nucleotide. All AATAAA sequences are within boxes and the three most highly conserved regions are within brackets. The two 9 nucleotide repeats which resemble the Mattaj and Zeller "consensus sequence" (50) (discussed in the text) are designated by a bar. The length of the poly A tails in the cDNA clones NB3 and pcI5 (Fig. 1) is given in subscripts.

and, therefore, the complete mouse α1(IV) 3' UTR of 1278 nucleotides.

Surprisingly, seven AATAAA sequences are found in the mouse UTR and four of these are represented in the human DNA (Fig. 4). The additional three murine sites (two are overlapping) are located within tandem A/T rich repeats of which only one remains in the human gene. Overall, the alignment

TABLE I: COMPARISON OF HUMAN AND MOUSE α1(IV) PROCOLLAGEN SEQUENCES

REGION	NUCLEOTIDES	DIVERGENCE
3' Gly-X-Y (GG of Gly incl.)	495	.178
3' Gly-X-Y (GG of Gly excl.)	389	.226
5' Half C-propeptide	342	.155
3' Half C-propeptide	345	.098
3' UTR (5' QUARTER)	300 (M-1) 323 (M-2)	.160 .207
3' UTR (5' MIDDLE)	300 (M-1) 323 (M-2)	.253 .374
3' UTR (3' MIDDLE)	300 (M-1) 323 (M-2)	.273 .291
3' UTR (3' QUARTER)	300 (M-1) 323 (M-2)	.173 .207
AATAAA Sequence #1 (5'→3')	90 (M-1) 90 (M-2)	.200 .277
AATAAA Sequence #2 (5'→3')	90 (M-1) 90 (M-2)	.210 .277
AATAAA Sequence #3 (5'→3')	90 (M-1) 90 (M-2)	.200 .211
AATAAA Sequence #4 (5'→3')	90 (M-1) 90 (M-2)	.055 .055

Nucleotides coding for the human and mouse α1(IV) Gly-X-Y and COOH-propeptide regions have been reported (29,31,32). The 3' UTR sequences are shown in Fig. 4. Sequences #1-4 (90 bases) include the AATAAA, 73 positions 5' and 11 positions 3'. Calculations for divergence (%) were based on the methods of Miyata et al. (1). In M-1 all gaps were excluded from the comparison and in M-2 a gap was counted a "mismatch". Those sites with more than 10 consecutive gaps were excluded from both analyses. M-1 equal M-2 values for the nucleotides encoding Gly-X-Y and COOH-propeptide regions.

of the uncommonly large 3' UTRs shows remarkable interspecies conservation with the exception of several groups of insertions and deletions. However, since more variability was noted overall in the central portion, we calculated the divergence of each quarter separately according to the methods of Miyata et al. (1). In Method 1 (M-1), only positions represented in both genes were counted, while in Method 2 (M-2), all mismatches and gaps, except for deletions greater than 10 consecutive bases, were included. As shown in Table 1, the divergence of the 5' and 3' quadrants is

clearly lower than that of the two middle quadrants and, moreover, is the same or less than the divergence of the nucleotides coding for the Gly-X-Y region and 5' half of the COOH-terminal propeptide. When we focused on the four AATAAA hexanucleotides common to both UTRs, it was apparent that the distal and primarily utilized poly A attachment signal is embedded in the most conserved area (Sequence #4 in Table 1). Strikingly, 5' to this AATAAA there are 73 nucleotides in which only 5 substitutions have occurred, while 3' there are 11 identical nucleotides preceding the poly A tail. These 90 bases were then compared to the corresponding ones around the three additional AATAAA signals (Sequences #1, #2 and #3). The M-1 and M-2 values are equal only in Sequence #4 and, more significantly, are four to five times lower than the values for Sequences #1, #2 and #3.

