Volume 11 Number 9 1983 Nucleic Acids Research

A conserved nucleotide sequence, coding for a segment of the C-propeptide, is found at the same location in different collagen genes

Yoshihiko Yamada, Klaus Kühn* and Benoit de Crombrugghe

Laboratory of Molecular Biology, National Cancer Institute, National Institutes of Health, Bethesda, MD 20205, USA

Received 13 January 1983; Revised and Accepted 5 April 1983

ABSTRACT

The nucleotide sequence of a segment of the chick α1 type III collagen gene which codes for the C-propeptide was determined and compared with the corresponding sequence in the α1 type I and α2 type I collagen genes. As in the α2 type I gene the coding information for the C-propeptide of the type III collagen gene is subdivided in four exons. Similarly, the amino proximal exon contains sequences for both the carboxy terminal end of the α-helical segment of collagen and for the beginning of the C-propeptide in both genes. Therefore, this organization of exons must have been established before these two collagen genes arose by duplication of a common ancestor. In several subsegments the deduced amino acid sequence for the C-propeptide of type III collagen shows a strong homology with the corresponding amino acid sequence in α1 and α2 type I. For one of these homologous amino acid sequences, however, the nucleotide sequence is much better conserved than for the others. It is possible that a mechanism of gene conversion has maintained the homogeneity of this nucleotide sequence among the interstitial collagen genes. Alternatively, the conserved nucleotide sequence may represent a regulatory signal which could function either in the DNA or in the RNA.

INTRODUCTION

The collagens are a family of chemically and structurally related proteins that are found in the connective tissues of many organisms (for reviews, see 1, 2). Because the five or more collagen types that are found in higher vertebrates show a tissue specific distribution pattern, it is likely that their synthesis is controlled by tissue specific developmental programs. Type III collagen synthesis occurs simultaneously with type I in many connective tissues but the relative concentrations of these two collagens vary greatly in different tissues.

The polypeptide chains of the interstitial collagens are synthesized as precursor molecules, preprocollagens. Beginning at the amino terminal end they contain a signal peptide, a N-propeptide, the helical portion of the protein with the repeating Gly-X-Y motif and a C-propeptide. Whereas the signal peptide is removed within the cell, the amino and carboxy propeptides are

cleaved off by specific proteases after the molecule has been secreted into the extracellular matrix. Several functions have been proposed for the propeptides. The C-propeptide has been postulated to play a prominent role in the correct positioning of the chains during formation of the collagen triple helix (3). The assembly of a correctly paired triple helix probably requires some anchor point outside the helix to ensure a proper alignment of the polypeptide chains. The N-propeptide, on the other hand, is thought to play a role in translational feedback control of collagen biosynthesis (4, 5).

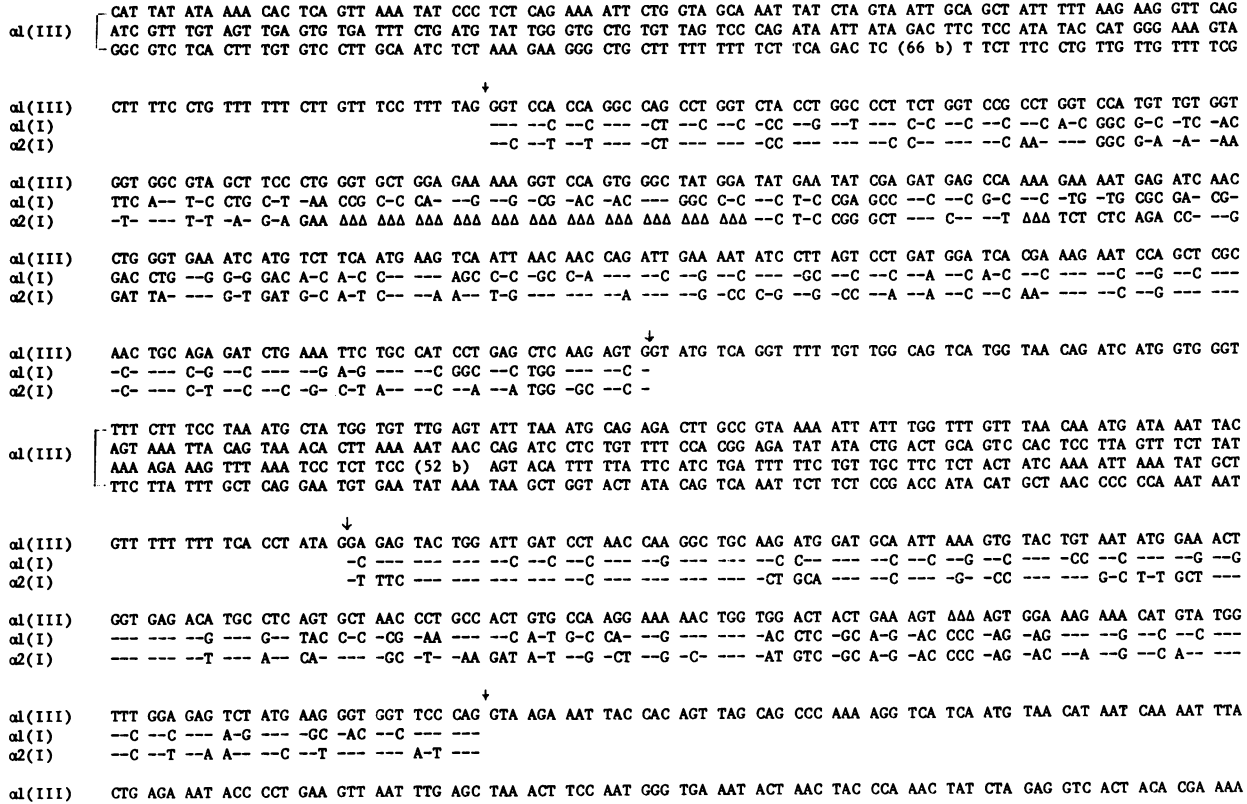
We have recently isolated the entire gene for chick α1 type III collagen (6, Y. Yamada, unpublished results) and wish to compare its structure, evolutionary assembly and regulation with that of the α2 type I gene which we, and others, previously isolated and characterized (7, 8, 9). In this paper we report the nucleotide sequence of the DNA segment that codes for the C-propeptide of α1 type III collagen and compare it with that of the α1 and α2 type I collagen genes. Several segments of the C-propeptide reveal amino acid sequence homologies in all three collagens. In these conserved segments the nucleotide sequence is often altered in the silent third base position. For one of the homologous amino acid segments the nucleotide sequence is clearly better conserved than in the others. This conserved nucleotide sequence contains about 50 bp and occurs at the same place in all interstitial collagen genes including chick type II (Upholt and Sandell, personal communication) and human α2 type I (10). Possible mechanisms for maintaining the homogeneity of this nucleotide sequence are discussed.

