Volume 12 Number 10 1984 Nucleic Acids Research

Isolation of cDNA and genomic DNA clones encoding type II collagen

Marian F.Young[1], Gabriel Vogeli[2], Anne Marie Nunez[1], M.Pilar Fernandez[1], Marjorie Sullivan[3] and Mark E.Sobel[1]*

[1]Laboratory of Developmental Biology and Anomalies, National Institute of Dental Research, [2]National Eye Institute, National Institutes of Health, Bethesda, MD 20205 and [3]Genex, Rockville, MD, USA

Received 13 January 1984; Revised and Accepted 18 April 1984

ABSTRACT

A cDNA library constructed from total chick embryo RNA was screened with an enriched fraction of type II collagen mRNA. Two overlapping cDNA clones were characterized and shown to encode the COOH propeptide of type II collagen. In addition, a type II collagen clone was isolated from a Charon 4A library of chick genomic fragments. Definitive identification of the clones was based on DNA sequence analysis. The 3' end of the type II collagen gene appears to be similar to that of other interstitial collagen genes. Northern hybridization data indicates that there is a marked decrease in type II collagen mRNA levels in chondrocytes treated with the dedifferentiating agent 5-bromodeoxyuridine. The major type II collagen mRNA species is 5300 bases long, similar to that of other interstitial collagen RNAs.

INTRODUCTION

The collagen family of proteins are differentially expressed in various connective tissues (1). A subset of this family, the interstitial collagens (types I, II, and III), are composed of three alpha chains which form a stable triple helical molecule, characterized by repeating triplets of amino acids with the sequence gly-X-Y (2,3). Procollagen molecules are precursor forms which contain NH_2 terminal and COOH terminal propeptides in addition to the helical portion of the protein. The interstitial collagen genes of several species have been shown to be remarkably complex, yet similar, in their exon-intron structures (4-8). The chick $\alpha 2(I)$ gene, for example, is 38 kilobases (kb) in length and its coding sequences are interrupted by approximately 50 introns (4,5). There is a clear delineation of functional domains in the interstitial collagen genes. Exons encoding the helical portion of the collagen molecule can be distinguished from those encoding the propeptides. The former exons are most commonly 54 base pairs in length, corresponding to 6 multiples of 9 base pair units which code for gly-X-Y triplets (4-10).

Type II collagen is a specific phenotypic marker of cartilage. When chondrocytes are cultured in vitro, either for prolonged periods or in the presence of a variety of agents, the cartilage phenotype is altered and the

cells appear to "dedifferentiate" (12-15). In particular, it has been demonstrated that 5-bromodeoxyuridine-treated chondrocytes cease expression of type II collagen and instead synthesize type I collagen. This change in gene expression is correlated with altered mRNA levels for type I and type II collagens(15-17). Changes in mRNA levels have also been reported in Rous sarcoma virus-transformed chondrocytes (18).

To investigate the molecular basis for changes in type II collagen gene expression, we are in the process of isolating cDNA and genomic clones encoding chick type II collagen. We report a novel protocol for the construction and isolation of type II collagen cDNA plasmids using total embryo RNA as a source of cDNA template, followed by counterscreens with RNAs from different tissues expressing the various collagen genes. Further, we report the isolation of a chick genomic fragment containing the 3' end of the type II collagen gene. Preliminary analysis suggests that the type II collagen gene is similar in structure to the other interstitial collagens. Our results agree with and extend recent reports of other type II collagen recombinant clones which appeared while this work was in progress (17,19,20).

MATERIALS AND METHODS

Preparation of RNA.

Total cellular RNA was extracted from ten day old chick embryos and from 16 day old chick embryo sterna, crops, calvaria and fibroblasts by an adaptation of the guanidine hydrochloride procedure described previously (21). Poly(A)-containing RNA was isolated by chromatography on oligo (dT) cellulose (Type T3, Collaborative Research) as described (22). Fractionation of poly(A) containing sterna RNA according to molecular weight was carried out by density centrifugation on 10%-36% linear sucrose gradients in an SW28 Beckman rotor for 21 hours at 26,000 rpm as described (23). Gradient fractions were analyzed by cell free translation in a micrococcal nuclease treated reticulocyte lysate (Bethesda Research Laboratories) as described (24) and the fractions containing type II collagen RNA were pooled.

Construction of recombinant plasmids.

Poly(A) RNA from ten day old chick embryos was used as template to synthesize double stranded cDNA as described previously (25). The double stranded cDNA was digested with S1 nuclease as described and tailed with dC residues using alpha [^{32}P]-labeled dCTP tracer (26). The reaction products were electrophoresed in a 0.8% low melting agarose gel in E buffer (27) and

all tailed cDNAs greater than 1000 bp in length were eluted from gel slices by heating at 65° in an equal volume of 1M NaCl, 40 mM Tris pH 7.5, 10 mM EDTA pH 7.5. After phenol extraction, the tailed double stranded cDNA was ethanol precipitated, dissolved in TEN buffer (300 mM NaCl, 1 mM EDTA, 10 mM Tris pH 7.5) and annealed with PstI-cleaved pBR322 that had previously been tailed with dG residues as described (26). The resulting recombinant plasmids were recovered by transformation of Escherichia coli strain N38 (C600 r^-m^-).

Colony hybridization.

Bacterial colonies were transferred to nitrocellulose filters (Schleicher and Schuell) and the filters were placed face up for 10 minutes onto Whatman 3MM paper which had been soaked in 0.5 N NaOH. Filters were subsequently placed for ten minutes onto 3MM paper soaked in 1M Tris, pH 8.0, 0.6M NaCl, and for 30 minutes onto paper soaked in 1M Tris pH 8.0, 0.6M NaCl, 40 µg proteinase K. The filters were then dipped into chloroform and placed face down on 3MM paper, covered with Whatman paper and pressed to leave bacterial debris on the bottom filter paper. The nitrocellulose filters containing the denatured DNA were then air dried and baked at 80° in vacuo for 2 hours. For relaxed hybridization conditions filters were hybridized in 10% Dextran sulphate, 0.6 M NaCl, 50% formamide at 30° overnight after which filters were washed in 2X SSC at 25°. For stringent conditions, filters were hybridized at 41° overnight in 10% Dextran sulphate, 40% formamide, 5X SSC, 0.25% Denhardt's, 100 µg/ml denatured salmon sperm DNA. After hybridization, filters were washed at 25° in 2XSSC, 0.1% SDS and at 52° in 0.1xSSC, 0.1% SDS. Filters were blotted dry and placed against preflashed X-Ray film.

Hybridization probes were prepared by isolation of inserts from recombinant plasmids pCg45 and pCg54 (obtained from H. Boedtker, Harvard U.) Plasmid pCg54 contains an 1100 bp HindIII-KpnI insert encoding chick α1(I) procollagen (28). Plasmid pCg45 contains a 2500 bp insert in the HindIII site of pBR322 encoding chick α2(I) procollagen (29). After restriction, plasmid DNA was electrophoresed on 1.5% agarose gels, electrophoresed onto DEAE membrane (Schleicher and Schuell) using a Bio Rad Trans Blot Apparatus. The collagen cDNA inserts were extracted from the DEAE by incubation in 6.5M urea, 1M NaCl for 10 minutes at 65° and ethanol precipitated. Purified inserts were nick translated with ^{32}P-deoxynucleotides as described (30), and denatured by boiling for 5 minutes before addition to the colony hybridization filters.

