ranged for maximal electron transfer. Further detailed experiments might reveal the nature of the reactive groups involved.

Acknowledgments

We appreciate the gift of a preparation of purified cytochrome aa_3 from Dr. C. R. Hartzell and the purification of cytochrome *c* samples by George McLain.

References

Davies, H. C., Pinder, P. B., Nava, M. E., & Smith, L. (1976) *Fed. Proc., Fed. Am. Soc. Exp. Biol. 35*, 1598.

Erecinska, M., Vanderkooi, J. M., & Wilson, D. (1975) *Arch. Biochem. Biophys. 171*, 108–116.

Errede, B., & Kamen, M. D. (1978) *Biochemistry 17*, 1015–1027.

Ferguson-Miller, S., Brautigan, D. L., & Margoliash, E. (1976) *J. Biol. Chem. 251*, 1104–1115.

Ferguson-Miller, S., Brautigan, D. L., & Margoliash, E. (1978a) *J. Biol. Chem. 253*, 149–159.

Ferguson-Miller, S., Brautigan, D. L., Chance, B., Waring, A., & Margoliash, E. (1978b) *Biochemistry 17*, 2246–2249.

Ferguson-Miller, S., Brautigan, D. L., & Margoliash, E. (1979) in *The Porphyrins* (Dolphin, D., Ed.) Academic Press, New York (in press).

Hartzell, C. R., & Beinert, H. (1974) *Biochim. Biophys. Acta 368*, 318–338.

Lee, C. P., & Ernster, L. (1967) *Methods Enzymol. 10*, 543–548.

Margoliash, E., & Frohwirt, N. (1959) *Biochem. J. 71*, 570–572.

Margoliash, E., & Walasek, O. (1967) *Methods Enzymol. 10*, 339–348.

Minnaert, K. (1961) *Biochim. Biophys. Acta 50*, 23–34.

Mochan, E., & Nicholls, P. (1972) *Biochim. Biophys. Acta 267*, 309–319.

Nicholls, P., & Chance, B. (1974) in *Molecular Mechanisms of Oxygen Activation* (Hayashi, O., Ed.) pp 479–534, Academic Press, New York.

Smith, H. T., Staudenmayer, N., & Millett, F. (1977) *Biochemistry 16*, 4971–4974.

Smith, L., & Conrad, H. (1956) *Arch. Biochem. Biophys. 63*, 403–413.

Smith, L., & Camerino, P. W. (1963a) *Biochemistry 2*, 1428–1432.

Smith, L., & Camerino, P. W. (1963b) *Biochemistry 2*, 1432–1439.

Smith, L., Nava, M. E., & Margoliash, E. (1973) in *Oxidases and Related Redox Systems* (King, T. E., Mason, H. S., & Morrison, M., Eds.) Vol. 2, pp 629–638, University Park Press, Baltimore, MD.

Smith, L., Davies, H. C., & Nava, M. E. (1974) *J. Biol. Chem. 249*, 2904–2910.

Smith, L., Davies, H. C., & Nava, M. E. (1976) *Biochemistry 15*, 5827–5831.

Smith, L., Davies, H. C., & Nava, M. E. (1978a) *Fed. Proc., Fed. Am. Soc. Exp. Biol. 37*, 1326.

Smith, L., Davies, H. C., & Nava, M. E. (1978b) in *Japanese–American Symposium on Cytochrome Oxidase*, Kobe, Japan.

Staudenmayer, N., Smith, M. B., Smith, H. T., Spies, F. K., & Millett, F. (1976) *Biochemistry 15*, 3198–3205.

Vanneste, W. H. (1966) *Biochim. Biophys. Acta 113*, 175–178.

Williams, J. N. (1968) *Biochim. Biophys. Acta 162*, 175–181.

Yonetani, T., & Ray, G. S. (1965) *J. Biol. Chem. 240*, 3392–3398.

Construction and Characterization of Pro α1 Collagen Complementary Deoxyribonucleic Acid Clones†

Hans Lehrach,‡ Anna Maria Frischauf,‡ Douglas Hanahan, John Wozney, Forrest Fuller, and Helga Boedtker*

ABSTRACT: Double-stranded cDNA to embryonic chick procollagen mRNAs was synthesized by using the avian myeloblastosis viral reverse transcriptase. After ligation to chemically synthesized decanucleotides containing the *Hin*dIII restriction site, these double-stranded cDNA sequences were inserted into the *Hin*dIII site of pBR322. The recombinant plasmids were then used to transform *Escherichia coli* χ1776 and recombinants containing procollagen cDNA sequences identified by colony hybridization to ^{32}P-labeled procollagen cDNA. In addition to the three pro α2 collagen cDNA clones described recently (Lehrach et al., 1978) three additional recombinant plasmids pCg26, pCgl, and pCg54 with inserts 640, 850, and 1100 base pairs long have been identified. Their sequence homology has been determined by restriction mapping and by DNA sequencing. pCg54 has been positively identified as a pro αl collagen cDNA clone by partial DNA sequencing of its ends: it has a sequence coding for residues 811–858 in the chick α1 chain near one end. pCg54 overlaps pCgl by 250 nucleotides and together these extend about 1500 nucleotides from the poly(A) end of pro α1 collagen messenger RNA.

Collagen is the most abundant protein in vertebrates, constituting 30% of the protein in mammals. The primary and secondary structure of collagen, its biosynthesis, and its assembly into fibrils have been well characterized (Fessler & Fessler, 1978; Miller, 1976). Little is known, however, of the mechanisms regulating the expression of the five or more collagen genes despite the fact that the onset of collagen synthesis is an essential step in differentiation and aberrations in collagen gene expression are the cause of some human diseases such as Osteogenesis Imperfecta and Ehlers Danlos

† From the Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachusetts 02138. *Received February 5, 1979.* This research was supported by National Institutes of Health Grant HD-01229 and a grant from the Muscular Dystrophy Association, Inc.

‡ Present address: European Molecular Biology Laboratory, Heidelberg, West Germany.