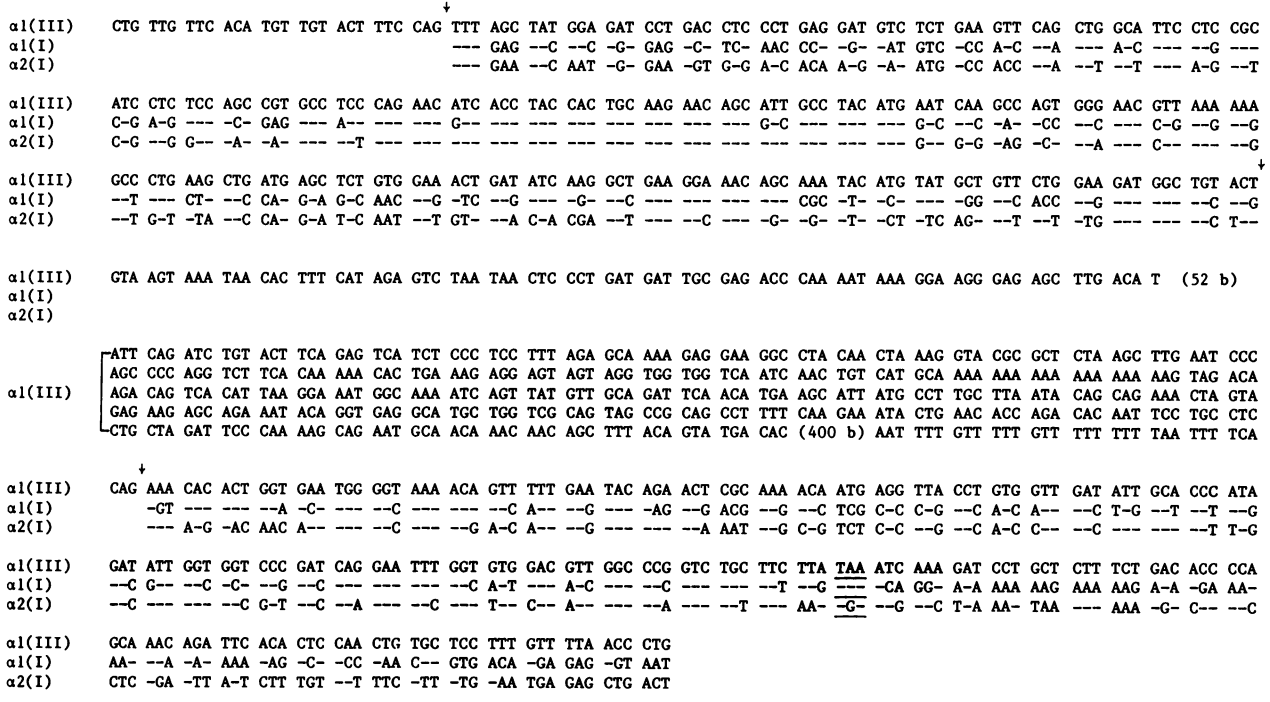
MATERIALS AND METHODS

The isolation of a genomic clone, λC3-C1-24, encoding a segment of the chick α1 type III collagen gene has been described (6). This clone contains about 9 kb of the gene. Three segments of the 3' portion of the gene were subcloned in plasmid pBR322. DNA restriction fragments were isolated from these plasmids by electrophoresis on agarose and acrylamide gels and were labeled with [γ-^{32}P]ATP by T4 kinase. Nucleotide sequence of the DNA fragments was determined by the method of Maxam and Gilbert (11).

