Johnson, M. L., Halvorson, H. R., & Ackers, G. K. (1976) *Biochemistry 15*, 5363-5371.
Johnson, M. L., Turner, B. W., & Ackers, G. K. (1984) *Proc. Natl. Acad. Sci. U.S.A.* (in press).
Kilmartin, J. V., Fogg, J. H., & Perutz, M. F. (1980) *Biochemistry 19*, 3189-3193.
Kwiatkowski, L. D., & Noble, R. W. (1982) *J. Biol. Chem. 257*, 8891-8895.
Makino, J., & Sugita, Y. (1982) *J. Biol. Chem. 257*, 163-168.
Matthew, J. B., Hanania, G. I. H., & Gurd, F. R. N. (1979a) *Biochemistry 18*, 1919-1927.
Matthew, J. B., Hanania, G. I. H., & Gurd, F. R. N. (1979b) *Biochemistry 18*, 1928-1935.
McDonald, M. J., & Noble, R. W. (1972) *J. Biol. Chem. 247*, 4282-4287.
Mills, F. C., & Ackers, G. K. (1979a) *Proc. Natl. Acad. Sci. U.S.A. 76*, 273-277.
Mills, F. C., & Ackers, G. K. (1979b) *J. Biol. Chem. 254*, 2881-2887.
Mills, F. C., Johnson, M. L., & Ackers, G. K. (1976) *Biochemistry 15*, 5350-5362.
Mills, F. C., Ackers, G. K., Gaud, H., & Gill, S. J. (1979) *J. Biol. Chem. 254*, 2875-2880.
Olson, J., & Gibson, Q. (1973) *J. Biol. Chem. 248*, 1623-1630.
Pettigrew, D. W., Romeo, P. H., Tsapis, A., Thillet, J., Smith, M. L., Turner, B. W., & Ackers, G. K. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 1849-1853.
Poyart, C., Bursaux, E., & Bohn, B. (1978) *Eur. J. Biochem. 87*, 75-83.
Rollema, H. S., deBruin, S. H., Janssen, L. H. M., & van Os, G. A. J. (1975) *J. Biol. Chem. 250*, 1333-1339.
Rollema, H. S., deBruin, S. H., & van Os, G. A. J. (1976) *FEBS Lett. 51*, 148-150.
Rossi-Bernardi, L., & Roughton, F. J. W. (1967) *J. Biol. Chem. 242*, 784-792.
Russu, I. M., Ho, N. T., & Ho, C. (1982) *Biochemistry 21*, 5031-5043.
Smith, F. R., & Ackers, G. K. (1983) *Biophys. J. 41*, 415.
Szabo, A., & Karplus, M. (1972) *J. Mol. Biol. 72*, 163-197.
Turner, B. W., Pettigrew, D. W., & Ackers, G. K. (1981) *Methods Enzymol. 76*, 596-628.
Valdes, R., & Ackers, G. K. (1977a) *J. Mol. Biol. 252*, 74-81.
Valdes, R., & Ackers, G. K. (1977b) *J. Biol. Chem. 252*, 88-91.
Valdes, R., & Ackers, G. K. (1978a) *Proc. Natl. Acad. Sci. U.S.A. 75*, 311-314.
Valdes, R., & Ackers, G. K. (1978b) in *Biochemical and Clinical Aspects of Hemoglobin Abnormalities* (Caughey, W. S., Ed.) pp 527-532, Academic Press, New York.
van Assendelft, O. W., & Zijlstra, W. G. (1975) *Anal. Biochem. 69*, 43-48.
Williams, R. C., Jr., & Tsay, K. Y. (1973) *Anal. Biochem. 54*, 137-145.
Wyman, J. (1964) *Adv. Protein Chem. 18*, 223-285.

Structure of the Carboxyl Propeptide of Chicken Type II Procollagen Determined by DNA and Protein Sequence Analysis†

Yoshifumi Ninomiya, Allan M. Showalter, Michel van der Rest, Nabil G. Seidah, Michel Chrêtien, and Bjorn R. Olsen*

ABSTRACT: The complete amino acid sequence of the carboxyl propeptide of chicken type II procollagen has been determined by nucleotide sequencing of three recombinant plasmids harboring inserts complementary to type II collagen messenger RNA (mRNA). In addition, we have characterized a recombinant plasmid containing sequences from the 3'-nontranslated region of type II collagen mRNA. Since the nucleotide sequences did not correspond to regions of chicken type II procollagen for which protein sequence data exist, we have also purified the physiologically cleaved type II carboxyl propeptide from organ cultures of chick embryo sternal cartilages and determined its amino-terminal amino acid sequence by automated Edman degradation. A comparison of the nucleotide-derived sequence with the sequence obtained by Edman degradation of the type II carboxyl propeptide provides definitive proof that the recombinant plasmids contain sequences specific for type II procollagen and allows for the elucidation of the cleavage site for procollagen C-protease within type II procollagen. The results of our sequence analysis indicate that the type II carboxyl propeptide contains 246 amino acid residues. When the peptide is compared with the homologous region of pro α1(I) chains, the type II carboxyl propeptide appears to have an inserted amino acid residue in position 7 (counted from the C-protease cleavage site) and a deleted amino acid residue at position 101. The type II carboxyl propeptide is similar to that of pro α1(I) chains in that it contains eight cysteinyl residues in the same positions, but it is different from the pro α1(I) peptide in that it contains two potential sites for N-linked oligosaccharide side chains while the pro α1(I) peptide contains only one such site.

The collagens are members of a class of structural proteins that are encoded by at least 10 gene loci [for a review, see Bornstein & Sage (1980)]. The modulated expression of these different loci appears to be crucial in normal embryonic development and in tissue repair processes. Isolation of different collagen genes and studies on their regulation are therefore of considerable interest.

The major collagen genes expressed by fibroblasts and osteoblasts are those of type I collagen. This collagen type contains triple-helical molecules composed of two α1(I) chains and one α2(I) chain, and it represents the major collagen type in skin, tendon, ligaments, and bone. In contrast, normal

† From the Department of Biochemistry, UMDNJ-Rutgers Medical School, Piscataway, New Jersey 08854 (Y.N., A.M.S., and B.R.O.), the Genetics Unit, Shriners Hospital, Montreal, Canada H3G 1A6 (M.v.d.R.), and the Clinical Research Institute of Montreal, Canada H2W 1R7 (N.G.S. and M.C.). *Received July 26, 1983.* This study was supported by Research Grant AM 21471 from the National Institutes of Health and by a Johnson & Johnson graduate student fellowship (to A.M.S.).