Analysis of Human and Avian Types I, II and III Procollagen 3' Untranslated Regions

To determine if other procollagen 3' UTRs shared preferential nucleotide conservation in close proximity to polyadenylation sites, we examined the sequence of the human and avian types I, II and III noncoding regions. In the $\alpha 2$(I) 3' UTRs, multiple AATAAA hexanucleotides are found at similar positions, i.e., +278, +292, +795 and +816 versus +274, +279, +728 and +751 for human and avian, respectively, (+1 designates the first nucleotide of the termination codon). Both sets, and probably the similar ones occurring in the mouse 3' UTR, appear from nuclease S_1 experiments and Northern blot profiles to be almost equally utilized (23, 25 and Fig. 3). We aligned the 3' noncoding regions of human and avian $\alpha 2$(I) to determine the extent of homology previously noted when only the 5' several hundred nucleotides prior to the AATAAA's were compared (47). In three discrete areas exhibiting 84%, 88% and 88% homology (indicated by brackets in Fig. 5), minimal deletions have taken place. The first area occurs in the 5' part of the UTR, the second is located 3' to the initial AATAAA set and the third contains two closely spaced 3' AATAAA hexanucleotides common to both the human and avian $\alpha 2$(I) genes. Similarity deteriorates rapidly following the site of poly A attachment for the larger $\alpha 2$(I) RNAs (23,25).

An apparently different situation exists with $\alpha 1$(I). Human $\alpha 1$(I) AATAAA sequences are present at +248, +252 and +1253 and these designate poly A addition in the 4.8 and 5.8 kb RNAs (28 and Fig. 3). The 3' UTR of the 4.8 kb human RNA (28,48) does not show any recognizable homology with the beginning of the chick UTR which consists mostly of large A tracts and includes an AATAAA of unknown significance at +92 (42). However, this

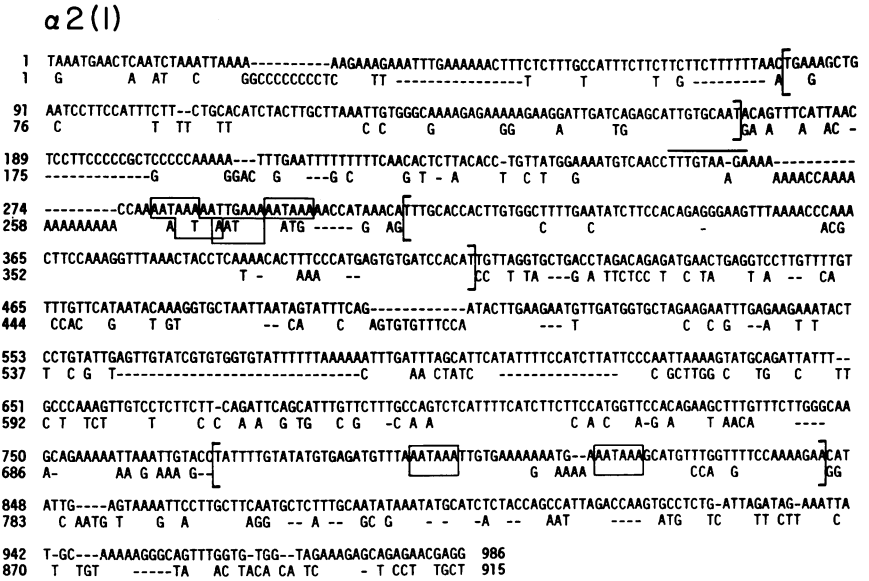

FIGURE 5. DNA Sequence of the Human and Avian α2(I) 3' UTRs. All human α2(I) 3' UTR nucleotides (top line) but only those which differ in the avian (bottom line) are shown (23,25,42). Dashes indicate lack of a corresponding nucleotide. The AATAAA sequences are within boxes and the three highly conserved regions (discussed in the text) are enclosed by brackets. A bar is drawn above the nucleotides resembling the "consensus sequence" reported by Mattaj and Zeller (50). The polyadenylation sites for the larger α2(I) RNAs are 25-30 bases after the 3' AATAAA sequences (23,25).

observation may be premature since only after the initial 81 base insertion in the avian α1(II) UTR can homology with the human α1(II) UTR be detected (Fig. 6). The two major avian α1(I) RNAs of 4.7 and 4.9 kb are electrophoretically similar to the 4.8 kb human and mouse species but the minor 6.7 kb avian α1(I) RNA differs greatly both in size and representation from the abundant human and mouse 5.8 kb transcripts (28,49 and Fig. 3).