0006-2960/79/0418-3146$01.00/0

Type 4 syndrome (Lapière & Nusgens, 1976). A prerequisite to understanding how collagen expression is controlled is the isolation of the genes coding for collagen. The recent advances in recombinant DNA technology have made the isolation of eukaryotic genes a realizable goal. The first step in isolating type I collagen genes was the isolation of the mRNAs coding for chick pro α1 and pro α2 collagen chains (Boedtker et al., 1976). The mRNAs were then reverse transcribed into complementary DNA by using the AMV[1] reverse transcriptase and the cDNA used to determine their purity (Frischauf et al., 1978). To obtain sufficiently pure probes to screen for clones containing collagen genes, we prepared double-stranded cDNA and amplified it in *E. coli* χ1776, using a modification of the procedure used to clone rat proinsulin cDNA (Ullrich et al., 1977). The characterization of three pro α2 collagen cDNA clones obtained has been reported recently (Lehrach et al., 1978). We report here a description of how the recombinant clones were constructed as well as the identification and partial characterization of three pro α1 collagen cDNA clones.

Materials and Methods

Procollagen Double-Stranded cDNA Synthesis. Procollagen mRNAs were partly purified by binding embryonic chick calvaria poly(A)-containing RNA to Sepharose 4B (Frischauf et al., 1978) and then used as a template to prepare double-stranded cDNA by using a procedure developed by Wickens et al. (1978). The 250-mL reaction mixture for the synthesis of the first strand contained 50 mM Tris-HCl, pH 8.3, 10 mM $MgCl_2$, 100 mM KCl, 10 mM DTT, 60 μg/mL oligo(dT_{12-18}) (Collaborative Research, Inc., Waltham, MA), 1 mM each of dATP, dGTP, and dTTP, 200 μM [^{3}H]dCTP (20 Ci/mmol, ICN), 50 units of AMV reverse transcriptase (a gift from Dr. James Beard, Life Sciences, Inc., St. Petersburg, FL), and 50 μg of RNA. After incubation for 1 h at 42 °C, the mixture was diluted with an equal volume of 5 mM DTT, 5 mM Tris-HCl, pH 8.3, and 200 μM dCTP (W. Keller, personal communication) and incubated for another hour at 42 °C. The cDNA reaction mixture was then boiled for 2 min, chilled in ice water, and brought to a final concentration of 100 mM each of dATP, dGTP, dCTP and dTTP, 100 mM Hepes, pH 6.9, and 70 mM KCl. Either 33 Kornberg units of *E. coli* DNA polymerase I or 50 units of the large fragment of polymerase I (both gifts from Yvonne Chow) was added, and the mixture was incubated for 6 h at 15 °C. Double-stranded cDNA was purified by chromatography over Sephadex G-150 in 20 mM NaCl. Excluded fractions were pooled, lyophilized, and ethanol precipitated, and the pellet was dissolved in 100 μL of 10 mM NaCl, 1 mM Tris-HCl, pH 7.6, and 0.05 mM Na_2EDTA. Approximately 5 μg of double-stranded cDNA was obtained from 50 μg of RNA.

Construction of HindIII Restriction Sites. Double-stranded procollagen cDNA was converted into perfect duplex molecules with blunt ends, by digesting with single-stranded specific nuclease S_1 from *Aspergillus oryzae* (purified from amylase (Sigma) as described by Vogt (1973)) and *E. coli* DNA polymerase I (Seeburg et al., 1977). Two micrograms of double-stranded cDNA in 50 μL of 0.3 M NaCl, 0.03 M NaAc, pH 4.5, and 3 mM $ZnCl_2$ was digested with 12 Vogt units of nuclease S_1 for 30 min at 37 °C and 30 min at 23 °C. Three microliters 0.1 M Na_2EDTA was then added, and the mixture was phenol extracted. The organic phase was reextracted with 30 μL of 1 mM Tris-HCl, pH 7.6, 10 mM NaCl, and 0.05 mM Na_2EDTA, and the combined aqueous phases were reextracted with ether and ethanol precipitated. The pellet was dissolved in 8 μL of 60 mM Tris-HCl, pH 7.6, and 8 mM $MgCl_2$; 10 mM mercaptoethanol, 1 mM ATP, 200 μM each of dATP, dCTP, dGTP, and dTTP, and 0.8 units of *E. coli* DNA polymerase I were added, and the mixture was incubated for 15 min at 14 °C.

The product of S_1–DNA polymerase reaction was brought to a final volume of 20 μL containing 0.72 μmol of Tris-HCl, pH 7.6, 120 nmol of $MgCl_2$, 120 nmol of mercaptoethanol, 10 nmol of ATP, and 120 pmol of *Hin*dIII linker (a gift from Richard Scheller) which had been end-labeled with ^{32}P by polynucleotide kinase as described by Heyneker et al. (1976). Twenty-four units of DNA ligase T_4 prepared as described by Panet et al. (1973) was added, and the sample was incubated for 30 min at 20 °C. It was then heated at 80 °C for 7 min, adjusted to 50 mM KCl, and digested with 30 units of *Hin*dIII (Boehringer Mannheim) for 4 h at 37 °C. After adding 4 μL of 50 mM Na_2EDTA, pH 7.0, the sample was phenol and ether extracted as described above.

To separate the ligated double-stranded cDNA from unligated linkers, the sample was sedimented on a 5–20% sucrose gradient in 10 mM Tris-HCl, pH 8.0, 1 mM Na_2EDTA, and 0.1 M NaCl on a 30% sucrose cushion for 8 h at 50 000 rpm in a Beckman SW 56 rotor. Aliquots were Cl_3CCOOH precipitated and counted. Large, medium, and small size fractions were pooled.

Construction of Chimeric Plasmids. pBR322 (obtained from H. Boyer), purified by ethidium bromide–CsCl centrifugation, was extensively digested with *Hin*dIII. After phenol extraction and ethanol precipitation, 5 μg of DNA was dissolved in 50 μL of 10 mM Tris-HCl, pH 8.0, and incubated with 1 unit of calf intestine alkaline phosphatase (Boehringer Mannheim) for 30 min at 37 °C. The DNA was again phenol extracted and alcohol precipitated. It was dissolved in 10 mM Tris-HCl, pH 8.0, and 0.5 mM Na_2EDTA. A two- to tenfold molar excess of this *Hin*dIII cut phosphatase treated pBR322 DNA was added to each of the size fractions of double-stranded cDNA linked to *Hin*dIII half-sites. The molarity of the cDNA was estimated from the chain length and the specific activity of the [^{3}H]dCTP used in the synthesis of the first strand. The DNAs were ethanol precipitated, dissolved in water, adjusted to a final concentration of 8 nM plasmid ends, 60 mM Tris-HCl, pH 7.6, 10 mM $MgCl_2$, 10 mM mercaptoethanol, and 1 mM ATP (ligase buffer), and ligated with 36 units of T_4 DNA ligase/mL. After incubation for 1 h at 14 °C, each solution was diluted tenfold with ligase buffer, the same amount of fresh ligase was added, and the mixture was reincubated for 2 more h (Dugaiczyk et al., 1975).

