Proc. Natl. Acad. Sci. USA
Vol. 78, No. 2, pp. 712–716, February 1981
Biochemistry

Fine structural analysis of the chicken pro α2 collagen gene

[α2 (type I) collagen gene/Southern blot restriction endonuclease mapping/DNA sequence determination of exons and introns]

JOHN WOZNEY*, DOUGLAS HANAHAN, RICHARD MORIMOTO, HELGA BOEDTKER, AND PAUL DOTY†

Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachusetts 02138

Contributed by Paul Doty, September 23, 1980

ABSTRACT **Forty-two kilobase pairs of cloned chicken DNA containing 80% of the pro α2 (type I) collagen gene and 8 kilobase pairs of 3′ flanking sequences have been isolated. Detailed analysis of these clones indicates that this collagen gene spans approximately 40 kilobase pairs of DNA and contains on the order of 50 introns. The fine structure of 40% of the pro α2 gene, including its 3′ end, was determined by Southern blot restriction endonuclease mapping using a 2.6-kilobase pair procollagen cDNA clone, pCg45, as a probe, and by DNA sequence determination of more than 2 kilobase pairs of this part of the genome. Exons in the triple-helical coding region are all multiples of the 9 base pairs coding for the Gly-X-Y triplet and vary in size from 45 to 108 base pairs. The sequences of all six exons in a 3.8-kilobase pair *Eco*RI fragment were determined. One of these, a 249-base pair exon, joins the collagen domains; it codes for the last 15 amino acids of the triple-helical coding region, the telopeptide, and the first 53 amino acids of the carboxy-terminal propeptide.**

Collagen is the major structural protein in vertebrates, being found in small amounts in almost all tissues and constituting as much as 60% of the protein in such connective tissues as bones and cartilage. There are at least five different types of collagen encoded by at least seven different genes; the structure and function of these collagens have been reviewed recently (1–3). The tightly controlled spatial and temporal expression of each of these genes is regulated by mechanisms that have yet to be elucidated (4, 5). Collagen also has a long evolutionary history, being found in such primitive invertebrates as fresh water sponges (1). Establishing the fine structure of the collagen genes and the regulatory signals in their flanking sequences is a prerequisite to eventually understanding their role in development and evolution.

The development of recombinant DNA technology has made it possible to study unique eukaryotic genes coding for specific proteins and has already produced some remarkable revelations. The coding sequences in many genes in higher vertebrates have been found to be interrupted by noncoding intervening sequences (6), sometimes with a very high frequency. For example, the chicken ovalbumin gene has 7 intervening sequences (7, 8); the conalbumin gene has 16 intervening sequences (9); and the 6-kilobase-pair (kb) vitellogenin gene has a startling 33 intervening sequences (10).

Two-thirds of the procollagen chain consists of 338 amino acid triplets of Gly-X-Y in which X and Y are often proline, hydroxyproline, or alanine. Because these amino acids are encoded by G+C-rich codons and account for 70% of the amino acids in collagen, one might expect collagen gene fragments, like the G+C-rich silk fibroin gene fragments, to be shifted from the major DNA species on actinomycin D/CsCl gradients (11). We found, however, that chicken pro α2 collagen gene fragments banded with total chicken DNA on such gradients (12). This immediately suggested that the G+C-rich coding sequences were embedded in multiple intervening sequences so that the composition and buoyant density were the same as those of total chicken DNA. Recent reports on the structure of part of the chicken pro α2 gene (13) and part of the sheep pro α2 gene (14, 15) have confirmed this prediction: electron microscope studies of collagen gene clones hybridized to procollagen mRNA present a striking picture of multiple coding regions separated by many, often very large, intervening sequences.

A similar picture of the structure of the collagen gene can be derived by Southern blot restriction endonuclease mapping of collagen gene clones and by limited DNA sequence determination. We have taken advantage of having the DNA sequence and detailed restriction map of more than half of the coding sequence of the pro α2 gene in two collagen cDNA clones, pCg13 and pCg45 (16, 17) to locate coding sequences in collagen gene clones by Southern blot analysis. In addition, to establish important fine structural details of this gene, we have determined the sequences of seven exons in the triple-helical-coding regions and four exons in the COOH-terminal propeptide.