To prepare labeled RNA probes, 2.0 µg sucrose gradient purified RNA were boiled in 50mM Tris pH 9.5 for 30 minutes after which half of the sample was

boiled for an additional 30 minutes. The two fractions were pooled and kinased with polynucleotide kinase as described (31). The ^{32}P-labeled RNA fragments were chromatographed on Sephadex G-50 (fine) and void volume fractions were pooled and used directly for stringent hybridization.

Plasmid DNA Isolation.

Supercoiled plasmid DNA was recovered from E. coli supernatant after equilibrium centrifugation in ethidium bromide-cesium chloride as previously described (25).

Northern Hybridization.

Total cellular RNA was extracted from a variety of tissues as described above and 5 µg of each RNA preparation were electrophoresed on a 1% agarose horizontal slab gel containing 6mM methylmercuric hydroxide and transferred to diazobenzyloxymethyl paper (Schleicher and Schuell) as described previously (30). To ensure that equal amounts of RNA were loaded and transferred in each lane, the gel was stained with ethidium bromide and quantities of 18S and 28S rRNA were assessed. The filter was hybridized to ^{32}P-labeled insert from pCs1 at 45^{o}C for 18 hour in 5 X SSC, 50% formamide, 0.1% SDS, 1% glycine, 1x Denhardt's without bovine serum albumin, and 1 mg/ml carrier yeast RNA as described (30). The filter was washed and autoradiographed as described previously (30).

Electronmicroscopic Analysis.

To form cDNA-R loops, pCs2 DNA was converted to a linear molecule with Hind III and hybridized to poly(A) RNA from chick embryo sterna in 70% formamide, 100mM Tris pH 8.0, 0.5M NaCl, 10mM EDTA at 52^{o} for 3 hours as described (25). The concentration of DNA was 10 µg/ml and of RNA was 100 µg/ml.

To analyze cDNA-cDNA heteroduplexes, equimolar quantities of linearized pCs1 and pCs2 were denatured at 85^{o} in 10 mM Tris pH 7.5, 1 mM EDTA, 50% formamide, hybridized at 45^{o} for 2 hours and then at 25^{o} overnight.

To analyze exon-intron structure of λCs7, R loops were prepared as described above. Heteroduplexes between the DNA of recombinant phage λCs7 and λCharon 4A were formed in 50% formamide, 0.1M tricine, pH 8.0 and 0.5M NaCl for 2 hours at 52^{o}, and then RNA was hybridized to the duplex at 52^{o} for 3 hours in 70% formamide, 0.5M NaCl, 0.1 M tricine pH 8.0. Samples were mounted for electron microscopy by the method described by Tilghman et al (31). Double stranded circular SV40 DNA was included as internal standard. The

preparation of the sample for microscopy, photography and analysis of data was performed as described (32).

Screening of Chick Genomic Library.

A library of chick genomic fragments in λCharon 4A (33), provided by D. Engel (Northwestern U.), was screened as previously described (34). A 300 base pair AluI-PstI fragment was isolated from the 5' end of pCs2, labeled by nick translation (30), and used as a hybridization probe. Hybridization was performed at 68° in 6 X SSC, 10 X Denhardt's and 0.1% SDS. DNA filters were washed three times with 1 X SSC and 0.1% SDS at room temperature.

DNA Sequencing.

DNA fragments were isolated, labeled with gamma ^{32}P-labeled ATP by polynucleotide kinase, and sequenced as described (31).

RESULTS AND DISCUSSION

Construction and screening of a chick embryo cDNA library.

The stage 36 (10 day) chick embryo expresses most known types of collagen (35). To construct a cDNA library containing a representative sample of collagen sequences, we prepared RNA from whole ten day old chick embryo and synthesized cDNA molecules by established procedures. We enriched for collagen cDNA molecules by selecting double stranded cDNAs greater than 1000 bp in length. The cDNA molecules were inserted into the PstI site of pBR322 by homopolymer tailing and recombinant plasmids were transformed into E. coli.

Four thousand plasmids were screened for type II collagen cDNA sequences by colony hybridization. Sucrose gradient purified [^{32}P]-labeled chick sternal mRNA known to be enriched in type II collagen mRNA was used as a probe and forty positive colonies were obtained. [^{32}P]-labeled lens RNA was used to counterscreen the filters since lens does not produce type II collagen but instead expresses other collagen types. Of the forty clones selected by the initial sternal RNA screen, only six appeared to contain cartilage-specific DNA sequences since they did not hybridize to the crude mixture of lens RNAs.

The colony hybridization filters were also screened using purified [^{32}P]-labeled cDNA inserts from plasmids pCg54 (28) and pCg45 (29), which code for pro α1(I) and pro α2(I) collagen, respectively. Relaxed hybridization conditions were used to allow for cross-hybridization to any collagen-like DNA sequences. Initially, three clones were identified using relaxed washing conditions but only two of these clones were visualized on the radioautographs of the filters after stringent washes. These same two clones, named pCs1 and

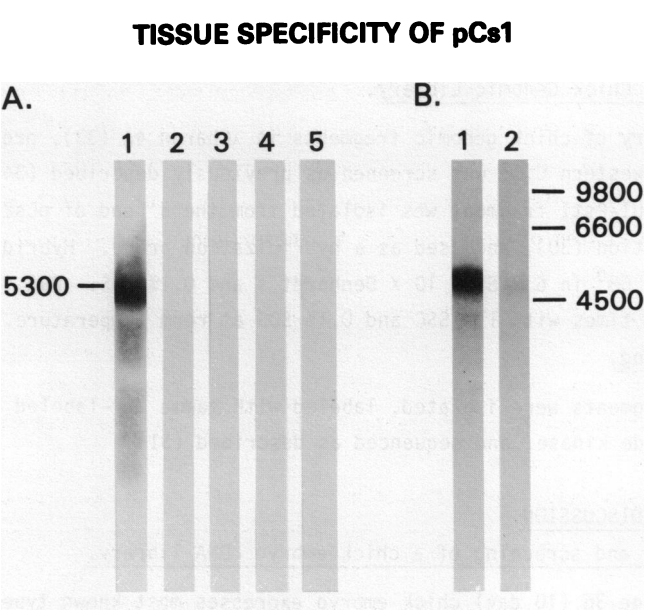


Figure 1. Northern hybridization. Total cellular RNA was extracted from a variety of chick embryo tissues, separated on methylmercury agarose gels and transferred to diazobenzyloxymethyl paper as described in Methods. The filter was hybridized to ^{32}P-labeled insert from pCs1. The length of the hybridized mRNA species was determined by comparison with the known sizes of λDNA restricted with HindIII. (A) Lane 1, sterna; lane 2, fibroblasts; lane 3, Rous sarcoma virus-transformed fibroblasts; lane 4, crops; lane 5, calvaria. (B) Lane 1, chondrocytes cultured in vitro for 5 days; lane 2, parallel culture of chondrocytes grown in the presence of 5-bromodeoxyuridine.

pCs2, were of the original six clones which hybridized to sternal RNA but not lens RNA. It thus appeared that pCs1 and pCs2 contained collagen-like sequences which were homologous but not identical to type I collagen DNA.