0006-2960/84/0423-0617$01.50/0

chondrocytes express a set of collagen gene loci distinctly different from those of fibroblasts. The major collagenous protein in hyaline cartilage is type II collagen containing molecules composed of three α1(II) chains (Miller, 1976), but in addition to type II, the extracellular matrix of hyaline cartilage contains several minor collagenous polypeptides. These include 1α, 2α, and 3α chains (Burgeson & Hollister, 1979), G-collagen (Gibson et al., 1982), or short-chain collagen (Schmid & Conrad, 1982a,b) as well as others (Butler et al., 1977; Shimokomaki et al., 1980; Reese & Mayne, 1981; Ayad et al., 1981, 1982; Reese et al., 1982; Ricard-Blum et al., 1982; von der Mark et al., 1982; Gibson et al., 1983). It appears, therefore, that differentiation of chondrocytes involves the synthesis and secretion of collagen polypeptides that are not expressed by most other cell types.

To obtain structural information about cartilage collagens and their procollagen precursors as well as probes for studies on collagen gene expression in chondrocytes, we have synthesized and cloned complementary DNAs (cDNAs)[1] that contain sequences coding for cartilage collagens. Here we report on the isolation and structure of cDNAs specific for chick type II procollagen. A preliminary report has been published elsewhere (Ninomiya et al., 1983).

Materials and Methods

Materials. Phenol, ultrapure urea (enzyme grade), agarose (gel electrophoresis grade), sucrose (ultrapure), CsCl, restriction endonucleases, T4 DNA polynucleotide kinase, S1 nuclease, T4 DNA ligase, and human placental RNase inhibitor were all purchased from Bethesda Research Laboratories. λ-Exonuclease was from New England Biolabs. Oligo(dT)–cellulose (10–18-mer average), poly(A), and *Hin*dIII linkers were obtained from Collaborative Research. [γ-^{32}P]ATP (3000 Ci/mmol), [^{35}S]methionine, and reticulocyte lysate were purchased from Amersham Corp. Avian myeloblastosis virus (AMV) reverse transcriptase was supplied by Dr. J. W. Beard, Life Sciences, Inc., St. Petersburg, FL. Oligo(dA)–cellulose (type 7), *Escherichia coli* RNase H, calf thymus terminal deoxynucleotidyltransferase, *E. coli* DNA ligase, and unlabeled deoxynucleotide triphosphates were purchased from P-L Biochemicals, Inc. [^{3}H]dTTP, [^{3}H]-dGTP, and deoxynucleotide [α-^{32}P]triphosphates were obtained from New England Nuclear. *E. coli* DNA polymerase I was from Boehringer Mannheim, and SeaKem ME agarose was from FMC Corp.

The two pBR322–SV40 recombinants that provided vector and linker DNA fragments were kindly provided by Drs. Hiroto Okayama and Paul Berg (Stanford University). One recombinant contained a SV40 DNA segment corresponding to map position 0.71–0.86 cloned between the *Pvu*II and *Hin*dIII sites of pBR322 DNA. The second recombinant contained a segment from map position 0.19–0.32 of SV40 DNA inserted between the *Bam*HI and *Hin*dIII sites of the pBR322 DNA. The expression plasmid pGW134 and the host *E. coli* K802(λ) were generously provided by Dr. Geoffrey Wilson (New England Biolabs, Beverly, MA). *E. coli* MC1061 was donated by Drs. Keith Mostov and Gunther Blobel (Rockefeller University).

Purification of the Type II Carboxyl Propeptide. The carboxyl propeptide (C-peptide) of type II procollagen was isolated from 17-day-old chick embryo sternal cartilages as described (Pesciotta et al., 1982). The sterna were incubated in Dulbecco's modified Eagle's medium containing [^{35}S]-methionine. After incubation for 24 h, the medium was separated from the sterna by centrifugation, protease inhibitors were added, and the medium was dialyzed at 4 °C against 10 mM Tris-HCl, pH 8.6 (at 25 °C), containing 2 M urea. The dialyzed medium was chromatographed on a DEAE-cellulose column, and the fractions containing the C-peptide were pooled. After rechromatography of the peptide on the DEAE-cellulose column, the C-peptide was desalted by dialysis against 0.5 M acetic acid and lyophilized. The C-peptide was purified for amino acid analysis and amino acid sequencing by preparative polyacrylamide gel electrophoresis in the presence of $NaDodSO_4$. After reduction with 5% 2-mercaptoethanol, the peptide was electrophoresed on a 12% slab gel. The protein bands were localized in the gel by staining with Coomassie Brilliant Blue for 5 min. The band containing the C-peptide was cut out of the gel and placed in a dialysis bag, and the peptide was recovered from the gel slice by electroelution at 400 V for 100 min. For electroelution, the buffer both inside and outside the bag was 12.5 mM Tris–glycine, pH 8.6, containing 0.025% $NaDodSO_4$. The electroeluted material was exhaustively dialyzed against 50 mM NH_4HCO_3, pH 7.8, and 0.025% $NaDodSO_4$. After lyophilization to remove the NH_4HCO_3, the peptide was stored at −20 °C.

Amino Acid Composition and Sequence Analysis. Samples were prepared for amino acid analysis by hydrolyzing peptides in 6 N HCl at 110 °C for 20 h under an atmosphere of nitrogen. The analyses were performed on a Dionex D-500 amino acid analyzer. Tryptophan, cysteine, and methionine were not quantitated.

The amino acid sequence of the purified C-peptide was determined by automated Edman degradation in the Beckman Model 890 C protein/peptide sequencer as described (Seidah et al., 1981). The phenylthiohydantoin derivatives of the amino acids were identified by high-pressure liquid chromatography as described by Lazure et al. (1983).

Preparation of RNA. RNA was extracted from 17-day-old chick embryo sternal cartilages by a modified guanidine hydrochloride method (Adams et al., 1977). Poly(A+) RNA was obtained by oligo(dT)–cellulose chromatography, and the poly(A+) RNA was fractionated by sucrose gradient centrifugation. In a typical experiment, about 30–40 μg of high molecular weight poly(A+) RNA (>28 S) was obtained from 36 dozen sterna.

Cell-Free Translation, cDNA Synthesis, and Transformation. To monitor the isolation of mRNA that directed the synthesis of type II procollagen, the activity of the mRNA was assayed by cell-free translation using a commercial rabbit reticulocyte lysate. Translation products, labeled with [^{35}S]methionine, were analyzed by polyacrylamide slab gel electrophoresis in the presence of $NaDodSO_4$.

In initial experiments, cDNA was synthesized by using a modification of the methods described by Goodman & MacDonald (1979). Briefly, sternal cartilage mRNA was transcribed with AMV reverse transcriptase using oligo(dT) as a primer. The single-stranded cDNA was fractionated on an alkaline sucrose gradient, precipitated with ethanol, and used for second-strand synthesis with AMV reverse transcriptase. The double-stranded cDNA was trimmed with S1 nuclease, fractionated on a sucrose gradient, blunt-end ligated to synthetic *Hin*dIII linkers, and inserted between two *Hin*dIII sites of the expression plasmid pGW134. The recombinant DNA

[1] Abbreviations: C-peptide, carboxyl propeptide; cDNA, complementary DNA; DEAE, diethylaminoethyl; DTT, dithiothreitol; EDTA, ethylenediaminetetraacetic acid; mRNA, messenger RNA; $NaDodSO_4$, sodium dodecyl sulfate; oligo(dT), oligo(thymidylic acid); poly(A), poly(adenylic acid); PTH, phenylthiohydantoin; Tris, tris(hydroxymethyl)aminomethane; bp, base pair(s).

was used to transform *E. coli* K802(λ), and transformants were selected on Luria agar plates supplemented with 50 μg/mL ampicillin.