MATERIALS AND METHODS

Library Screening. The λ Charon 4A recombinant library containing chicken DNA fragments (constructed by J. Slightom, M. Sung, and O. Smithies) was screened as described (18, 19) with minor modifications. Phage were plated at 2–3 × 10^4 per 15-cm petri dish, using *Escherichia coli* DP50SupF or 803-8 as the bacterial host. Phage DNA was adsorbed to nitrocellulose filters (Millipore), denatured, and neutralized *in situ*, and then baked onto the filters at 80°C for 2 hr in a vacuum oven. The filters were incubated at 68°C in 4× SET (SET is 0.15 M NaCl/2 mM Na_2EDTA/30 mM Tris·HCl, pH 8) and 5× Denhardt's solution (20) for 3–4 hr. Hybridizations were carried out at 68°C for 15 hr in 3× SET, 1× Denhardt's solution, 0.1% $NaDodSO_4$, 10 μg of denaturated salmon sperm DNA per ml, and 4 × 10^6 to 10^7 cpm per filter of ^{32}P-labeled probe. To prepare probes, specific restriction endonuclease fragments of the pro α2 collagen cDNA clone, pCg45, were eluted from a polyacrylamide gel. These were then labeled with ^{32}P by nick translation to a specific activity of 1–5 × 10^8 cpm/μg. After hybridization, the filters were washed at 68°C for 3 hr with three changes of 2× SET/0.1% $NaDodSO_4$. In some cases, they were subsequently washed at room temperature for 1/2 hr with two changes of 3 mM Tris base.

Restriction enzymes [purchased from Bethesda Research Laboratories (Rockville, MD) or New England BioLabs] were used under the conditions specified by the suppliers. Restriction enzyme analysis and Southern (21) transfer analyses were

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviations: kb, kilobase pair(s); bp, base pair(s); SET, 150 mM NaCl/2 mM Na_2EDTA/30 mM Tris·HCl, pH 8.

* To whom requests for details of DNA sequence analysis should be addressed.
† To whom reprint requests should be addressed.

accomplished by using standard techniques.

DNA Sequence Analysis. In order to facilitate the production of large quantities of collagen gene fragments for sequence analysis, portions of the DNA from recombinant λ clones were subcloned in plasmid pBR322 DNA that had been extensively digested with *Eco*RI restriction endonuclease and treated with phosphatase. This ligation mix was used to transform *E. coli* DH1 (*recA1, nalA*$^-$, r_k^-, m_k^+, *endA1, thi-1*). Recombinant plasmids were detected by colony hybridization, direct restriction analysis, or both. The nucleotide sequences of the DNA fragments from these recombinant plasmids were determined by the method of Maxam and Gilbert (22).

Electron Microscopy. R-loops were formed by a procedure based on that of Thomas *et al.* (23) as modified by Kaback *et al.* (24). DNA (50 μg/ml) was irradiated with long-wave ultraviolet light in trioxsalen at 1 μg/ml to obtain 0.2 crosslink per kb. The crosslinked DNA was incubated for 12 hr at 57°C with total chicken calvaria RNA at 200 μg/ml in 80% (vol/vol) formamide/0.3 M NaCl/2 mM Na_2EDTA/10 mM piperazine-*N,N'*-bis(2-ethanesulfonic acid) (pH 6.5). The sample was diluted 1:100 in a solution containing 0.1 M Tris·HCl at pH 8.3, 20 mM Na_2EDTA, 50% (vol/vol) formamide, and cytochrome *c* at 40 μg/ml, and spread on distilled water. Plasmid pBR322 and phage φX174 DNA were included as double- and single-stranded DNA standards. Samples of the film produced were picked up on parlodion-coated slides, stained with uranyl acetate in 90% (vol/vol) ethanol, and rotary shadowed with platinum/paladium. Molecules were visualized and photographed with a Philips 300 electron microscope.

RESULTS

Restriction Map of the Pro α2(I) Collagen Gene. The 2.6-kb cDNA clone, pCg45, was previously shown to contain coding sequences for the 3' half of the pro α2(I) collagen gene (16). This clone was used to identify and isolate 35 gene clones containing nine different chicken DNA fragments in a Charon 4A library obtained from J. Slightom, University of Wisconsin. This library was constructed by a partial *Eco*RI digestion of DNA isolated from Leghorn chicken reticulocytes and selecting fragments in the 8- to 20-kb size range for ligation into Charon 4A. *Eco*RI fragments containing collagen gene sequences were identified by their hybridization to pCg45, and to its fragments, by Southern gel blotting.