Tissue specificity of pCs1 and pCs2.

Since type II collagen is found in cartilage but not in other tissues such as skin, bone, and muscle, we performed a Northern blotting experiment. (Figure 1A). [^{32}P]-labeled insert from pCs1 hybridized to a mRNA species approximately 5300 bases long in chick sterna (lane 1) but not in fibroblasts (lane 2), crops (lane 4), or calvaria (lane 5). There was also no hybridization to RNA from Rous sarcoma virus-transformed chick embryo fibroblasts (lane 3), which produce low levels of type I collagen (29). The same RNA preparations in lanes 2 and 5 hybridized to type I collagen cDNA and the crop RNA used in lane 4 hybridized to type III collagen genomic clones

(not shown). A similar experiment showed the same tissue specificity profile when pCs2 insert was used as the hybridization probe. Two minor cartilage RNA species, approximately 7200 and 6000 bases long, were identified in the sterna RNA after extended exposure of the radioautograph (not shown). Polymorphism in the lengths of mRNAs appears to be a general feaure of the collagen gene family. The mechanism appears to be variability in the length of the 3' untranslated region of the mRNA due to the presence of multiple poly (A) attachment sites (8,11).

Previous studies have demonstrated that chondrocytes cultured in vitro express type II collagen (12-14). However, when treated with the thymidine analog 5-bromodeoxyuridine, the cells no longer synthesize type II collagen and have decreased levels of translatable type II collagen RNA (15-17). We therefore performed a Northern blotting experiment in which [^{32}P]-labeled insert from pCs1 was hybridized to a filter containing RNA from control chondrocytes (Figure 1B, lane 1) and from 5-bromodeoxyuridine-treated cells (Figure 1B, lane 2). As would be expected using a type II collagen probe, there was a marked reduction in the amount of hybridization of the pCs1 cDNA insert to the RNA from the treated chondrocytes.

Electromicroscopic analysis of pCs1 and pCs2.

We prepared a Southern blot to determine if pCs1 and pCs2 contain overlapping DNA sequences (Figure 2 insert). pCs1 and pCs2 DNAs were digested with PstI and then resolved by electrophoresis through a bed of 1% agarose. DNA was then transferred to nitrocellulose and hybridized to [^{32}P]-labeled purified insert from pCs2. The pCs2 insert hybridized to itself (lane 2) and also to the pCs1 insert (lane 1). The sizes of the inserts of pCs1 and pCs2, determined by comparison to known molecular weight standards, were 780 and 1250 base pairs, respectively.

Preliminary restriction digests of plasmids pCs1 and pCs2 suggested that the cDNA inserts were oriented in the same direction in pBR322 (data not shown). It was therefore possible to determine the precise extent of homology between pCs1 and pCs2 by electronmicroscopic heteroduplex analysis (Figure 2). Equimolar quantities of linearized pCs1 and pCs2 were denatured, allowed to reanneal, and were then visualized by electronmicroscopy. Analysis of 13 hybrid molecules revealed a loop of single stranded DNA of 474 $\pm$ 85 bases. The short arm of the double stranded hybrid molecules was measured to be 1331 $\pm$ 210 base pairs. Taking into consideration the known distance between the PstI site and the EcoRI restriction site used to linearize the plasmids (752 base pairs), we determined that pCs1 and pCs2 inserts contained 759 $\pm$ 210 base

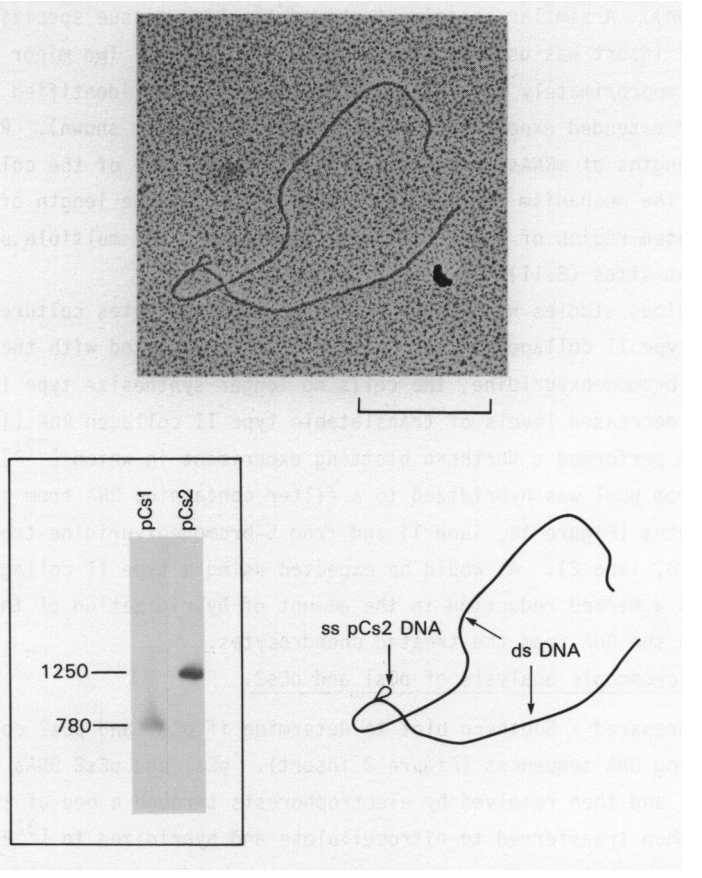


Figure 2. Cross hybridization of pCs1 and pCs2. (Top) Electron micrograph of a representative heteroduplex structure of pCs1 and pCs2 DNA. Heteroduplexes were formed between pCs1 and pCs2 and were analyzed as described in Methods. The bar below the photograph represents 1 kilobase. (Bottom) Schematic diagram of micrograph. (Inset) Southern hybridization. PstI-digested DNAs from pCs1 (lane 1) and pCs2 (lane 2) were electrophoresed and transferred to nitrocellulose paper by the technique of Southern (41). The blot was hybridized to ^{32}P-labeled inserts from pCs1 and pCs2. Sizes of inserts were determined by comparison to the known lengths of fragments of lambda DNA restricted with HindIII.

pairs in common. In some electronmicrographs, an additional loop (of less than 100 bases) was visualized approximately 600 base pairs from the large D loop, at a distance consistent with the junction of the long arm of pBR322 and the cDNA inserts. This indicated that there might be a small difference in length between the pCs1 and pCs2 inserts at one end, in addition to a