Transformation, Identification, and Amplification of Recombinant Clones. The construction of the chimeric plasmids described above, the transformation of *E. coli* χ1776 by these plasmids, the identification of recombinant clones, and the amplification of recombinants containing procollagen cDNA sequences were carried out in a P3 physical containment laboratory first at Cold Spring Harbor, NY, and most recently in the Biological Laboratories at Harvard University, in compliance with the *NIH Guidelines* for *Recombinant DNA Research* (1976).

E. coli χ1776 was transformed with fractions of the ligated chimeric plasmids by a procedure suggested by Al Bothwell and first described by Villa-Komaroff et al. (1978) as follows: 200-mL cultures of χ1776 were grown to an optical density

[1] Abbreviations used: AMV, avian myeloblastosis virus; Na_2EDTA, disodium ethylenediaminetetraacetate; $NaDodSO_4$, sodium dodecyl sulfate; Pipes, piperazine-*N*,*N*′-bis(2-ethanesulfonic acid); DTT, dithiothreitol.

at 600 nm of 0.2/mL, centrifuged down, resuspended in 20 mL of 10 mM NaCl, repelleted, and resuspended in 10 mL of 70 mM $MnCl_2$/40 mM sodium acetate (pH 5.6)/30 mM $CaCl_2$. A 0.2-mL portion of the suspension was then added to sterile tubes containing 25 or 100 ng of ligated plasmid. Recombinant plasmids containing procollagen cDNA sequences were identified as described previously (Lehrach et al., 1978).

The colonies which hybridized most strongly to the ^{32}P-labeled procollagen cDNA probe were streaked out on fresh plates; small overnight cultures were grown from single colony isolates and stored in LB containing Hogness' freezing media. DNA was isolated and purified following a modification of the procedure of Tanaka & Weisblum (1975). One liter of *E. coli* χ1776 broth (25 g of Bactotryptone, 7.5 g of yeast extract, 100 g of diaminopimelic acid, 50 g of thymine, 20 mM $MgCl_2$, 50 mM Tris-HCl, pH 7.6) in a 4-L flask was inoculated with 50 mL of a fresh unsaturated culture of χ1776 (grown overnight at room temperature) carrying the recombinant collagen plasmid. The culture was grown at 37 °C with shaking until the OD_{550} reached 0.5–0.8 (about 5 h). Fifty milligrams of chloramphenicol was then added, and the culture was incubated at 37 °C for another 15–20 h with vigorous shaking. The cells were then spun down and resuspended in 5 mL of 50 mM Tris-HCl, pH 8.0, and 25% sucrose, 1 mL of lysozyme (5 mg/mL) was added and the cells were placed on ice for about 10 min. Two milliliters of 0.25 M Na_2EDTA, pH 8.0, was added, and the cells were allowed to remain on ice for an additional 10 min. Finally 2.5 mL of 5 M NaCl, 1 mL of 10% $NaDodSO_4$, and 0.4 mL of 20% poly(ethylene glycol) were added, and the mixture was shaken by hand for 10 s and then incubated at 4 °C for at least 12 h. The viscous suspension was centrifuged for 30 min at 27000 rpm in a Beckman angle 30 rotor. The nonviscous supernatant was carefully removed with a transfer pipet and centrifuged in an ethidium bromide–CsCl gradient. A typical yield was 100 μg of supercoiled plasmid/L of cell culture.

Restriction Mapping. Inserts were released from recombinant plasmids by digestion with *Hin*dIII as described previously (Lehrach et al., 1978). Restriction digestion was carried out with *Hae*III, *Hpa*II, *Eco*RI, *Bam*HI, and *Hin*fI as described previously (Lehrach et al., 1978). Restriction with *Kpn* (New England Biolabs) was carried out in 6 mM Tris-HCl, pH 7.6, 6 mM $MgCl_2$, 6 mM NaCl, 6 mM β-mercaptoethanol, and 100 μg/mL bovine serum albumin; restriction with *Sal*I (Bethesda Research Laboratories) was carried out in the same buffer containing 150 mM NaCl, and restriction with *Ava*I (Boehringer Mannheim) was carried out in 50 mM Tris-HCl, pH 7.6, 6 mM $MgCl_2$, 6 mM β-mercaptoethanol, and 100 μg/mL bovine serum albumin.

Messenger RNA Protection of Procollagen Inserts. Both pCg54 and pCg45 were linearized with *Hin*dIII, ethanol precipitated, resuspended in distilled water, and heated at 80 °C for 1 min to denature the DNA. The length of the DNA protected by hybridization to procollagen mRNA was then determined following a modification of a procedure first described by Berk & Sharp (1977). Two micrograms of chick calvaria poly(A)-containing RNA was hybridized to 0.02 μg of plasmid DNA in the presence of 8 μg of chick rRNA carrier in 80% deionized formamide (Eastman), 0.3 M NaCl, 0.4 mM Na_2EDTA, and 0.01 M Pipes buffer, pH 6.4, for 10–18 h at 54 °C. At the end of the hybridization, an equal volume of ice-cold 100 mM sodium acetate, pH 4.0, 0.3 M NaCl, and 6 mM $ZnSO_4$ was added and the solution incubated at 37 °C for 5 min with about 10 Vogt units of nuclease S_1 (Vogt, 1973).