Stoker et al. (39) recently reported the sequence of the human α1(II) 3' UTR and pointed out two specific areas, inclusive of AATAAA or related hexanucleotides, highly homologous to the avian 3' UTR determined by Ninomija et al. (40) and Sandell et al. (41) from cDNA and genomic clones. We compared the entire 3' UTRs to ascertain how limited the conservation was and how the two areas compared with analogous regions in α2(I) and α1(IV). The 5' halves of the α1(II) UTRs are characterized mainly by large deletions and insertions interdispersed among very short regions of

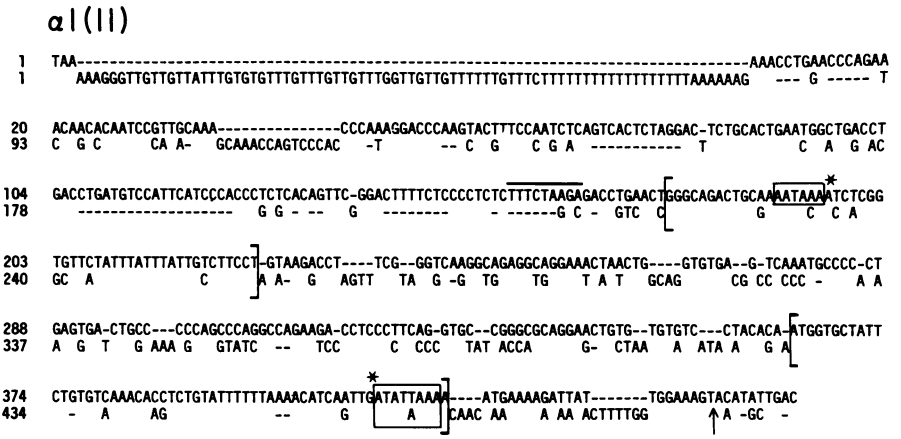

FIGURE 6. DNA Sequence of the Human and Avian α1(II) 3' UTRs. Human sequences are on the top line and differences in the avian are below (39-41). Location of the poly A tail in the avian cDNA clone is indicated by an arrow. Brownlee-Proudfoot hexanucleotides or variations (*) are within boxes. The two conserved areas are designated by brackets and the sequence found both in human α1(II) and α1(IV) is shown by a bar.

similarity. In the 3' halves are the two prominent areas of 51 and 59 nucleotides which display 84% and 86% homology, respectively, (indicated by brackets in Fig. 6). In the first conserved region, these sequences begin 13 bases 5' to an AATAAA in the human gene and extend mostly 3', while in the second region, the homology pertains exclusively to the sequences preceding a poly A attachment signal utilized in the avian system (40). The polyadenylation site(s) for the human α1(II) RNA(s) have not yet been identified.

We derived the 3' UTR of the 4.9 kb human α1(III) RNA from the cDNA

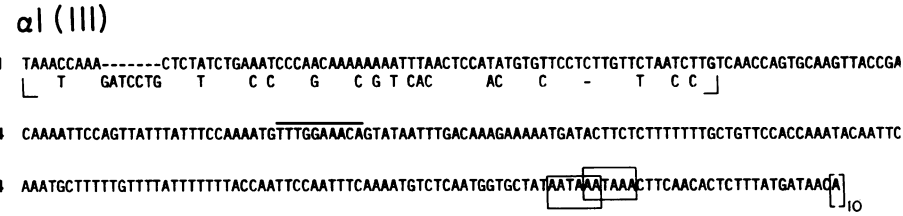

FIGURE 7. DNA Sequence of the 3' UTR of the Smaller Human α1(III) RNA. The human sequence (top line) was derived from the α1(III) cDNA clone E6 (Methods). The 3' 78 nucleotides of the avian UTR (38) are delineated by corner symbols (bottom line). Specific sequences discussed in the text are shown by a bar and overlapping AATAAA sites are enclosed.

clone E6 (27) that terminates in a poly A tail 22 bases following overlapping AATAAA sequences at +254-263 (Fig. 7). Although only the 5' 78 nucleotides of the avian α1(III) 3' UTR are currently known (38), a high degree of conservation can already be detected. From the data of Chu et al. (30), the larger human α1(III) UTR extends to +870 and contains an unusual variation of the canonical sequence. The major avian α1(III) RNA of 5.5 kb (49) is analogous to the 5.4 kb human species in contrast to the mouse α1(III) RNA which corresponds to the 4.9 kb human transcript (Fig. 3).