RESULTS

We have isolated a chick genomic DNA clone which specifies the 3' part of the gene for chick type III collagen (6). The clone was identified by determining the DNA sequence of a segment coding for the helical part of the protein and by comparison of the deduced amino acid sequence with the previously de-


α1(III) CAT TAT ATA AAA CAC TCA GTT AAA TAT CCC TCT CAG AAA ATT CTG GTA GCA AAT TAT CTA GTA ATT GCA GCT ATT TTT AAG AAG GTT CAG
ATC GTT TGT AGT TGA GTG TGA TTT CTG ATG TAT TGG GTG CTG TGT TAG TCC CAG ATA ATT ATA GAC TTC TCC ATA TAC CAT GGG AAA GTA
GGC GTC TCA CTT TGT GTC CTT GCA ATC TCT AAA GAA GGG CTG CTT TTT TTT TCT TCA GAC TC (66 b) T TCT TTC CTG TTG TTG TTT TCG
α1(III) CTT TTC CTG TTT TTT CTT GTT TCC TTT TAG GGT CCA CCA GGC CAG CCT GGT CTA CCT GGC CCT TCT GGT CCG CCT GGT CCA TGT TGT GGT
α1(I) --- --C --C --- -CT --C --C -CC --G --T --- C-C --C --C --C A-C GGC G-C -TC -AC
α2(I) --C --T --T --- -CT --- --- -CC --- --- --C C-- --- --C AA- --- GGC G-A -A- -AA
α1(III) GGT GGC GTA GCT TCC CTG GGT GCT GGA GAA AAA GGT CCA GTG GGC TAT GGA TAT GAA TAT CGA GAT GAG CCA AAA GAA AAT GAG ATC AAC
α1(I) TTC A-- T-C CTG C-T -AA CCG C-C CA- --G --G -CG -AC -AC --- GGC C-C --C T-C CGA GCC --C --C G-C --C -TG -TG CGC GA- CG-
α2(I) -T- --- T-T -A- G-A GAA ΔΔΔ ΔΔΔ ΔΔΔ ΔΔΔ ΔΔΔ ΔΔΔ ΔΔΔ ΔΔΔ ΔΔΔ ΔΔΔ --C T-C CGG GCT --- C-- --T ΔΔΔ TCT CTC AGA CC- --G
α1(III) CTG GGT GAA ATC ATG TCT TCA ATG AAG TCA ATT AAC AAC CAG ATT GAA AAT ATC CTT AGT CCT GAT GGA TCA CGA AAG AAT CCA GCT CGC
α1(I) GAC CTG --G G-G GAC A-C A-C C-- --- AGC C-C -GC C-A --- --C --G --C --- -GC --C --C --A --C A-C --C --- --C --G --C ---
α2(I) GAT TA- --- G-T GAT G-C A-T C-- --A A-- T-G --- --- --A --- --G -CC C-G --G -CC --A --A --C --C AA- --- --C --G --- ---
α1(III) AAC TGC AGA GAT CTG AAA TTC TGC CAT CCT GAG CTC AAG AGT GGT ATG TCA GGT TTT TGT TGG CAG TCA TGG TAA CAG ATC ATG GTG GGT
α1(I) -C- --- C-G --C --- --G A-G --- --C GGC --C TGG --- --C -
α2(I) -C- --- C-T --C --C -G- C-T A-- --C --A --A TGG -GC --C -
α1(III) TTT CTT TCC TAA ATG CTA TGG TGT TTG AGT ATT TAA ATG CAG AGA CTT GCC GTA AAA ATT ATT TGG TTT GTT TAA CAA ATG ATA AAT TAC
AGT AAA TTA CAG TAA ACA CTT AAA AAT AAC CAG ATC CTC TGT TTT CCA CGG AGA TAT ATA CTG ACT GCA GTC CAC TCC TTA GTT TCT TAT
AAA AGA AAG TTT AAA TCC TCT TCC (52 b) AGT ACA TTT TTA TTC ATC TGA TTT TTC TGT TGC TTC TCT ACT ATC AAA ATT AAA TAT GCT
TTC TTA TTT GCT CAG GAA TGT GAA TAT AAA TAA GCT GGT ACT ATA CAG TCA AAT TCT TCT CCG ACC ATA CAT GCT AAC CCC CCA AAT AAT
α1(III) GTT TTT TTT TCA CCT ATA GGA GAG TAC TGG ATT GAT CCT AAC CAA GGC TGC AAG ATG GAT GCA ATT AAA GTG TAC TGT AAT ATG GAA ACT
α1(I) -C --- --- --- --C --C --C --- --G --- --- --C C-- --- --C --C --G --C --- -CC --C --- --G --G
α2(I) -T TTC --- --- --- --- --C --- --- --- --- -CT GCA --- --C --- -G- -CC --- --- G-C T-T GCT ---
α1(III) GGT GAG ACA TGC CTC AGT GCT AAC CCT GCC ACT GTG CCA AGG AAA AAC TGG TGG ACT ACT GAA AGT ΔΔΔ AGT GGA AAG AAA CAT GTA TGG
α1(I) --- --- --G --- G-- TAC C-C -CG -AA --- --C A-T G-C CA- --G --- --- -AC CTC -GC A-G -AC CCC -AG -AG --- --G --C --C ---
α2(I) --- --- --T --- A-- CA- --- -GC -T- -AA GAT A-T --G -CT --G -C- --- -AT GTC -GC A-G -AC CCC -AG -AC --A --G --C A-- ---
α1(III) TTT GGA GAG TCT ATG AAG GGT GGT TCC CAG GTA AGA AAT TAC CAC AGT TAG CAG CCC AAA AGG TCA TCA ATG TAA CAT AAT CAA AAT TTA
α1(I) --C --C --- A-G --- -GC -AC --C --- ---
α2(I) --C --T --A A-- --C --T --- --- A-T ---
α1(III) CTG AGA AAT ACC CCT GAA GTT AAT TTG AGC TAA ACT TCC AAT GGG TGA AAT ACT AAC TAC CCA AAC TAT CTA GAG GTC ACT ACA CGA AAA


