Proc. Natl. Acad. Sci. USA
Vol. 77, No. 12, pp. 7059–7063, December 1980
Biochemistry

Isolation and characterization of overlapping genomic clones covering the chicken α2 (type I) collagen gene

(intervening sequences/recombinant DNA)

HIROAKI OHKUBO*, GABRIEL VOGELI*, MARIA MUDRYJ*, V. ENRICO AVVEDIMENTO*, MARGERY SULLIVAN†, IRA PASTAN*, AND BENOIT DE CROMBRUGGHE*

*Laboratory of Molecular Biology, National Cancer Institute, and †Laboratory of Molecular Genetics, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20205

Communicated by Joseph E. Rall, August 20, 1980

ABSTRACT **A series of overlapping recombinant clones, which cover the α2 (type I) collagen gene, have been isolated by stepwise screening of two libraries of chicken genomic DNA fragments. The first genomic clone was isolated by using a cloned cDNA containing α2 collagen DNA sequences as hybridization probe. The other clones were obtained by a sequence of screenings using defined fragments of the successive genomic clones as hybridization probes. Several types of experiments indicated that the DNA of these clones are truly overlapping and span 55 kilobase pairs of contiguous DNA sequences in the chicken genome. Sequence analysis of small DNA segments of some of these clones confirm that they contain coding sequences which specify α2 collagen. Electron microscopic analysis of hybrids between type I α2 collagen mRNA and the overlapping genomic clones indicates that the chicken α2 collagen gene has a length of at least 37 kilobases, about 7.4 times longer than the corresponding translatable cytoplasmic mRNA. The coding information for α2 collagen is distributed in more than 50 coding sequences which are interrupted by intervening sequences of various sizes. The structure of the gene implies that the conversion of precursor RNA to mature mRNA for α2 collagen includes at least 50 splicing events.**

The collagens belong to a family of proteins that constitute the major component of the extracellular matrix of many animal tissues. At least five genetically distinct types of collagen (1, 2) are found in higher vertebrates. It is likely that these different collagen types play an important role in embryonic development and morphogenesis and that the synthesis of each collagen species responds to a tissue-specific developmentally regulated genetic program (3).

We have been studying the synthesis of type I collagen by chicken embryo fibroblasts (CEF) to examine questions related to the differentiation program of these cells. α1 and α2 collagen are the constituent subunits of type I collagen, the principal collagen synthesized by CEF. At least two types of events strongly decrease the synthesis of type I collagen in CEF. One is transformation by the product of the *src* gene of Rous sarcoma virus (4). Another is the administration of phorbol esters (5). The decrease in type I collagen synthesis caused by these agents is due to a coordinate decrease in the levels of α1 and α2 collagen RNA (refs. 6–8; M. Sobel, personal communication).

We and others have constructed cDNA clones for chicken α1 and α2 (type I) collagens (9–12). These cDNA clones were used to measure changes in the levels of collagen RNA in CEF (8). In order to study the expression of the type I collagen genes in appropriate *in vivo* and *in vitro* systems, it was essential to isolate one of these collagen genes, particularly the segment lying at the 5′ end of the gene, which could play an important role in the control of its expression. We recently described the isolation of a genomic clone which contains 6.8 kilobases (kb) at the 3′ end of the α2 collagen gene (13). Similar clones containing the 3′ end of the sheep α2 collagen gene have been isolated (14).

Here we report the isolation and characterization of several additional overlapping clones. Together these clones span about 55 kb of contiguous genomic DNA sequences and cover the α2 collagen gene. The gene is at least 37 kb long and contains more than 50 different coding segments.

MATERIALS AND METHODS

Hybridization Probes. A 1.5-kb *Hin*dIII fragment of λ COL-204, carrying chicken genomic sequences from the 3′ end of the α2 collagen gene, was subcloned into pBR322, and this DNA fragment was isolated from the plasmid as described (13). Similarly, a 3.2-kb *Eco*RI fragment of λCOL-271, carrying α2 collagen genomic sequences, was subcloned in pBR313; its insert was purified by sucrose gradient centrifugation. Other DNA fragments used as hybridization probes were purified from agarose slab gels (15). DNA was labeled by nick translation (16).