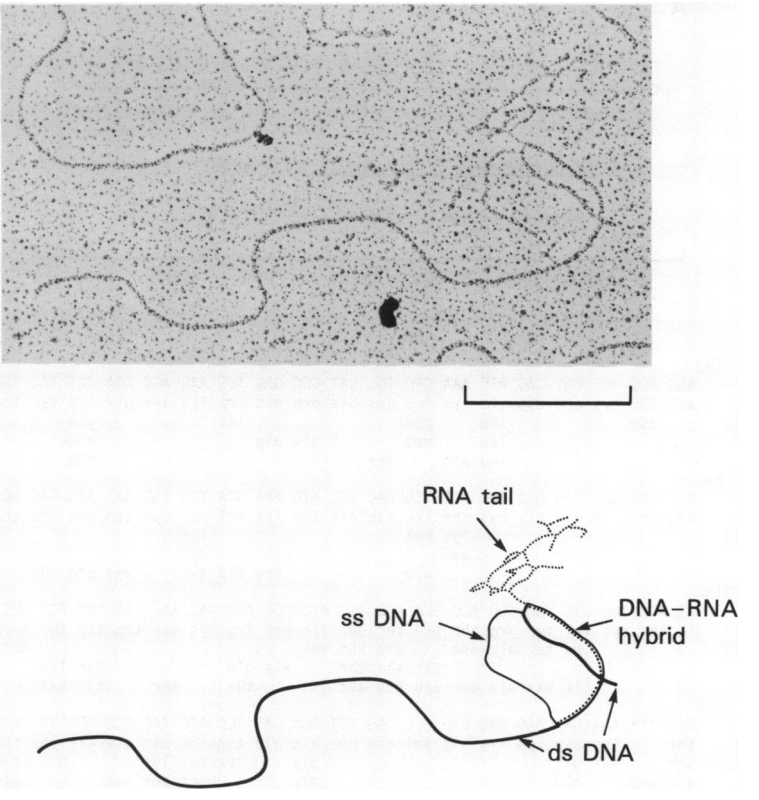
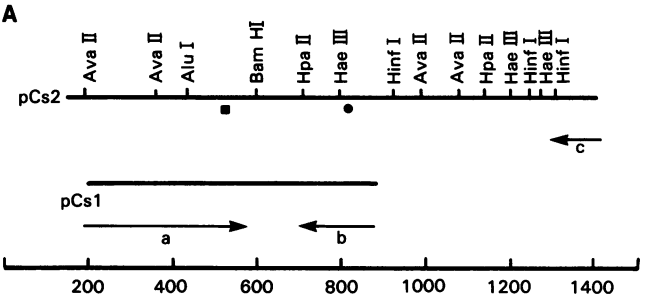


Figure 3. Electronmicrograph of R loop between pCs2 and sterna RNA. Linearized pCs2 DNA was hybridized to poly(A) RNA from chick embryo sterna as described in Methods. In the schematic drawing of the micrograph, the stippled line represents the mRNA. The bar underneath the micrograph represents 1 kilobase.

relatively large difference at the other.

To determine the extent of homology of pCs2 insert with type II collagen mRNA, cDNA-R loops were analyzed. Figure 3 shows a representative R loop in which linearized pCs2 DNA was hybridized to sucrose gradient purified sterna RNA enriched for type II collagen mRNA molecules. A single stranded loop of the expected size (1137 $\pm$ 146 bases) was visualized in the 9 molecules studied, indicating that pCs2 insert contained faithful complementary sequences to type II collagen mRNA. The large tail of unhybridized mRNA visualized at the junction of the R loop and the short arm of the plasmid most likely represents 5' mRNA sequences. The absence of a 3' tail suggested that



B

```
      193
         AGG ACC TGC CGC GAC ATC AAA CTC TGC CAT CCC GAG TGG AAG AGC GGA GAT TAC TGG ATT
α1(II)   arg thr cys arg asp ile lys leu cys his pro glu trp lys ser gly asp tyr trp ile
α1(III)      asn             leu     phe             leu             glu
α1(I)                        leu     met         gly asp             glu
α2(I)                        leu arg     ser                 ser     phe
      253
         GAC CCG AAC CAG GGC TGC ACC TTG GAC GCC ATC AAA GTA TTC TGC AAC ATG GAG ACA GGC
α1(II)   asp pro asn gln gly cys thr leu asp ala ile lys val phe cys asn met glu thr gly
α1(III)                      lys met                 tyr
α1(I)                        asn                     tyr
α2(I)                            ala             arg ala tyr     asp phe ala
      313
         GAG ACC TGC GTC TAC CCG ACC CCC AGC AGC ATC CCC AGG AAG AAC TGG TGG ACC AGC AAG
α1(II)   glu thr cys val tyr pro thr pro ser ser ile pro arg lys asn trp trp thr ser lys
α1(III)          leu ser ala asn     ala thr val                         thr glu
α1(I)                        gln ala thr     ala gln             tyr leu
α2(I)            ile his ala ser leu glu asp         thr     thr     tyr val
      373
         ACG *** AAA GAC AAG AAG CAC GTC TGG TTT GCA GAG ACC ATC AAC GGC GGT TTC CAC TTC
α1(II)   thr     lys asp lys lys his val trp phe ala glu thr ile asn gly gly phe his phe
α1(III)  ser     ser gly                         gly     ser met lys             gln
α1(I)    asn pro     glu                         gly         met ser asp         gln
α2(I)    asn pro                      ile         gly                     thr gln
      433
         AGC TAC GGC GAT GAG AAC CTG TCC CCC AAC ACC GCC AGC ATC CAG ATG ACC TTC CTG CGC
α1(II)   ser tyr gly asp glu asn leu ser pro asn thr ala ser ile gln met thr phe leu arg
α1(III)              pro asp     pro glu asp val ser glu val     leu ala
α1(I)    glu         gly     gly ser asn     ala asp val ala         leu
α2(I)    glu     asn gly     gly val thr thr lys asp met ala thr     leu ala     met
      493
         CTC CTG TCC ACC GAG GGC TCC CAG AAC GTC ACC TAC CAC TGC AAG AAC AGC ATC GCC TAC
α1(II)   leu leu ser thr glu gly ser gln asn val thr tyr his cys lys asn ser ile ala tyr
α1(III)  ile         ser arg ala             ile
α1(I)        met             ala thr                             val
α2(I)            ala asn his ala ser         ile
      553
         ATG
α1(II)   met
α1(III)
α1(I)
α2(I)
      724
         ACC TCG CGC CTG CCC ATT GTA GAT ATT GCA CCT ATG GAC ATT GGC GGA GCC GAT CAG GAG
α1(II)   thr ser arg leu pro ile val asp ile ala pro met asp ile gly gly ala asp gln glu
α1(III)      met             val                 ile             pro
α1(I)                            ile     leu             val     ala pro
α2(I)    pro                     leu     leu         leu
      784
         TTT GGC GTG GAT ATT GGC CCA GTC TGC TTC TTG
α1(II)   phe gly val asp ile gly pro val cys phe leu
α1(III)              val
α1(I)            ile
α2(I)            leu his                     lys
```

pCs2 might contain cDNA sequences close to the 3' terminus of the type II collagen mRNA molecule.

Physical mapping of pCs1 and pCs2.

We constructed a restriction map of the cDNA clones by a combination of partial restriction endonuclease digestions and by restriction digestion of isolated fragments of pCs1 and pCs2 (Fig 4A). There were no restriction sites for EcoRI, HindIII, TaqI, KpnI, ClaI, SalI or EcoRV. pCs1 and pCs2 have nearly identical 5' termini. As predicted by the heteroduplex (Fig 2), pCs1 and pCs2 have 700 base pairs in common. pCs2 extends approximately 500 bp 3' to pCs1, corresponding to the D-loop of 474 $\pm$ 85 bases seen in the heteroduplex. There was a striking similarity with other sternal cDNA clones, pCar1 and pCar2, previously described by Vuorio et al. (19). pCs2 appears to contain all the sequences of both pCar1 and pCar2 with an additional 100 base pairs at the 5' end. pCs2 contains a single BamHI site, confirming the suggestion by Vuorio et al. (19) that the BamHI site at the 3' end of pCarI was artificially generated during homopolymer tailing.