In later experiments, cDNA was synthesized by using the method of Okayama & Berg (1982). Briefly, primer and linker DNAs were prepared as described (Okayama & Berg, 1982) and tailed with oligo(dT) and oligo(dG), respectively, by using terminal transferase. The length of the primer oligo(dT) tail was about 50 nucleotides while that of the linker oligo(dG) tail was about 15 nucleotides. In typical experiments, 0.2–10 μg of poly(A+) RNA from the sucrose gradient was incubated with 1.3 μg of primer DNA in 15–20 μL of 50 mM Tris-HCl, pH 8.3, 30 mM KCl, 8 mM $MgCl_2$, 0.3 mM DTT, 40 μCi of [α-^{32}P]dCTP, 2 mM each of dATP, dGTP, dCTP, and dTTP, and 10 units of AMV reverse transcriptase. To prevent RNA degradation, human placental RNase inhibitor was added to the reaction mixture at a final concentration of 100 units/mL. cDNA synthesis was initiated by addition of reverse transcriptase and continued at 37 or 42 °C for 30–120 min. The reaction was stopped by adding 2 μL of 0.25 M EDTA, pH 8.0. Following phenol/chloroform/isoamyl alcohol (25:24:1) extraction, the cDNA/RNA hybrid was precipitated with ethanol. The resulting pellet was dissolved in 15 μL of 0.14 M sodium cacodylate, 30 mM Tris-HCl, pH 6.8, 1.5 mM $CoCl_2$, 0.1 mM DTT, 0.2 μg of poly(A), 67 μM [α-^{32}P]dCTP, and 15 units of terminal transferase. The reaction mixture was incubated at 37 °C for 5 min to permit the addition of 10–15 residues of dCMP per 3′ end. This was followed by cleavage with *Hin*dIII, addition of oligo(dG)-tailed linker, and replacement of the RNA strand by using RNase H and DNA polymerase I as described (Okayama & Berg, 1982). Transformation was carried out by using the procedure described by Kushner (1978). *E. coli* K-12 strains HB 101 or MC 1061 were used as hosts, and transformants were selected on Luria agar plates supplemented with 50 μg/mL ampicillin.

Primer-Extension Synthesis of cDNA with λ-Exonuclease. In an attempt to elongate the insert of one of the type II cDNA clones that we isolated (pYN509), we linearized the plasmid with *Bam*HI, a restriction endonuclease that cut at a single site within the 5′ region of the insert. The linearized plasmid was then treated with λ-exonuclease to expose a region of single-stranded DNA at the *Bam*HI site. This was done by incubating 50 μg of linearized plasmid DNA in 500 μL of 67 mM glycine/KOH, pH 9.4, and 25 mM $MgCl_2$, containing 50 units of λ-exonuclease, on ice for 60 min. The reaction was stopped by addition of EDTA and neutralization followed by extraction with phenol and precipitation with ethanol. The λ-exonuclease-treated DNA was then used for cDNA synthesis with poly(A+) RNA from sternal cartilage essentially as described above for the oligo(dT)-tailed primer. The cDNA/RNA hybrid was digested with *Hin*dIII and *Bam*HI, tailed with oligo(dC), and added to oligo(dG)-tailed linker. For the screening of elongated clones, we used a restriction fragment probe that was derived from a region 5′ to the *Bam*HI site in pYN509.

Screening of Recombinant Clones. In initial experiments, DNA prepared from liquid broth cultures of transformants (see below) was digested with appropriate restriction enzymes and electrophoresed on 1% agarose gels to screen for the presence of inserts. DNA from transformants that harbored plasmids with inserts was then further screened by the dot hybridization method of Kafatos et al. (1979) using alkali-fragmented ^{32}P-labeled poly(A+) RNA from 17-day-old chick sterna and calvaria. Alkali degradation of these RNAs was performed by treating RNA with 0.1 N NaOH on ice for 1 h. After neutralization with Tris and precipitation with ethanol, the RNAs were kinased by using [γ-^{32}P]ATP and polynucleotide kinase. The labeled RNA was separated from the free nucleotide by gel filtration on a Sephadex G-75 column.

In subsequent experiments, transformants were screened by colony hybridization as described (Hanahan & Meselson, 1980). Positive colonies were amplified in liquid broth culture, and DNA that was isolated from the amplified transformants was analyzed by dot hybridization or Southern blot analysis.

Isolation of DNA, Restriction Analysis, and Nucleotide Sequencing. For rapid isolation of plasmid DNA, we grew transformants in liquid broth culture for 3–15 h. Plasmid DNA was then isolated by the rapid-boiling method of Holmes & Quigley (1981). For large-scale preparation of DNA, the same method was used in a scaled-up version that included further purification of the DNA by centrifugation on CsCl gradients as described (Holmes & Quigley, 1981).

Digestion of DNA with restriction endonucleases was carried out according to the conditions described by Davis et al. (1980). Sequence analysis of recombinant clones was performed by using the method of Maxam & Gilbert (1980). Sequence data from various experiments were compared, assembled into a concatenated sequence, and examined by using the computer programs of Staden (1979) which were revised for use on the Data General Nova 3 computer.

Results

Purification and Partial Sequence Analysis of the Type II Procollagen C-Peptide. The C-peptide of type II procollagen has been isolated from organ cultures of 17-day-old chick embryo sternal cartilage as a disulfide-bonded trimer of three identical subunits (Curran & Prockop, 1984), each with an apparent molecular weight (M_r) of 35 000 when compared with globular molecular weight standards. We used the same organ culture system here to provide material for amino acid analysis and amino acid sequencing, but we found that purification of the peptide by previously published procedures did not yield a peptide that was pure enough for sequence analysis. To obtain peptide material of sufficient purity, we therefore used slab gel electrophoresis as the final preparative step.

Amino acid sequence analysis of the C-peptide was performed twice. In the first experiment, about 3 nmol of peptide (M_r 35 000) was subjected to 19 degradation cycles (Figure 1). Because of a technical problem during the high-pressure liquid chromatography, we failed to identify the amino acid residue of cycle 6 in this experiment. Therefore, in a second experiment, about 0.7 nmol of peptide (M_r 35 000) was subjected to nine degradation cycles (Figure 1). This second analysis allowed identification of residue 6 as glycine and confirmed the initial sequence determination.