Inter- and Intraspecies Correlation of Nucleotides Preceding Poly A Attachment Signals

Mattaj and Zeller identified a homologous region 3' to the coding sequence of U1 and U2 sn RNA genes that is partially shared by a number of RNA polymerase II transcribed genes (50). Although this unit is usually located 3' to AATAAA hexanucleotides, in the human leukocyte interferon α-1 gene it is positioned 5' and in the Drosphilia hsp 70-1 gene, the sequence is found both 5' and 3'. Within α1(IV) Sequence #4 are two direct repeats (Figs. 4 and 8), TTTC/TTAAGA and TTCTTAAGA that closely resemble this "consensus sequence". Surprisingly, the identical human α1(IV) sequence, TTTCTAAGA, is also seen just 5' to the AATAAA in human α1(II) (Fig. 6). The fact that neither sequence is conserved in avian α1(II) may be significant. Furthermore, in the human and avian α2(I) UTRs, the nucleotides TTTGTAAGA and TTTGTAAAGA, respectively, are in close proximity 5' to the first set of AATAAAs (Fig. 5) and a variation GGTTTAAGA occurs in the human α2(V) UTR 44-52 bases preceding two overlapping AATAAA sites at +286 and +290 (33).

Although sequences in human α1(III) and α1(I) are less similar, they do show an unexpected relationship to each other. The α1(III) sequence TTTGGAAACA found 131 bases 5' to the AATAAA (Fig. 7) is paralleled in α1(I) 121 bases 5' to the AATAAA as TTTGGAAAATA (28,48). A second set of sequences, CCACCAAA in α1(III) and CCAACCGAA in α1(I), are located 46 and 47 nucleotides, respectively, 3' to the first sequence. The homology for the 5' sequence is 82% versus 78% for the second set. The sizes of the UTRs in the smaller RNAs are approximately equal since the AATAAAs are at +254 and +248 in α1(III) and α1(I) (Fig. 7 and ref. 28).

Because of these observations and the localized interspecies homology, we compared the α1(IV) Sequence #4 with the regions 5' to the utilized poly A attachment signals in the types I, II, III and V procollagen genes. With the exception of α1(I), which is therefore not included, the nucleotide

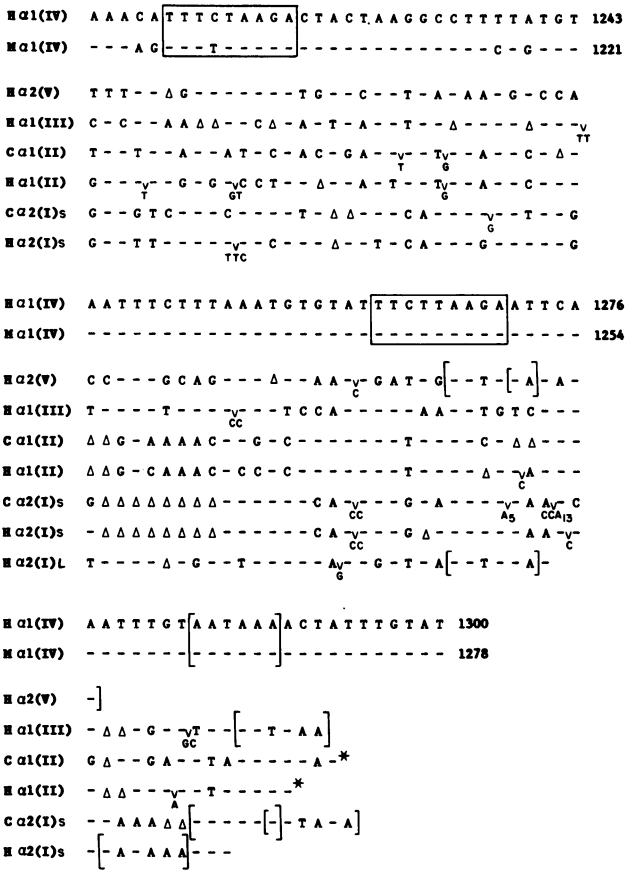

FIGURE 8. Procollagen Sequences 5' to Poly A Attachment Signals. The 3' 90 nucleotides of the human (H) and mouse (M) 3' UTRs (Fig. 4) are shown on the top two lines. Mattaj and Zeller "consensus sequences" (50) in α1(IV) are within boxes. Location of the human α1(III), human and chick (C) α1(II) and human and chick α2(I) sequences within the 3' UTRs (Figs. 5-7) are shown by the position of the AATAAA (brackets) or variations (*). Dashes indicate the same base as found in α1(IV). Open triangles denote lack of a corresponding nucleotide. The letters S and L refer to nucleotides preceding poly A attachment signals for the smaller and larger α2(I) RNAs (Fig. 5). The 31 bases shown in the human α2(I) large (L) UTR are identical to the chick. The human α2(V) AATAAA sequences for the smaller RNA occur 286 bases 3' to the end of translation (33).

adjustments shown in Fig. 8 could generally be equated with those made in order to align parts of the UTRs of the same gene in different species. In this alignment, the α2(V) "consensus sequence" discussed above corresponds to the 5' α1(IV) repeat, and the α2(I) "consensus sequence" corresponds to the 3' α1(IV) repeat.