Initially, three groups of nonoverlapping clones were identified: group I clones contained coding sequences located within 10 kb of the 3' end of the pro α2 collagen gene; group II clones contained coding sequences in the middle of the gene; group III clones contained coding sequences 5' to pCg45 sequences. None of the three groups had any *Eco*RI fragments in common with each other.

In order to obtain clones containing *Eco*RI restriction fragments that would provide links between the three groups of collagen gene clones, the library was rescreened and three new clones, α2CG 651, α2CG 653, and α2CG 657, were identified. Each contained a 0.23-kb *Eco*RI fragment that links group I and group II clones. One of the clones, α2CG 653, also contained a 5.4-kb *Eco*RI fragment that is 5' to the group II clones previously identified but does not overlap the group III clones. The arrangement of the *Eco*RI fragments in 10 of the 12 collagen gene clones we have analyzed is shown in Fig. 1.

α2CG 657 best demonstrates the overlap because it contains both the 4.0-kb and 1.5-kb *Eco*RI fragments of group II as well as the 1.8-kb fragment of group I clones linked by the 0.23-kb *Eco*RI fragment located at the 3' end of α2CG 653 and the 5' end of α2CG 651. It is worth noting that α2CG 657 also has two *Eco*RI fragments at the 3' end that do not hybridize to pCg45 and must have originated from another part of the genome. Cloning artifacts of this type, which were seen in 5 of the 12 clones we studied, may have resulted from the ligation of two smaller fragments in the relatively broad size distribution (8–20 kb) of fragments used in cloning. Charon 4A has a clear preference for accepting 11- to 15-kb fragments.

The group I clones, labeled A and B in Fig. 1, must be alleles because they have identical *Eco*RI fragments, as well as *Hin*dIII and *Bam*HI fragments, except that α2CG 291 (IB) has a single extra *Eco*RI site 0.3 kb from the 5' end of the 3.8-kb *Eco*RI fragment in α2CG 241 and is missing a *Hin*dIII site near the

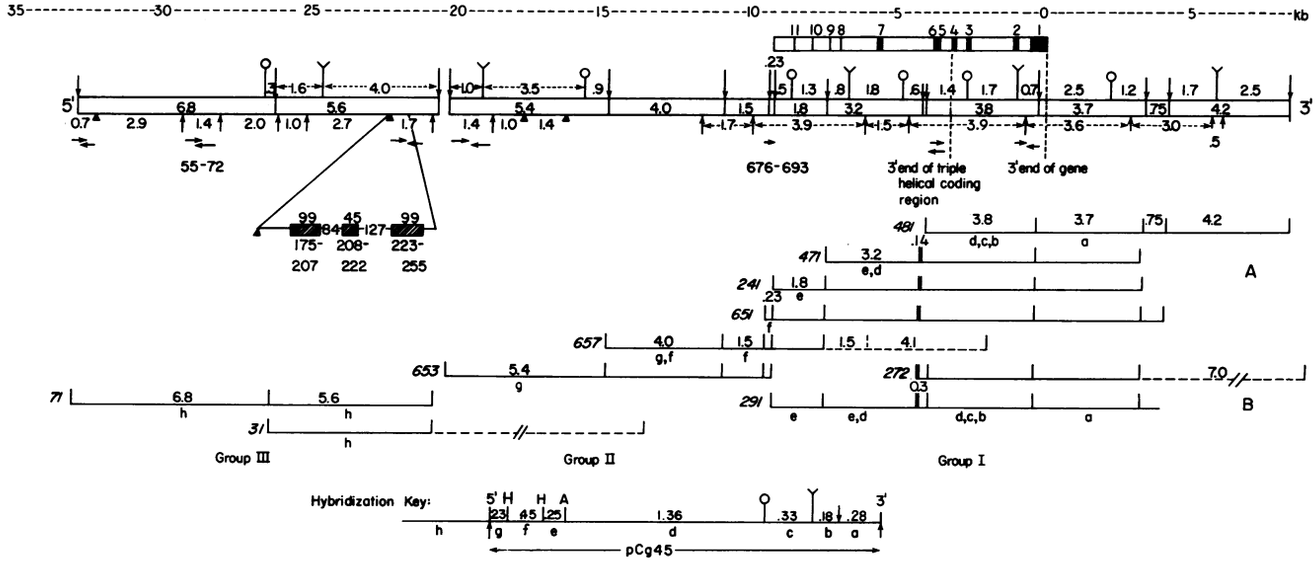

FIG. 1. Restriction endonuclease map of pro α2 collagen gene clones. The location and orientation of each of the clones were determined by hybridization of the clones digested with restriction endonucleases *Eco*RI (↓), *Bam*HI (Y), *Hin*dIII (↑), *Pst* I (▽), and *Hpa* II (▲) to nick-translated fragments of pCg45 shown in the hybridization key. In addition to the exons whose DNA sequences were determined and are identified by residue numbers, the sequences of six other exons in the 3.8-kb *Eco*RI fragment were determined and are shown in Fig. 2.

