Isolation and partial characterization of the entire human proα1(II) collagen gene

Frank O.Sangiorgi, Virginia Benson-Chanda, Wouter J.de Wet*, Mark E.Sobel§, Petros Tsipouras+ and Francesco Ramirez

Departments of Obstetrics and Gynecology and +Pediatrics, University of Medicine and Dentistry of New Jersey-Rutgers Medical School, Piscataway, NJ 08854, USA, *Department of Biochemistry, Potchefstroom University, Potchefstroom, 2520 South Africa, and §Laboratory of Pathology, National Institutes of Health, Bethesda, MD 20205, USA

Received 4 February 1985; Accepted 7 March 1985

ABSTRACT

Using a cDNA probe specific for the bovine Type II procollagen, a series of overlapping genomic clones containing 45 kb of contiguous human DNA have been isolated. Sequencing of a 54 bp exon, number 29, provided direct evidence that the recombinant clones bear human Type II collagen sequences. Localization of the 5' and 3' ends of the gene indicated that the human Type II collagen gene is 30 kb in size. This value is significantly higher than that of the homologous avian gene. The segregation of a polymorphic restriction site in informative families conclusively demonstrated that the Type II gene is found in a single copy in the human haploid genome. Finally, sequencing of a triple helical domain exon has confirmed that a rearrangement leading to the fusion of two exons occurred in the proα1(I) gene, following the divergence of the fibrillar collagens.

INTRODUCTION

Type II collagen, a genetically distinct member of the fibrillar collagens (Types I, II and III) is the major structural component of hyaline cartilage (1). This protein is composed of three identical α1(II) chains and it shares with the other fibrillar collagens a similar molecular structure and biosynthetic pathway. The nascent chains of all three collagen types are synthesized as precursor molecules, procollagens, with a long α chain domain and two short terminal extensions which are removed extracellularly after secretion of the assembled triple helix (1). The central α-chain domains, approximately one thousand residues long, are composed of a repetitive tripeptide structure $(\text{Gly-X-Y})_{340\pm2}$(1).

The temporal expression of the Type II and Type I collagen genes is believed to play a fundamental role in the normal process of cartilage differentiation and bone formation (2). Defects at different levels of the ossification process have been postulated for the chondrodysplasias (3), a highly heterogeneous group of disorders. A structural defect in Type II collagen has been reported in at least one case of diastrophic dysplasia (4). Other studies have suggested that biosynthetic changes in Type II collagen are related to the degeneration of the articular cartilage in the affected joints of osteoarthritic individuals (5). However, because of the complexity and sequential nature of the ossification process, very little is known about the pathogenesis of these acquired and inherited disorders.

In order to investigate the molecular basis of Type II collagen gene expression in normal and diseased states, several investigators have isolated cDNA and genomic clones coding for the proα1(II) collagen chain (6-13). The most extensive studies have been carried out on the chicken gene because of the relative ease of obtaining embryonic material in sufficient amounts for molecular cloning (6-11). More recently, however, some information has been obtained for portions of the proα1(II) gene from different mammalian species (12-13). In particular, we have isolated a cDNA clone (Bc-7) coding for the calf proα1(II) collagen chain and in turn used it for the identification of two bovine genomic clones covering the 3' end of the gene (Sangiorgi, F. et al., manuscript submitted). Here we report the use of Bc-7 for the isolation of a series of overlapping genomic clones encompassing the entire human proα1(II) collagen gene. Sequencing of selected areas has established that the gene is 30 kb long. Familial segregation of a restriction enzyme length polymorphism (RFLP) has demonstrated that the proα1(II) gene is present in a single copy in the human haploid genome. Finally, comparison of our data with those reported for a gene previously characterized as bearing "α1-like" collagen sequences (14) has revealed that the latter clone is indeed the human Type II procollagen gene.

MATERIALS AND METHODS

Gene Isolation

The probe used for the initial screening of the human library was Bc-7, a recombinant molecule containing 1011 bp of bovine proα1(II) collagen sequences (Sangiorgi, F. et al., manuscript submitted). The phage libraries utilized in these studies were both constructed in Charon 4A; one contained 15-20 kb partially digested Eco RI human genomic fragments (kindly provided by Dr. A. Bank, Columbia University), whereas the other was obtained by insertion of Alu I/Hae III partially digested DNA (kindly provided by Dr. T. Maniatis, Harvard University). Screening and isolation of the DNA from the positive phage clones was performed as described (15). Chromosomal walking was achieved by screening of the libraries with the appropriate genomic fragments subcloned in pBR322.

Southern and Northern blotting hybridizations

Transfer of nucleic acids onto nitrocellulose paper was done as previously described (15). Cross-hybridizations of the bovine probe to the human DNA were carried out at 40°C in 2x SSC (1x SSC:0.15M NaCl, 0.015 Sodium Citrate pH 6.8) 1x Denhart's solution, 100 μg/ml of sheared salmon sperm DNA and 50% formamide. Washing of the filters was done by sequentially decreasing the salt concentration from 2x SSC to 0.1x SSC and by increasing the temperature from 20°C to 60°C. For the cross-hybridization of the rat cDNA probe specific for the N-prepropeptide region of the proα1(II) gene, the salt concentration was increased to 4x SSC, whereas the temperature of the incubation was decreased to 37°C. In this case, washing of the filters was terminated at 0.2x SSC at 55°C. Hybridization and washing conditions for the other experiments were as previously detailed (15). *In vitro* labelling of the DNA probes was obtained by nick translation (16). Following mild alkaline hydrolysis (17) the RNA was labelled with $[\gamma^{32}P]$ ATP in the presence of T4 Kinase.

Primer extension

The synthetic oligonucleotide used in the primer extension experiments was derived from the heptapeptide Leu-Leu-Thr-Leu-Leu-Ile-Ala located between amino acid residues 12 and 18 in the rat proα1(II) N-propeptide (12). The oligonucleotide was used to prime cDNA synthesis using poly(A+) RNA isolated from a rat chondrosarcoma generously provided by Dr. V. Hascall (National Institutes of Health). The reaction was carried out in the presence of $[\alpha^{32}P]$ labelled deoxynucleotides according to a modification of the procedures of Agarwal et al. (18) and Tate et al. (19). The product of the reaction was approximately 200 nucleotides long, as determined on a denaturing polyacrylamide gel, and it exhibited a specific activity of 0.8×10^7 cpm/μg.

RNA purification

Total poly(A+) RNA was extracted from a rat chondrosarcoma using the guanidinium HCl procedure (20), from frozen fetal bovine articular cartilage using a modification of the previously described method (21) and from cultured fibroblasts according to the published procedure (15).