Initial cDNA Synthesis and Cloning Using Oligo(dT) Priming. In initial experiments, cDNA was synthesized and amplified by molecular cloning using methods described by Goodman & MacDonald (1979). For screening, plasmid DNA was isolated from 179 transformants and analyzed by agarose gel electrophoresis. Of those screened, 48 recombinants with inserts of 150–800 base pairs were selected for further analysis by dot hybridization. As probes, we used ^{32}P-labeled alkali-degraded poly(A+) RNA from chick sterna and calvaria. Four recombinants hybridized strongly to sternal RNA but showed no detectable hybridization with calvarial RNA. These four recombinants were analyzed further by detailed restriction endonuclease mapping and nucleotide sequencing. The analysis showed that the four recombinant

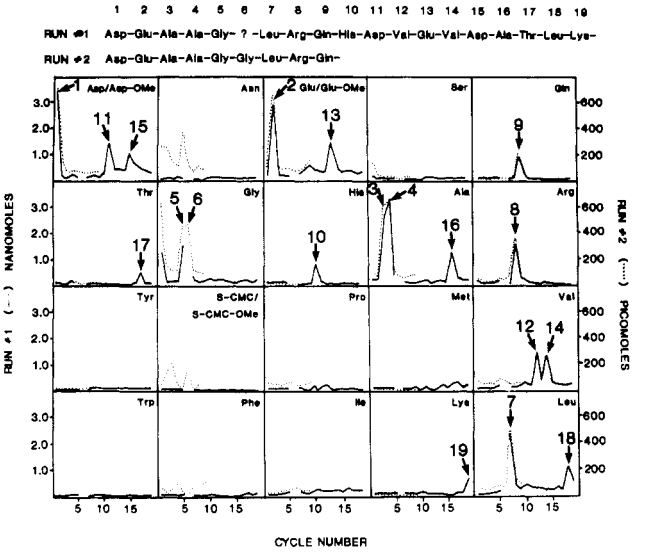

FIGURE 1: Amino acid sequence analysis of the chick type II C-peptide. As outlined under Results, two sequence determinations were performed. In the first determination, about 3 nmol of purified C-peptide was subjected to 19 degradation cycles in the amino acid sequencer. In the second determination, about 0.7 nmol of peptide was subjected to nine degradation cycles. The sequences determined from the two runs are indicated at the top of the figure, together with the cycle numbers. Because of a technical problem, the amino acid residue in cycle 6 could not be determined in run 1. In the figure, the recovery (in nanomoles) of PTH derivatives of each of 20 amino acid residues is given for each cycle. For run 1, the amounts of the PTH derivatives of each residue are indicated along the left-hand side of the diagrams for each amino acid. For run 2, the amounts of the PTH derivatives of each residue are indicated along the right-hand side of the diagrams for each amino acid. The cycle numbers are indicated along the abscissas of the diagrams for each kind of amino acid residue. For each kind of amino acid residue, the points indicating the amount of that residue in each degradation cycle are connected by a solid line (run 1) and by a dotted line (run 2). The assignment of different residues to each of the 19 cycles is indicated by the numbers placed above the peaks in the diagrams for each of the 20 different amino acids.

clones were identical. Only one of them, pYN40 (Figure 2), was therefore used in further experiments.

As reported previously (Ninomiya et al., 1983), several observations led us to tentatively conclude that the insert of pYN40 represented a partial copy of α1(II) collagen mRNA. First, pYN40 hybridized by dot hybridization to a species of RNA present in sternal cartilage (a type II collagen-synthesizing tissue) but absent in calvarial bone (a type I collagen-synthesizing tissue). Second, Northern blot analysis showed that the size of the mRNA that hybridized to pYN40 was consistent with its coding for a polypeptide of pro α1(II) size. Third, pYN40 did not hybridize to RNA from mesenchymal cells derived from 4-day-old chick embryo limb buds, but it did hybridize to RNA extracted from limb bud cells allowed to differentiate into chondroblasts in mass culture (Scott Argraves and Paul Goetinck, unpublished results). For these reasons, we used pYN40 DNA as well as ^{32}P-labeled sternal RNA in subsequent experiments to screen for longer type II collagen cDNAs.

Synthesis of cDNA Using a Plasmid-Primer Method. Synthesis of cDNA according to the method of Okayama & Berg (1982) provided a large number of transformants. In several independent experiments, the number of transformants obtained per microgram of poly(A+) RNA was between 5000 and 50 000. To determine the fraction of transformants that contained inserts, we isolated plasmid DNA, from about 3000 transformants, and estimated the size of their inserts by agarose gel electrophoresis. The results showed that more than

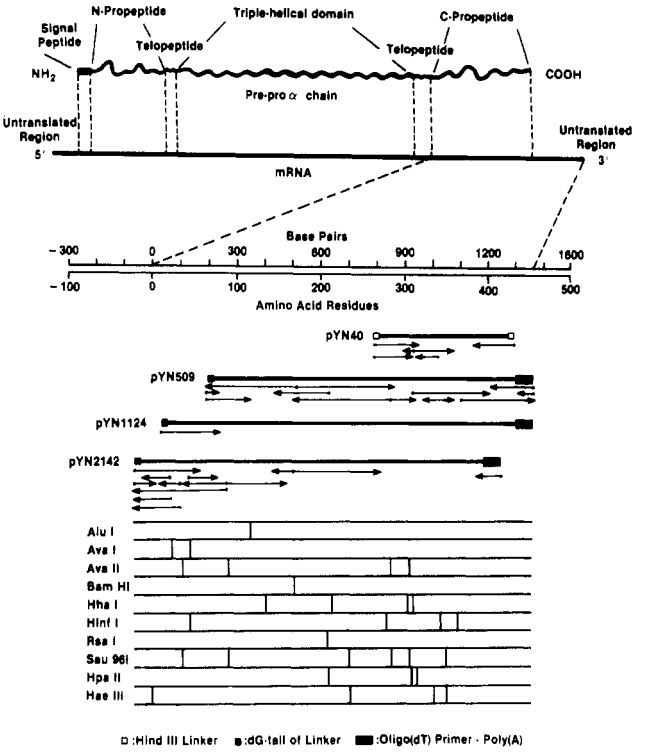

FIGURE 2: Composite restriction endonuclease map of the overlapping pro α1(II) cDNA inserts. The direction of transcription is from left to right, as indicated by the 5′ and 3′ notations. The four clones pYN40, pYN509, pYN1124, and pYN2142 are shown in their positions relative to the restriction map. The strategy of nucleotide sequencing for each insert is indicated with dots representing the position of 5′ end labeling and the arrows showing the direction and length of the sequence analysis. The cDNA inserts produced by the method of Okayama & Berg (pYN509, pYN1124, and pYN2142) were released from the vector by digesting the recombinant plasmids with *Pst*I and *Pvu*II. The insert of pYN40 was removed from the vector with *Hin*dIII. On the basis of the DNA and amino acid sequence analyses, the +1 coordinate was assigned to the first nucleotide/amino acid residue within the C-peptide adjacent to the C-protease cleavage site. Thus, positive numbers refer to nucleotides/amino acids in the C-peptide, the 3′-nontranslated region, and the poly(A) region, while zero and negative numbers refer to nucleotides/amino acids within the carboxyl telopeptide and the triple-helical domain of pro α1(II) chains. At the top of the figure, the positions of the different domains within the translation product [pre-pro α1(II)] of type II mRNA are indicated relative to the mRNA.