DISCUSSION

The Northern blot profiles and DNA sequences of the 3' UTRs presented here show that the multiple human and mouse α1(IV) AAUAAA hexanucleotides 5' to the most distal and conserved site (Sequence #4) are rarely, if at all, utilized in the fibroblast cell lines. The minor human α1(IV) RNAs appear to result from poly A attachment designated by AAUAAA Sequences #1 and #2; weak hybridization of the 3' clones to the 5.9 kb RNA implies that Sequence #3 is not functional (Fig. 2). This same representation of the three α1(IV) transcripts is found in poly A+ RNA isolated from cultured human umbilical endothelial cells (unpublished data). In the mouse fibroblast RNA, only the 5' and 3' α I(IV) hexanucleotides seem to be utilized. The five middle AAUAAA sequences are bypassed even though three of these occur within a span of 21 nucleotides (Fig. 4). Major and minor mouse α1(IV) RNAs of 6.7 and 5.4 kb, respectively, have also been detected in a parietal yolk sac cell line by Oberbaumer et al. (29).

A number of genes other than the collagens contain more than one poly A attachment signal in the 3' UTR resulting in multiple transcripts from differential utilization without additional internal 3' RNA splicing variations. These genes include the mouse (51,52) and human (53) dihydrofolate reductase, the mouse α-amylase (54), the chicken ovomucoid (55), the chicken vimentin (56), the eel calmodulin (57), the rat β_{2U}-globulin (58), the mouse β-$_2$ microglobulin (59) and the human β-tubulin (60). Tissue specific differences in the distribution of the vimentin and calmodulin RNAs have been described (61,57).

Evidence for a tissue specific relationship in the type I (α1 and α2) avian and human procollagen RNAs is suggested by several reports. Merlino et al. (24) detect a shift towards the lower molecular weight avian α2(I) RNA accompanied by an increase in gene transcription during embryonic development. Gerstenfeld et al. (49) state that they have seen greater quantities of both the smaller avian α1(I) and α2(I) RNAs in fibroblasts and calvaria than in myoblasts. Barsh and co-workers (62) have correlated the presence of DNase I hypersensitive sites in the 3' end of the human α1(I) procollagen gene with preferential representation of the α1(I) 4.8 kb transcript in placenta as compared to skin fibroblasts. Future studies, facilitated by the recent isolation of clones encoding other procollagen chains, should yield more comprehensive results.

The factors involved in polyadenylation and in accurate and efficient 3' RNA processing have been addressed in a number of elegant experiments.

Deletions and point mutations strongly implicate sequences in addition to the Brownlee-Proudfoot hexanucleotide or variations (3-16). DNA sequencing has revealed the presence of poly A attachment signals in regions distant from the 3' termini of mRNAs. Furthermore, mutations in the AAUAAA sequence greatly affect processing in some systems, while in others, the same variations occur naturally with efficient utilization (10). Pertinent to the procollagen genes, the poly A attachment signals for the avian α1(II) (40) and the larger human α1(III) RNAs (30) are so unusual that within the groups AUAUAAAA and AUUAUAAUA, respectively, the precise sequences are difficult to recognize. In the UTRs analyzed here, most of the highly conserved "subsegments" (34) involve sequences inclusive of or adjacent to utilized poly A attachment signals.

To fully perceive the extent of conservation in the human and mouse α1(IV) UTRs, the divergence of the coding nucleotides is shown in Table I. Generally, the first two nucleotides of each codon are subject to intense selective pressure to maintain the same residue or an allowable substitution. However, the third position is comparatively free of evolutionary constraints except when an unacceptable amino acid conversion results. In the third position of the Gly-X-Y and 5' COOH-terminal propeptide codons, 44% and 40% divergence have occurred while only 7.9% and 4.3% of the amino acids have changed. The 3' symmetrical half of the COOH-terminal propeptide (29,31,32) appears to be a more recent duplication since 27% of the third position nucleotides and 1.7% of the 115 amino acids have diverged.