DNA sequencing

The chemical modification method of Maxam and Gilbert (22) was used for the sequencing of the genomic clones. Sequencing of both strands was performed for most of the sequences presented in this paper.

RESULTS

Gene isolation and mapping

For the initial screening of the human genomic libraries the cDNA clone

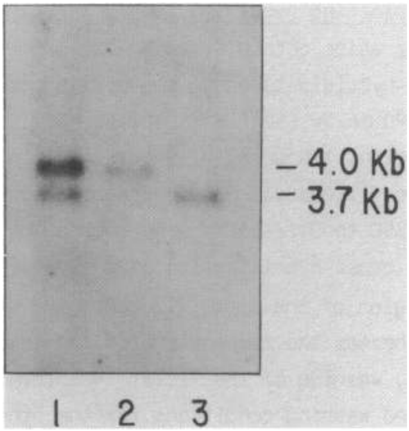

Fig. 1. Southern blotting analysis of human genomic DNA digested with Eco RI and hybridized to the entire bovine probe Bc-7 (lane 1), to the 3' segment of Bc-7 (lane 2) and to the 5' segment of Bc-7 (lane 3).

Bc-7 was used. This recombinant molecule contains a eucaryotic insert of 1 kb, coding for almost 75% of the C-propeptide domain of the bovine Type II collagen. Interspecies comparison has shown that the sequences of Bc-7 exhibit an 85% homology with those determined for the homologous avian chain (Sangiorgi, F. et al., manuscript submitted). Prior to the screening of the genomic libraries we optimized the cross-hybridization conditions of the bovine probe to the corresponding human sequences by Southern blotting analysis of Eco RI digested total human nuclear DNA. These experiments showed a unique pattern even under stringent conditions of hybridization and washing (Fig. 1). More precisely, by this analysis two bands, 3.7 kb and 4.0 kb in size, were identified in the human genome. It should be noted that in Bc-7 a unique Eco RI site, 35 bp from the termination codon, generates two subfragments of almost equal length, one specific for the 5' coding region and the other for the 3' untranslated region. Hence, we used these two subfragments as probes to assess whether the human gene possessed the same Eco RI site and to orient the two genomic bands. As is shown in Figure 1, the 5' subfragment of Bc-7 hybridized with the 3.7 kb Eco RI band, whereas the 4.0 kb was recognized by the 3' subfragment of Bc-7. Based on these results we proceeded to screen the Eco RI genomic library which yielded two positive clones, Pis 8 and Pis 36. Eco RI digestion of the DNA from Pis 8 generated four fragments: 4.5 kb, 4.0 kb, 3.7 kb and 2.5 kb. The DNA of Pis 36 showed a slightly different Eco RI pattern: 4.5 kb, 3.8 kb, 3.7 kb and 2.5 kb. Con-

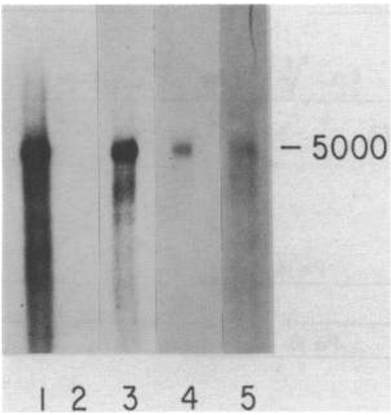

Fig. 2. Northern blotting analysis of poly(A+) RNA isolated from cartilagenous and non-cartilagenous tissues and hybridized to Bc-7 (lanes 1 and 2), the 3.7 kb Eco RI subclone of Pis 8 (lane 3), the 4.0 kb Eco RI subclone of Pis 8 (lane 4) and the 3.8 kb Eco RI subclone of Pis 36 (lane 5). Fibroblast RNA is in lane 2, whereas bovine fetal cartilage RNA is in the other four lanes. The size of the mRNA is expressed in nucleotides.

sistent with the previous Southern blotting analysis of total genomic DNA, the 5' end segment of Bc-7 hybridized to the 3.7 kb fragments of Pis 8 and Pis 36. On the other hand, the 3' segment of Bc-7 recognized both the 4.0 kb fragment of Pis 8 and the 3.8 kb fragment of Pis 36.

Several possibilities existed to explain the difference observed between the Eco RI patterns of Pis 8 and Pis 36. First, one could have argued for the existence of a polymorphic Eco RI site in Pis 36, which shortened the length of the 4.0 kb fragment to 3.8 kb. We excluded this possibility because Pis 36 did not contain an additional 0.2 kb Eco RI fragment and because after extensive restriction mapping the 4.5 kb and 2.5 kb Eco RI fragments were placed downstream to the 4.0 kb (Fig. 3).

A second alternative was the existence of two distinct genes with slightly different Eco RI patterns, a notion, however, not substantiated by our previous Southern blotting analysis of total genomic DNA. In addition, Northern blotting hybridizations of the 4.0 kb and 3.8 kb as well as the 3.7 kb fragments to RNA extracted from bovine articular cartilage, exhibited a pattern identical to that obtained with Bc-7 (Fig. 2). This result demonstrated that the genomic fragments behaved in a manner consistent with the size, tissue specificity and quantitative representation predicted for a proα1(II) collagen species. Finally, partial sequencing of the 4.0 kb and 3.8 kb DNA from the two Eco RI sites showed complete identity between the two fragments.

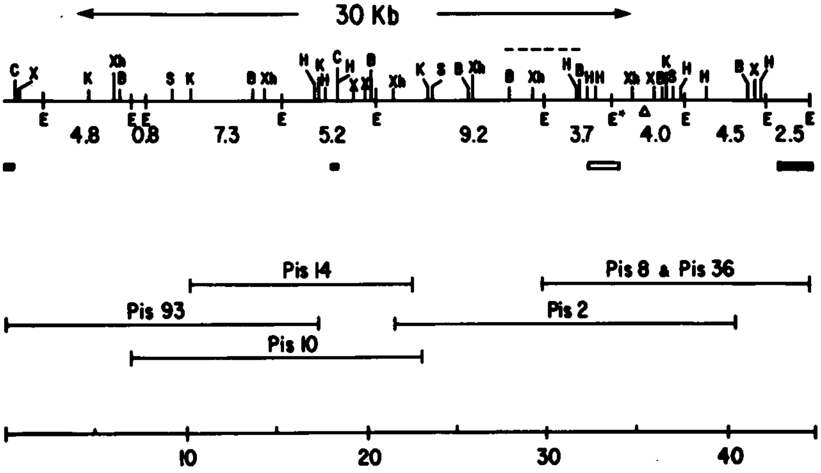

Fig. 3. Restriction map of the 45 kb of chromosomal DNA contained within the overlapping genomic clones. The sites shown refer to the following enzymes: B: Bam HI; C: Cla I; E: Eco RI; H: Hind III; K: Kpn I; S: Sph I; X: Xba I; Xh: Xho I. The numbers beneath refer to the size of the Eco RI fragments expressed in kilobases. The asterisk identifies the Eco RI site common to both the genomic clones and the cDNA Bc-7, whose position relative to the genomic clones is signified by the open bar below the restriction map. Similarly, the open triangle locates the 200 bp deletion of Pis 36, whereas the black bars indicate the presence of repetitive sequences. The dotted line above the restriction map shows the location of the Bam HI fragment used in the RFLP experiments. The relative position of the Type II gene with respect to the 45 kb of chromosomal DNA is indicated by the double arrow on top of the map. At the bottom of the map are the relative positions of the genomic clones and the scale expressed in kilobases.