Sequence Analysis.

We sequenced three regions of pCs1 and pCs2 to determine if they encode type II procollagen (fragments a, b, and c, in Fig. 4). The DNA sequences were compared to the known sequences of other interstitial chick procollagen genes. An optimal alignment of the sequenced regions of pCs1 and pCs2 with

Figure 4A. Restriction map and sequencing strategy of pCs1 and pCs2. A restriction map of the insert of pCs2 was constructed as described in the text. The restriction map of pCs1 was identical to the corresponding region of pCs2 and is aligned accordingly. The number bar below indicates the distance in base pairs from the end of the helical coding region of the α1(I) collagen molecule as predicted by alignment of the derived sequences with the known sequences of other collagen molecules (see text). The (▪) indicates the highly conserved region surrounding the carbohydrate attachment site found in interstitial collagens. The (●) indicates the location of the stop codon. The arrows indicate the sequencing direction of fragments 'a' and 'b' of pCs1 and of fragment 'c' of pCs2.

Figure 4B. Sequence of fragments 'a' and 'b' of pCs1. Purified insert from pCs1 was kinased, restricted with HpaII, electrophoresed on 7% polyacrylamide gels and fragments 'a' and 'b' were isolated as described in Methods. DNA fragments were sequenced by the method of Maxam and Gilbert (31). The base numbering system is described in Figure 4A. The protein sequence derived from the DNA is shown below and is compared with published protein sequences for the other interstitial collagens (37). Only amino acids which differ from α1(II) are shown. The *** in fragment 'a' indicate the location of a deleted codon which is present in α1(I) and α2(I).

α1(I), α2(I), and α1(III) procollagens indicates that pCs1 and pCs2 contain sequences encoding the COOH-terminal propeptide of a procollagen chain. The translated pCs1 sequence is 72%, 63% and 70% homologous with the amino acid sequences of the COOH terminal propeptide of α1(I), α2(I) and α1(III), respectively, and the respective nucleotide sequence homologies are 67%, 61%, and 64%.

The sense strand DNA sequence of pCs1 and pCs2 fragments and their corresponding translation products are presented in Fig. 4B. In the base numbering system used in Fig. 4, COOH terminal collagen sequences are numbered from the end of the helical-coding region (coordinate 0) and correspond to the numbering system used by Fuller and Boedtker (36) for the α1(I) procollagen gene. We aligned the most 5' sequences of pCs1 (fragment a, Fig 4) with base number 193 of the α1(I) COOH terminal propeptide coding region. Fragment "a" contains sequences from bases 193 to 555. It should be noted that bases 376-378 (coding for proline) of the α1(I) COOH terminal coding region are not present in the corresponding region of pCs1. These three bases have also been shown to be missing in αI(III) (6,37).

By analogy with the known exon-intron splice sites of the α2(I) and α1(III) collagen genes, in which exons are numbered progressively from the 3' end of the gene, fragment "a" of pCs1 contains part of exon 4 (ending at base number 238 within the gly codon), all of exon 3 (from bases 239-429), and the 5' end of exon 2.

Fragment "a" contains 5 regions (corresponding to bases 193-213, 235-270, 277-321, 385-414, 508-555) of the COOH propeptide which have been previously identified as showing homologies in their amino acid sequence between α1(I), α2(I) and α1(III) collagens (37). The amino acid sequence of these segments is also highly conserved in pCs1. The fifth segment (bases 508-555) has also been identified as being highly conserved not only in primary protein structure but also in nucleotide sequence (37). In this segment, 42 nucleotide positions contain the identical base sequence in pCs1 and the other 3 collagen genes described above. Furthermore 3 of the 6 nucleotide positions that differ among these 4 genes occur at base positions 509, 511 and 513. This suggests an even tighter conservation of nucleotide sequence between bases 514-555, coding for a 14 amino acid long region of the COOH-propeptide that surrounds the carbohydrate attachment site (bases 517-525).

By analogy with the exon-intron structure of the α1(III) gene (6), fragment "b" contains part of exon 1 of an interstitial collagen gene. An ochre termination codon (TAA) was identified in fragment "b" at bases 817-

819. The 3' flanking sequence appears to be relatively long since the 3' end of fragment "c" is approximately 600 bp from the stop codon. Although R-looping data (see Fig. 3) suggests that pCs2 contains cDNA sequences close to the 3' terminus of the type II collagen mRNA molecule, it is unlikely that the 3' end of pCs2 actually represents the end of the noncoding portion of the mRNA. The canonical pAATAAA sequence that is usually present 10-20 nucleotides 5' to the poly(A) addition site of eukaryotic mRNA is not discernible in fragment "c" (data not shown).

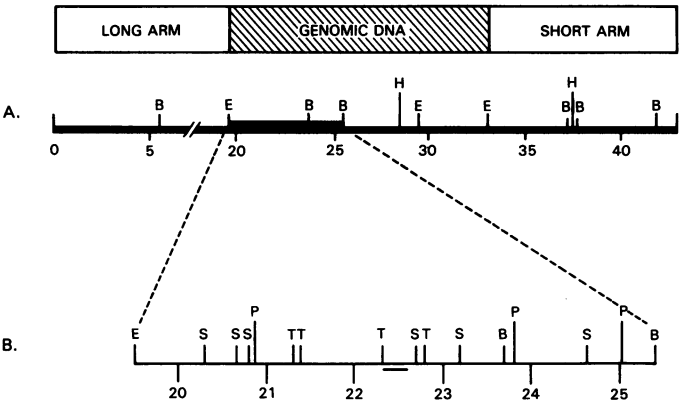
Isolation of a Genomic Clone.

Sequence analysis of pCs1 and pCs2 demonstrated that the recombinant plasmids encoded an interstitial collagen molecule different from α1(I), α2(I), and α1(III). Furthermore, the cDNA clones hybridized specifically to RNA from sterna, suggesting that they encoded type II collagen. Amino acid sequence of the type II collagen COOH propeptide is not available. Therefore, we screened a Charon 4A phage library of chick genomic fragments to obtain cloned DNA likely to encode the helical region of type II collagen, for which the sequence was available (Dr. W. Butler, U. Alabama). 5×10^5 plaques were screened using the more 5' PstI-AluI fragment of the pCs2 insert (see Fig 4) as a cross hybridization probe. One clone, λCs7, was isolated and characterized.

Restriction endonuclease mapping and Southern blot hybridization with different subfragments of pCs1 and pCs2 defined the region of λCs7 corresponding to the 3' end of the gene (see Fig 5).

We used two types of heteroduplex analysis to study the exon-intron structure of the 3' end of the gene cloned in λCs7. Heteroduplexes formed between the DNA of recombinant phage λCs7 and the DNA of Charon 4A were hybridized to sucrose gradient fractions of chick sternal RNA enriched for type II collagen mRNA. A representative structure is depicted in Fig. 6. At least 10 exons are found interspersed within a 4.5 kb DNA segment.