Screening of the Chicken DNA Libraries. Two libraries of random chicken genomic DNA fragments were screened (17). Both libraries contain fragments (15–20 kb long) introduced in the bacteriophage λ Charon 4A vector (18). One library was constructed by J. Dodgson, R. Axel, and D. Engel using a partial *Hae* III digest and a partial *Alu* I digest of chicken reticulocyte DNA (19). The other library was constructed by J. Slightom and colleagues using a partial *Eco*RI digest of chicken reticulocyte DNA.

Electron Microscopy. All of the electron microscopic data were obtained by described methods (13).

RESULTS

Isolation of Overlapping α2 Collagen Genomic Clones. To purify the α2 collagen gene we isolated a series of overlapping clones by successive screenings of a library of random genomic fragments. Between 150,000 and 300,000 plaques were screened at each passage through the library. Usually one or two positive signals were retained for further purification. The successive screening steps are illustrated in Fig. 1.

We first used as hybridization probe a 1.7-kb *Hin*dIII fragment which is located at the 5′ end of the chicken DNA insert in λCOL-204 (13). This fragment was subcloned in the unique *Hin*dIII site of pBR322. One of the new genomic clones, which was obtained from the genomic library by using this hybrid-

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: CEF, chicken embryo fibroblasts; kb, kilobase(s); bp, base pairs(s).

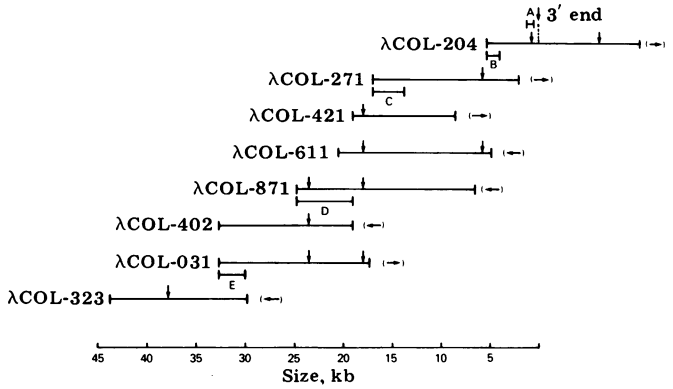

FIG. 1. Overlapping genomic clones of the chicken α2 collagen gene. Segment A, α2 collagen cDNA clone (9) used to isolate λCOL-204 (13); segments B, C, D, and E, genomic DNA fragments used as hybridization probes to isolate the other overlapping clones; vertical arrows, *Bam* HI sites; horizontal arrows pointing right, polarity of the α2 collagen segment is the same as the genes A–J of λ (i.e., the 5′ end of the coding strand is adjacent to the long arm of λ); arrows pointing to the left, the polarity of the cloned segment is inverted.

ization probe, λCOL-271, was chosen for further studies. A 3.2-kb *Eco*RI fragment (fragment C of Fig. 1) located at the 5′ end of λCOL-271 was subcloned in pBR313 and used as a hybridization probe to isolate other overlapping genomic clones. Of these, λCOL-421, λCOL-611, and λCOL-871 were retained for further studies. Fragment D, a 5.7-kb *Eco*RI fragment isolated from λCOL-871, was used to isolate λCOL-402 and λCOL-031. Fragment E, a 2.7-kb *Eco*RI/*Xba* I fragment located at the 5′ end of λCOL-031, was used to isolate λCOL-323.

To determine the orientation of the α2-collagen gene within these clones, pairs of overlapping clones were hybridized to each other and analyzed by electron microscopy. The orientation of collagen DNA was deduced from the previously determined orientation of the collagen gene in λCOL-204 (13). These orientations were further confirmed by DNA sequence analysis of some of the coding segments (see below). In clones λCOL-204, λCOL-271, λCOL-421, and λCOL-031, the orientation of the collagen gene within the λ vector is the same. In these clones the segment of the cloned DNA nearest to the long arm of λ is also closest to the 5′ end of the gene. In clones λCOL-611, λCOL-871, λCOL-402, and λCOL-323, this orientation is inverted.

We wanted to exclude the possibility that, by successive screenings of a genomic library, we had isolated fragments of different collagen genes which could contain sequences in common with the hybridization probe we used. Several types of experiments convincingly demonstrated that the clones we isolated were truly overlapping. Heteroduplex analysis between pairs of overlapping clones, in which the chicken DNA was in the same orientation, clearly showed that the 5′ portion of the chicken DNA sequence in one clone was always homologous to the 3′ segment of the chicken DNA insert in the overlapping clone. The same analysis also indicated that the overlapping segments were homologous over the entire length of the overlap and were not interrupted by nonhomologous regions (data not shown). The sizes of the overlapping segments of various pairs of clones as determined by electron microscopic analysis of heteroduplexes are presented in Table 1.