3′ end of the 3.7-kb *Eco*RI fragment (data not shown). α2CG 291 contains sequences in common with the α2 collagen gene clone described by Vogeli *et al.* (13), but it extends 3 kb further in the 5′ direction and is missing 2.5 kb of 3′ flanking sequence.

None of the clones shown in Fig. 1 hybridized to the pro α1 cDNA clones, pCg1 and pCg54 (25), and we could clearly distinguish clones that hybridized to different small restriction fragments of pCg45 on Southern blots. We are therefore confident that our hybridization conditions were sufficiently stringent to rule out the possibility that any group I or II clones contained collagen gene sequences other than pro α2 collagen gene sequences. The actual placement of the group III clones within the pro α2 collagen gene was accomplished only after direct DNA sequence analysis, as described below.

The restriction map of 28 kb of contiguous gene sequences (including 8 kb of 3′ flanking gene sequences) and of 12.4 kb of collagen gene sequences 5′ to these sequences is shown in Fig. 1. The pro α2 collagen gene is seen to span at least 33 kb. It has 12 *Eco*RI sites, 4 *Bam*HI sites, 11 *Hin*dIII sites, and at least 5 *Pst* I sites. Hybridization of ordered restriction fractions of pCg45, shown at the bottom of Fig. 1, allowed the location of all 2.6 kb of coding sequences in pCg45. These extend from the 3′ end of gene located at the 5′ end of the 3.7-kb *Eco*RI fragment of α2CG 241 (and other group I gene clones), into the 5.4-kb *Eco*RI fragment of α2CG 653. Therefore the 3′ half of the α2 collagen gene has 2.6 kb of coding sequences distributed over more than 15 kb of genomic DNA.

The nonoverlapping clone α2CG 71, (left of Fig. 1), was assigned a location 5′ to the group I and II clones because it hybridized only weakly to pCg45 but strongly to procollagen mRNA. To definitively identify this gene clone, and to establish its 5′-3′ orientation, the sequences of three regions of the clone were determined, indicated by arrows under the restriction map. No coding sequences were found at the end of α2CG 71; three small exons, 99 base pairs (bp), 45 bp, and 99bp, separated by short introns, 84 bp and 127 bp long, were located in the 1.7-kb *Hpa* II-*Hin*dIII fragment at the other end of α2CG 71. These three exons code for amino acids 175–207, 208–222, and 223–255, measured from the NH_2-terminus of the triple-helical region. The 1.4-kb *Hin*dIII fragment located near the middle of the 6.8-kb *Eco*RI fragment contains an exon 54 bp long that codes for residues 55–72 of the α2 collagen chain. This exon and hence the 1.4-kb *Hin*dIII fragment must be 5′ to the 1.7-kb *Hpa* II-*Hin*dIII fragment of α2CG 71, thereby establishing its 5′-3′ orientation. The four exons whose sequences were determined are all multiples of 9 bp, and each begins with the codon for glycine and ends with that for Y in the Gly-X-Y repeating unit of collagen. There are 5.7 kb of genomic DNA between the two regions of α2CG 71 in which exons were located. This region must code for residues 73–174 and hence contains a total of 306 bp of coding sequences, or 54 bp of exons per 1000 bp of genomic DNA. We expeet to find other exons 5′ to the 1.4-kb *Hin*dIII fragment because this region hybridizes to procollagen mRNA. On the basis of the exon density in this part of the collagen gene, an additional 162 bp of coding sequences might be located here. This would suffice to code for the remaining 54 residues in the triple-helical region but not for the telopeptide, NH_2-terminal propeptide, or 5′ noncoding region of the mRNA.