α1(III) CTG TTG TTC ACA TGT TGT ACT TTC CAG TTT AGC TAT GGA GAT CCT GAC CTC CCT GAG GAT GTC TCT GAA GTT CAG CTG GCA TTC CTC CGC
α1(I) --- GAG --C --C -G- GAG -C- TC- AAC CC- -G- -AT GTC -CC A-C --A --- A-C --- --G ---
α2(I) --- GAA --C AAT -G- GAA -GT G-G A-C ACA A-G -A- ATG -CC ACC --A --T --T --- A-G --T
α1(III) ATC CTC TCC AGC CGT GCC TCC CAG AAC ATC ACC TAC CAC TGC AAG AAC AGC ATT GCC TAC ATG AAT CAA GCC AGT GGG AAC GTT AAA AAA
α1(I) C-G A-G --- -C- GAG --- A-- --- --- G-- --- --- --- --- --- --- G-C --- --- --- G-C --C -A- -CC --C --- C-G --G --G
α2(I) C-G --G G-- -A- -A- --- --T --- --- --- --- --- --- --- --- --- --- --- --- --- --- G-- G-G -AG -C- --A --- C-- --- --G
α1(III) GCC CTG AAG CTG ATG AGC TCT GTG GAA ACT GAT ATC AAG GCT GAA GGA AAC AGC AAA TAC ATG TAT GCT GTT CTG GAA GAT GGC TGT ACT
α1(I) --T --- CT- --C CA- G-A G-C AAC --G -TC --G --- -G- --C --- --- --- --- CGC -T- -C- --- -GG --C ACC --G --- --- --C --G
α2(I) --T G-T -TA --C CA- G-A T-C AAT --T GT- --A C-A CGA --T --- --C --- -G- -G- -T- -CT -TC AG- --T --T -TG --- --- --C T--
α1(III) GTA AGT AAA TAA CAC TTT CAT AGA GTC TAA TAA CTC CCT GAT GAT TGC GAG ACC CAA AAT AAA GGA AGG GAG AGC TTG ACA T (52 b)
α1(I)
α2(I)
α1(III) ATT CAG ATC TGT ACT TCA GAG TCA TCT CCC TCC TTT AGA GCA AAA GAG GAA GGC CTA CAA CTA AAG GTA CGC GCT CTA AGC TTG AAT CCC
AGC CCC AGG TCT TCA CAA AAA CAC TGA AAG AGG AGT AGT AGG TGG TGG TCA ATC AAC TGT CAT GCA AAA AAA AAA AAA AAA AAG TAG ACA
AGA CAG TCA CAT TAA GGA AAT GGC AAA ATC AGT TAT GTT GCA GAT TCA ACA TGA AGC ATT ATG CCT TGC TTA ATA CAG CAG AAA CTA GTA
GAG AAG AGC AGA AAT ACA GGT GAG GCA TGC TGG TCG CAG TAG CCG CAG CCT TTT CAA GAA ATA CTG AAC ACC AGA CAC AAT TCC TGC CTC
CTG CTA GAT TCC CAA AAG CAG AAT GCA ACA AAC AAC AGC TTT ACA GTA TGA CAC (400 b) AAT TTT GTT TTT GTT TTT TTT TAA TTT TCA
α1(III) CAG AAA CAC ACT GGT GAA TGG GGT AAA ACA GTT TTT GAA TAC AGA ACT CGC AAA ACA ATG AGG TTA CCT GTG GTT GAT ATT GCA CCC ATA
α1(I) -GT --- --- --A -C- --- --C --- --- --C A-- --G --- -AG --G ACG --G --C TCG C-C C-G --C A-C A-- --C T-G --T --T --G
α2(I) --- A-G -AC AAC A-- --- --C --- --G A-C A-- --G --- --- --A AAT --G C-G TCT C-C --G --C A-C C-- --C --- --- --T T-G
α1(III) GAT ATT GGT GGT CCC GAT CAG GAA TTT GGT GTG GAC GTT GGC CCG GTC TGC TTC TTA TAA ATC AAA GAT CCT GCT CTT TCT GAC ACC CCA
α1(I) --C G-- --C -C- --G --C --- --- --- --C A-T --- A-C --- --C --- --- --T --G --- -CA GG- A-A AAA AAG AAA AAG A-A -GA AA-
α2(I) --C --- --- --C G-T --C --A --- --C --- T-- C-- A-- --- --A --- --T --- AA- -G- --G --C T-A AA- TAA --- AAA -G- C-- --C
α1(III) GCA AAC AGA TTC ACA CTC CAA CTG TGC TCC TTT GTT TTA ACC CTG
α1(I) AA- --A -A- AAA -AG -C- -CC -AA C-- GTG ACA -GA GAG -GT AAT
α2(I) CTC -GA -TT A-T CTT TGT --T TTC -TT -TG -AA TGA GAG CTG ACT