Restriction enzyme analysis also indicated that the overlapping clones contained identical restriction sites at the same location in their homologous segments. A detailed restriction map of each of the clones shown in Fig. 1 will be presented elsewhere. We conclude that the overlapping clones depicted in Fig. 1 cover a 55,000-base-pair (bp) segment of contiguous DNA sequences in the chicken genome.

Table 1. Size of homologous region in pairs of overlapping clones

Heteroduplex pairs	Size of homologous region, base pairs
λCOL-204/λCOL-271	3,062 ± 230 (7)
λCOL-271/λCOL-421	9,265 ± 1,614 (8)
λCOL-421/λCOL-031	1,776 ± 106 (10)
λCOL-611/λCOL-871	16,290 ± 407 (3)
λCOL-871/λCOL-402	6,101 ± 1,074 (8)
λCOL-402/λCOL-323	2,902 ± 93 (5)

The DNAs of pairs of overlapping clones were hybridized, and the resultant heteroduplexes were examined by electron microscopy. Circular simian virus 40 DNA was used as size standard. Results are shown as mean ± SEM; number of molecules measured is indicated in parentheses.

Structural Organization of the Chicken α2 Collagen Gene. To examine for the presence of intervening sequences we hybridized the cloned genomic DNAs to α2 collagen mRNA and analyzed the resulting hybrids by electron microscopy. We first formed heteroduplexes between the recombinant clone and its parent λ Charon 4 in order to produce single-stranded DNA in the cloned segment of the phage. In a second step, α2 collagen mRNA was added to this heteroduplex and the mixture was incubated under optimal conditions for RNA·DNA hybrid formation. This method allows a better visualization of coding and intervening sequences than does direct R-loop formation with double-stranded DNA.

Fig. 2 shows typical examples of such analysis for representative overlapping clones. It is obvious that the collagen clones contained numerous intervening sequences of various sizes. Measurements made on such pictures provided estimates of the sizes of the coding and intervening sequences. These measurements also were useful for aligning the overlapping clones. Indeed, the patterns of coding and intervening sequences were identical in the overlapping segments. Table 2 presents a summary of the measurements made on electron micrographs. The coding sequences were numbered beginning at the 3′ end of the gene. Some measurements reported as exon measurements clearly represent the sum of two coding sequences with a small intervening sequence between the coding segments. Such small intervening sequences appear as a small pimple which interrupts the continuity of the coding sequences. When this was seen reproducibly in several molecules, we counted two coding sequences, although we did not attempt to measure the size of the pimple or the size of the two coding sequences flanking the small intervening sequence. Comparison of the measurements of the different clones in Table 2 shows that the sizes of coding and intervening sequences within the overlapping segments of the different clones are similar. This observation confirms that the recombinant clones that we isolated are truly overlapping.

Fig. 3 is a graphic representation of the measurements reported in Table 2. In this figure, each vertical bar represents a coding sequence or exon, and the lines between the bars correspond to intervening sequences. Note that the coding sequences at the 3′ end of the gene are larger than those in the rest of the gene except for the coding sequence nearest the 5′ end. We conclude that the coding information for α2 collagen is divided into more than 50 coding segments which are interspersed by intervening sequences of various lengths. Indeed, the size of the intervening sequences varies from <100 bp to >2000 bp.

At the 5′ end of λCOL-323 there is a region of about 8.5 kb which does not hybridize with collagen mRNA and probably