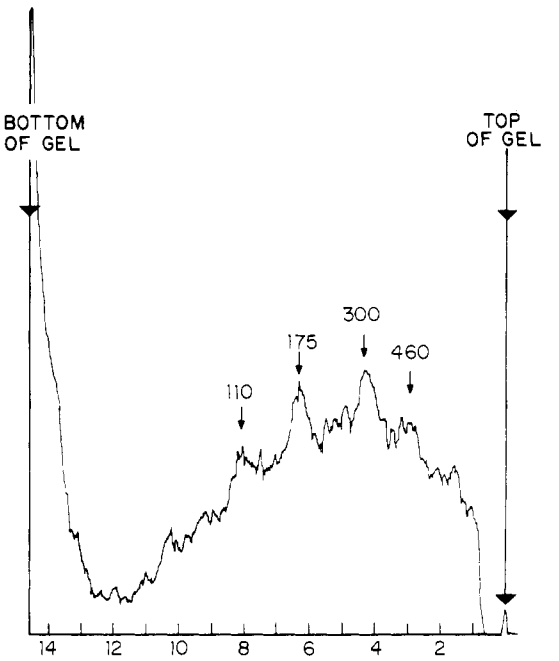

FIGURE 1: *Hae*III restriction fragments of double-stranded procollagen cDNA. Densitometer tracing of autoradiogram of ^{32}P-labeled double-stranded procollagen cDNA digested with *Hae*III endonuclease and electrophoresed on a 10% polyacrylamide slab gel.

The reaction was stopped by fast cooling to 0 °C and bringing the Na_2EDTA concentration to 12 mM. Calf thymus DNA (2 μg) was added, and the mixture was extracted with phenol–chloroform–isoamyl alcohol (50:49:1) and ethanol precipitated. The nucleic acids were resuspended in 0.06 M NaOH and 4 mM Na_2EDTA, electrophoresed on alkaline agarose gels, neutralized, and blotted onto nitrocellulose sheets (Southern, 1975). DNA fragments protected from digestion by nuclease S_1 were identified by hybridization of the nick translated plasmid ($\sim 10^8$ cpm/μg) to the nitrocellulose sheet for 16 h at 65 °C. After repeated washing the sheets were air-dried and autoradiographed.

DNA Sequence Determination. DNA fragments were sequenced following the procedure of Maxam & Gilbert (1977). The pCg1 and pCg26 inserts were 5′ end labeled with polynucleotide kinase and digested with *Hin*fI, and the resultant fragments were isolated by preparative polyacrylamide gel electrophoresis. pCg54 was labeled at its unique *Hin*dIII or *Hin*fI end and then restricted with *Ava*I or *Kpn*, respectively, and the resultant uniquely end labeled fragment was isolated by preparative polyacrylamide gel electrophoresis. The fragments were sequenced by using the following chemicals for cleavage: G, dimethyl sulfate/piperidine; A + G, formic acid, pH 2.0, depurination/piperidine (Alan Maxam, personal communication); C + T, hydrazine/piperidine; C, hydrazine plus NaCl/piperidine.

Results

Transformation of E. coli χ1776 by Recombinant of pBR322 and Double-Stranded Procollagen cDNA Sequences. Double-stranded procollagen cDNA was synthesized with AMV reverse transcriptase and *E. coli* DNA polymerase I by using partly purified embryonic chick calvaria procollagen mRNAs as templates, as described under Materials and Methods. Double-stranded cDNA in which the second strand was labeled with ^{32}P was digested with *Hae*III and the product analyzed on polyacrylamide slab gels. Four distinct bands corresponding to fragments 460, 300, 175, and 110 base pairs long were clearly visible against a heterogeneous background

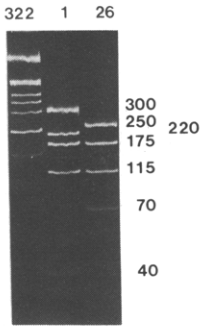

FIGURE 2: *Hae*III restriction fragments of pCg1 and pCg26. Photograph of ethidium bromide stained 8% polyacrylamide gel. Left lane: *Hin*fI fragments of pBR322, in order of size, 1631, 516, 506, 396, 344, 298, 220, 154 (Sutcliffe, 1978).

on autoradiograms of the gels, as shown in Figure 1. These fragments are certainly likely to have been derived from the only high abundance sequences in the cDNA, namely, procollagen sequences. Single chain fragments having the same number of bases could also be seen when single-stranded cDNA was digested with *Hae*III (data not shown); the presence of these *Hae*III fragments was in fact used as an assay to optimize the conditions to be used for the specific synthesis of the second strand.

Double-stranded cDNA enriched for procollagen sequences as described above was treated with nuclease S_1 and *E. coli* DNA polymerase I to open the hairpin structure at one end of these molecules and to make perfect duplexes with blunt ends. The double-stranded cDNAs were then ligated with DNA ligase T_4 to chemically synthesized decanucleotides containing *Hin*dIII restriction sites and end labeled with ^{32}P. The ligated cDNAs were digested with *Hin*dIII to create cohesive single-stranded ends on the double-stranded cDNA, and to digest the excess polymerized ligated linkers. Next, the double-stranded cDNA was separated from excess linker molecules by sucrose gradient centrifugation. Three size fractions ranging in size from 100 to 3000 base pairs were pooled and inserted into pBR322 which had been cut at its single *Hin*dIII site and treated with alkaline phosphatase to remove the 5′ terminal phosphates and prevent self-ligation. The resultant recombinant DNA plasmids were then used to transform *E. coli* χ1776, as described under Materials and Methods.

The most strongly hybridizing colonies obtained from transformations with the largest size fraction of nuclease S_1–*E. coli* polymerase I treated double-stranded cDNA consisted of two different classes: the larger ones, pCg10, pCg13, and pCg45, which contain pro α2 collagen cDNA sequences (Lehrach et al., 1978), and two smaller ones, pCg1 and pCg26, which had inserts 850 and 640 base pairs long. Since nick-translated pCg10 and pCg45 did not hybridize to Southern gel blots of these inserts, the latter have no sequence homology with the pro α2 collagen clones. Therefore, we assumed that either they were derived from the 5′ end of pro α2 mRNA or they contained pro α1 cDNA sequences.

As an initial step in identification of pCg1 and pCg26, the inserts were released by digestion with *Hin*dIII, and the isolated inserts were then restricted with *Hae*III. The photograph of the ethidium bromide stained gel obtained after polyacrylamide gel electrophoresis is shown in Figure 2. pCg1 has five *Hae*III fragments, 300, 220, 175, 115, and 40 base pairs long, while pCg26 has four *Hae*III fragments, 250, 175, 115, and 70 base pairs long. Both, thus, have in common two fragments, 175 and 115 base pairs long, very similar in size to two of the prominent *Hae*III fragments in the double-stranded cDNA which was cloned (see Figure 1). Since pro α1 and pro α2 collagen cDNA sequences were the only abundant sequences in the double-stranded cDNA that was cloned (Frischauf et al., 1978), pCg1 and pCg26 can be expected to contain collagen sequences.