Fine Structure Mapping of α2CG 241. α2CG 241 contains gene sequences coding for the triple-helical region, the telopeptide, the COOH-terminal propeptide, and noncoding sequences in the pro α2 collagen mRNA. It was therefore of interest to obtain a detailed analysis of the structure of this part of the gene. It was also the region of the gene most readily analyzed because it contains the coding sequences in the cDNA clones, pCg45 and pCg13, whose DNA sequences have been determined (17).

Southern blots of *Eco*RI and *Eco*RI-*Hin*dIII fragments of clone α2CG 241 were hybridized to fragments of pCg45, labeled a to g in the map shown at the bottom of Fig. 1. This enabled us to determine the approximate location of coding sequences in the gene clone, and by comparing distances between restriction sites in the gene clone with those in pCg45 we could determine the size and location of intervening sequences.

We first located the 3′ end of the gene. Because the *Eco*RI-*Hin*dIII fragment at the 3′ end of pCg45 hybridizes only to the 3.7-kb *Eco*RI fragment, as shown in Fig. 1, the 3′ end of the gene must be in this fragment. There are 335 bp between the *Eco*RI site in pCg13 and the poly(A) end of the cDNA clone (17); this suggests that the *Eco*RI site between the 3.8-kb and 3.7-kb gene fragments is likely to be the same as that in the cDNA clone. This was confirmed by sequence analysis from this *Eco*RI site and finding that this sequence was the same as the one obtained for the cDNA clone.

The *Eco*RI-*Bam*HI fragment is 183 bp long in the cDNA clone, whereas the distance between the *Eco*RI site and the nearest 5′ *Bam*HI site is 700 bp in the gene. This gene fragment also contains a *Hin*dIII site not found in the coding sequence. This immediately suggests an intron about 500 bp long. Similarly, there must be one or more introns between the *Bam*HI site and the *Pst* I site 5′ to it, because these sites are separated by 322 bp in pCg45 and by 1.7 kb in the gene. This suggests a total intron size of 1.4 kb.

To obtain a precise location of exons and introns in the 3.8-kb fragment, it was digested with restriction endonucleases with *Sau*3A, *Alu* I, and *Hpa* II. The sequences of numerous fragments were determined as indicated by arrows in the upper part of Fig. 2. By this means five exons were located in the 3.8-kb fragment as well as the 3′-most exon. The two exons near the 5′ end are 54 and 108 bp long and code for the amino acid residues 946–963 and 964–999, both of which lie in the helical region of the pro α2 collagen chain. Two other exons, toward the 3′ end, are 189 and 243 bp long and code for the COOH-terminal propeptide of the pro α2 collagen chain. Exon 4 (d in Fig. 3) lies between these two sets of exons and, interestingly enough, its 249 bp code for the last 15 residues in the helical region, the 15 residues of the telopeptide, and the first 53 residues of the COOH-terminal propeptide. Its DNA sequence is identical to that in pCg45 (17).

The *Ava* I site in pCg45 is 679 bp from the end of the triple-helical coding region and is located 0.3 kb from the 5′ end of the 3.2-kb *Eco*RI gene fragment. We have located the *Ava* I site between residues 788 and 789 of the chicken α2 chain (17), and this places an exon containing these residues at this position in the gene clone.

The hybridization results, summarized in Fig. 1, suggest that the 1.8-kb *Eco*RI fragment at the 5′ end of α2CG 241 does not contain some 680 bp at the 5′ end of pCg45 that code for residues 478–705. The exact coding content of this clone was established by determining the sequence of the entire 0.23-kb *Eco*RI fragment linking group I and group II gene clones. It contains a single 54-bp exon coding for residues 676–693. Therefore the exon nearest the 5′ end of α2CG 241 begins with the nucleotides coding for residue 694.

The exons and introns in α2CG 241 can be seen directly by electron microscopy of this clone R-looped to procollagen mRNA. Fig. 3 is a representative photograph, showing 9 introns separating 10 exons, labeled a–j in the schematic sketch in the *Inset*. There also appears to be a small intron in the fifth exon (labeled e), and this was confirmed by determining the se-