The size of exons and introns in λCs7 was measured by electronmicroscopic scanning and is shown in Table I. An interesting feature of the gene is that exons within the carboxy propeptide encoding region are larger than those in the helical region. A similar observation has been noted for the chick α2(I) and α1(III) collagen genes (37). Another interesting feature of the α1(II) gene is that compared to the α2(I) and α1(III) collagen genes, the total length of intron (1-9) is substantially (50%) smaller. Because the resolution of electron microscope analysis does not allow detection of introns smaller than 100 bp, it is possible that the region of the α1(II) gene contained



C.

1

ser ala phe ala gly leu gly gln thr glu
TCT GCT TTT GCT GGA CTG GGT CAG ACG GAG

19

lys gly pro asp pro ile arg tyr met
AAG GGC CCC GAC CCC ATC CGC TAC ATG

Figure 5A. Restriction map of λCs7. The restriction map of λCs7 was derived from a series of single and double digestions of the recombinant clone by EcoRI (E), BamHI (B), and HindIII (H). A southern blot of the restriction digests was hybridized to ^{32}P-labeled PstI-AluI fragments of pCs2 insert to determine the orientation of genomic DNA within λCs7. The numbers below the map refer to kilobases from the end of the long arm of the phage. The heavy bar between 19.5 and 25.4 kb delineates the region of λCs7 that hybridized to pCs2 DNA.

Figure 5B. Restriction map of the 5.9 kb genomic fragment. A restriction map was derived from a series of single and double digestions by SmaI (S), PvuII (P), and TaqI (T).

Figure 5C. Nucleotide sequence of CB 14 of α1(II) procollagen. The TaqI-SmaI fragment delineated by the bar in Figure 5B was kinased and restricted with HinfI which cleaves close to the 3' SmaI site. A portion of the sequence of the TaqI-HinfI fragment corresponding to cyanogen bromide peptide 14 of α1(II) collagen (commencing several codons after the end of the helical coding region and ending two codons before the carboxypeptidase cleavage site) is shown. Derived amino acid residues are numbered starting with the first codon of CB 14.

within λCs7 may contain more than 10 introns and exons. Confirmation of the precise intron-exon structure of λCs7 must await complete DNA sequence analysis.

If the chick genomic sequences in λCs7 are complementary to the cDNA clones pCs1 and pCs2, then heteroduplexes between λCs7 and pCs1 should

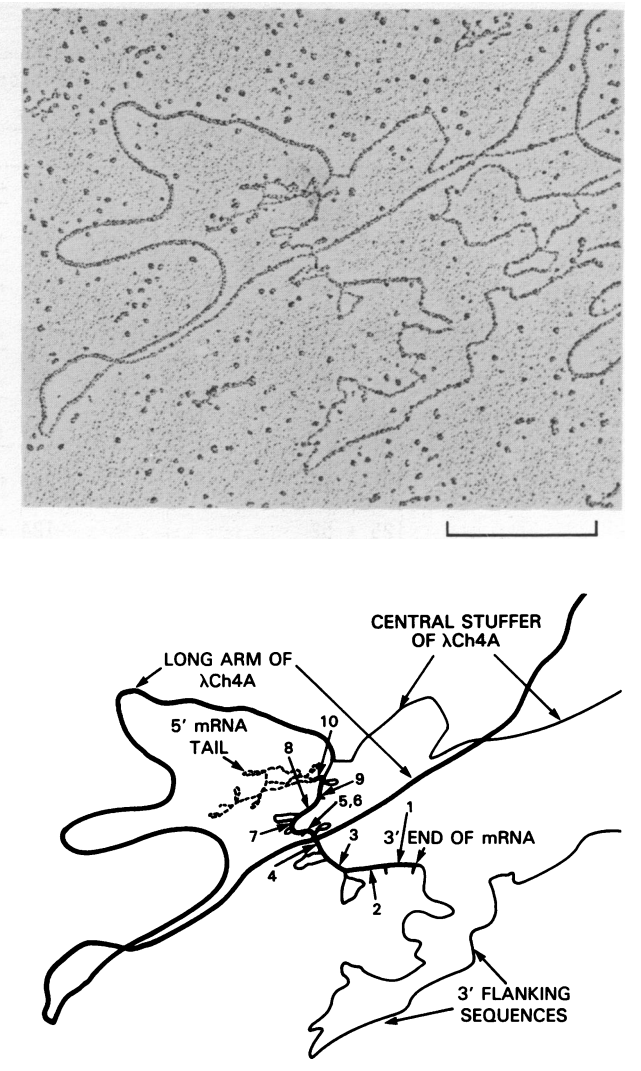


Figure 6. Electronmicrograph of heteroduplex between λCs7 and sterna RNA. Heteroduplexes between recombinant phage λCs7 and Charon 4A were formed as described in Methods, and poly(A) RNA from chick embryo sterna was hybridized to the duplexes. The bar beneath the micrograph designates 1 kilobase. The numbers in the schematic diagram below correspond to exons beginning at the 3' end of the gene.

specifically indicate the exon-intron structure of the 3' end of the presumptive type II collagen gene. A representative molecule of such a duplex is shown in Fig. 7. DNA sequence analysis of pCs1 (Fig. 4B) indicated that it

Table I: Sizes of exons and introns in λCs7

Exon[a]	Exon length (bp)	Intron length (bp)
1	764 ± 59[b]	97 ± 41
2	316 ± 61	634 ± 95
3	186 ± 45	299 ± 73
4	299 ± 81	133 ± 44
5	104 ± 50	83 ± 31
6	121 ± 44	109 ± 39
7	131 ± 53	283 ± 47
8	125 ± 52	124 ± 58
9	97 ± 40	175 ± 119
10	108 ± 55	

[a] Exons are numbered from the 3' end of the gene.
[b] Standard deviation
The 3' flanking region downstream from exon 1 which extends to the short arm of λCharon 4A is about 9kB. The 5' end of the DNA insert of λCs7 is an intervening sequence containing part or all of intron 10 of the gene.

coded for sequences homologous to exons 1-4 of the other interstitial collagen genes. The three single stranded DNA loops visualized in the electronmicrograph in Fig. 7 most likely represent introns which delineate exons 1, 2, 3, and 4 of the gene. Moreover the sizes of the DNA loops correspond well to those of the introns separating exons 1-4 in the R loop in Fig. 6. The size of exons 2 and 3 in both forms of heteroduplex analysis also correspond well. Exons 1 and 4 in the cDNA genomic heteroduplex, however, are shorter than those analyzed for the genomic R loop seen in Fig. 6. This is consistent with DNA sequence analysis of pCs1, which showed that its 3' end includes some, but not all, of the untranslated region of the 3' end of its corresponding mRNA. In addition, the 5' end of pCs1 encodes only a small portion of the 3' end of exon 4.

Sequence of the COOH telopeptide of the type II collagen gene.