Table I: Restriction Fragments of pCg1 and pCg26 Inserts[a]

	pCg1	pCg26		pCg1	pCg26
*Hae*III	300	250	*Hpa*II	250	
	220*	175		240*	210 (2)
	175	115		210	110
	115	70*		110	100*
	40*	<25*		40*	<25*
*Hin*f1	500	430	*Kpn*	650	510
	350	210		200	130

[a] *Hae*III and *Hpa*II 5′ end fragments are marked with an asterisk.

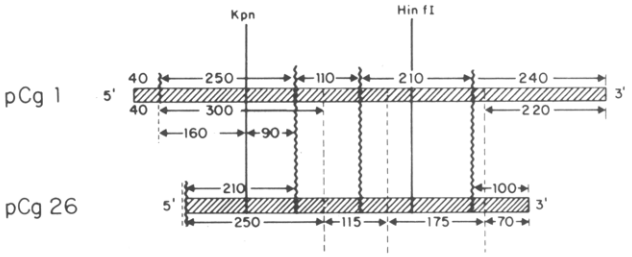

FIGURE 3: Restriction map of pCg1 and pCg26. 5′–3′ orientation determined from primary sequence data given in the text. (§) *Hpa*II sites; (¦) *Hae*III sites.

Restriction Mapping of pCg1 and pCg26. The isolated inserts from pCg1 and pCg26 were end labeled with ^{32}P and restricted with *Hpa*II, *Hin*fI, *Eco*RI, *Bam*HI, and *Kpn*. Both inserts had a single *Hin*fI and *Kpn* site but were not digested by either *Eco*RI or *Bam*HI. Both also appeared to have four *Hpa*II sites. The sizes of the fragments as well as the identification of the end fragments are given in Table I. pCg1 and pCg26 inserts have two internal *Hpa*II fragments in common, 210 and 110 base pairs long. This suggests that about half of pCg26 may have sequences in common with pCg1. Only one end-labeled fragment was detected in both the *Hae*III and the *Hpa*II digests of pCg26. Therefore, we suspected that this insert had a *Hae*III and *Hpa*II site very close to one end, resulting in an end fragment of less than 25 base pairs which ran off the gel.

By codigesting both inserts with *Hin*fI and *Kpn*, and with *Hin*fI and either *Hae*III or *Hpa*II, the location of each fragment was established, resulting in the restriction map shown in Figure 3. The 5′–3′ orientation and the definitive identification of a *Hae*III and *Hpa*II site at the 5′ end of pCg26 were determined from DNA sequence data described below. Since there is no *Hae*III or *Hpa*II site at this location in pCg1, there is a real discrepancy between the sequences of these two inserts.

Primary Sequence of the 5′ Ends of pCg1 and pCg26. To determine the 5′–3′ orientation of the two recombinant plasmids and to investigate the discrepancy at one of their ends, the DNA sequences at the ends of pCg1 and pCg26 were determined. After end labeling with ^{32}P and digesting with *Hin*fI, the fragments obtained from both inserts were separated by polyacrylamide gel electrophoresis, eluted from the gel and sequenced as described by Maxam & Gilbert (1977). The 350 base pair *Hin*fI fragment of pCg1 has 38 uninterupted T's at its 5′ end and this must be derived from the poly(A) at the 3′ end of the mRNA (data not shown). The sequence at the 5′ end of the 500 base pair *Hin*fI fragment of pCg1, corresponding to coding strand, is shown in Figure 4 (sequence gel supporting Figure 4 can be seen in supplementary material).

```
AGCTTGG   ATC   CGC   AGC   CCC   GAA   GGC   ACC
            10                    20
          ile   arg   ser   pro   glu   gly   thr
             pro   gln   pro   arg   arg   his   pro
                                 ↓
          CGC   AAG   AAC   CCG   GCC   CGC   ACC
           30                   ↑
          arg   lys   asn   pro   ala   arg   thr
             gln   glu   pro   gly   pro   his   leu

          TGC   CGG   GAC   CTG   AAG   ATG   TGC
          50    ↑             60                70
          cys   arg   asp   leu   lys   met   cys
             pro   gly   pro   glu   asp   val   pro

          CAC   GGC   GAC   TGG   AAG   AGC   GGC
                              80                90
          his   gly   asp   trp   lys   ser   gly
             arg   arg   leu   glu   glu   arg
```

FIGURE 4: DNA Sequence at 5′ end of the coding strand of pCg1 and two amino acid sequences for which it might code. (↓) *Hae*III sites; (↑) *Hpa*II sites.

This sequence contains two *Hpa*II sites, following the 39th and 53rd residue, and one *Hae*III site, following the 41st nucleotide. Although only one *Hpa*II and one *Hae*III site were expected in this sequence, the 13 base pair *Hpa*II fragment is too small to have been detected on the gel. Seventy percent of the bases in this sequence are G or C, making it surprisingly GC rich.

While the reading frame of this sequence is not known, one of the three possibilities can be ruled out because nucleotides 60–62 are the termination codon UGA. One of the other two polypeptides for which it could code, shown in Figure 4, is part of the C-terminal propeptide whose sequence has not been determined (P. Fietzig, private communication).

The sequence at the 5′ end of the pCg26 insert had two *Hae*III sites, 7 and 16 base pairs from the end, and two *Hpa*II sites, 16 and 18 base pairs from the end. No *Hae*III or *Hpa*II sites are present in the corresponding pCg1 sequence. Therefore, the 5′ end of one of these inserts cannot be colinear with pro α1 collagen mRNA. The coincidence of the other eight restriction sites, shown in Figure 3, certainly strongly suggests that the pCg1 and pCg26 inserts have sequence homology from their *Kpn* site to their 3′ ends.