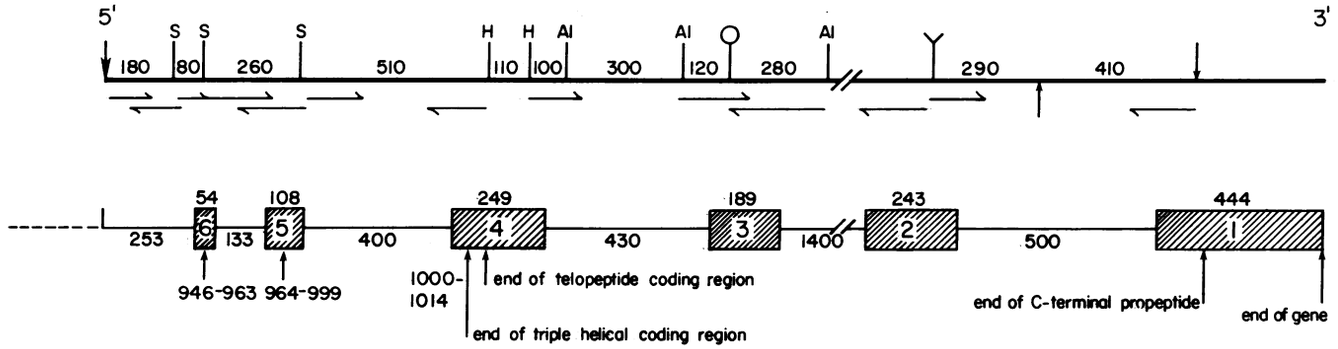

FIG. 2. Fine structure map of the 3.8-kb *Eco*RI fragment of α2CG 241. Distances are bp. Arrows identify regions whose sequences were determined. Symbols for *Eco*RI, *Bam*HI, *Hin*dIII, and *Pst* I restriction sites are the same as in Fig. 1; others are *Sau*3A (S), *Hpa* II (H), and *Alu* I (Al). Exon and small intron sizes are based on DNA sequence determination; intron sizes larger than 300 bp were determined by restriction mapping. The absence of introns in the exon at the end of the gene was determined by electron microscope analysis shown in Fig. 3.

quence of the region of the clone discussed above and shown schematically in Fig. 2.

We have located the exons and introns of α2CG 241 above the restriction map in Fig. 1. The *Eco*RI, *Bam*HI and *Pst* I sites, near the 3' end of pCg45, are located in exons 1, 2, and 3, respectively; these exons are 400 bp, 250 bp, and 200 bp long, in excellent agreement with sizes determined by DNA sequence analysis shown in Fig. 2. By way of contrast, our electron microscopic measurements of exons 4 and 5/6 yielded lengths of only 200 bp, compared with 249 and 295 (54 + 133 + 108) bp from sequence determination, suggesting that substantial inaccuracies can attend electron microscopic measurements in this range.

Overall, the distribution of exons and introns in the 3.8-kp *Eco*RI fragment of α2CG 241 are very similar to those reported for λgCOL 204 (15), but they differ for the next exon located in the 1.5-kb *Hin*dIII fragment. In this case we find a single 250-bp exon by electron microscopy (exon f in Fig. 3), whereas Vogeli *et al.* (13) report three exons based on small "nubs" in the electron micrograph.

The remaining four exons are all about 100 bp in size. The *Ava* I site is located in exon 9. In all, the 11 exons contain 1.9 kb of pro α2 mRNA sequences, on the basis of electron microscope measurements. This is in excellent agreement with the 2.05 kb of coding sequence placed in this clone by DNA sequence determination. The 735 bp coding for the COOH-terminal propeptide are located in about 3.2 kb of genomic DNA, including three introns totaling 2.5 kb. The 960 bp coding for the triple-helical region of collagen occupy almost 6 kb of genomic DNA, including six introns containing more than 5 kb of DNA. The density of coding sequence is somewhat lower in the triple-helical coding region than in the COOH-terminal propeptide coding region.

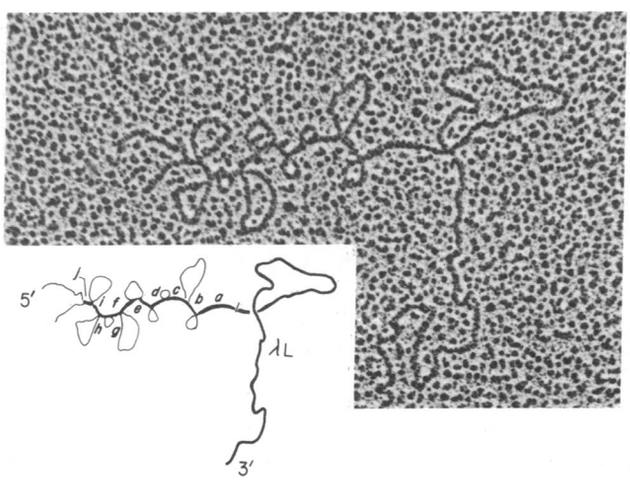

FIG. 3. Electron microscope photograph of the collagen gene clone α2CG 241 hybridized to pro α2 collagen mRNA prepared from calvaria of 16-day-old embryonic chickens as described (26). (*Inset*) Interpretative drawing. Ten exons, labeled a to j going from 3' to 5', correspond to exons 1–11 in Fig. 1. Exon e corresponds to exon 5 and 6 in Fig. 1 because this exon is interrupted by a 133-bp intron determined by DNA sequencing and shown in Fig. 2.