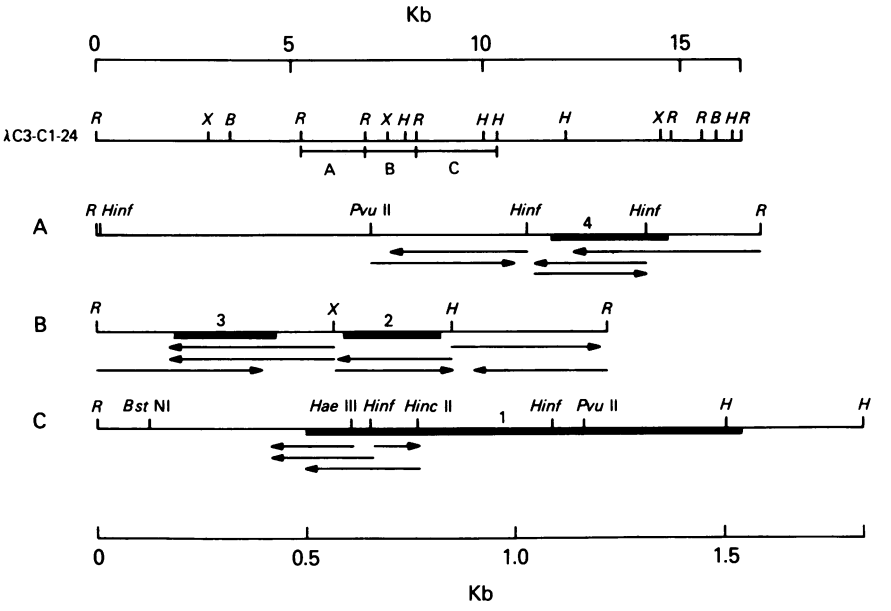


Fig. 2. Sequencing strategy for the chick α1 type III gene. The top map represents the 17 kb fragment derived from genomic clone λC3-C1-24 that contains the 3' portion of the chick α1 type III collagen gene. (A), (B) and (C) represent restriction fragments showing the various sites used for the sequence. The arrows indicate the direction of the sequence. The exons represented by the solid boxes are numbered from 1 to 4 beginning at the 3' end of the gene. B, BamHI; H, HindIII; R, EcoRI; X, XbaI.

termined protein sequence of calf and human α1 type III collagen (12, 13). The nucleotide sequence of the segment of the α1 type III collagen gene that specifies the C-propeptide is presented in Fig. 1. Fig. 2 shows a map of the DNA restriction fragments that were used to determine this DNA sequence and indicates the direction of sequencing for each of these fragments. Fig. 1 also compares the nucleotide sequence of the α1 type III collagen gene with the cor-

Fig. 1. Nucleotide sequence of the 3' portion of the chick α1 type III collagen gene and comparison with the equivalent sequence in the α1 type I and α2 type I genes. The comparison is made only for exon sequences. Intron sequences are only given for the type III collagen gene. The nucleotide sequence alignment was based on an optimal amino acid alignment of the three collagens. Nucleotide sequences of α1 type I and α2 type I which differ from α1 type III are shown. Arrows indicate splicing sites of exons. △, represent deletions; --, represent bases that are identical to α1 type III. The first stop codon in exon 1 is underlined. The nucleotide sequence of α1 type I is taken from ref. 14, that of α2 type I from ref. 15.

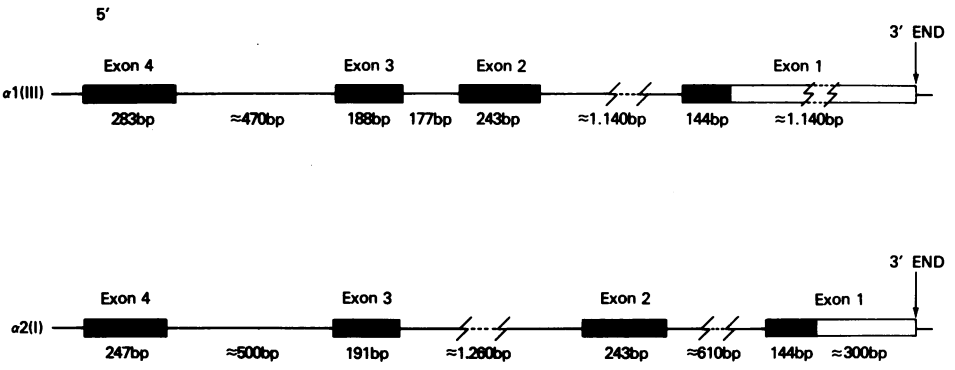


Fig. 3. Exon-intron organization of the α1 type III and α2 type I collagen genes for segments coding for their respective C-propeptide. The exon structure for α2 type I is from ref. 15.

responding nucleotide sequence for α1 and α2 type I collagen (14). Whereas a genomic clone for the α1 type I collagen gene is not yet available, the gene for the α2 type I collagen has been isolated (7, 9) and the nucleotide sequence of the exons encoding the C-propeptide of this collagen has been determined (15). A small number of differences exists between the cDNA sequence and the sequence of the exons.

Exon organization

As in the α2 type I collagen gene, the coding information for the C-propeptide of α1 type III collagen is subdivided in four exons, which we will designate exons 1 to 4, beginning with the exon coding for the carboxy terminal part of the C-propeptide (see Fig. 3). Exon 4 contains 36 bp less in the α2 type I collagen gene than in the α1 type III collagen gene. Exon 3 contains 3 bp less in the α1 type III collagen gene than in the α2 type I collagen gene. We note also that as in the α2 type I collagen gene, exon 3 contains only the first two bases of the glycine codon located at its 5' end. The coding sequences of exon 2 and exon 1 are identical in length in both genes although the 3' untranslated segment of the α1 type III collagen gene is much longer than in the α2 type I collagen gene (Yamada, unpublished data). The length of the 3'-untranslated sequence was determined by R-loop analysis (6) and by S1 mapping experiments (data not shown). The size of the introns separating these exons is, by contrast, very different in the two genes.