A third possibility was that the Pis 8 and Pis 36 were derived from two alleles with a short polymorphic 200 bp DNA deletion. We proved that this was indeed the case by extensive restriction mapping of the 4.0 kb and 3.8 kb fragments in conjunction with Southern blotting analysis of the DNA from several individuals. These experiments located the deletion between an Xho I and an Xba I site of Pis 8 (Fig. 3). The nature of the deletion and its frequency in the population will be discussed elsewhere (Tsipouras, P. et al., in preparation). Direct confirmation of the identity of the genomic clones with Bc-7 was obtained by selective sequencing of the 3.7 kb and 4.0 kb Eco RI fragments of Pis 8 (Figs. 4A, 5).

The 3.7 kb Eco RI subclone of Pis 8 was used for a second genomic screening which resulted in the isolation of Pis 2 from the Alu I/Hae III

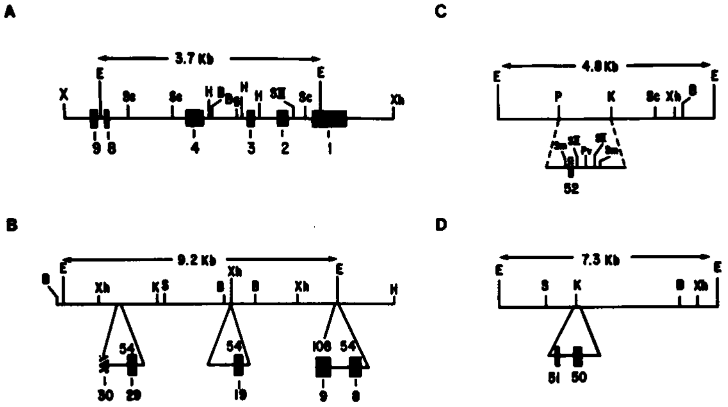

Fig. 4. Detailed restriction maps of the regions comprising the Eco RI fragments (A,B,C,D) whose areas have been selectively sequenced (see Figs. 3,6, and 8). The letters indicate the following restriction sites: B: Bam HI; Bg: Bgl II; E: Eco RI; H: Hind III; K: Kpn I; P: Pst I; Pv: Pvu II; S: Sph I; Sc: Sac I; SII: Sac II; Sm: Sma I; Xh: Xho I. The exons are identified by the black boxes and they are numbered underneath in a sequential order beginning from the 3' end of the gene. The sizes of four triple helical exons are indicated above them. The cross hatched box of exon 1 signifies the non-coding sequences, whereas in exons 30 and 52 it indicates that they have been partially sequenced.

library. The DNA of Pis 2 was subjected to extensive restriction mapping in parallel with hybridization of its different genomic fragments to Southern blotted human DNA cleaved with various restriction enzymes. From this analysis a highly polymorphic Hind III site was identified (Tsipouras, P. et al., in preparation). This finding allowed us to demonstrate the single copy nature of the human proα1(II) gene by analyzing the segregation of the Hind III polymorphism in informative families. In these experiments, we hybridized total human DNA digested with Hind III to the 2 kb Bam HI:Bam HI subclone of Pis 2, which overlaps the 9.2 kb and 3.7 kb Eco RI fragments (Fig. 3). When the Hind III site was absent in the nuclear DNA as in Pis 2, the probe hybridized to a unique band approximately 13 kb in size. On the other hand, the presence of the Hind III site resulted in two comigrating fragments, 6.7 kb in length (Fig. 6).