DISCUSSION

The detailed analysis of 33 kb of genomic DNA containing 4 kb of pro α2 collagen gene sequences from chicken indicates that the structure of this gene is striking with respect to the size of the gene and the number and arrangement of exons. It is certainly one of the largest genes identified thus far. The density of exons in the 80% of the collagen gene characterized to date is about 1 kb of exon per 8 kb of gene, predicting a size of approximately 40 kb for the collagen gene. This is about twice the size of the 21-kb *Xenopus laevis* vitellogenin gene (10) and is on the order of the 42-kb size postulated for the mouse dihydrofolate reductase gene (27).

Another remarkable aspect of the pro α2 collagen gene is its very large number of exons, and their correspondingly small size. The sequences of seven exons coding for the triple-helical region have been determined; all are multiples of nine base pairs coding for the collagen repeating unit, Gly-X-Y. Each exon begins with the codon for glycine and ends with the codon for the Y residue. The small size of the exons, 45–108 bp, is similar to the 60- to 116-bp exons found in the conalbumin gene (9). From the average size of these seven exons coding for the triple-helical region, we estimate 40–45 exons are required to code for this part of the gene. An additional 4 account for the COOH-terminal propeptide and the 3' untranslated region of the mRNA, and possibly another 3–4 account for the NH_2-terminal propeptide (the size of which has not been unambiguously determined) and the 5' untranslated region. This predicts a total of 47–53 exons. Thus 5 kb of coding information is imbedded in 40 kb of DNA through the presence of about 50 introns. Each 40-kb primary transcript of the collagen gene must therefore undergo about 50 precise splicing reactions to generate a correct mRNA, a tribute to the accuracy and efficiency of the RNA splicing mechanism.

Electron microscopy of R-loops, while confirming the overall structure of this gene, appears to be inadequate to reveal accurate and detailed information on the arrangement and size of the coding regions. This is particularly applicable to the part of the gene coding for the triple-helical region, which contains multiple small exons not infrequently separated by small introns. Three of the seven triple-helical coding exons whose sequences were determined were 54 bp, four were not. It has been suggested (28) that the ancestral collagen gene was 54 bp and that during evolution this gene was duplicated [presumably by recombination between introns (6)] to give rise to the present structure. Our finding that all triple-helical coding exons are not 54 bp would indicate that if the ancestral collagen gene were 54 bp (or some other multiple of 9 bp), then duplication was accomplished in large part by some type of recombination (such as unequal crossing over) between exons rather than between introns. Alternatively, it is possible that the introns were inserted after creation of the gene for stabilization of the repetitive sequences or for other structural or functional reasons. Examination of the structure of a collagen gene in a primitive invertebrate such as a fresh water sponge should answer this question in part.

The detailed structural analysis of 40% of the pro $\alpha 2$ collagen gene provided another interesting result with respect to intron–exon placements. Rather than finding introns between the most apparent domains of procollagen, we identified a single exon that codes for the last 15 residues of the helical region, the 15 residues of the telopeptide, and the first 53 residues of the COOH-terminal propeptide. This is not in obvious agreement with a commonly held proposal that introns separate regions that code for functionally and structurally distinct parts of the protein. However, it could be argued that these structural domains constitute a single functional domain. The COOH-terminal propeptide is essential for the formation of the triple helix but must be cleaved prior to fibrinogenesis. The gene segment coding for the COOH-terminal propeptide cleavage site may be fused to the gene segment coding for the end of the triple helical region because this structure is obligatory for collagen expression and thus is a functional domain.

We thank Jerry Slightom for providing us with his Charon 4A chicken DNA library, which he and Drs. Michael Sung and Oliver Smithies constructed. We also thank Joe Sambrook for giving us *E. coli* strain 803-8. We gratefully acknowledge the help of Forrest Fuller in reading sequencing gels, Thomas McDonnell in making electron microscope length determinations, and Daniel Finley for his help in some of the restriction mapping. Julia Smith was both patient and efficient in preparing this manuscript. This research was supported by grants from the National Institutes of Health and the Muscular Dystrophy Association.