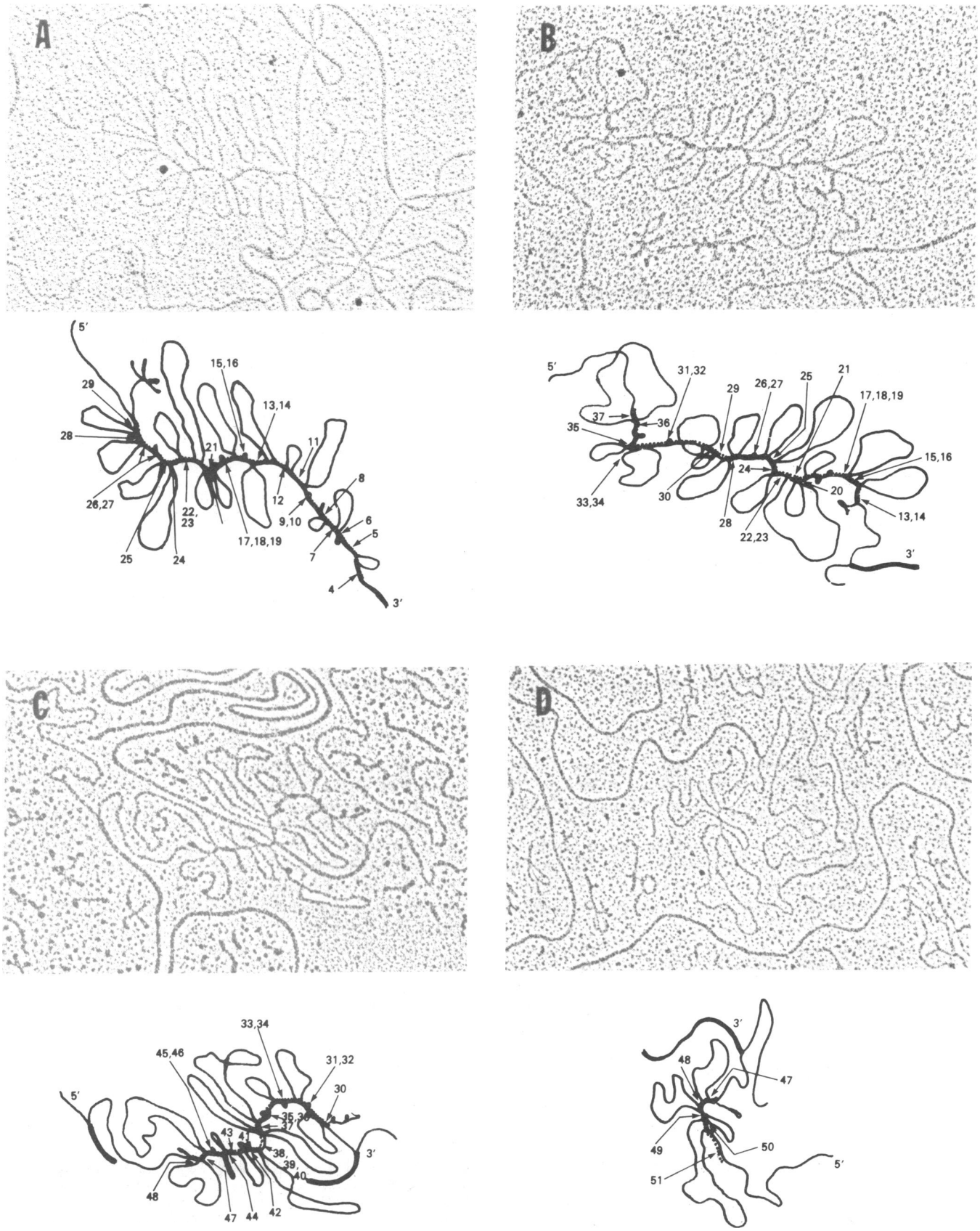

FIG. 2. Electron microscopic pictures and diagrams of the hybrids between the DNAs of four representative overlapping clones and chicken $\alpha 2$ collagen mRNA. Recombinant clones used were: *A*, λCOL-271; *B*, λCOL-871; *C*, λCOL-031; *D*, λCOL-323. The numbers correspond to the coding sequences, beginning at the 3′ end of the gene. When two numbers are shown, the indicated segment contains two coding segments separated by a small intervening sequence often seen as a small pimple.

Table 2. Size of the coding and intervening sequences in overlapping clones of α2 collagen gene