At this point we came to the realization that the restriction map of Pis 2 closely resembled that of an "α1-like" collagen gene previously isolated by Weiss et al. from a human cosmid library using a chicken proα1(I) cDNA probe (14). In their report, the authors excluded the possibility that this cosmid clone was the proα1(II) collagen gene based on the sequencing of a selected

```
 1000                                                   1014 1c
                   C   C   T   T   C       T       C       C     A C                   T
▼GGT CCT CCT GGA AAT CCT GGC CCC CCT GGT CCC CCT GGT CCC CCT|GGC CCT GGC ATC GAC ATG TCC
 Gly Pro Pro Gly Asn Pro Gly Pro Pro Gly Pro Pro Gly Pro Pro|Gly Pro Gly Ile Asp Met Ser
                                                                 Thr

               A C G   T  A   CG                         A C  GC         A     A     G
 GCT TTT GCT GGC TTA GGC CCG AGA GAG AAG GGC CCC GAC CCC CTG CAG TAC ATG CGG GCC↓GAC CAG
 Ala Phe Ala Gly Leu Gly Pro Arg Glu Lys Gly Pro Asp Pro Leu Gln Tyr Met Arg Ala Asp Gln
                         Gln Thr                         Ile Arg                     Glu
 30c
           A   G     C G       C      TG           C       C       A               T
 GCG GCC GGT GGC CTG AGA CAG CAT GAC GCC GAG GTG GAT GCC ACA CTC AAG TCC CTC AAC AAC CAG
 Ala Ala Gly Gly Leu Arg Gln His Asp Ala Glu Val Asp Ala Thr Leu Lys Ser Leu Asn Asn Gln
                                     Val

                                 60c
                                         AAG               C A G         C C     A C
 ATT GAG AGC ATC CGC AGC CCC GAG GGC TCC CGC AAG AAC CCT GCT CGC ACC TGC AGA GAC CTG AAA
 Ile Glu Ser Ile Arg Ser Pro Glu Gly Ser Arg Lys Asn Pro Ala Arg Thr Cys Arg Asp Leu Lys
                                         Lys                                     Ile

                                                                 90c
           T   C               C       T                   G       G
 CTC TGC CAC CCT GAG TGG AAG AGT GGA GAC TAC TGG ATT GAC CCC AAC CAA GGC TGC ACC TTG GAC
 Leu Cys His Pro Glu Trp Lys Ser Gly Asp Tyr Trp Ile Asp Pro Asn Gln Gly Cys Thr Leu Asp

       C   A   A                       A           C               G  CC   C AGC  G  A C
 GCC ATG AAG GTT TTC TGC AAC ATG GAG ACT GGC GAG ACT TGC GTC TAC CCC AAT CCA GCA AAC GTT
 Ala Met Lys Val Phe Cys Asn Met Glu Thr Gly Glu Thr Cys Val Tyr Pro Asn Pro Ala Asn Val
     Ile                                                             Thr     Ser Ser Ile

         120c
      G                   C           CG   A   C       G     G            C    G
 CCC AAG AAG AAC TGG TGG AGC AGC AAG AGC AAG GAG AAG AAA CAC ATC TGG TTT GGA GAA ACC ATC
 Pro Lys Lys Asn Trp Trp Ser Ser Lys Ser Lys Glu Lys Lys His Ile Trp Phe Gly Glu Thr Ile
     Arg                 Thr         Thr     Asp             Val         Ala

   C   C   T       C
 AAT GGT GGG TTC CAT▼
 Asn Gly Gly Phe His

                     150c
           C   C       G         T C                 AG  A                             C
▽TTC AGC TAT GGA GAT GAC AAC CTG GCT CCC AAC ACC GCC GAC GTC CAG ATG ACC TTC CTG CGC CTG
 Phe Ser Tyr Gly Asp Asp Asn Leu Ala Pro Asn Thr Ala Asp Val Gln Met Thr Phe Leu Arg Leu
                     Glu         Ser                 Ser Ile

                                                     180c
         A                       G                                 C
 CTG TCC GCC GAG GGC TCC CAG AAC ATC ACC TAC CAC TGC AAG AAC AGC ATT GCC TAC ATG GAC GAG
 Leu Ser Ala Glu Gly Ser Gln Asn Ile Thr Tyr His Cys Lys Asn Ser Ile Ala Tyr Met Asp Glu
         Thr                     Val
                             -----------

                                                                                     210c
  A  A G           G       A     A C               A       C                 A A   C
 GCG GCT GGC AAC CTC AAG AAG GCC CTG CTC ATC CAG GGC TCC AAT GAC GTG GAG ATC CGG GCA GAG
 Ala Ala Gly Asn Leu Lys Lys Ala Leu Leu Ile Gln Gly Ser Asn Asp Val Glu Ile Arg Ala Glu
 Glu Thr

       C               C      GC  T  T   G     C                   C   T   C   A
 GGC AAT AGC AGG TTC ACG TAC ACT GCC CTG AAG GAT GGC TGC ACG▽AAA CAT ACC GGT AAG TGG GGC
 Gly Asn Ser Arg Phe Thr Tyr Thr Ala Leu Lys Asp Gly Cys Thr Lys His Thr Gly Lys Trp Gly
                             Ser Val     Glu

                             240c
       G   G                   G               G      .G       T G A   T           T
 AAG ACT GTT ATC GAG TAC CGG TCA CAG AAG ACC TCA CGC CTC CCC ATC ATT GAC ATT GCA CCC ATG
 Lys Thr Val Ile Glu Tyr Arg Ser Gln Lys Thr Ser Arg Leu Pro Ile Ile Asp Ile Ala Pro Met
                                                                 Val

                                                             270c
       T   C   A G     T       G   T   C       T   T   C   A
 GAC ATA GGA GGG CCC GAG CAG GAA TTC GGT GTG GAC ATA GGG CCT GTC TGC TTC TTG TAA
 Asp Ile Gly Gly Pro Glu Gln Glu Phe Gly Val Asp Ile Gly Pro Val Cys Phe Leu ***
                 Ala Asp
```

Fig. 5. DNA and amino acid sequences of the four exons coding for the C-propeptide domain of the proα1(II) chain. Only the differences with the avian gene are shown. Underlined are the cysteinyl residues and the carbohydrate attachment site (dotted line). The triangles demarcate the exons, the vertical bar separates the triple helical domain from the C-terminal telopeptide. The arrow identifies the C-proteinase cleavage site. The three asterisks emphasize the termination codon for translation.

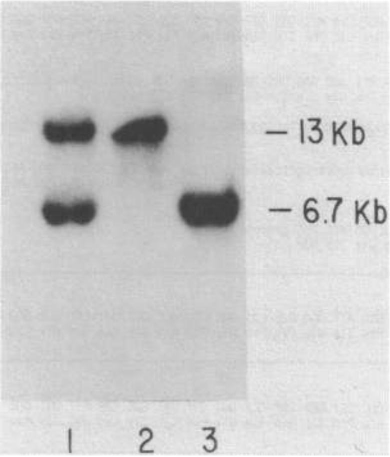

Fig. 6. Southern blotting analysis of the segregation of the Hind III RFLP in an informative family. Lane 1: offspring heterozygous for the presence of the site; lane 2: parent homozygous for the absence of the site, and lane 3: parent homozygous for the presence of the site. The approximate sizes of the hybridizing bands were estimated by comparison with λ DNA digested with Hind III and run in a parallel lane.

area. To demonstrate identity between our clone and the "α1-like" gene we sequenced the same chromosomal region and found complete homology of two exons (numbers 8 and 9) as well as of the intervening sequences (Figs. 4A, B and 7).