1. Bornstein, P. & Sage, H. (1980) *Annu. Rev. Biochem.* **49,** 957–1003.
2. Eyre, D. (1980) *Science* **207,** 1315–1322.
3. Prockop, D. J., Kivirikko, K. I., Tuderman, L. & Guzman, N. A. (1979) *N. Engl. J. Med.* **301,** 13–23.
4. Linsenmayer, T. F., Toole, B. P. & Trelstad, R. L. (1973) *Dev. Biol.* **35,** 232–239.
5. Von der Mark, H., Von der Mark, K. & Gay, S. (1976) *Dev. Biol.* **48,** 237–249.
6. Gilbert, W. (1979) *Eucaryotic Gene Regulation* (Academic, New York).
7. Dugaiczyk, A., Woo, S. L. C., Lai, E. C., Mace, M. L., Jr., McReynolds, L. & O'Malley, B. W. (1978) *Nature (London)* **274,** 328–333.
8. Gannon, F., O'Hare, K., Perrin, F., LePennec, J. P., Benoist, C., Cochet, M., Breathnach, R., Royal, A., Garapin, A., Cami, B. & Chambon, P. (1979) *Nature (London)* **278,** 428–434.
9. Cochet, M., Gannon, F., Hen, R., Maroteaux, L., Perrin, F. & Chambon, P. (1979) *Nature (London)* **282,** 567–574.
10. Wahli, W., Dawid, I. B., Wyler, T., Weber, R. & Ryffel, G. U. (1980) *Cell* **20,** 107–117.
11. Ohshima, Y. & Suzuki, Y. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5363–5367.
12. Wozney, J., Hanahan, D., Fuller, F. & Boedtker, H. (1979) *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **38,** 617 (abstr.).
13. Vogeli, G., Avvedimento, E. V., Sullivan, M., Maizel, J. V., Jr., Lozano, G., Adams, S. L., Pastan, I. & de Crombrugghe, B. (1980) *Nucleic Acids Res.* **8,** 1823–1837.
14. Boyd, C. D., Tolstoshev, P., Schafer, M. P., Trapnell, B. C., Coon, H. C., Kretschmer, P. J., Nienhuis, A. W. & Crystal, R. G. (1980) *J. Biol. Chem.* **255,** 3212–3220.
15. Schafer, M., Boyd, C., Tolstoshev, P. & Crystal, R. (1980) *Nucleic Acids Res.* **8,** 2241–2253.
16. Lehrach, H., Frischauf, A. M., Hanahan, D., Wozney, J., Fuller, F., Crkvenjakov, R., Boedtker, H. & Doty, P. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 5417–5421.
17. Fuller, F. & Boedtker, H. (1980) *Biochemistry,* in press.
18. Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K. & Efstratiadis, A. (1978) *Cell* **15,** 687–701.
19. Benton, W. D. & Davis, R. W. (1977) *Science* **196,** 180–182.
20. Denhardt, D. T. (1966) *Biochem. Biophys. Res. Commun.* **23,** 641–646.
21. Southern, E. M. (1975) *J. Mol. Biol.* **98,** 503–517.
22. Maxam, A. & Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560.
23. Thomas, M., White, R. L. & Davis, R. W. (1976) *Proc. Natl. Acad. Sci. USA* **73,** 2294–2298.
24. Kaback, D. B., Angerer, L. M. & Davidson, N. (1979) *Nucleic Acids Res.* **6,** 2499–2517.
25. Lehrach, H., Frischauf, A. M., Hanahan, D., Wozney, J., Fuller, F. & Boedtker, H. (1979) *Biochemistry* **18,** 3146–3152.
26. Boedtker, H., Frischauf, A. M. & Lehrach, H. (1976) *Biochemistry* **15,** 4765–4770.
27. Nunberg, J. H., Kaufman, R. J., Chang, A. C. Y., Cohen, S. N. & Schimke, R. T. (1980) *Cell* **19,** 355–364.
28. Yamada, Y., Awedimento, V. E., Mudryj, M., Ohkubo, H., Vogeli, G., Irani, M., Pastan, I. & de Crombrugghe, B. (1980) *Cell* **22,** 887–892.