Identification of a New Procollagen Clone. The discrepancies between the 5′ sequences of the pCg1 and pCg26 inserts required our finding a third recombinant clone containing an overlapping pro α1 collagen cDNA sequence to enable us to establish which sequence was the correct one. We also wanted to obtain a pro α1 recombinant clone that extends into the triple helical coding region of collagen. To accomplish this, the nitrocellulose filters on which the collagen transformants had been replicated were rescreened with ^{32}P-labeled procollagen cDNA that was over 80% pure (Frischauf et al., 1978) which had been prehybridized to saturating amounts of unlabeled pCg45 and pCg1. One strongly hybridizing colony was located; it represented a clone previously selected and named pCg54 but not investigated further because the insert could not be released by *Hin*dIII. When hybridized to either nick translated pCg1 or pCg45, only the former had any sequence homology with pCg54. Therefore, pCg54 was believed to be a pro α1 collagen clone and one expected to extend further toward the 5′ end of the mRNA because of the way it was selected.

Restriction Mapping of pCg54. Since pBR322 has a single *Sal*I site 621 base pairs from its only *Hin*dIII site, the pCg54 insert was released and its size measured by digesting with *Hin*dIII and *Sal*I. A fragment about 1700 base pairs long was obtained, suggesting that the insert is about 1100 base pairs long. pBR322 also has a single *Ava*I site, 2966 and 1395 base pairs from its *Hin*dIII site (Sutcliffe, 1978). When pCg54 was digested with *Hin*dIII and *Ava*I, three fragments, 2966, 1800, and 700 base pairs long, were obtained as shown in Figure 5, lane 4. The existence of two smaller fragments places an *Ava*I site in the pCg54 insert, about 700 base pairs from its *Hin*dIII end. The 1800 base pair *Ava*I fragment, therefore, contains 1395 base pairs of pBR322 and about 400 base pairs of the pCg54 insert. This confirms a length of about 1100 base pairs for the pCg54 insert. If, as predicted, most of this sequence extends beyond the 5′ end of pCg1, this sequence should contain some of the triple helical coding region of the pro α1 chain.

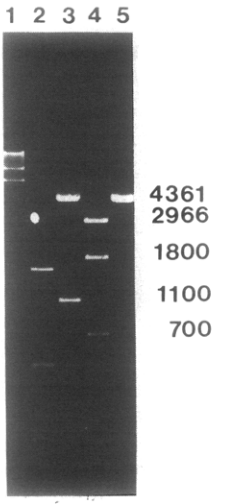

FIGURE 5: pCg54 insert released by *Hin*dIII and *Kpn* restriction. (Lane 1) *Hin*dIII fragments of λ as size markers. The six visible fragments are 23.7, 9.5, 6.6, 4.4, 2.3, and 2.0 kb. (Lane 2) *Hin*f fragments of pBR322. Sizes are as in the legend to Figure 2. (Lane 3) pCg54 digested with *Hin*dIII and *Kpn*I. (Lane 4) pCg54 digested with *Hin*dIII and *Ava*I. (Lane 5) pBR322 digested with *Hin*dIII and *Kpn*I. One percent agarose gel.

To determine the extent of overlap between the pCg1 and pCg54 inserts, pCg54 was digested with *Hin*dIII and *Kpn*. Since pCg1 has a single *Kpn* site 200 base pairs from its 5′ end, pCg54 might have a *Kpn* site near its 3′ end. We therefore expected that either a very small fragment or one about 1100 base pairs long would be released. The latter proved to be correct, as shown in Figure 5, lane 3. The *Hin*dIII–*Kpn* fragment includes almost all of the procollagen cDNA sequence in pCg54 since it is 1100 base pairs long. The *Kpn* site must be at the 3′ end of the pCg54 insert, because it has a single *Hin*fI site 1000 base pairs from its *Kpn* site. Both pCg1 and pCg26 have a *Hin*fI site 300 base pairs from their *Kpn* site (see Figure 3).

The location of the *Kpn* site in pCg54 was also confirmed by codigesting this plasmid with *Kpn* and either *Hae*III or *Hpa*II. In both cases, a 160 base pair fragment is generated which is identical in size with the 160 base pairs between the *Kpn* site and the *Hae*III site and *Hpa*II site at the 5′ end of pCg1 (see Figure 3); it is larger, however, than the 120 base pairs separating these sites at the 5′ end of pCg26 and, therefore, this fact confirms that the sequence at the 5′ end of pCg1 is correct, while that of pCg26 is not.

When the insert released from pCg54 by restricting with *Hin*dIII and *Kpn* was digested with *Hpa*II and *Hae*III, the 160 base pair fragment described previously was obtained together with multiple smaller fragments, each of which was about 9 base pairs or a multiple of 9 base pairs larger than the next. A similar set of fragments had previously been obtained after *Hae*III digestion of *Hpa*II fragments of pCg45 coding for the triple helical region of the pro α2 collagen chain

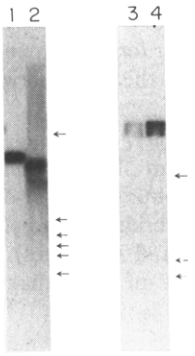

FIGURE 6: Autoradiogram of pCg54 fragment and pCg45 insert protected from S_1 nuclease by hybridization to procollagen mRNA. (Lane 1) pCg54 *Hin*dIII–*Kpn* insert, control; (lane 2) pCg54 fragment protected by mRNA; (lane 3) pCg45 fragment protected by mRNA; (lane 4) pCg45 insert, control. Arrows indicate location of *Hin*fI fragments of pBR322 from top to bottom, 1631, 516/501 doublet, and 396 bases. Lanes 1 and 2: 1.4% alkaline agarose gel. Lanes 3 and 4: 1.0% alkaline agarose gel. Electrophoresis was for 15 h at 27 V.

(Lehrach et al., 1978) and appears to be a fingerprint of this region which contains Gly-X-Y triplets resulting in a *Hae*III site every time a glycine codon with a C in its third position is followed by a proline codon.

The *Hin*dIII end of the *Hin*dIII–*Kpn* fragment was expected to contain the triple helical coding region of the chick pro α1 chain and was therefore sequenced. Contrary to our expectations, the sequence obtained did not have the characteristic pair of guanines separated by seven nucleotides required by the $(Gly\text{-}X\text{-}Y)_n$ collagen sequence. The only explanation for this result is that the pCg54 insert was not wholly colinear with pro α1 collagen mRNA. This raised a major question of whether enough of the insert was colinear to make this a useful clone. To test this, messenger RNA protection experiments were carried out.