Previously, we identified Bc-7 as a Type II collagen specific clone because of its high degree of sequence homology with the C-propeptide domain of the avian gene (Sangiorgi, F. et al., manuscript submitted). Similarly, Strom et al. used this type of evidence to prove that their genomic clone coded for the 3' end of the human proα1(II) gene (13). Although highly suggestive, neither of the investigations provided the nucleotide sequences for those regions of the α1(II) chain whose amino acid sequences have been determined (23,24). Analysis of the 5' foremost section of Pis 2 identified a 54 bp exon, number 29, coding for amino acid residues 361 to 378 (Figs. 4B and 7). The amino acid translation of exon 29 was compared to the published amino acid sequences of the bovine α1(II) chain and found to be identical (24). Moreover, sequencing 5' from exon 29 identified a 60 bp open reading frame, exon 30, whose sequences exhibited a high level of homology with the corresponding segment of the proα1(I) gene. This observation was well in agreement with evidence suggesting that the proα1(I) and proα1(II) collagens are more homologous than are the proα1(I), proα2(I) and proα1(III) collagens

```
       856
tgacag GGT AAT CGT GGT GAA ACC GGT GCT GTG GGA GCT CCT GGA ACC CCT GGG CCC CCT GGC TCC CCT
       Gly Asn Arg Gly Glu Thr Gly Ala Val Gly Ala Pro Gly Thr Pro Gly Pro Pro Gly Ser Pro

                                                        891
GGC CCC GCT GGT CCA ACT GGC AAG CAA GGA GAC AGA GGA GAA GCT gtaagtatcctggaattcactaaaagcc
Gly Pro Ala Gly Pro Thr Gly Lys Gln Gly Asp Arg Gly Glu Ala

gccttcccctgcgcggtggggctgaggcagttcctgggttttcccagtgtctggactaaggagcagtggccccagatgcagaggaggc

                                               892
ccccacctgtcctgcttttctctagcctgcgctcactctctcctcag GGT GCA CAA GGC CCC ATG GGA CCC TCA GGA
                                               Gly Ala Gln Gly Pro Met Gly Pro Ser Gly

                            909
CCA GCT GGA GCC CGG GGA ATC CAG gtgagta
Pro Ala Gly Ala Arg Gly Ile Gln
```

```
   586                                                                603
ag GGC CTG ACA GGT CCC ATT GGC CCC CCT GGC CCA GCC GGT GCT AAC GGC GAG AAG gtgagt
   Gly Leu Thr Gly Pro Ile Gly Pro Pro Gly Pro Ala Gly Ala Asn Gly Glu Lys
```

```
   361                 *                   *                          378
ag GGT CTC ACT GGC CGC CCT GGT GAT GCT GGT CCT CAA GGC AAA GTT GGT CCC TCC gtaagt
   Gly Leu Thr Gly Arg Pro Gly Asp Ala Gly Pro Gln Gly Lys Val Gly Pro Ser
```

```
    341                                                                            360
... CCC AAG GGA GCC AAC GGT GAC CCT GGC CGT CCT GGA GAA CCT GGC CTT CCC GGA GCC CGG gtaagt
... Pro Lys Gly Ala Asn Gly Asp Pro Gly Arg Pro Gly Glu Pro Gly Leu Pro Gly Ala Arg
```

Fig. 7. DNA and amino acid sequences of five triple helical exons. The amino acid residues are numbered above and some of the intervening sequences are shown (small caps). The asterisks signify possible ambiguities.

(25). Thus, these data conclusively proved that the gene coded for the α1 chain of human Type II procollagen.

The same 5' subclone of Pis 2 was used for a further genomic screening which led to the isolation of a clone from the Alu I/Hae III library, Pis 10, and a clone from the Eco RI library, Pis 14. The restriction map of the two recombinants and their relationship to Pis 2 is shown in Fig. 3. Comparison with the map of the "α1-like" gene (14) revealed one major difference. Contrary to the placement by Weiss et al. of the 0.8 kb Eco RI fragment immediately 5' to the 9.2 kb fragment, our restriction mapping suggested that this small Eco RI fragment was 5' to the 7.3 kb Eco RI fragment. In order to resolve this discrepancy, the 0.8 kb Eco RI subclone of Pis 10 was used for the screening of the Alu I/Hae III library which led to the isolation of Pis 93. Southern blotting hybridizations showed that Pis 93 contained the same 0.8 kb fragment of Pis 10, but it did not overlap with the 9.2 kb Eco RI fragment. Thus, this analysis amended the incorrect location of the 0.8 kb fragment as it appeared in the restriction map of the "α1-like" gene (14). Finally, the relative locations of three repetitive sequences within the 45 kb of the overlapping clones were identified by hybridization to nick translated total DNA (Fig. 3).

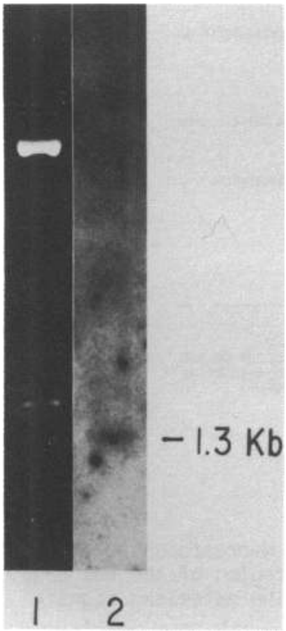

Fig. 8. Identification of exon 52. Left, ethidium bromide staining of the 4.8 kb Eco RI subclone cleaved with Pst I and Kpn I (see also Fig. 3C) and right, autoradiography of the same DNA after Southern blotting hybridization to the primer-extended rat cDNA probe.

Determination of the 5' end of the gene

In order to locate the 5' end of the gene we first hybridized the Eco RI digests of the DNA from Pis 10, Pis 14 and Pis 93 to in vitro labelled total poly(A+) RNA purified from bovine fetal articular cartilage. While the 7.3 kb, 5.2 kb and 4.8 kb fragments gave a positive signal, the 2.0 kb and 0.8 kb fragments were negative (data not shown). Based on the mapping of the three genomic clones we concluded that the 0.8 kb was negative because it contained intervening sequences. On the other hand, the lack of hybridization of the 2.0 kb Eco RI fragment could have been explained by the presence of either intervening or 5' flanking sequences. To discriminate between these two alternatives and to establish the possible location of the 5' most segment of the gene within the 4.8 kb Eco RI subclone, we used a single stranded cDNA coding for the last 200 nucleotides of the rat proα1(II) collagen (12). The rationale for this approach was based on the high sequence homology of the human and bovine proα1(II) collagens as well as the human and the avian genes. Thus, we constructed an artificial 21-mer coding for the heptapeptide which Kohno et al. have found to be located between amino acid residues 12 to 18 in the last exon of the rat proα1(II) gene (12). Hybridization of the rat probe to the 4.8 kb subclone cleaved with different restriction enzymes suggested

```
   29n                34n
    A A   TT                                              * *
ag GG CAA CCA GGA CCA AAG gtaagggctttcttctttttctttttttcgtgtttttttggctttgtgtttcgct
  Arg Gln Pro Gly Pro Lys
      Lys Leu

 ***
cggggcaatgctatgttaatccagtctgtgatttttggacatcgggggggtgtctgtttcgg...4 bases...aacac

aggggctgctggcagggttttagactaggggcttagtgggctactcggcttaatcctgtgaatgtttcatgtttc
                           ----------------------------

   35n                                    45n
    G                             T     A
ag GGA CAG AAA GGA GAA CCT GGA GAC ATC AAG GAT gtaagt
   Gly Gln Lys Gly Glu Pro Gly Asp Ile Lys Asp
```

```
     7n
                                            A       A G     A   AA
..T CCC CAG TCG CTG GTG CTG CTG ACG CTG CTC GTC GCC GCT GTC CTT CGG TGT CAG GGC
... Pro Gln Ser Leu Val Leu Leu Thr Leu Leu Val Ala Ala Val Leu Arg Cys Gln Gly
                                            Ile     Thr             Gln

        28n
         C
CAG GAT GTC C gtaagt
Gln Asp Val
        Ala
```