80% of transformants obtained with this cloning method harbored inserts.

The initial screening of the clones was by dot hybridization using ^{32}P-labeled poly(A+) RNA from sterna and calvaria or nick-translated pYN40 insert DNA (see above) as probes. Plasmid DNA was isolated from several hundred transformants harboring inserts >500 bp as judged by agarose gel electrophoresis and then used for dot hybridization. One of the clones that hybridized strongly to pYN40 and ^{32}P-labeled sternal cartilage RNA, pYN509, was analyzed further by detailed restricted endonuclease mapping and nucleotide sequencing (Figure 2). The results demonstrated that pYN509 and pYN40 represented partial copies of the same mRNA and moreover that pYN509 contained a sequence that coded for about two-thirds of a procollagen C-peptide, based on homologous amino acid sequence alignment to the pro α1(I) chain sequence (Figure 3).

In order to screen for recombinants with inserts longer than that of pYN509, we used a fragment of pYN509 as a probe in the high-density colony hybridization method of Hanahan & Meselson (1980). Positive colonies were picked and amplified in liquid culture, and plasmid DNA was then isolated

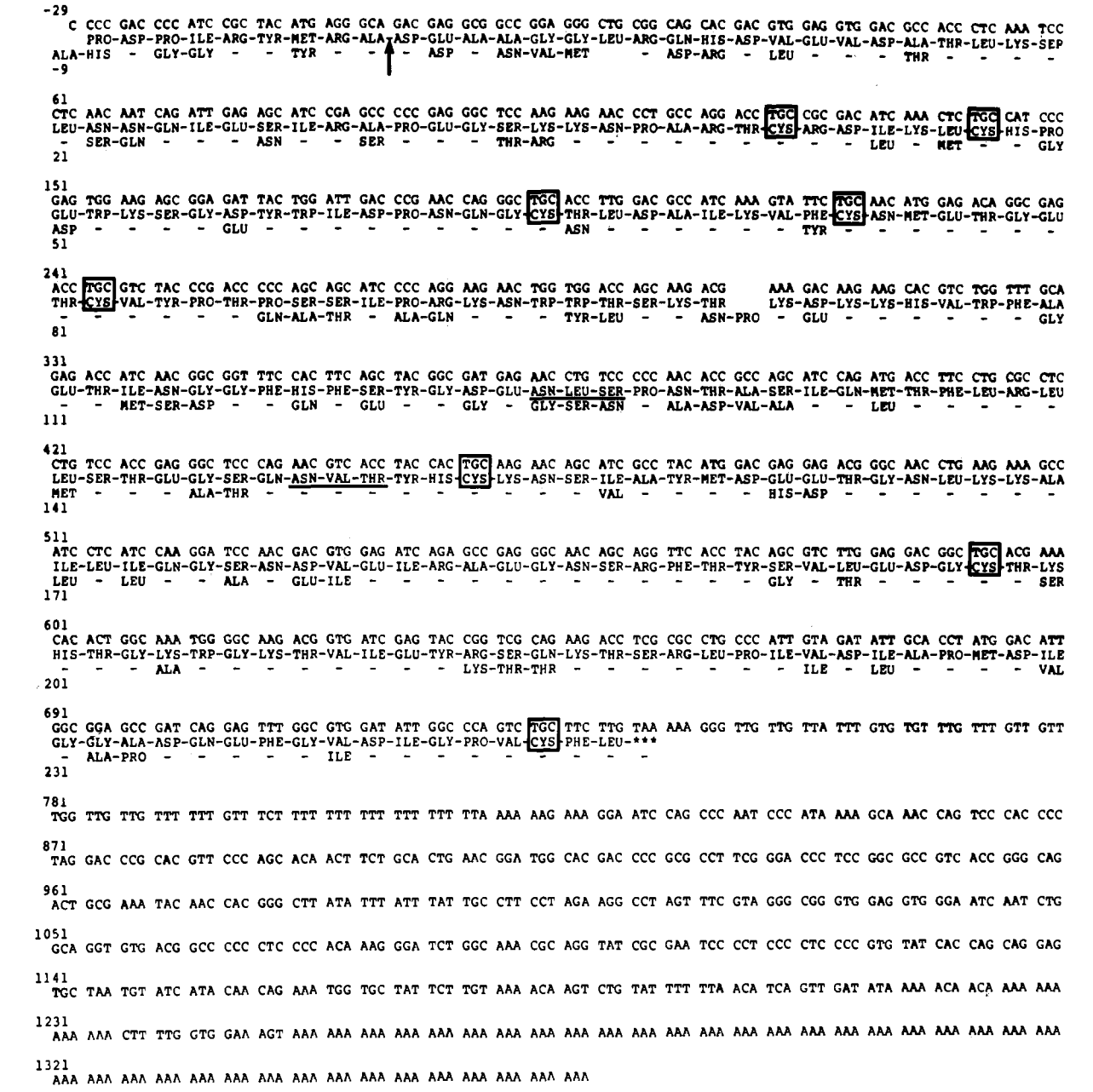

FIGURE 3: Composite nucleotide sequence and the corresponding amino acid sequence of the coding strands of the overlapping pro α1(II) cDNA inserts. For comparison, the homologous pro α1(I) amino acid sequence is also shown (Fuller & Boedtker, 1981). The coordinates are as defined in the legend to Figure 2. The location of cysteinyl residues is indicated by (□). The C-protease cleavage site is indicated by (↑). Potential acceptor sites for N-linked oligosaccharide side chains are underlined. The first in-phase translational termination codon is indicated by (***).

from these cultures. The DNA was used for dot hybridization with a restriction fragment derived from pYN509 as a probe. With this approach, several additional clones were isolated. One of these, pYN2142, was found to extend into the carboxyl telopeptide of α1(II) chains (Figure 2).

Synthesis of cDNA Using λ-Exonuclease and Primer Extension. In an attempt to synthesize cDNAs that would extend into the type II triple-helical region, pYN509 DNA was cleaved at the single *Bam*HI site of its insert and treated with λ-exonuclease, and the linearized exonuclease-treated DNA was then used as a primer for cDNA synthesis. One of the clones isolated by this method, pYN1124, was about 200 bp longer than pYN509, but unfortunately, it did not extend beyond the junction between the carboxyl telopeptide and the C-peptide (Figure 2).