Exon	Exon length, bp	Intron length, bp	Exon	Exon length, bp	Intron length, bp	Exon	Exon length, bp	Intron length, bp
	λCOL-204			λCOL-271			λCOL-611	
1	327 ± 106	491 ± 82	4	117 ± 53	395 ± 53	11–24	ND	1301 ± 111
2	208 ± 38	1010 ± 101	5,6	137 ± 41	593 ± 68	25	31 ± 10	898 ± 111
3	162 ± 35	362 ± 46	7	108 ± 34	298 ± 54	26, 27	194 ± 56	807 ± 47
4	203 ± 49	365 + 58	8	68 ± 9	163 ± 55	28	57 ± 22	804 ± 127
5,6	146 ± 33	538 ± 47	9,10	186 ± 50	1329 ± 158	29	96 ± 23	343 ± 28
7	108 ± 29	250 ± 61	11	108 ± 33	437 ± 100	30	66 ± 14	798 ± 44
8	109 ± 86	162 ± 46	12	112 ± 36	1047 ± 151	31,32	148 ± 28	901 ± 73
9,10	171 ± 37	885 ± 115	13,14	146 ± 41	899 ± 166	33,34	215 ± 46	602 ± 22
			15,16	122 ± 17	968 ± 163	35	70 ± 26	ND
	λCOL-031		17,18,19	213 ± 42	537 ± 51			
30	78 ± 11	700 ± 50	20	72 ± 23	587 ± 88		λCOL-871	
31,32	123 ± 11	818 ± 129	21	77 ± 25	1536 ± 203	13–29	ND	342 ± 38
33,34	161 ± 38	579 ± 91	22,23	113 ± 28	764 ± 160	30	71 ± 31	805 ± 99
35	115 ± 28	101 ± 27	24	117 ± 34	1251 ± 140	31,32	130 ± 31	912 ± 120
36	113 ± 34	1763 ± 232	25	79 ± 36	854 ± 104	33,34	151 ± 27	593 ± 84
37	57 ± 24	2276 ± 244	26,27	178 ± 52	825 ± 133	35	104 ± 24	106 ± 22
38,39,40	137 ± 41	1840 ± 274	28	76 ±15	798 ± 77	36	123 ± 22	1926 ± 129
41	63 ± 16	200 ± 55	29	92 ± 24	ND	37	78 ± 38	2240 ± 268
42	55 ± 17	493 ± 120						
43	59 ± 22	371 ± 82		λCOL-323				
44	54 ± 23	276 ± 39	47	89 ± 32	1283 ± 120			
45	53 ± 13	81 ± 27	48	65 ± 30	1562 ± 162			
46	75 ± 10	1891 ± 137	49	50 ± 15	608 ± 63			
47	61 ± 22	1256 ± 172	50	48 ± 17	2143 ± 218			
48	83 ± 11	1153 ± 135	51	154 ± 37				

Double-stranded simian virus 40 DNA circles were used as length standards. Measurements of introns correspond to intron immediately adjacent to the 5′ side of each exon. Based on the size measurements of the cloned chicken inserts obtained by gel electrophoresis, the electron microscope measurements appeared to be overestimates by about 10% and were corrected accordingly. These overestimates are probably due to the single-stranded intron measurements because the length standard was double stranded. When two or more exon numbers appear on the same line, a small intron, which separates two adjacent exons, was detected. In such cases, the exon length reported corresponds to two (or three) exons plus the intervening sequence(s) between them. The measurements for λCOL-204 have been reported (13). The data are shown as mean ± SD. ND, not determined.

represents sequences flanking the 5′ end of the gene. If this is the case, the size of the α2 collagen gene is approximately 37 kb, which is about 8 times more than the total length of mature α2 collagen mRNA (8).

DNA Sequence Analysis Confirms the Identity of the α2 Collagen Genomic Clones. We previously determined the DNA sequence of a small portion of λCOL-204 and identified the residues in the α2 collagen protein encoded by this segment (13). The segment of λCOL-204 that we analyzed is present at the 3′ end of λCOL-271. Fig. 4 shows a restriction map of a 3.2-kb *Eco*RI fragment located at the 5′ end of λCOL-271. We determined the sequence of the DNA adjacent to the *Eco*RI site at the 5′ end of this fragment because electron microscopic analysis indicated that there was an exon at the 5′ proximal portion of the insert in λCOL-271. Fig. 4 indicates that this segment encodes a portion of the α2 collagen protein from residue 391 on (21). The sequence of another DNA segment in λCOL-031 has been determined (Y. Yamada, personal communication). The amino acid sequence in the part of chicken α2 collagen that is specified by this segment in λCOL-031 has not yet been determined. However, the amino acid sequence that was derived from the DNA sequence in λCOL-031 is identical to the known amino acid sequence of bovine α2 collagen at positions 55–72 (22).