Fig. 9. DNA and nucleotide sequences of the three N-prepropeptide exons. Shown are only the differences with the corresponding region of the rat gene. The vestigial exon is underlined by the dotted line. The asterisks signify possible ambiguities.

that some sequence similarities were present within the 1.3 kb Pst I:Kpn I fragment of the 4.8 kb subclone (Figs. 4C and 8). Sequencing of this area identified a 67 bp open-reading frame potentially coding for 22 1/3 amino acid residues. Comparison with the rat sequences showed that the human peptide corresponded to amino acid residues 7 to 28 of the rodent N-prepropeptide domain (Figs. 4C and 9). This notion was further supported by the high level of nucleotide (90%) and amino acid (82%) homology observed between the two species in this region of the gene. Based on the remarkable conservation of the exon arrangement of the fibrillar collagen genes, we believe that this exon represents the last coding unit of the proα1(II) gene, for which we tentatively assigned the number 52. From these data we therefore concluded that the 45 kb of contiguous DNA of the overlapping phage clones contained the entire human Type II procollagen gene, which is approximately 30 kb long.

Additional information about the arrangement of the exons coding for the N-propeptide domain was gathered from sequencing around the unique KpnI site of the 7.3 kb Eco RI fragment common to Pis 10 and Pis 93 (Figs. 4D and 9). Two small open-reading frames were found at approximately 120 bp 5' to the Kpn I site and 84 bp 3' to the Kpn I site (Figs. 4D and 9). These two exons, 17 bp and 33 bp respectively, code for a segment corresponding to amino acid

TABLE I

Exon-intron arrangement of the C-propeptide region of the human fibrillar collagen genes

	Exon 1	Intron 1	Exon 2	Intron 2	Exon 3	Intron 3	Exon 4	Exon/Intron Ratio	
								Human	Chicken
Proα1(I)	144	170	243	280	191	210	238	1.5	N.D.
Proα2(I)	144	770	243	900	185	500	214	0.36	0.34
Proα1(II)	144	500	243	350	188	600	244	0.56	0.73
Proα1(III)	144	600	243	300	188	1000	235	0.42	0.45

The sizes of exon 1 refer only to the coding elements. The sizes of exon 4 do not include the C-terminal portion of the triple helical domain. The sizes of most of the introns have been approximated on the basis of R-looping analysis, partial DNA sequencing and restriction mapping of the genomic fragments. The data presented in the table were extracted from several investigations (8,10,15,28,31,24,42).

residues 29 to 45 of the rat N-propeptide. Comparison between the human and rodent gene showed only 7 nucleotide changes resulting in two amino acid replacements. Thus, the interspecies homology for these two exons is identical to that previously discussed for exon 52. In line with those assumptions, the two exons were identified as numbers 50 and 51.

Interestingly, 20 bp upstream from exon 50 we found another open-reading frame, 29 bp long, coding for 3 Gly-X-Y repeats and ending with a split codon (Fig. 9). Intronic sequences coding for collagenous elements have also been identified in the mouse and human proα1(I) collagen genes (26, Ramirez, F., unpublished data). Monson et al. have reasoned that these unusual elements may represent remnants of the complex evolutionary history of the collagen genes (26). Unlike those found in the proα1(I), the vestige of the Type II gene is flanked by accepted variations of the intron splice sequences (27,28). However, although the proα1(II) vestige has the potential to be processed into the mature mRNA, its amino acid sequence translation does not correspond to any segment of the rat N-prepropeptide (12). Furthermore, the translation of this element would put the reading frames of both exon 51 and 50 out of phase. We presently do not have any conclusive explanation for this unusual finding, although one could argue for the lack of other DNA elements which may be necessary for proper exon processing.

DISCUSSION

Proα1(II) C-propeptide

As originally noted by Wozney et al. for the chicken proα2(I) (29), the C-propeptide domain of all fibrillar procollagen genes is divided into four

TABLE II

Codon usage for the human fibrillar collagen α-chains

Third Base	Gly				Pro				Ala			
	α1(I) (261)	α1(II) (38)	α2(I) (166)	α1(III) (58)	α1(I) (179)	α1(II) (32)	α2(I) (92)	α1(III) (35)	α1(I) (32)	α1(II) (12)	α2(I) (55)	α1(III) (12)
U	50%	37%	56%	47%	57%	53%	66%	65%	74%	58%	84%	58%
C	30%	37%	20%	17%	40%	38%	21%	14%	21%	33%	3%	16%
A	18%	24%	13%	31%	3%	9%	10%	25%	5%	9%	7%	25%
G	2%	2%	4%	5%	0%	0%	0%	0%	0%	0%	0%	0%

The values in parentheses indicate the number of codons examined.

exons coding distinct functional or conformational segments of this region of the protein molecule (28). This notion is supported by several lines of experimental evidence suggesting that the C-propeptides play a key role in the selection and correct alignment of the proα chains to form the procollagen trimers (30). In Table I is shown a summary of our data relating to the exon-intron arrangement of the region coding for the C-propeptide domains of the human fibrillar procollagens. Regardless of the size differences of exons 3 and 4, the most striking dissimilarities are seen in the relative exon/intron ratios of the four genes (Table I). Comparison between the human and avian genes revealed almost identical exon/intron ratios in the proα2(I) and proα1(III), but not in the proα1(II) genes. In this context, it should be noted that Strom et al. (13) have reported that in the human Type II collagen gene the distance between a triple helical exon coding for amino acid residues 694 to 711 and the 3' end of exon 4 spans approximately 7.6 kb. This value differs from the 2.9 kb of the avian counterpart and the 4.0 kb we estimated by restriction mapping and DNA sequencing. We believe that this discrepancy is due to a mapping error on the part of these investigators.

Interspecies comparison between the human and the homologous sequences from the bovine and chicken genes indicated a high degree of evolutionary conservation, 96% and 88% respectively. An interesting amino acid change was observed at position 153 (Ser to Ala) in the mammalian C-propeptides when compared to the avian chain (8, Sangiorgi, F. et al., manuscript submitted). This change eliminates the formal possibility that a second glycosylation site (Asn-Leu-Ser) may be present in the Type II C-propeptide. Sequencing of 158 bp of the 3' untranslated region indicated an almost complete identity with the homologous segment of the bovine gene. Based on this finding and on the mRNA sizing by Northern blotting hybridization (Fig. 2), we extrapolated that the total length of exon 1 should not significantly differ from the 580 bp we reported for the calf gene. Interestingly, the proα1(II) gene is the only one among the fibrillar collagens which does not exhibit polymorphic mRNA transcripts (31-34).