DNA Sequence Analysis. To minimize errors in the nucleotide sequence of the region of type II mRNA that codes for the C-peptide, we determined the sequences from four overlapping cDNA clones. In addition, sequences were determined along both strands of the cDNAs (Figure 2). We did not observe any nucleotide sequence differences among the overlapping clones.

Identification of the Cleavage Site for Procollagen C-Protease. By comparing the amino-terminal amino acid sequence of the type II C-peptide with that derived from nucleotide sequences of the cDNAs, it is evident that the cDNAs described here indeed contain sequences that code for part of type II procollagen. Since the C-peptide isolated from the sternal cartilage organ culture is the product of the physiological conversion of type II procollagen to collagen, we suggest that the amino-terminal sequence of the isolated peptide defines the cleavage site for C-protease. The conversion of type II procollagen to collagen involves, therefore, cleavage of an alanyl–aspartyl bond.

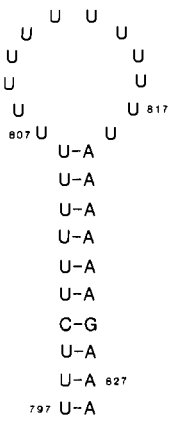

FIGURE 4: Hypothetical stem–loop structure in the 3′-nontranslated region of the chick pro α1(II) sequence. The nucleotide position numbers are defined as for Figures 2 and 3. The stem–loop structure was calculated to possess a ΔG (37 °C) value of −2.3 kcal for the corresponding RNA transcript (Tinoco et al., 1973).

3′-Nontranslated Sequence and Poly(A) Tail. The length of the 3′-nontranslated region of the type II mRNA is 510 nucleotides. In the cDNA, a T-rich sequence starting 52 nucleotides downstream from the translational stop codon contains an internal homology and can be arranged into a stem–loop structure. The location of this potential stem–loop structure corresponds to the 5′ end of the inserts in a number of cDNA clones that we have isolated (data not shown), suggesting that the secondary structure of the mRNA in this region (Figure 4) may cause the reverse transcriptase to "pause" at this point during cDNA synthesis.

The four type II collagen cDNAs reported here all contained a poly(A) tail at their 3′ ends. The length of these tails varied, probably as a result of different sites of hybridization between the oligo(dT) primer and the mRNA during cDNA synthesis.

Discussion

Amino-Terminal Sequence of the Type II Procollagen C-Peptide and Identification of the cDNA Clones. A comparison of the amino-terminal amino acid sequence of the type II C-peptide with those of chicken type I and type III C-peptides (Figure 5) shows that there is a striking variability among the different types of procollagen within this region. Of particular interest is the insertion of an extra amino acid residue (Leu) in position 7 of the type II sequence as compared with those of types I and III. Because of this sequence variability within the first 10–12 amino acid residues of C-peptides, the matching of DNA-derived sequences with protein sequences in this region provides strong evidence that the cDNAs described here are specific for type II procollagen.

What is the significance of the sequence variability in the amino-terminal region of the different types of procollagen C-peptides? Although experimental data are not yet available to definitively answer this question, two extreme possibilities may be considered. One possibility is that the sequence connecting the carboxyl end of the triple-helical domain and the carboxyl propeptide has little or no structural significance and simply serves as a flexible linker between these two domains. Consequently, the carboxyl telopeptide and the first 10–12 amino acid residues of the C-peptide of different procollagen chains may together form a flexible and highly variable domain containing the C-protease cleavage site. The alternate possibility is that the carboxyl telopeptides and the first 10–12 amino acid residues of different procollagen C-peptides have type-specific functions and are therefore different in different procollagens. These sequences could, for example,

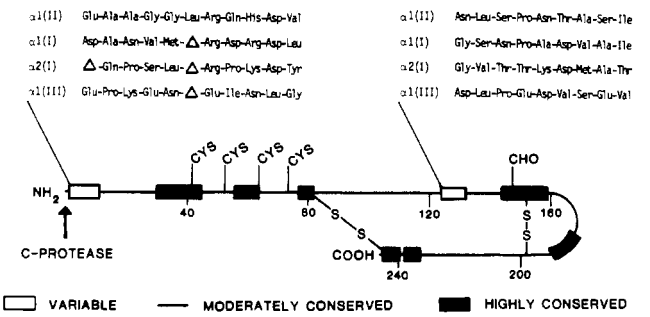

FIGURE 5: Diagram showing variable and conserved amino acid sequence domains within the C-peptides of pro α1(II), pro α1(I), pro α2(I), and pro α1(III) chains. The amino acid sequence data for the pro α1(I) and pro α2(I) C-peptides are from Fuller & Boedtker (1981), and the sequence for the pro α1(III) peptide is from Yamada et al. (1983). Within highly conserved domains, sequences of five or more amino acid residues are identical among the four different peptides. The amino acid sequences for all four peptides within two highly variable domains are shown. The locations of two intrachain disulfide bonds have been determined for the C-peptide of pro α1(I) chains (Showalter et al., 1980; Olsen & Dickson, 1981). Note that the cysteinyl residues are conserved in all the C-peptides except the cysteinyl residue in position 48 which is replaced by a seryl residue in the C-peptide of pro α2(I) chains. The attachment site for an N-linked oligosaccharide common to all four C-peptides is indicated by CHO. The numbers indicate amino acid residues counted from the C-protease cleavage site. In the comparison of the different C-peptide sequences, maximum homology was obtained by assuming amino acid deletions (Δ) in some sequences.

contain information important for the specific associations between procollagen chains during assembly of procollagen molecules in cells.

It has been shown for both type I and type II procollagens that the folding of the collagen triple helix follows the formation of interchain disulfide bonds between the C-peptides (Schofield et al., 1974; Uitto & Prockop, 1974). In addition, if the synthesis of procollagen polypeptides is prematurely terminated with puromycin, it has been observed (Rosenbloom et al., 1976) that polypeptides lacking the cysteinyl residues of the C-peptides are not assembled into triple-helical molecules. It is likely, therefore, that association of C-peptides and formation of interchain disulfide bonds between them are required before the triple-helical domain of procollagen can be formed. With this in mind, it is difficult to understand how a flexible random-coil peptide whose only function is to link the C-peptide with the triple-helical domain can effectively "transmit" the necessary structural information from the assembled C-peptide to the triple-helical domain so that proper alignment of the three polypeptide chains at the carboxyl end of the triple helix is obtained. Therefore, we favor the view that the carboxyl telopeptide and the amino-terminal 10–12 amino acid residues of C-peptides have specific structural roles in the molecular assembly of different types of procollagen.

Cleavage Site of Type II Procollagen C-Protease. The 5′ portion of the insert in pYN2142 contains the sequence that codes for the cleavage site of procollagen C-protease. The site contains the amino acid sequence -Arg-Tyr-Met-Arg-Ala-Asp-Glu-Ala-Ala-Gly. The amino-terminal sequence of the C-peptide, Asp-Glu-Ala-Ala-Gly, defines, therefore, the alanyl–aspartyl bond as the bond cleaved by the enzyme. This bond is the same as the bond cleaved by C-protease in type I procollagen, but different from the arginyl–aspartyl bond that is presumably cleaved by a similar enzyme in type III procollagen.