Discrimination among first, second and third positions is presumably not operative in the α1(IV) 3' UTRs unless these nucleotides are also translated. This event seems unlikely in that the reading frame would be changed by interdispersed deletions and insertions. Therefore, the divergence of the 3' UTR should parallel or exceed that of the third codon position in the absence of a requirement to maintain specific nucleotides. In the human and mouse α1(IV) 3' UTRs, three areas, in particular, are far more conserved (within brackets in Fig. 4). Area One, consisting of 198 bases with 84% homology (4 deletions), occurs in the 5' quadrant and terminates 3' subsequent to AATAAA Sequence #1. Area Two, consisting of 74 bases with 85% homology (no deletions), is centrally located and contains AATAAA Sequence #2. Area Three with 95% homology (no deletions) consists of the 3' terminal 90 bases including the primarily utilized AATAAA Sequence #4.

Similar regions containing 95, 109 and 76 nucleotides with 84%, 88% and 88% homology, respectively, are found in human and avian α2(I) (indicated by brackets in Fig. 5). The first occurs in the 5' segment, the second closely follows the 5' AATAAA set and the third extends both 5' and 3' from the second set of AATAAA sequences. Notably in this last set, the 24 nucleotides prior to the first AATAAA are 100% homologous and can be aligned with sequences in α1(IV) (Fig. 8). The human and avian α1(II) 3' UTRs are extremely divergent explaining why the two conserved areas are so striking (Fig. 6). These relatively short regions of 51 and 59 nucleotides are 84% and 86% homologous, respectively. The first extends 5' and 3' to the AATAAA in human α1(II) and the second is 5' to and inclusive of a variation recognized at least in avian α1(II) (40).

The interspecies homology in the procollagen 3' UTRs is in accord with that found for the actins (17 and refs. therein). However, major differences between these two gene families reside in the intraspecies nucleotide homology 5' to poly A attachment signals shown in Fig. 8 and the intraspecies conservation of the lengths of the UTRs, especially in the smaller RNAs. In all 5 procollagen genes from which multiple RNAs are transcribed, the 5' AATAAA sequences are found at extremely similar positions: human α1(I) + 248, +252; human α2(I) +278, +292 (chick +274, +279); human α1(III) +254, +258; human α1(IV) +239 (mouse +267) and human α2(V) +286, +290. Also except for α1(IV) is the unusual occurrence of these being overlapping or in tandem. Although the location of the 3' signals are more diverse, those in human α1(I) are at +1253 versus +1284 in human α1(IV) (mouse +1262), and those in human α1(III) are at +855 versus +795 and +816 in human α2(I) (chick +728 and +751).

The significance of conservation in the 3' untranslated regions awaits the results of alternative approaches to DNA sequencing such as those of Krowczynska and Brawerman (63). Their investigation of the "complex configuration at the 3' end of the globin mRNAs" provides evidence for a double stranded structure around the poly A junction and, furthermore, one which probably results from association with another region of the same molecule. Similar studies using RNAs with multiple polyadenylation attachment signals may distinguish differences in secondary structure at sites frequently or infrequently utilized and whether the conserved sequences in the 3' UTRs are involved in these interactions.

ACKNOWLEDGEMENTS

We thank Maryann Mason for excellent typing of the manuscript and sequences and Pamela Howard for assistance with the photography. These studies were supported by NIH Grants AM20553, HL34005 and NCI Grant PO1-22427. L.J.G. is an E.I. of the American Heart Association.

*To whom correspondence should be addressed:
Connective Tissue Research Institute, University of Pennsylvania
3624 Market Street, Philadelphia, PA 19104

