and Gurd, F. R. N. (1975), *Biochemistry 14,* 5336.
Dwulet, F. E., and Gurd, F. R. N. (1976), *Anal. Biochem. 76,* 530.
Dwulet, F. E., Jones, B. J., Lehman, L. D., and Gurd, F. R. N. (1977), *Biochemistry 16,* 873.
Edmundson, A. B. (1965), *Nature (London) 205,* 389.
Fitch, W. M., and Markowitz, E. (1970), *Biochem. Genet. 4,* 579.
Fontana, A. (1972), *Methods Enzymol. 25,* 419.
Fontana, A., Marchiori, F., Rocchi, R., and Pajetta, P. (1966), *Gazz. Chim. Ital. 96,* 1301.
Hapner, K. D., Bradshaw, R. A., Hartzell, C. R., and Gurd, F. R. N. (1968), *J. Biol. Chem. 243,* 683.
Hunter, M. J., and Ludwig, M. L. (1962), *J. Am. Chem. Soc. 84,* 3491.
Jones, B. N., Vigna, R. A., Dwulet, F. E., Bogardt, R. A., Lehman, L. D., and Gurd, F. R. N. (1976), *Biochemistry 15,* 4418.
Kluh, J., and Bakardjieva, A. (1971), *FEBS Lett. 17,* 31.
Lehman, L. D., Dwulet, F. E., Bogardt, R. A., Jones, B. N., and Gurd, F. R. N. (1977), *Biochemistry 16,* 706.
Omenn, G. S., Fontana, A., and Anfinsen, C. B. (1970), *J. Biol. Chem. 245,* 1895.
Romero Herrera, A. E., and Lehman, H. (1974), *Biochim. Biophys. Acta 336,* 318.
Rothgeb, T. M., and Gurd, F. R. N. (1978), *Methods Enzymol.* (in press).
Stark, G. R., and Smyth, D. G. (1963), *J. Biol. Chem. 238,* 214.
Teale, F. W. J. (1959), *Biochim. Biophys. Acta 35,* 543.
Wang, C. C., Garner, W. H., and Gurd, F. R. N. (1977), *Fed. Proc., Fed. Am. Soc. Exp. Biol. 36,* 890.

Covalent Structure of Cartilage Collagen. Amino Acid Sequence of Residues 363–551 of Bovine α1(II) Chains†

William T. Butler,* John Edward Finch, Jr., and Edward J. Miller

ABSTRACT: The covalent structures of α1(II)-CB11-C6, a chymotryptic peptide from the COOH terminus of α1(II)-CB11, and of α1(II)-CB8 from bovine nasal cartilage collagen are reported. These structures represent residues 363–551 of the bovine α1(II) chain. The sequence displays the repeating Gly-X-Y sequence characteristic of the triple helical portions of all α chains. Another phenomenon observed here, which is also true for other collagen α chains, was the occurrence of phenylalanyl and leucyl residues exclusively in the X positions of the repetitive triplet structure. When the amino acid sequence for this segment of α1(II) was compared with that of α1(I), the level of identity was 73%, a figure slightly lower than that for residues 1–162 at the NH_2-terminal triple-helical region (Butler, W. T., Miller, E. J., and Finch, J. E., Jr. (1976), *Biochemistry 15,* 3000). Three sites of occurrence of glycosylated hydroxylysines in the α1(II) chain were identified by the present studies. Two of these are galactosylhydroxylysines while the other site is a mixture of glucosylgalactosylhydroxylysine and galactosylhydroxylysine. One of the monosaccharides (at residue 408) and the mixture of mono- and disaccharides (at position 531) occur in positions occupied by lysines in the α1(I) chain. The other monosaccharide occurs in a site (residue 420) present as arginine in α1(I). A comparison of the sequences reported for residues 360–660 of α1(I), α2, and α1(III) chains with that of α1(II) reported here along with other unpublished data for α1(II) revealed 65 residues which are identical in these chains and, thus, are possibly "invariant." The frequency of occurrence of lysine, arginine, glutamic acid, and phenylalanine as invariant residues was higher than expected from their overall contents in collagen.