Triple helical domain exons

Although the aim of this study was to identify the exact nature of the overlapping genomic clones, sequencing of selected areas of the region coding for the triple helical domain led to an interesting structural finding related to the evolution of the fibrillar collagen genes. The molecular analysis of this subfamily of genes has revealed that they share a remarkably similar arrangement of the numerous exons coding for the triple helical domain (28). Thus far, only one major difference has been identified: the presence of a single 108 bp exon in the proα1(I) gene coding for amino acid residues 568 to 603 (exon 19) which, in the proα2(I) and proα1(III) genes, is encoded by two 54 bp exons (15,26,28,35 and de Wet et al. in preparation). We have now determined that in the human proα1(II) gene the size of exon 19 is also 54 bp. Thus, we concluded that the exon fusion observed in the proα1(I) gene occurred sometime after the duplication and divergence of the four fibrillar collagen genes from the ancestral multiexon unit. In this context we should add that a similar recombinational event has also occurred in the triple helical segment of exon 4 of the proα1(III) gene. In fact, in the human gene and unlike in any of the other collagens, including the chicken proα1(III) gene, this portion of exon 4 has an extra Gly-X-Y triplet resulting in an unusual coding segment 63 bp long (35).

In the various collagens there is a constant pattern, maintained throughout evolution, for the third position preference of the proline, alanine and glycine codons of the triple helical domain exons (28). Thus, we analyzed this feature in the available sequences of the α1(II) chain and compared it to those previously determined for the human α1(I), α2(I) and α1(III) chains, (Table II) (35). Although the number of α1(II) codons examined was relatively small, it is clear that unlike the proα1(III) gene, the proα1(II) does not significantly differ from the proα1(I) and proα2(I) collagens for the first and second choice in the wobble position.

N-terminal domain

The N-terminal portion of the fibrillar collagens has been divided into four distinct subdomains: the signal peptide, an N-terminal globular region, a triple helical segment and a short non-helical part (1). Variations in the size and composition of the N-terminal domains of the collagen molecules have been reported for the different chains as well as for the same chain in different species (1). For example, in the globular domain of the human, mouse and chicken proα1(I) N-terminal region [pNα1(I)], 12, 9 and 7 acidic amino acid residues are present, respectively (15). Similar variations have been reported for the avian and bovine proα1(III) chains (36). In the pNα2(I) the globular region is almost absent, resulting in a significantly shorter N-terminal propeptide (1). As in the proα2(I) chain, Type II collagen lacks the cluster of cysteinyl residues in the globular domain which in the pNα1(I) and pNα1(III) are believed to participate in intrachain disulfide bonds (1,12). Furthermore, the Gly-X-Y repeats of the collagenous region of the pNα1(II), which is much larger than that of the other fibrillar collagen chains, are interrupted once between residues 44 and 47 (12). At the gene level it has been noted that, regardless of these structural differences, six exons are present in the region coding for the collagen N-prepropeptides (15,28). Moreover, the transitions between the four N-terminal subdomains and between the N-terminal propeptide and the signal peptide as well as the large triple helical domain of the collagen, are encoded by junction exons (15,28). Although preliminary, our data seem to suggest that a similar exon distribution is present in the Type II gene. Interestingly, the short non-collagenous segment which interrupts the Gly-X-Y repeats of the pNα1(II) is split between two exons. Investigations are currently in progress aimed at defining in more detail the structure of this portion of the proα1(II) collagen gene.

Gene size and copy number

Although the proα1(I), proα2(I), proα1(II) and proα1(III) collagen genes have maintained a complex structural similarity during evolution, they have diverged from each other both at the level of their nucleotide sequences and in the length of their introns (15,28). It has been argued that, as in the α-like and β-like globin gene clusters (37), the size differences may be the result of the distinct chromosomal origins of the four genes (38, Huerre-Jeanpierre, C. et al., manuscript submitted). Comparative analysis of the proα1(I), proα2(I) and part of the proα1(III) collagens has shown that the sizes of these genes do not seem to vary significantly in different species (15,31,35). It was therefore surprising to find that the total length of the human proα1(II) gene is similar to that of the proα2(I) gene (28, de Wet, W. et al., in preparation), and greater than that reported for the avian counterpart (8). To the best of our knowledge, a similar phenomenon has been observed only in the mammalian dihydrofolate reductase (DHFR) gene (39). The five introns of the DHFR gene are located in homologous positions and show a strong length divergence in man, mouse and Chinese hamster. However, and unlike the proα1(II) collagen, the overall length of the DHFR gene is approximately the same (25-30 kb) in the three species (39).

The familial segregation of a highly polymorphic Hind III site has proved that the proα1(II) gene, which is located on the segment 12q131→12q132 of chromosome 12 (Huerre-Jeanpierre, C. et al., manuscript submitted), is present in a single copy in the haploid human genome. This notion does not seem to substantiate the idea that the minor Type II collagen found in bovine nasal cartilage (40) is indeed a separate gene product, unless one postulates that its sequences significantly diverged from those of the major Type II collagen gene. Conceivably, the biochemical heterogeneity of the α1(II) chains (2,40,41) could also be explained by alternative splicing of a common mRNA precursor. A more detailed comparative analysis of the proα1(II) gene structure in different vertebrates and in different tissues in the same species will provide some insights into this important question.

SUMMARY

In conclusion, we have isolated the entire human proα1(II) collagen gene and found that it is 30 kb in length and present in a single copy in the human haploid genome. Sequencing of selected areas has determined the primary structure and the exon/intron arrangement of the regions coding for the entire C-propeptide domain and part of the N-prepropeptide domain. Pairwise and interspecies comparisons have confirmed some of the evolutionary features of the collagen multigene family. Experiments are currently in progress aimed at completing the characterization of the proα1(II) collagen gene, including its 5' promoter region. These studies are an important prerequisite for elucidating the nature of those factors regulating Type II expression in physiological and pathological conditions.

ACKNOWLEDGEMENT

The authors wish to thank Drs. A. Bank, V. Hascall and T. Maniatis for providing us with some of the materials used in these studies, Dr. Y. Yamada for communicating results prior to publication, and Drs. Chu, Olsen and Weil for many helpful discussions and suggestions. We are also very much in debt to Messrs. M. Bernard, R. Reid and D. Woodbury for their help during the course of this work and Ms. L. Bertha and Ms. K. Valentine for typing the manuscript. This work was supported in part by grants from the National Institutes of Health (AM-32380), the Human Growth Foundation, the Southern African Council for Scientific and Industrial Research and the March of Dimes Birth Defects Foundation.