REFERENCES

1. Miyata, T, Yasunaga, T. and Nishida, T. (1980) Proc. Natl. Acad. Sci. USA 77, 7328-7332.
2. Proudfoot, N.J. and Brownlee, G.G. (1976) Nature 263, 211-214.
3. Fitzgerald, M. and Shenk, T. (1981) Cell 24, 251-260.
4. Montell, C., Fisher, E.F., Caruthers, M.H. and Berk, A.J. (1983) Nature 305, 600-605.
5. Higgs, D.R., Goodbowin, S.E.Y, Lamb, J., Clegg, J.B. and Weatherall, D.J. (1983) Nature 306, 398-400.
6. Moore, C.L. and Sharp, P.A. (1984) Cell 36, 581-591.
7. Woychik, R.P., Lyons, R.H., Post, L. and Rottman, F.M. (1984) Proc. Natl. Acad. Sci. USA 81, 3944-3948.
8. Sadofsky, M. and Alwine, J.C. (1984) Mol. Cell. Biol. 4,1460-1468.
9. Gil, A. and Proudfoot, N.J. (1984) Nature 312, 473-474.
10. Wickens, M. and Stephenson, P. (1984) Science 226, 1045-1051.
11. Birnsteil, M.L., Busslinger, M. and Strub, K. (1985) Cell 41, 349-359.
12. Conway, L. and Wickens, M. (1985) Proc. Natl. Acad. Sci. USA 82, 3949-3953.
13. Manley, J.L., Yu, H. and Ryner, L. (1985) Mol. Cell. Biol. 5, 373-379.
14. Moore, C.L. and Sharp, P.A. (1985) Cell 41, 845-855.
15. Bhat, B.M. and Wold, W.S.M. (1985) Mol. Cell Biol. 5, 3183-3193.
16. Hart, R.P., McDevitt, M.A. and Nevins, J.R. (1985) Cell 43, 677-683.
17. Yaffe, D., Nudel, U., Mayer, Y. and Neuman, S. (1985) Nucleic Acids Res. 13, 3723-3737.
18. Boedtker, H., Fuller, F. and Tate, V. (1983) Int. Rev. Connect. Tissue Res. 10, 1-63.
19. Reddi, A.H. (1984) in Extracellular Matrix Biochemistry (Piez, K.A. and Reddi, A.H., eds.) pp. 375-412.
20. Cheah, K.S.E. (1985) Biochem. J. 229, 287-303.
21. Miller, E.J. and Gay, S. (1986) Methods Enzymol., in press.
22. Vuorio, E., Sandell, L., Kravis, D., Sheffield, V.C., Vuorio, T., Dorfman, A. and Upholt, W.B. (1982) Nucleic Acids Res. 10, 1175-1192.
23. Aho, S., Tate, V. and Boedtker, H. (1983) Nucleic Acids Res. 11, 5443-5450.
24. Merlino, G.T., McKeon, C., deCrombrugghe, B. and Pastan, I. (1983) J. Biol. Chem. 258, 10041-10048.
25. Myers, J.C., Dickson, L.A., deWet, W.J., Bernard, M.P., Chu, M-L., Di Liberto, M., Pepe, G., Sangiorgi, F. and Ramirez, F. (1983) J. Biol. Chem. 258, 10128-10135.
26. Ninomija, Y. and Olsen, B.O. (1984) Proc. Natl. Acad. Sci. USA 81, 3014-3018.