Procollagen mRNA Protection Experiments. pCg54 and pCg45 were linearized by digestion with *Hin*dIII and hybridized to procollagen mRNA as described under Materials and Methods. After digestion with nuclease S_1, the hybrid was electrophoresed on alkaline agarose gels, blotted to nitrocellulose and hybridized to the appropriate nick translated recombinant plasmid. The autoradiogram of the results obtained with pCg54 is shown in Figure 6. One thousand base pairs or almost 90% of the pCg54 insert was protected by hybridization to procollagen mRNA. This must be compared with the 100% protection of pCg45 obtained in a parallel experiment (shown in Figure 6). These results suggest that the *Hin*dIII end of pCg54 is derived from some other part of procollagen cDNA, but that most of this insert is colinear with procollagen mRNA and the collagen sequence should be found 100–200 base pairs from the *Hin*dIII end. Since we have previously located a single *Hin*fI site 140 base pairs from the *Hin*dIII end of the pCg54 insert, the *Hin*dIII–*Kpn* insert was digested with *Hin*fI, and the large 1000 base pair fragment was isolated and its 5′ end sequenced.

The nucleotide sequence obtained is shown in Figure 7 (sequence gels supporting Figure 7 can be seen in supplementary material), together with the amino acids for which it codes. These correspond to residues 811–858 in the chick α1 collagen chain (Fietzek & Kuhn, 1976). Therefore, 600 nucleotides of pCg54, or more than 50% of this insert, contain sequences coding for the triple helical region of the chick α1 collagen chain. This sequence contains multiple *Hae*III and *Hpa*II sites as expected. It is also remarkably GC rich, three-fourths of its nucleotides being either G or C.

GGT	GAA	CGC	GGT	CCT	CCC
gly	glu	arg	gly	pro	hyp↑
811					
GG↓C	CCC	ATG	GGX	CCC	CCC
gly	pro	met	gly	pro	hyp↑
	818				
↓GGC	CTT	GCT	↓GGC	CCC	CCT
gly	leu	ala	gly	pro	hyp
		825			
GGT	GAA	GCT	GGA	CGT	GAG
gly	glu	ala	gly	arg	glu
			832		
GGT	GCT	CCC↑	GGT	GCC	GAA
gly	ala	hyp	gly	ala	glu
				839	
GGT	GCC	CCC↑	GGT	CGC	GAC
gly	ala	hyp	gly	arg	asp
					846
GGT	GCT	GCC↑	GGT	CCC	AAG
gly	ala	ala	gly	pro	lys
GGT	GAC	CGT	GGT	GAA	ACY
gly	asp	arg	gly	glu	thr
853					858

FIGURE 7: DNA sequence following *Hin*fI restriction site near 5′ end of pCg54 and the amino acid sequence for which it codes, corresponding to residues 814–858 in the chick α1 collagen chain. (↓) *Hae*III site; (↑) *Hpa*II site; Y is a pyrimidine. Sequence is 30 base pairs to the right of *Hin*fI site and 170 base pairs from the *Hin*dIII 5′ end of the insert.

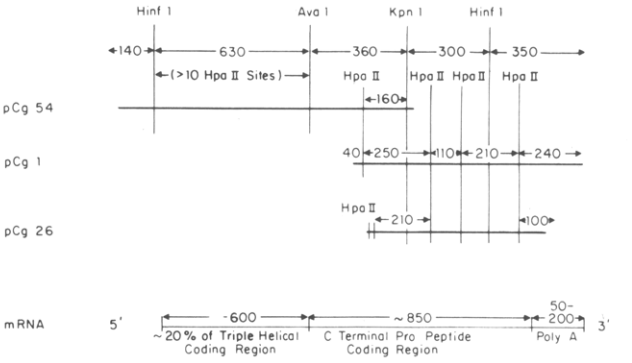

FIGURE 8: Restriction map of pro α1 collagen clones. Orientation of pro α1 collagen cDNA sequences relative to procollagen mRNA sequence was determined from primary sequence data given in Figure 7. Approximately 170 base pairs at the 5′ end of pCg54 are not colinear with procollagen mRNA.

The location of one end of pCg54 in the triple helical coding region of the chick α1 collagen chain and the results obtained from restriction digests lead to the restriction map of the three α1 collagen clones shown in Figure 8. More than 200 base pairs at the 3′ end of pCg54 overlap the 5′ end of pCg1. This has been confirmed by sequencing the 360 base pair *Ava*I–*Kpn* fragment of pCg54 and identifying an 83 base pair sequence identical with that shown in Figure 4.

An interesting aspect of the map shown in Figure 8 is the location of the *Ava*I site, C↓PyCGPuG, very near the end of the triple helical coding region of the α1 collagen chain. This region ends with five Gly-X-Y triplets in which X and Y are either Pro or Hyp. Therefore, the sequence GGXCCXCCX will be repeated five times, offering a high probability of containing the *Ava*I site, C↓CCGGG.

Discussion

Although the identification of chick pro α2 collagen cDNA clones has been reported recently (Lehrach et al., 1978; Sobel et al., 1978), clones containing pro α1 cDNA sequences had not been identified previously. We have now confirmed the identity of three pro α1 clones which together contain about

1500 base pairs complementary to the 3′ end of pro α1 collagen mRNA including 600 base pairs which code for the 200 amino acids at the carboxy-terminal end of the chick α1 chain. The pro α1 clones, like the pro α2 clone, pCg45, have less than 1000 nucleotides to code for the C-terminal propeptide and the noncoding region that precedes the poly(A) at the 3′ end. Messenger protection experiments have established the colinearity of the pro α1 and pro α2 inserts, ruling out internal deletions which might have occurred if either reverse transcription or subsequent amplification of the double-stranded cDNA had not faithfully reproduced the collagen sequence. It is very likely, therefore, that the C-terminal propeptide will be somewhat smaller than expected (Fessler & Fessler, 1978).