The heteroduplex analysis (Fig. 7) suggested that λCs7 encoded genomic

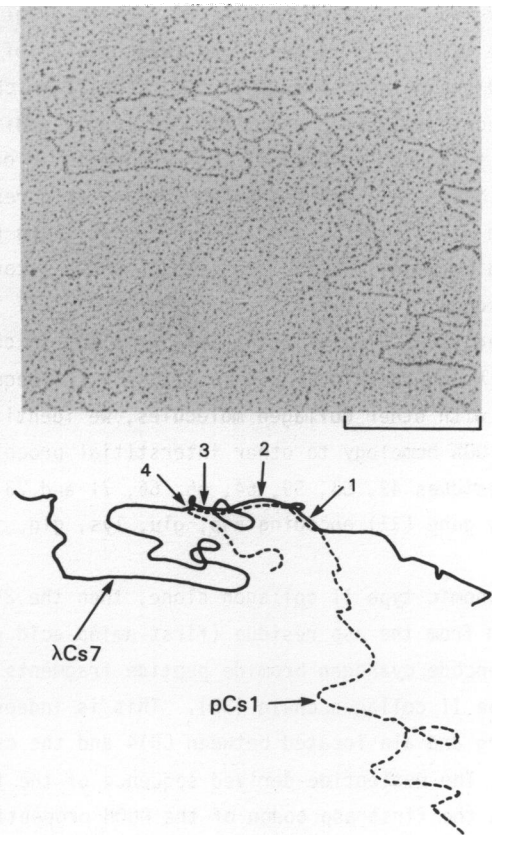


Figure 7. Electronmicrograph of heteroduplex between pCs1 and λCs7 DNA. Heteroduplexes were formed between pCs1 (----) and λCs7 (——) as described in Methods. Exons 1-4 correspond to those shown in Figure 6. The bar below the micrograph designates 1 kilobase.

DNA both 3' and 5' to the cDNA inserts of pCs1 and pCs2. The exon structure of the gene encoded by λCs7 appeared to be similar to that of other interstitial collagen genes. In such genes, exon 4 encodes the end of the helical region as well as the COOH telopeptide region of the procollagen molecule. The latter is specific for each collagen chain. We predicted that DNA sequence analysis of exon 4 would reveal a type II collagen-specific telopeptide coding region.

We therefore isolated and sequenced a TaqI-HinfI restriction fragment of λCs7 corresponding to a region immediately 5' to (and including) the cDNA insert of pCs1 (shown as a bar on the diagram in Fig. 5B). We identified a

region approximately 175 base pairs downstream from the TaqI site which corresponded exactly with the 5' sequence of fragment 'a' of pCs1 (see Fig. 4B). This confirmed the heteroduplex analysis (Fig. 7) which indicated that λCs7 contains sequences including, and 5' to, pCs1. In addition, the twelve bases upstream from fragment 'a' of pCs1 (corresponding in our cDNA base numbering system to bases 181-192) coded for 4 amino acid residues which were 100% homologous with α1(I), α2(I), and α1(III) procollagen protein sequences. The residues encoded by bases 181-192 are lys-asn-pro-ala corresponding to resudues 76-79 of exon 4 of the α1(III) collagen gene (37). This region was previously identified as an area of amino acid homology in collagen genes by Yamada et al (37). As an additional aid in aligning the sequences of the TaqI-HinfI fragment with other collagen molecules, we identified 8 other amino acid residues with 100% homology to other interstitial procollagen chains (37). These were residues 42, 53, 59, 64, 65, 66, 71 and 73 of exon 4 of the α1(III) collagen gene (37) encoding asp, glu, lys, gln, ile, glu, pro, and gly respectively.

If λCs7 is a genomic type II collagen clone, then the 21 codons immediately upstream from the asp residue (first amino acid of the COOH propeptide) should encode cyanogen bromide peptide fragments 14 and 15 (CB14 and CB15) of the type II collagen chain (38). This is indeed the case. CB15 is a dipeptide of arg and ala located between CB14 and the carboxypeptidase cleavage site (38) . The nucleotide-derived sequence of the two residues immediately prior to the first asp codon of the COOH propeptide of λCs7 is pAGGGCA, coding for arg-ala.

The sense strand DNA sequence and translated amino acid sequence of CB14 is presented in Fig. 5C. The deduced amino acid composition of CB14 shown in Fig 5C is 100% in agreement with the known amino acid composition of CB14 (38). Residue 18 is tyr, which is also found 4 residues upstream from the carboxypeptidase cleavage site in the other three interstitial collagen chains (37). The wobble position of this codon, however, is a T in the α1(III) gene. No other amino acid residue in CB14 is completely homologous with all the other collagen chains. Overall amino acid homology between the type II collagen sequence and α1(I), α2(I), and α1(III) collagens was 32%, 11%, and 42%, respectively. This is consistent with the expected uniqueness of the telopeptide. This region of the α1(II) chain contains one more amino acid residue than do the α1(I) and α1(III) collagens.

We also compared the translated amino acid sequence shown in Fig. 5C with a protein sequence provided to us by Dr. W. Butler (U. Alabama) which aligned

with the first 12 residues of CB14. We found two discrepancies between the protein sequence and the nucleotide-derived sequence. Whereas the first translated residue in Fig. 5C is ser, the protein sequence predicted glu in this position. Possibly, this reflects an allelic difference. We are convinced that our DNA sequence is correct since our gels are unambiguous and the amino acid composition of CB14 (38) predicts 1 ser and cannot account for a glu in residue 1 in addition to the glu and gln at residues 8 and 10. Furthermore, Miller (38) found that at a noticeable frequency, CB14 was not cleaved from the CB7 fragment (more NH_2 terminal to it). The presence of a ser at residue 1 is consistent with inefficient cleavage of met-ser peptide bonds which has been reported by others (39). There was also a discrepancy at residue 9, where a pro was predicted by the available protein sequence. The amino acid composition of CB14 (38) predicted 1 thr and 2 pro residues, consistent with our DNA-derived sequence. Sandell et al (20) who recently identified a type II collagen genomic clone, also found a thr in this position.

SUMMARY

In conclusion, we constructed and isolated two recombinant cDNA clones and a genomic clone containing the 3' end of the α1(II) collagen gene. It has been widely assumed that, because of the secondary structure of the collagen message, long cDNA transcripts could best be obtained by using highly enriched sources of collagen RNA. We have demonstrated that at least in the case of the type II collagen gene, it is possible to obtain reasonable size transcripts and to clone the collagen gene from a highly heterogeneous message source, the whole chick embryo. However, the success of our approach was dependent upon counterscreening with a variety of other collagen probes in which relaxed and stringent hybridization conditions were used in tandem.

Conclusive identification of the clones was complicated by the limited availability of amino acid sequence data. We therefore depended upon sequencing a specific telopeptide region of a genomic DNA clone. Colinearity of the genomic clone with the cDNA clones was demonstrated by a combination of heteroduplex mapping, Southern hybridization, and DNA sequencing.

The cDNA clones pCs1 and pCs2 appear to overlap with the cartilage-specific cDNA clones described by Vuorio et al (19) and by Lukens et al (16). In the latter case, the clones were identified as encoding type II collagen by the method of hybrid-selected translation. The clones of Vuorio et al (18) were identified by the method used in the present report, i.e. DNA

sequencing of the telopeptide region of a genomic clone. It should be noted that pCs1 and pCs2 and the clones of Vuorio et al (19) and Lukens et al (17) are quite different in restriction fragment analysis and DNA sequence from a presumptive α1(II) collagen cDNA clone described by Duchene et al (40). In our hands, Northern hybridization experiments show that the latter clone cross hybridizes with type I collagen mRNA even under the most stringent conditions (data not shown). We conclude that the clone of Duchene et al (40) is in fact not a type II collagen DNA probe.