27. Loidl, H.R., Brinker, J.M., May, M., Pihlajaniemi, T., Morrow, S., Rosenbloom, J. and Myers, J.C. (1984) Nucleic Acids Res. 112 9383-9394.
28. Chu, M-L., deWet, V., Bernard, M. and Ramirez, F. (1985) J. Biol. Chem 260, 2315-2320.
29. Oberbaumer, I., Laurent, M., Schwarz, U., Sakurai, Y., Yamada, Y., Vogeli, G., Voss, T., Siebold, B., Glanville, R.W. and Kuhn, K. (1985) Eur. J. Biochem. 147, 217-224.
30. Chu, M-L., Weil, D., deWet, W., Bernard, M., Sippola, M. and Ramirez, F. (1985) J. Biol. Chem. 260, 4357-4363.
31. Brinker, J.M., Gudas, L.J., Loidl, H.R., Wang, S-Y., Rosenbloom, J., Kefalides, N.A. and Myers, J.C. (1985) Proc. Natl. Acad. Sci. USA 82, 3649-3653.
32. Pihlajaniemi, T., Tryggvason, K., Myers, J.C., Kurkinen, M., Lebo, R., Cheung, M.-C., Prockop, D.J. and Boyd, C.D. (1985) J. Biol. Chem. 260, 7681-7687.
33. Myers, J.C., Loidl, H.R., Seyer, J.M. and Dion, A.S. (1985) J. Biol. Chem. 260, 11216-11222.
34. Gunning, P., Mohun, T., Ng, S-Y., Ponte, P. and Kedes, L. (1984) J. Mol. Evol. 20, 202-214.
35. Wang, S-Y. and Gudas, L.J. (1983) Proc. Natl. Acad. Sci USA 80, 5880-5884.
36. Messing, J. (1983) Methods Enzymol. 101, 20-78.
37. Maxam, A.J. and Gilbert, W. (1980) Methods Enzymol. 65, 499-560.
38. Yamada, Y., Kuhn, K. and deCrombrugghe, B. (1983) Nucleic Acids Res. 11, 2733-2744.
39. Stoker, N.G., Cheah, K.S.E., Griffin, J.R., Pope, F.M. and Solomon, E. (1985) Nucleic Acids Res. 13, 4613-4622.
40. Ninomya, Y., Showalter, A.M., vanderRest, M., Seidah, N.G., Cretien, M., and Olsen, B.R. (1984) Biochemistry 23, 617-624.
41. Sandell, L.J., Prentice, H.L., Kravis, D. and Upholt, W.B. (1984) J. Biol. Chem. 259, 7826-7834.
42. Fuller, F. and Boedtker, H. (1981) Biochemistry 20, 996-1006.
43. Levine, R., LaRosa, G. and Gudas, L.J. (1984) Mol. Cell. Biol. 4, 2142-2150.
44. Pihlajaniemi, T. and Myers, J.C. (1986) Methods Enzymol., in press.
45. Cheah, K.S.E., Stoker, N.G., Griffin, J.R., Grosveld, F.G. and Solomon, E. (1985) Proc. Natl. Acad. Sci. USA 82, 2555-2559.
46. Sangiorgi, F.O., Benson-Chanda, V., deWet, W.J., Sobel, M.E., Tsipouras, P. and Ramirez, F. (1985) Nucleic Acids Res. 13, 2207-2225.
47. Bernard, M.P., Myers, J.C., Chu, M-L., Ramirez, F., Eikenberry, E.F. and Prockop, D.J. (1983) Biochemistry 22, 1139-1145.
48. Bernard, M.P., Chu, M-L., Myers, J.C., Ramirez, F., Eikenberry, E.F. and Prockop, D.J. (1983) Biochemistry, 22, 5213-5223.
49. Gerstenfeld, L.C., Crawford, D.R., Boedtker, H. and Doty, P. (1984) Mol. Cell. Biol. 4, 1483-1492.
50. Mattaj, I.W. and Zeller, R. (1983) Embo. J. 2, 1883-1891.
51. Setzer, D.R., McGrogan,M., Nunberg, J.H. and Schimke, R.T. (1980) Cell 22, 361-370.
52. Setzer, D.R., McGrogan, M. and Schimke, R.T. (1982) J. Biol. Chem. 257, 5143-5147.
53. Morandi, C., Masters, J.N., Mottes, M. and Attardi, G. (1982) J. Mol. Biol. 156, 583-607.
54. Tosi, M., Young, R.A., Hagenbuchle, O. and Schibler, V. (1981) Nucleic Acids Res. 10, 2313-2323.
55. Gerlinger, P., Krust, A., LeMeur, M., Perrin, F., Cochet, M., Gannon, F., Dupret, D. and Chambon, P. (1982) J. Mol. Biol. 162, 345-364.

56. Zehner, Z.E. and Paterson, B.M. (1983) Proc. Natl. Acad. Sci. USA 80, 911-915.
57. Lagace, L., Chandra, T., Woo, S.L.C. and Means, A.R. (1983) J. Biol. Chem. 258, 1684-1688.
58. Unterman, R.D., Lynch, K.R., Nakhasi, H.L., Dolan, K.P., Hamilton, J.W., Cohn, D.V. and Feigelson, P. (1981) Proc. Natl. Acad. Sci. USA 78, 3478-3482.
59. Parnes, J.R., Robinson, R.R. and Seidman, J.G. (1983) Nature 302, 449-452.
60. Lee, M.G-S., Lewis, S.A., Wilde, C.D. and Cowan, N.J. (1983) Cell 33, 477-487.
61. Capetanaki, Y.G., Ngai, J., Flytzanis, C.N. and Lazarides, E. (1983) Cell 35, 411-420.
62. Barsh, G.S., Roush, C.L. and Gelinas, R.E. (1984) J. Biol. Chem. 259, 14906-14913.
63. Krowczynska, A. and Brawerman, G. (1986) J. Biol. Chem. 261, 397-402.