DISCUSSION

By successive screenings of a library of random chicken genomic DNA fragments we have isolated a series of overlapping clones which contain the α2 collagen gene. The first genomic clone (13) was isolated by using as hybridization probe a α2

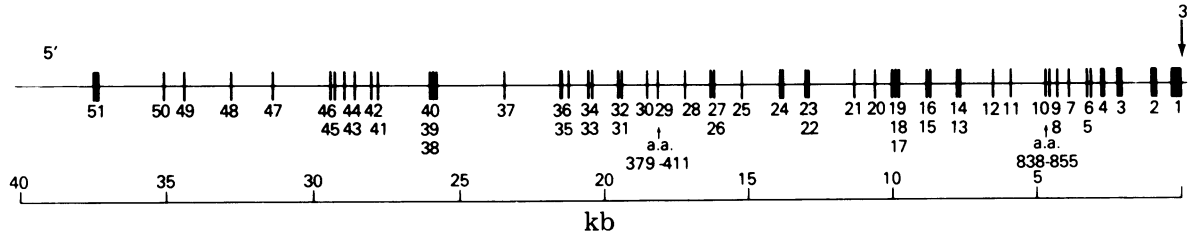

FIG. 3. Graphic representation of measurements in Table 2. Vertical bars correspond to exons. Horizontal lines between the vertical bars represent introns or intervening sequences.

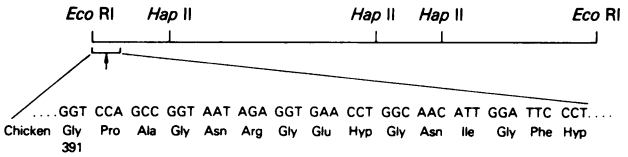

FIG. 4. Sequence analysis of a small DNA segment (bracket at vertical arrow) of λCOL-271. The *Eco*RI fragment represented is a 3.2-kb fragment located adjacent to the long λ arm of λCOL-271. This fragment was subcloned in pBR313. The fragment was 5′ end-labeled, digested with endonuclease *Hap* II, and fractionated by 5% acrylamide electrophoresis. DNA sequence was determined by the method of Maxam and Gilbert (20). The deduced amino acid sequence corresponds to that reported by Dixit *et al.* (21).

collagen cDNA clone which we had constructed (9). A fragment at the 5′ end of this first genomic clone was used as hybridization probe in the next step, then a fragment at the 5′ end of the next clone, etc. These successive steps allowed us to walk down the α2 collagen gene. Several types of experiments demonstrated that the clones are truly overlapping and that our screening procedure did not result in the isolation of other collagen genes. First, heteroduplex analysis between pairs of overlapping clones showed that their homologous segments are large and uninterrupted and occur at the expected location within the recombinant clones. Second, the sizes and succession of coding and intervening sequences were identical in the overlapping segments. Third, the same restriction enzyme sites were found at identical locations in the overlapping segments of the clones. Hence, the clones that we isolated cover 55 kb of sequences that are contiguous in chicken genomic DNA. Sequence analysis of three small DNA segments in three different clones showed that these clones contained sequences coding for amino acid residues in α2 collagen which were found at the expected locations.

Of the 55 kb that we isolated, the α2 collagen gene occupies at least 37 kb. Electron microscopic examination of hybrids between α2 collagen RNA and the different clones shows that the coding information of the α2 collagen gene is subdivided into more than 50 coding segments. These exons are interrupted by intervening sequences of various sizes. Of all genes isolated so far, the α2 collagen gene contains the largest number of exons. The electron microscopic measurements, summarized in Fig. 3, show that the coding sequences at the 3′ end of the gene—and hence in the region that corresponds to the C peptide of proα2 collagen—are larger than those in the rest of the gene. Except for exon 51 and exons 1–4, many other coding sequences are probably smaller than 100 bp. It should be noted that electron microscope measurements are not accurate for DNA segments smaller than 100 bp.

We believe that the 7-kb segment that extends beyond the last coding sequence (number 51) contains the 5′ flanking sequence of the gene. This region does not hybridize to α2 collagen mRNA, and no mRNA tail could be visualized by electron microscopy beyond the last coding region. However, we have not yet excluded the possibility that this 7-kb noncoding region could be a large intron located close to the 5′ end of the gene. We have previously reported that clone λCOL-204 probably contains the 3′ end of this gene (13).

The complex structure of the α2 collagen gene implies that the conversion of the primary transcript of this gene to cytoplasmic translatable mRNA includes at least 50 precise excision and splicing events. In fact, this number is an underestimate because the removal of at least some of the intervening sequences occurs in more than one step (23).