Like the triple helical coding region of the pro α2 insert in pCg45, the pro α1 insert in pCg54 contains numerous *Hae*III sites. The guanine–cytosine content of the latter appears to be considerably higher, however. The 146 nucleotides that were sequenced are almost 75% GC, while a 200 nucleotide sequence at the 5′ end of pCg45 is only 58% GC (unpublished data). Both the higher GC content and the relative higher frequency of Hyp in the Y position of the Gly-X-Y triplet may explain the multiple *Hpa*II sites in pCg54 which were not found in pCg45. $^{5'}CG^{3'}$ sequences appear to be less rare in collagen gene sequences than in other eukaryotic genes (Konkel et al., 1978).

During the initial characterization of these three pro α1 clones, we have found that two of them contained dimers. These probably were formed during the ligation of the double-stranded cDNA to chemically synthesized linker molecules. Such dimer formation may have occurred because the double-stranded cDNA molecules were not fractionated according to size to remove very short duplex molecules before being ligated to the linkers.

The explanation of the other anomaly, the absence of a *Hin*dIII site at one end of pCg54, awaits sequencing data. It is difficult to conceive of a mechanism whereby a double-stranded cDNA could be inserted into the *Hin*dIII site of pBR322 without a *Hin*dIII end. It seems likely that this site was modified during transformation or amplification.

In spite of pCg54's odd ends, the *Kpn–Hin*fI fragment of the insert in this plasmid contains about 1000 base pairs of contiguous pro α1 cDNA sequences. Moreover the 300 nucleotides at the 5′ end code for a highly conserved part of the triple helical region of collagen. There are only 6 residues that differ in the chick α1 chain from those found in the calf α1 chain between residues 811 and 909 (Fietzek & Kuhn, 1976). Since each of these could have resulted from a single base change, the pCg54 insert should prove to be an excellent hybridization probe for type I α1 collagen genes of other species as well as those of chick.

Acknowledgments

We express our deep appreciation to James D. Watson and Joe Sambrook for allowing us to use the Cold Spring Harbor P3 Laboratory. For sending us his EK2 strain of *E. coli* K12, χ1776, we thank Roy Curtiss, and for the plasmid, pBR322, we thank Herb Boyer. We are indebted to James Beard, Yvonne Chow, and Greg Sutcliffe for gifts of enzymes and to Richard Scheller for his gift of chemically synthesized decanucleotides. For helpful discussions, advice, and making available unpublished methods, we thank Gray Crouse, Radomir Crkvenjakov, Allan Maxam, Gary Buell, Marvin Wickens, Walter Keller, and Lydia Villa-Komaroff. Finally, we thank Tricia Bredbury for her assistance in RNA preparations, Cynthia Rosner for her help in preparing cDNA, and Doris Boger for typing this manuscript.

Supplementary Material Available

A supplementary figure showing the fluorograph of labeled wheat germ product obtained after translation of mRNA positively selected by hybridization to pCg1, and sequence gels in support of Figures 4 and 7 (4 pages). Ordering information is given on any current masthead page.

References

Berk, A. J., & Sharp, P. A. (1977) *Cell 12*, 721–732.
Boedtker, H., Frischauf, A. M., & Lehrach, H. (1976) *Biochemistry 15*, 4765–4770.
Dugaiczyk, A., Boyer, H. W., & Goodman, H. M. (1975) *J. Mol. Biol. 96*, 171–184.
Fessler, J. H., & Fessler, L. I. (1978) *Annu. Rev. Biochem. 47*, 129–162.
Fietzek, P. P., & Kuhn, K. (1976) *Int. Rev. Connect. Tissue Res. 7*, 1–60.
Frischauf, A. M., Lehrach, H., Rosner, C., & Boedtker, H. (1978) *Biochemistry 17*, 3243–3249.
Heyneker, H. L., Shine, J., Goodman, H. M., Boyer, H. W., Rosenberg, J., Dickerson, R. E., Narang, S. A., Itakura, K., Lin, S., & Riggs, A. D. (1976) *Nature (London) 263*, 748–752.
Konkel, D. A., Tilghman, S., M., & Leder, P. (1978) *Cell 15*, 1125–1132.
Lapière, C. M., & Nusgens, B. (1976) *Biochemistry of Collagen* (Ramachandran, G. N., & Reddi, A. H., Eds.) pp 377–447, Plenum Press, New York.
Lehrach, H., Frischauf, A. M., Hanahan, D., Wozney, J., Fuller, F., Crkvenjakov, R., Boedtker, H., & Doty, P. (1978) *Proc. Natl. Acad. Sci. U.S.A. 75*, 5417–5421.
Maxam, A. M., & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 560–564.
Miller, E. J. (1976) *Mol. Cell. Biochem. 13*, 165–192.
NIH Guidelines for Recombinant DNA Research (1976) *Fed. Regist. 41*, 27902–27943.
Panet, A., van de Sande, J. H., Loewen, P. C., Khorana, H. G., Raae, A. J., Lillehaug, J. R., & Kleppe, K. (1973) *Biochemistry 12*, 5045–5050.
Rigby, P. W. J., Dieckman, M., Rhodes, C., & Berg, P. (1977) *J. Mol. Biol. 113*, 237–251.
Seeburg, P. H., Shine, J., Martial, J. A., Baxter, J. P., & Goodman, H. M. (1977) *Nature (London) 270*, 486.
Sobel, M. E., Yamamoto, T., Adams, S. L., DiLauro, R., Avvedimento, V. E., de Crombrugghe, B., & Pastan, I. (1978) *Proc. Natl. Acad. Sci. U.S.A. 75*, 5846–5850.
Southern, E. M. (1975) *J. Mol. Biol. 98*, 503–517.
Sutcliffe, J. G. (1978) *Nucleic Acid Res. 5*, 2721–2728.
Tanaka, T., & Weisblum, B. (1975) *J. Bacteriol. 121*, 354–362.
Ullrich, A., Shine, J., Chirgwin, J., Pictet, R., Tischer, E., Rutter, W. J., & Goodman, H. M. (1977) *Science 196*, 1313–1319.
Villa-Komaroff, L., Efstratiadis, A., Broome, S., Lomedico, P., Tizard, F., Naber, S. P., Chick, W. L., & Gilbert, W. (1978) *Proc. Natl. Acad. Sci. U.S.A. 75*, 3727–3731.
Vogt, V. M. (1973) *Eur. J. Biochem. 33*, 192–200.
Wickens, M. P., Buell, G. N., & Schimke, R. T. (1978) *J. Biol. Chem. 253*, 2483–2495.