The amino acid sequence (Fig. 4) that is deduced from the nucleotide sequence of exon 1 reveals several features that are characteristic of type III collagens. First, two successive cysteine residues are found at the end of

the α-helical portion. These two cysteine residues form interchain disulfide bonds in type III collagen (12, 13); they are not found in α1 or α2 type I collagens (14, 15) or in α1 type II collagen (W. Upholt, personal communication). Second, five tandem repeats of the Gly-Pro-Pro tripeptide are found at the carboxy terminal end of the α-helical domain in both α1 and α2 type I collagen (14, 15). Such repeats do not occur in the α1 type III collagen (12, 13). It is of interest to note that both in α2 type I and α1 type III, the amino terminal portion of exon 4 contains 6 Gly-X-Y units. These 18 amino acids are encoded by 54 bp, a length that corresponds to the conserved size of collagen gene exons. It is likely, therefore, that exon 4 evolved by fusion between a 54 bp exon and an exon coding for the beginning of the nonhelical peptide.

The amino acid sequence of the telopeptide and the beginning of the C-propeptide diverge considerably between type I and type III collagen. It has been proposed that the recognition site for the procollagen C-protease in α1 and α2 type I procollagen includes the sequence Tyr-Tyr-Arg-Ala-Asp-Glu, the Ala-Asp bond being cleaved by the enzyme (14, 15). The sequence of the corresponding site in α1 type III collagen appears to be Tyr-Glu-Tyr-Arg-Asp-Glu. It is possible that the C-protease would cleave the bond between Arg and Asp.

A conserved nucleotide sequence in exon 2

There are several stretches of amino acids that are conserved in all three collagens (see Fig. 4). For instance, residues 75 to 85 are identical in all three collagens except for one difference in α1 type III. Residues 94 to 105 are identical in all three genes except for one residue in α2 type I. Residues 108 to 122 and residues 143 to 152 are conserved between α1 type I and α1 type III. The most conserved segment spans residues 184 to 199 in which 16 amino acids are identical in α2 type I and α1 type III collagen. In this region, 13 amino acids in α1 type I are also found at the same place as in α2 type I and α1 type III. Overall in the C-propeptide, 145 of 246 amino acids or 59%, are conserved between α1 type I and α1 type III collagen. Similarly, 150 of 246 amino acids, or 61%, are conserved in α1 type I and α2 type I. On the other hand, less homology of amino acid sequence (50%) was found between α2 type I and α1 type III.

The overall nucleotide sequence divergence of the coding segment appears to be consistent with an evolutionary distance of at least 200 x 10^6 years. This period corresponds approximately to the time of speciation of the avians. In places where amino acid residues are unchanged, the nucleotide sequence often shows a change in the silent position of the codons. As far as can be

```
          1                                   10                              *   *   20                                  30
α1(III)   Gly Pro Pro Gly Gln Pro Gly Leu Pro Gly Pro Ser Gly Pro Pro Gly Pro Cys|Cys Gly Gly Gly Val Ala Ser Leu Gly Ala Gly Glu Lys Gly Pro
α1(I)                     Pro         Pro             Pro             Ser Gly Gly Phe Asp Phe Ser Phe Leu Pro Gln Pro Pro Gln         Ala His
α2(I)                     Pro         Pro             Pro         Asn Gly Gly Gly Tyr Glu Val     Phe Asp Ala Glu  Δ   Δ   Δ   Δ   Δ   Δ   Δ

                                  40                                      50                                      60
α1(III)   Val Gly Tyr Gly Try Glu Tyr Arg Asp Glu Pro Lys Glu Asn Glu Ile Asn Leu Gly Glu Ile Met Ser Ser Met Lys Ser Ile Asn Asn Gln Ile Glu
α1(I)     Asp     Gly Arg     Tyr Arg Ala     Asp Ala Asn Val Met Arg Asp Arg Asp Leu     Val Asp Thr Thr Leu         Leu Ser Gln
α2(I)      Δ   Δ   Δ   Δ      Tyr Arg Ala     Gln      Δ  Ser Leu Arg Pro Lys Asp Tyr     Val Asp Ala Thr Leu     Thr Leu
                                         Λ
                      70                                      80       *                       *      90                  ↓
α1(III)   Asn Ile Leu Ser Pro Asp Gly Ser Arg Lys Asn Pro Ala Arg Asn Cys Arg Asp Leu Lys Phe Cys His Pro Glu Leu Lys Ser Gly Glu Tyr Trp Ile
α1(I)             Arg         Glu     Thr                         Thr                     Met         Gly Asp Trp
α2(I)     Thr Leu     Thr     Glu         Lys                     Thr                 Arg Leu Ser             Trp Ser         Phe

          100                  *                  110              *                      120      *                          130
α1(III)   Asp Pro Asn Gln Gly Cys Lys Met Asp Ala Ile Lys Val Tyr Cys Asn Met Glu Thr Gly Glu Thr Cys Leu Ser Ala Asn Pro Ala Thr Val Pro Arg
α1(I)                         Asn Leu                                                         Val Tyr Pro Thr Gln         Ile Ala Gln
α2(I)                         Thr Ala         Arg Ala         Asp Phe Ala                     Ile His     Ser Leu Glu Asp Ile     Thr

                                      140                                     150                        ↓        160
α1(III)   Lys Asn Trp Trp Thr Thr Glu Ser  Δ  Ser Gly Lys Lys His Val Trp Phe Gly Glu Ser Met Lys Gly Gly Phe Gln Phe Ser Tyr Gly Asp Pro Asp
α1(I)                 Tyr Leu Ser Lys Asn Pro Lys Glu                                 Thr     Ser Asp                 Glu         Gly Glu Gly
α2(I)         Thr     Tyr Val Ser Lys Asn Pro Lys Asp             Ile                 Thr Ile Asn         Thr         Glu     Asn Gly Glu Gly

                              170                                     180                      †          190      *
α1(III)   Leu Pro Glu Asp Val Ser Glu Val Gln Leu Ala Phe Leu Arg Ile Leu Ser Ser Arg Ala Ser Gln Asn Ile Thr Tyr His Cys Lys Asn Ser Ile Ala
α1(I)     Ser Asn Pro Ala Asp Val Ala Ile         Thr             Leu Met     Thr Glu     Thr         Val                             Val
α2(I)     Val Thr Thr Lys Asp Met Ala Thr                 Met     Leu     Ala Asn His

                  200                                     210                                     220                                     230
α1(III)   Tyr Met Asn Gln Ala Ser Gly Asn Val Lys Lys Ala Leu Lys Leu Met Ser Ser Val Glu Thr Glu Ile Lys Ala Glu Gly Asn Ser Lys Tyr Met Tyr
α1(I)             Asp His Asp Thr         Leu                 Leu     Gln Gly Ala Asn     Ile         Arg                     Arg Phe Thr
α2(I)             Asp Glu Glu Thr         Leu             Val Ile     Gln Gly     Asn Asp Val     Leu Arg                     Arg Phe Thr Phe

                               *     ↓    240                                     250                                     260
α1(III)   Ala Val Leu Glu Asp Gly Cys Thr Lys His Thr Gly Glu Trp Gly Lys Thr Val Phe Glu Tyr Arg Thr Arg Lys Thr Met Arg Leu Pro Val Val Asp
α1(I)     Gly     Thr                     Ser             Ala                     Ile         Lys     Thr         Ser             Ile Ile
α2(I)     Ser         Val             Ser     Lys Asn asn Lys                 Ile Ile                 Asn     Pro Ser             Ile Leu

                                  270                                     280              *
α1(III)   Ile Ala Pro Ile Asp Ile Gly Gly Pro Asp Gln Glu Phe Gly Val Asp Val Gly Pro Val Cys Phe Leu
α1(I)     Leu         Met     Val     Ala                         Ile     Ile
α2(I)                 Leu                 Ala                     Leu His Ile                     Lys
```