REFERENCES

1. Bornstein, P. and Sage, H. (1980) Ann. Rev. Biochem. 49, 957-1003.
2. Mayne, R. and von der Mark, K. (1983) in Cartilage. Structure Function and Biochemistry, Hall, B.K., Ed. Vol. 1. pp. 181-214, Academic Press, New York.
3. Horton, W.A. (1984) in Practice of Pediatrics, Kelley, V.C., Ed. Vol. 7. pp. 1-15, Harper and Row, Philadelphia.
4. Stanescu, R., Stanescu, V. and Maroteaux, P. (1982) Collagen Rel. Res. 2, 111-116.
5. Nimni, M. and Deshmukh, K. (1973) Science 181, 751-753.
6. Vuorio, E., Sandell, L., Kravis, D., Sheffield, V.C., Vuorio, T., Dorfman, A. and Upholt, W.B. (1982) Nucleic Acids Res. 10, 1175-1192.
7. Sandell, L.J., Yamada, Y., Dorfman, A. and Upholt, W.B. (1983) J. Biol. Chem. 258, 11617-11621.
8. Sandell, L.J., Prentice, H.L., Kravis, D. and Upholt, W.B. (1984) J. Biol. Chem. 259, 7826-7834.
9. Ninomiya, Y., Showalter, A.M., van der Rest, M., Seidah, N.G., Chretien, M. and Olsen, B.R. (1984) Biochemistry 23, 617-624.
10. Young, M.F., Vogeli, G., Nunez, A.M., Fernandez, M.P., Sullivan, M. and Sobel, M.E. (1984) Nucleic Acids Res. 12, 4207-4228.
11. Lukens, L.N., Frischauf, A.M., Pawlowski, P.J., Brierley, G.T. and Lehrach, H. (1983) Nucleic Acids Res. 11, 6021-6039.
12. Kohno, K., Martin, G.R. and Yamada, Y. (1984) J. Biol. Chem. 259, 13668-13673.
13. Strom, C.M. and Upholt, W.B. (1984) Nucleic Acids Res. 12, 1025-1038.
14. Weiss, E.H., Cheah, K.S.E., Grosveld, F.G., Dahl, H.H.M., Solomon, E. and Flavell, R.A. (1982) Nucleic Acids Res. 10, 1981-1994.
15. Chu, M.-L., de Wet, W., Bernard, M., Ding, J.F., Morabito, M., Myers, J., Williams, C. and Ramirez, F. (1984) Nature 310, 337-340.
16. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) in Molecular Cloning. A Laboratory Manual. pp. 270-294, Cold Spring Harbor Laboratory, New York.
17. Maizels, N. (1976) Cell 9, 431-438.
18. Agarwal, K.L., Brunstedt, J. and Noyes, B.E. (1981) J. Biol. Chem. 256, 1023-1028.
19. Tate, V.E., Finer, M.H., Boedtker, H. and Doty, P. (1983) Nucleic Acids Res. 11, 91-103.
20. Chirgwin, J.M., Przybyla, A.E., MacDonald, R.J. and Rutter, W.J. (1979) Biochemistry 18, 5294-5299.
21. Ramirez, F., Natta, C., O'Donnell, J.V., Canale, V., Bailey, G., Sanguensermsri, T., Maniatis, G.M., Marks, P.A. and Bank, A. (1975) Proc. Natl. Acad. Sci. USA 72, 1550-1554.
22. Maxam, A.M. and Gilbert, W. (1980) Methods in Enzymol. 65, 499-560.
23. Bornstein, P. and Traub, W. in The Proteins. Neurath, H. and Hill, R.L., Eds. Vol. 4 pp. 411-632, Academic Press, New York.
24. Butler, W.T., Finch, J.E. and Miller, E.J. (1977) Biochemistry 16, 4981-4990.
25. Fuller, F. and Boedtker, H. (1981) Biochemistry 20, 996-1006.
26. Monson, J.M., Friedman, J. and McCarthy, B.J. (1982) Mol. Cell Biol. 2, 1362-1371.

27. Breathnach, R. and Chambon, P. (1981) Ann. Rev. Biochem. 50, 349-383.
28. Tate, V., Finer, M., Boedtker, H. and Doty, P. (1982) Cold Spring Harbor Symp. Quant. Biol. 47, 1039-1049.
29. Wozney, J., Hanahan, D., Tate, V., Boedtker, H. and Doty, P. (1981) Nature 294, 129-135.
30. Dickson, L.A., Pihlajaniemi, T., Deak, S., Pope, F.M., Nicholls, A., Prockop, D.J. and Myers, J.C. (1984) Proc. Natl. Acad. Sci. USA 81, 4524-4528.
31. Myers, 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.
32. Chu, M.-L., de Wet, W., Bernard, M. and Ramirez, F. (1985) J. Biol. Chem., in press.
33. Aho, S., Tate, V. and Boedtker, H. (1983) Nucleic Acids Res. 11, 5443-5450.
34. Merlino, G.T., McKeon, C., de Crombrugghe, B. and Pastan, I. (1983) J. Biol. Chem. 258, 10041-10048.
35. Chu, M.-L., Weil, D., de Wet, W., Bernard, M., Sippola, M. and Ramirez, F. (1985) J. Biol. Chem.,in press.
36. Timpl, R. and Glanville, W.R. (1981) Clin. Orthop. Relat. Res. 158, 224-241.
37. Smithies, O., Bleche, A.E., Shen, S., Slightom, J.L. and Vanin, E.L. (1981) in Levels of Genetic Control in Development, pp. 185-200, A. Liss Inc., New York.
38. Huerre, C., Junien, C., Weil, D., Chu, M.-L., Morabito, M., Foubert, C., Myers, J.C., Van Cong, N., Gross, M.S., Prockop, D.J., Boue, A., Kaplan, J.L., de la Chapelle, A. and Ramirez, F. (1982) Proc. Natl. Acad. Sci. USA 78, 6627-6630.
39. Yang, J.K., Masters, J.N. and Attardi, G. (1984) J. Mol. Biol. 176, 169-187.
40. Burgeson, R.E., Hebde, P.A., Morris, P.N. and Hollister, D.W. (1982) J. Biol. Chem. 257, 7852-7856.
41. Burgeson, R.E. and Hollister, D.W. (1979) Biochem. Biophys. Res. Comm. 87, 1124-1131.
42. Yamada, Y., Liau, G., Mudryj, M., Obici, S. and de Crombrugghe, B. (1984) Nature 310, 333-337.