Amino Acid Coding Sequences of Type II cDNAs. On the basis of the combined sequences obtained from the clones pYN2142, pYN1124, and pYN509, the C-peptide of type II

Table I: Amino Acid Composition of the C-Peptide of Chick Type II Procollagen

amino acid residue	residues/peptide: values predicted from nucleotide sequence	residues/peptide: values determined by amino acid analysis
Asp	29	30.1
Thr[a]	19	16.0
Ser[a]	19	17.8
Glu	25	31.0
Pro	11	13.4
Gly	20	26.5
Ala	13	13.8
Val	12	10.0
Ile	17	13.6
Leu	14	14.7
Tyr[a]	7	6.2
Phe	9	7.7
His	6	4.6
Lys	18	13.8
Arg	10	9.5
	229[b]	228.7[b]
Cys	8	ND[c]
Met	4	ND
Trp	5	ND
	246[d]	

[a] Uncorrected for loss during hydrolysis. [b] Subtotal. [c] ND, not determined. [d] Total.

procollagen contains 246 amino acid residues (Figure 3). The predicted amino acid composition of the peptide agrees well with the amino acid analysis of the C-peptide isolated from sternal cartilage organ culture (Table I). We conclude, therefore, that the coding sequences of the cDNAs we have isolated are accurate copies of the type II procollagen mRNA.

The type II C-peptide contains eight cysteinyl residues. The positions of these residues are conserved when comparing the α1(I), α1(II), and α1(III) C-peptides (Fuller & Boedtker, 1981; Yamada et al., 1983). We have previously shown (Showalter et al., 1980; Dickson & Olsen, 1981) that within the α1(I) peptide the cysteinyl residues in positions 82, 153, 198, and 245 form two intrachain disulfide bonds (Figure 5). The conservation of these residues probably reflects the importance of the intrachain disulfide bonds for the stabilization of the tertiary structure of the C-peptides. The only exception to the rule of cysteine conservation is the replacement of the cysteinyl residue in position 48 by a seryl residue in the pro α2(I) C-peptide (Fuller & Boedtker, 1981; Dickson et al., 1981). This residue may therefore be relatively unimportant for the stabilization of C-peptide structure.

The amino acid sequence predicted for the type II C-peptide contains two potential sites for attachment of an N-linked oligosaccharide side chain. One of the sites starts at position 125 with the sequence -Asn-Leu-Ser-; the second site starts at position 148 with the sequence -Asn-Val-Thr-. Although labeling of the type II C-peptide with radioactive mannose has been reported (Guzman et al., 1978), the degree of glycosylation of the chick type II C-peptide has not been determined. Therefore, we do not know whether both of these sites are utilized. However, since the site at position 148 is conserved among the chick pro α1(I), pro α2(I), pro α1(III), and pro α1(II) C-peptides (Pesciotta et al., 1981; Fuller & Boedtker, 1981; Dickson et al., 1981; Yamada et al., 1983), and this site has been shown to contain an oligosaccharide side chain in the pro α1(I) and pro α2(I) peptides (Pesciotta et al., 1981), it seems reasonable to assume that this site is glycosylated in the type II C-peptide. The potential site at position 125 is only found in the type II C-peptide and is not present in pro α1(I), pro α2(I), and pro α1(III). This site may therefore not be glycosylated in type II procollagen.

In Figure 3, we have compared the type II sequence with that reported for the chick α1(I) C-peptide. To optimize alignment between the α1(II) and the α1(I) sequences, we have inserted a codon triplet for Leu in position 7 of the type II C-peptide and inserted a codon triplet for Pro in position 101 of the type I C-peptide. The amino acid sequences predicted for the two C-peptides show a high degree of homology (71% identical amino acid residues), in agreement with the homology observed in the triple-helical region of pro α(I) and pro α1(II) chains [see Bornstein & Traub (1979)]. This homology is not evenly distributed along the C-peptides. In fact, a comparison of the type II C-peptide sequence with those of α1(I), α2(I), and α1(III) C-peptides indicates that the C-peptides contain several domains of highly conserved amino acid sequences separated by moderately conserved sequence domains (Figure 5). Two domains, one at the amino terminus of the C-peptide and one around amino acid residue 125, show a high degree of variability. As discussed above, it is possible that these variable domains may be important for the type-specific interactions between C-peptide subunits during procollagen assembly in cells.

Nucleotide Sequences of the Type II cDNAs. The composite sequence based on the sequences determined for the inserts of the overlapping clones pYN2142, pYN1124, pYN509, and pYN40 (Figure 3) contains an open reading frame of 766 nucleotides encoding 255 amino acid residues, while the 3′-nontranslated region consists of 510 nucleotides.

A comparison of the restriction endonuclease map of the four overlapping cDNA clones described here (Figure 2) with those of two cDNA clones, pCAR1 and pCAR2, previously reported by Vuorio et al. (1982) clearly indicates that pYN509 encodes a portion of type II procollagen mRNA that is covered by pCAR1 and pCAR2. Specifically, pCAR1 and pCAR2 respectively correspond to approximate nucleotide positions 220–720 and 570–1260 as numbered in Figure 3.

Whereas the length of the type II procollagen C-peptide is identical with that of the pro α1(I) chain, the 3′-nontranslated region of type II mRNA is much larger than that of pro α1(I) mRNA (Fuller & Boedtker, 1981; Showalter et al., 1980). The canonical poly(A) addition signal, AAUAAA, is not found in the type II sequence downstream of the translational stop codon. However, the tetranucleotide AUAA is found to occur twice in the type II 3′-nontranslanted region (nucleotide positions 844–847 and 1210–1213). This tetranucleotide sequence has previously been observed in the 3′-nontranslated region of the mouse dihydrofolate reductase gene, and Setzer et al. (1982) have suggested that this sequence may partly, but not solely, be responsible for polyadenylation. Alternatively, the hexanucleotides TATAAA or ATAAAA in nucleotide positions 1209–1214 or 1210–1215, respectively, could act as signals for polyadenylation. Directly downstream of the TAA termination codon an extremely G/T-rich region is found in the type II cDNAs (Figure 3). The significance of this region is unknown. An A/T-rich sequence involving nucleotides 797–828 (Figure 3) overlaps the G/T-rich region. The A/T-rich region contains an internal homology of 10 nucleotides and can potentially form a stem–loop structure. Since a large number of type II cDNA clones (data not shown) had inserts with 5′ ends ending at or close to this potential stem–loop structure, it is attractive to hypothesize that the secondary structure of the type II collagen mRNA in this region (Figure 4) may cause reverse transcriptase to "pause" and/or "fall off" during the synthesis of cDNA.