determined from the available but incomplete intron sequences, there is no detectable homology between the introns of α2 type I and α1 type III except for the splicing signals. The nucleotide sequences of the 3' untranslated regions show much less homology than the coding sequences in the three collagen genes.

In exon 2 the nucleotide sequence which codes for amino acid residues 184 to 199 is, however, highly conserved in different collagen genes. In this segment 48 bp are identical in the α2 type I and type III genes. The same nucleotide sequence is also conserved in the genes for α1 type I and for type II (W. Upholt, personal communication). An equivalent segment of the human α2 type I gene also shows an extraordinary nucleotide sequence homology with the chick α2 type I gene (10). This homology between equivalent segments of the chicken and human α2 type I collagen genes extends over an additional 100 bp towards the 5' ends of these genes.

DISCUSSION

We have determined the nucleotide sequence of a segment of the chick α1 type III collagen gene which codes for the C-propeptide of this collagen polypeptide. As in the gene for α2 type I collagen, four exons specify the C-propeptide of α1 type III collagen. The exon boundaries were determined by comparison with the known amino acid sequences of α1 type I and α2 type I collagens. The locations of the boundaries were confirmed by the presence of conserved splicing signals at each end of these exons (Fig. 1). The length of the coding region in exon 1 and the length of exon 2 are identical in the gene for α1 type III collagen and in the gene for α2 type I collagen. The size of exon 3 is 3 nucleotides shorter in α1 type III than in α2 type I. Exon 4 is 36 bp longer in the type III than in the α2 type I gene. A comparative analysis of the sequences suggests that this could be due to a 33 bp deletion that includes part of the sequences for the telopeptide of α2 type I plus an additional 3 bp deletion (see Fig. 4). Notwithstanding these differences, it is clear that the exon-intron organization of these two genes is very similar. Additional structural similarities are seen between these two genes. First,

Fig. 4. Amino acid sequence of the carboxy terminal part of α1 type III collagen and comparison with α1 type I and α2 type I. The amino acid sequence deduced from exon 1 to exon 4 of the α1 type III collagen gene is shown. Only amino acid sequences of α1 type I and α2 type I which differ from α1 type III are shown. Symbols used are: △, deletion; *, cysteine residue of α1 type III; ↓, splicing sites of exons; ∧, C-protease cleavage site; +, carbohydrate attachment site; |, end of the helical part of α1 type III collagen.

exon 4 contains 54 bp specifying the carboxy proximal portion of the collagen α-helical segment together with sequences for the telopeptide and the amino terminal part of the C-propeptide in both genes. It is likely, therefore, that this exon resulted from the fusion of at least two exons. Second, the 3' terminal codon in exon 4 is split between exon 4 and exon 3 in both genes. We conclude that the exon arrangement in the portion of these genes that specify the C-propeptide was established before these genes were duplicated from a common ancestor.