In the Northern hybridization experiments described in Figure 1, we used stringent hybridization conditions to achieve specific hybridization of our recombinant cDNA probes with sternal RNA. We confirmed our previous results and the results of others (15-17) that show that type II collagen mRNA levels are markedly decreased when the cartilage phenotype is altered by BUdR treatment of chondrocytes. No cross hybridization of pCs1 and pCs2 was demonstrated with the considerable amounts of type I collagen mRNA present in dedifferentiated (BUdR-treated) chondrocytes. We are currently in the process of using the cDNA and genomic clones to determine the molecular mechanisms involved in chondrocyte dedifferentiation.

ACKNOWLEDGEMENTS

We would like to thank Dr. Kwang-Sam Koh, Kenneth Galen and Frances Cannon for their technical assistance, William Upholt and Linda Sandell for preliminary sequence information of their type II chick cDNA clones and Bill Butler for protein sequence data. We are also grateful to George Martin and Yoshihiko Yamada for useful discussions and careful review of this manuscript. We also thank Helga Boedtker for permission to use cDNA probes developed in her laboratory.

*Current address: Laboratory of Pathology, National Cancer Institute, National Institutes of Health, Bethesda, MD 20205, USA

REFERENCES

1. Prockop, D.J and Champe, F.C., Eds (1980) Gene Families of Collagen and Other Proteins, Elservier/North Holland, New York.
2. Bornstein, P. and Sage, H. (1980) Annu. Rev. Biochem. 49, 959-1003.
3. Galloway, D. (1982) in Collagen in Health and Disease, Weiss, J.B. and Jayson, M.I.V. Eds, pp. 528-558, Churchill-Livingstone, Edinburgh.
4. Ohkubo, H., Vogeli, G., Mudryj, M., Avvedimento, V.E., Sullivan, M., Pastan, I. and de Crombrugghe, B. (1980) Proc Natl. Acad. Sci. USA 78, 7057-7063.

5. Wozney, J., Hanahan, D., Tate, V., Boedtker, H. and Doty, P. (1981) Nature 294, 129-135.
6. Yamada, Y., Mudryj, M., Sullivan, M. and de Crombrugghe, B. (1983) J. Biol. Chem. 258, 2758-2761.
7. Monsen, J. and McCarthy, B.J. (1981) DNA 1, 59-69.
8. Meyers, J.C., Dickson, L.A., de Wet, W.J., Bernard, M.P., Chu, M-L, Di Liberto, M., Pepe, G., Sangiorgi, F.O. and Ramirez, F. (1983) J. Biol. Chem. 258, 10128-10135.
9. Vogeli, G., Ohkubo, H., Avvedimento, V.E., Sullivan, M., Yamada, Y., Mudryj, M., Pastan, I., and de Crombrugghe, B. (1981) Cold Spring Harbor Symposia on Quantitative Biology. Volume XLV, pp 777-783.
10. Yamada, Y., Avvedimento, V., Mudryj, M.,Ohkubo, H., Vogeli, G., Irani, M., Pastan, I., and de Crombrugghe, B. (1980) Cell 22, 887-892.
11. Aho, S., Tate, V., and Boedtker, H. (1983) Nucl. Acids. Res. 11, 5443-5450.
12. Schiltz, J.R., Mayne, R.M. and Holtzer, H. (1973) Differentiation 1, 97-108.
13. Abbott, J. and Holtzer, H. (1968) Proc. Natl. Acad. Sci. USA 59, 1144-1151.
14. Mayne, R., Vail, M.S., and Miller, E.J. (1975) Proc. Natl. Acad. Sci. USA 72, 4511-4515.
15. Kaul, R., Hewitt, A.T., Varner, H., Somerman, M., Martin, G.R. and Sobel, M.E. (1981) Fed. Proc. 40, 1626.
16. Pawlowski, P.J., Brierley, G.T. and Lukens, L.N. (1981) J. Biol. Chem. 256, 7695-7698.
17. Lukens, L.N., Frischauf, A.M., Pawlowski, P.J., Brierley, G.T. and Lehrach, H. (1983) Nucl. Acids Res. 11, 6021-6039.
18. Adams, S.L., Boettiger, D., Focht, R.J., Holtzer, H. and Pacifici, M. (1982) Cell 30, 373-384.
19. Vuorio, E., Sandell, L., Kravis, D., Sheffield, V.C., Vuorio, T., Dorfman, A., Upholt, W.B. (1982) Nucl. Acids Res. 10, 1175-1192.
20. Sandell, L.J., Yamada, Y., Dorfman, A. and Upholt, W.B. (1983) J. Biol. Chem. 258, 11617-11621.
21. Gottesman, M.M. and Sobel, M.E. (1980) Cell 19, 449-455.
22. Aviv, H. and Leder, P. (1972) Proc. Natl. Acad. Sci. USA 69, 1408-1412.
23. Rosen, J.M., Woo, S.L.C. and Comstock, J.P. (1975) J. Biochem. 14, 2895-2903.
24. Sobel, M.E., Dion, L.D., Vuust, J., Colburn, N.H. (1983) Mol. and Cell Biol. 3, 1527-1532.
25. Sobel, M.E., Yamamoto, T., Adams, S.L., Dilauro, R., Avvedimento, V.E., de Crombugghe, B. and Pastan, I. (1978) Proc. Natl. Acad. Sci. USA 75, 5846-5850.
26. Nelson, T. and Brutlag, D. (1979) Methods in Enzymology 68, 41-49.
27. Loening, U.E. (1969) Biochem. J. 102, 251-257.
28. Lehrach, H., Frishauf, A.M., Hanahan, D., Wozney, J., Fuller, F., and Boedtker, H. (1979) Biochemistry 18, 3146-3152.
29. Lehrach, H., Frischauf, A.M., Hanahan, D., Wozney, J., Fuller, F., Crkvenjakov, R., Boedtker, H., and Doty, P. (1978) Proc. Natl. Acad. Sci. 75, 5417-5421.
30. Sobel, M.E., Yamamoto, T., de Crombrugghe, B. and Pastan, I. (1981) Biochemistry 20, 2678-2684.
31. Gilbert, W. and Maxam, A. (1973) Proc. Natl. Acad. Sci. USA 70, 3581-3584.
32. Tilghman, S.M., Curtis, P.J., Tiemeier, D.C., Leder, P., and Weissman, C. (1978) Proc. Nat. Acad. Sci. USA 75 1309-1313.
33. Dodgson, J.B., Strommer, J., and Engel, J.D. (1979) Cell 17, 879-887.
34. Maniatis, T., Hardison, R.C., Lacy, E., Lauer, J., O'Connell, C. and Quon, E. (1978) Cell 15, 687-701.

35. Merlino, G.T., McKeon, C., de Crombrugghe, B., and Pastan, I. (1983) J. Biol. Chem. 258, 10041-10048.
36. Fuller, F. and Boedtker, H. (1981) Biochemistry 20, 996-1006.
37. Yamada, Y., Kuhn, K. and de Crombrugghe, B. (1983) Nucl. Acids Res. 11, 2733-2744.
38. Miller, E.J. (1972) Biochemistry 11, 4903-4909.
39. Schroeder, W.A., Shelton, J.B. and Shelton, J.R. (1969) Arch. Biochem. Biophys. 130, 551.
40. Duchene, M., Sobel, M.E., and Muller, P.K. (1982) Exp. Cell Res. 142, 317-324.
41. Southern, E. (1979) Methods in Enzymology 68, 152-176.