Acknowledgments

We gratefully acknowledge the invaluable help of Dr. Eric F. Eikenberry in the computer analysis of nucleotide sequence data; we thank Dr. G. Wilson for the gift of the expression vector pGW134 and the host K802(λ), Dr. K. Mostov for providing the host MC1061, and Dr. H. Okayama for the generous gift of the plasmids used to prepare primer and linker DNA.

Registry No. Procollagen C-protease, 68651-95-6; collagen II (chicken embryo cartilage α-chain), 88343-59-3.

References

Adams, S. L., Sobel, M. E., Howard, B. H., Olden, K., Yamada, K. M., de Crombrugghe, B., & Pastan, I. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 3399–3403.

Ayad, S., Abedin, M. Z., Grundy, S. M., & Weiss, J. B. (1981) *FEBS Lett. 123*, 195–199.

Ayad, S., Abedin, M. Z., Weiss, J. B., & Grundy, S. M. (1982) *FEBS Lett. 139*, 300–304.

Bornstein, P., & Traub, W. (1979) *Proteins (3rd Ed.) 4*, 411–632.

Bornstein, P., & Sage, H. (1980) *Annu. Rev. Biochem. 49*, 957–1003.

Burgeson, R. E., & Hollister, D. W. (1979) *Biochem. Biophys. Res. Commun. 87*, 1124–1131.

Butler, W. T., Finch, J. E., & Miller, E. J. (1977) *J. Biol. Chem. 252*, 639–643.

Curran, S., & Prockop, D. J. (1984) *Biochemistry* (in press).

Davis, R. W., Botstein, D., & Roth, J. R. (1980) *Advanced Bacterial Genetics*, p 227, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Dickson, L. A., Ninomiya, Y., Bernard, M. P., Pesciotta, D. M., Parsons, J., Green, G., Eikenberry, E. F., de Crombrugghe, B., Vogeli, G., Pastan, I., Fietzek, P. P., & Olsen, B. R. (1981) *J. Biol. Chem. 256*, 8407–8415.

Fuller, F., & Boedtker, H. (1981) *Biochemistry 20*, 996–1006.

Gibson, G. J., Schor, S. L., & Grant, M. E. (1982) *J. Cell Biol. 93*, 767–774.

Gibson, G. J., Keilty, C. M., Garner, C., Schor, S. L., & Grant, M. E. (1983) *Biochem. J. 211*, 417–426.

Goodman, H. M., & MacDonald, R. J. (1979) *Methods Enzymol. 68*, 75–90.

Guzman, N. A., Graves, P. N., & Prockop, D. J. (1978) *Biochem. Biophys. Res. Commun. 84*, 691–698.

Hanahan, D., & Meselson, M. (1980) *Gene 10*, 63–67.

Holmes, D. S., & Quigley, M. (1981) *Anal. Biochem. 114*, 193–197.

Kafatos, F. C., Jones, C. W., & Efstratiadis, A. (1979) *Nucleic Acids Res. 7*, 1541–1552.

Kushner, S. R. (1978) in *Genetic Engineering* (Boyer, H. B., & Nicosia, S., Eds.) pp 17–23, Elsevier/North-Holland, Amsterdam.

Lazure, C., Seidah, N. G., Chrêtien, M., Lallier, R., & St-Pierre, S. (1983) *Can. J. Biochem. 61*, 287–292.

Maxam, A., & Gilbert, W. (1980) *Methods Enzymol. 65*, 499–560.

Miller, E. J. (1976) *Mol. Cell. Biochem. 13*, 165–192.

Ninomiya, Y., Showalter, A. M., & Olsen, B. R. (1983) in *Limb Development and Regeneration* (Kelley, R. O., Goetinck, P. F., & MacCabe, J. A., Eds.) Part B, pp 183–192, Alan R. Liss, New York.

Okayama, H., & Berg, P. (1982) *Mol. Cell. Biol. 2*, 161–170.

Olsen, B. R., & Dickson, L. (1981) in *The Chemistry and Biology of Mineralized Connective Tissues* (Veis, A., Ed.) pp 143–153, Elsevier/North-Holland, New York.

Pesciotta, D. M., Dickson, L. A., Showalter, A. M., Eikenberry, E. F., de Crombrugghe, B., Fietzek, P. P., & Olsen, B. R. (1981) *FEBS Lett. 125*, 170–174.

Pesciotta, D. M., Curran, S., & Olsen, B. R. (1982) in *Immunochemistry of the Extracellular Matrix* (Furthmayr, H., Ed.) Vol. 1, pp 91–109, CRC Press, Boca Raton, FL.

Reese, C. A., & Mayne, R. (1981) *Biochemistry 20*, 5443–5448.

Reese, C. A., Wiedemann, J., Kuhn, K., & Mayne, R. (1982) *Biochemistry 21*, 826–829.

Ricard-Blum, S., Hartmann, D. J., Herbage, D., Payen-Meyran, C., & Ville, G. (1982) *FEBS Lett. 146*, 343–347.

Rosenbloom, J., Endo, R., & Harsch, M. (1976) *J. Biol. Chem. 251*, 2070–2076.

Schmid, T. M., & Conrad, H. E. (1982a) *J. Biol. Chem. 257*, 12444–12450.

Schmid, T. M., & Conrad, H. E. (1982b) *J. Biol. Chem. 257*, 12451–12457.

Schofield, J. D., Uitto, J., & Prockop, D. J. (1974) *Biochemistry 13*, 1801–1806.

Seidah, N. G., Rochemont, J., Hamelin, J., Lis, M., & Chrêtien, M. (1981) *J. Biol. Chem. 256*, 7977–7984.

Setzer, D. R., McGrogan, M., & Schimke, R. T. (1982) *J. Biol. Chem. 257*, 5143–5147.

Shimokomaki, M., Duance, V. C., & Bailey, A. J. (1980) *FEBS Lett. 121*, 51–54.

Showalter, A. M., Pesciotta, D. M., Eikenberry, E. F., Yamamoto, T., Pastan, I., de Crombrugghe, B., Fietzek, P. P., & Olsen, B. R. (1980) *FEBS Lett. 111*, 61–65.

Staden, R. (1979) *Nucleic Acids Res. 6*, 2601–2610.

Tinoco, I., Jr., Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M., & Gralla, J. (1973) *Nature (London), New Biol. 246*, 40–41.

Uitto, J., & Prockop, D. J. (1974) *Biochemistry 13*, 4586–4591.

von der Mark, K., van Menxel, M., & Wiedemann, H. (1982) *Eur. J. Biochem. 124*, 57–62.

Vuorio, E., Sandell, L., Kravis, D., Sheffield, V. C., Vuorio, T., Dorfman, A., & Upholt, W. B. (1982) *Nucleic Acids Res. 10*, 1175–1192.

Yamada, Y., Kuhn, K., & de Crombrugghe, B. (1983) *Nucleic Acids Res. 11*, 2733–2744.