We have recently determined that many of the exons that specify the helical portion of α2 collagen have an identical length, 54 bp, although their sequences vary, implying that the ancestral gene for collagen arose by amplification of a single genetic unit containing an exon of unique size (24).

We thank J. Dodgson, R. Axel, D. Engel, J. Slightom, F. R. Blattner, and their colleagues for their generosity in sending us the chicken genomic libraries. We thank Sherryl Adams for a collagen RNA preparation. We are grateful to Linda Harris for secretarial assistance.

1. Eyre, D. R. (1980) *Science* **207,** 1315–1322.
2. Prockop, D. J., Kivirikko, K. I., Tuderman, L. & Guzman, K. A. (1979) *N. Engl. J. Med.* **301,** 13–23 and 77–85.
3. Reddi, A. H. (1976) in *Biochemistry of Collagen,* eds. Ramachandran, G. N. & Reddi, A. H. (Plenum, New York), pp. 449–478.
4. Levinson, W., Bhatnagar, R. S. & Liu, T. (1975) *J. Natl. Cancer Inst.* **55,** 807–810.
5. Delclos, K. B. & Blumberg, P. M. (1979) *Cancer Res.* **39,** 1667–1672.
6. Adams, S. L., Sobel, M. E., Howard, B. H., Olden, K., Yamada, K. M., de Crombrugghe, B. & Pastan, I. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 3399–3403.
7. Howard, B. H., Adams, S. L., Sobel, M. E., Pastan, I. & de Crombrugghe, B. (1978) *J. Biol. Chem.* **253,** 5869–5874.
8. Adams, S. L., Alwine, J. C., de Crombrugghe, B. & Pastan, I. (1979) *J. Biol. Chem.* **254,** 4935–4938.
9. Sobel, M. E., Yamamoto, T., Adams, S. L., DiLauro, R., Avvedimento, V. E., de Crombrugghe, B. & Pastan, I. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 5846–5850.
10. Yamamoto, T., Sobel, M. E., Adams, S. L., Avvedimento, V. E., DiLauro, R., Pastan, I., de Crombrugghe, B., Showalter, A., Pesciotta, D., Fietzek, P. & Olsen, B. (1980) *J. Biol. Chem.* **255,** 2612–2615.
11. 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.
12. Lehrach, H., Frischauf, A. M., Hanahan, D., Wozney, J., Fuller, F. & Boedtker, H. (1979) *Biochemistry* **18,** 3146–3152.
13. Vogeli G., Avvedimento, E. V., Sullivan, M. A., Maizel, J. V. Jr., Lozano, G., Adams, S. L., Pastan, I. & de Crombrugghe, B. (1980) *Nucleic Acids Res.* **8,** 1823–1837.
14. Boyd, G. D., Tolstoshev, P., Shafer, M. P., Trapnell, B. C., Coon, H. C., Kretshmer, P. J., Nienhuis, A. W. & Crystal, R. G. (1980) *J. Biol Chem.* **255,** 3212–3220.
15. Vogelstein, B. & Gillespie, D. (1979) *Proc. Natl. Acad. Sci. USA* **76,** 615–619.
16. Maniatis, T., Jeffrey, A. & Kleid, D. G. (1975) *Proc. Natl. Acad. Sci. USA* **72,** 1184–1188.
17. Benton, W. D. & Davis, R. W. (1977) *Science* **196,** 180–182.
18. Blattner, F. R., Williams, B. G., Blechl, A. E., Thompson, K. D., Farber, H. E., Furlong, L., Grunwald, D. J., Kiefer, D. O., Moore, D. D., Schumm, J. W., Sheldon, E. L. & Smithies, O. (1977) *Science* **196,** 161–169.
19. Dodgson, J. B., Strommer, J. & Engel, J. D. (1979) *Cell* **17,** 879–887.
20. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 560–564.
21. Dixit, S. N., Seyer, J. M. & Kang, A. H. (1977) *Eur. J. Biochem.* **73,** 213–221.
22. Fietzek, P. P. & Rexrodt, F. W. (1975) *Eur. J. Biochem.* **59,** 113–119.
23. Avvedimento, V. E., Vogeli, G., Yamada, Y., Maizel, J. V., Pastan, I. & de Crombrugghe, B. (1980) *Cell* **21,** 689–696.
24. Yamada, Y., Avvedimento, V. E., Mudryg, M., Ohkubo, H., Vogeli, G., Irani, M., Pastan, I. & de Crombrugghe, B. (1980) *Cell,* in press.