Several segments of the C-propeptide show homologies in their amino acid sequence between α1 type I, α2 type I, and type III collagens. In these conserved segments the corresponding nucleotide sequence often shows variations in the silent third base. However, in one of these homologous amino acid segments the corresponding nucleotide sequence is more highly conserved than in the others. One possible explanation for this conserved nucleotide sequence that is found in exon 2 in four different collagen genes is that it represents a common controlling element for these genes. This regulatory signal could be a DNA binding site for a control protein or a site that is critical for RNA processing, stability or transport. It would, however, be unusual for a regulatory sequence to be located in the middle of an exon. Furthermore, there is no obvious symmetrical element in this sequence as is often found in regulatory signals in DNA and RNA.

An alternative hypothesis is that the nucleotide sequence was conserved because the amino acid sequence in this segment plays a critical role in the biosynthesis or the assembly of the collagen molecule. This amino acid sequence contains the unique carbohydrate attachment site of the carboxy propeptide (15). It is possible that this conserved sequence is essential for the correct alignment of the collagen polypeptides during the formation of the triple helix. The reason why the nucleotide sequence is better conserved in one segment whereas only the amino acid sequence is conserved in others is not clear, unless the segment with the conserved nucleotide sequence has a more critical role.

If the conservation of the amino acid sequence is the reason why the nucleotide sequence is conserved, then its conservation must have occurred by a different mechanism than in segments where only the amino acid sequence is conserved. A possible mechanism to maintain nucleotide sequence homogeneity is gene conversion or double unequal crossing-over. The conservation of this same segment in four different collagen genes suggests that the mechanism that was responsible for maintaining the homogeneity of the sequences must have

occurred with a certain frequency. In yeast, gene conversion is more frequent than simple unequal crossing-over (16) and it probably occurs also at a high frequency in mouse cells under the appropriate selective pressure (17). Gene conversion is implicated in mating-type interconversion in yeast (18, 19) and has been postulated as a mechanism to maintain sequence identities in families of related genes (20). Such a mechanism could be important to conserve an ancestral function among the members of a family whereas other parts of the gene would be allowed to acquire independent functions.

The same nucleotide sequence which is conserved in different chick collagen genes is also maintained in the human α2 type I collagen gene (10). Here the homology between the human and chick α2 collagen nucleotide sequence extends approximately another 100 bp towards the 5' end of these genes. This interspecies homology could have been caused by a horizontal exchange of genetic information between chicken and man and may have been mediated by a retrovirus type vector.

ACKNOWLEDGMENT

We wish to thank W. Upholt for providing us information before publication. We thank J. Maizel for advice on the use of sequence analysis programs.

*Present address: Max-Planck-Institut für Biochemie, 8033 Martinsried, München, FRG

REFERENCES

1. Bornstein, P. and Sage, H. (1980) Ann. Biochem. 49, 954-1003.
2. Miller, E.J. and Gray, S. (1982) Methods in Enzymology 82, 3-32.
3. Rosenbloom, J., Endo, R. and Harsch, M. (1976) J. Biol. Chem. 251, 2070-2076.
4. Paglia, L., Wilczek, J., de Leon, L.D., Martin, G.R., Hörlein, D. and Müller, P.K. (1979) Biochemistry 18, 5030-5034.
5. Wiestner, M., Krieg, T., Hörlein, D., Glanville, R.W., Fietzek, P. and Müller, P.K. (1979) J. Biol. Chem. 245, 7016-7023.
6. Yamada, Y., Mudryj, M., Sullivan, M. and de Crombrugghe, B. (1982) J. Biol. Chem. in press.
7. Ohkubo, H., Vogeli, G., Mudryj, M., Avvedimento, V.E., Sullivan, M., Pastan, I. and de Crombrugghe, B. (1980) Proc. Natl. Acad. Sci. USA 78, 7059-7063.
8. Yamada, Y., Avvedimento, V.E., Mudryj, M., Ohkubo, H., Vogeli, G., Irani, M., Pastan, I. and de Crombrugghe, B. (1980) Cell 22, 887-892.
9. Wozney, J., Hanahan, D., Tate, V., Boedtker, H. and Doty, P. (1981) Nature 294, 129-135.
10. Bernard, M.P., Myers, J., Chu, M.-L., Ramirez, F., Eikenberry, E.F. and Prockop, D. (1982) Biochemistry in press.
11. Maxam, A.M. and Gilbert, W. (1977) Proc. Natl. Acad. Sci. USA 72, 3961-3965.
12. Fietzek, P.P., Allman, H., Rauterberg, J., Henkel, W., Wachter, E. and Kuhn, K. (1979) Hoppe-Seyler's Z. Physiol. Chem. 360, 861-868.

13. Sayer, J.M. and Kang, A. (1981) Biochemistry 20, 2621-2627.
14. Fuller, F. and Boedtker, H. (1981) Biochemistry 20, 996-1006.
15. Dickson, L.A., Ninomiya, Y., Bernard, M.P., Pesciotta, D.M., Parsons, J., Green, G., Eikenberry, E.F., de Crombrugghe, B., Vogeli, G., Pastan, I., Fietzek, P.P. and Olsen, B.R. (1981) J. Biol. Chem. 256, 8407-8415.
16. Klein, H.L. and Petes, T.D. (1981) Nature 289, 144-148.
17. Roberts, J.M. and Axel, R. (1982) Cell 29, 109-119.
18. Haber, J.E., Rogers, D.T. and McCusker, J.H. (1980) Cell 22, 277-289.
19. Klar, A.J.S., McIndoa, J., Strathern, J.N. and Hicks, J.B. (1980) Cell 22, 291-298.
20. Baltimore, D. (1981) Cell 24, 592-594.