The type II collagen of hyaline cartilages is made of three α1(II) chains, each containing approximately 1050 amino acids. The α1(II) chains are similar in amino acid composition to the α1(I) and α2 chains of type I collagen and to the α1(III) chains of type III collagen (Miller, 1976) but have much higher levels of hydroxylysine-bound glucose and galactose (Miller, 1971; Trelstad et al., 1970; Miller, 1976). Recent studies from our laboratory have also shown that the majority of α1(II) hydroxylysine glycosides occurs in positions which are occupied by lysyl residues in the α1(I) chains (Butler et al., 1974a, 1976). The α1(II) chains were shown to be identical with α1(I) chains in about 80% of the positions (Butler et al., 1974a, 1976). One unusual discovery was that at least two distinct α1(II) chains are present in the bovine nasal septum (Butler et al., 1977). It was observed that three positions of α1(II) display sequence heterogeneity; that is, each is occupied by two amino acids. The chains were tentatively called α1(II)Major and α1(II)Minor to reflect the relative amounts. At present no information is available on the distribution or significance of the two chains.

In this publication we present data showing the primary structure at the COOH-terminal end of α1(II)-CB11, representing residues 363–402[1] of the triple-helical portion of α1(II), and the complete structure of α1(II)-CB8 (residues 403–551) (see the review by Miller, 1976, for clarification of

† From the Institute of Dental Research and the Department of Biochemistry, University of Alabama in Birmingham, University Station, Birmingham, Alabama 35294. *Received June 7, 1977.* This research was sponsored by U.S. Public Health Service Grant DE-02670 from the National Institute of Dental Research.

[1] Numbering begins with the NH_2-terminal triple helical portion of the collagen α chains (Hulmes et al., 1973; Fietzek and Kühn, 1976).

the nomenclature and alignment of the CNBr[2] peptides). A preliminary report on the partial structure of α1(II)-CB8 was published (Butler et al., 1974a). These experiments along with information on residues 1–162 (Butler et al., 1976) and residues 551–660 (G. Francis, W. T. Butler, and J. E. Finch, Jr., unpublished) have revealed about 45% of the bovine α1(II) sequence.

Materials and Methods

Preparation of α1(II)-CB11 and α(II)-CB8. Insoluble bovine nasal cartilage collagen was obtained from young cattle as previously described (Miller and Lunde, 1973). One gram of collagen was suspended in 200 mL of 70% formic acid and treated with 1.5 g of CNBr by vigorously stirring the suspension for 4 h at 24 °C. Most (>90%) of the insoluble material was solubilized by the treatment; the small amount which remained undissolved was removed by centrifugation. The α1(II) CNBr peptides in the supernatant were desalted on 4.0 × 40 cm columns of Sephadex G-25 (coarse) eluted with 0.2 M acetic acid and lyophilized. The procedure for separating the CNBr peptides by CM-cellulose chromatography has been described (Miller and Lunde, 1973). The protein in appropriate zones of several chromatograms was pooled and desalted. Peptide α1(II)-CB8 was further purified by CM-cellulose chromatography at 40 °C in sodium acetate buffer, pH 4.8, and elution with a concave gradient between 0 and 0.14 M NaCl as described (Butler et al., 1967). Peptide α1(II)-CB11 was separated from α1(II)-CB12 and from other impurities by gel chromatography on a 1.5 × 140 cm column of Bio-Gel A-1.5m eluted with 0.05 M Tris-HCl buffer, pH 7.5, containing 1.0 M $CaCl_2$ according to the method of Piez (1968).

Proteolytic Cleavage of CNBr Peptides. Peptides (1–3 μmol) were cleaved with trypsin in 1 mL of 0.05 M Tris-HCl buffer (pH 7.4), 0.001 M $CaCl_2$ by adding 5% (w/w) of trypsin (Worthington Biochemical Corp., thrice crystallized) and incubating at 37 °C for 2–6 h. Peptide α1(II)-CB11 was treated with chymotrypsin by dissolving 2 μmol of the peptide in 1.5 mL of 0.05 M Tris-HCl buffer (pH 7.6), containing 0.005 M $CaCl_2$ and 0.2% sodium azide, adding 0.01 μmol of chymotrypsin (Sigma Chemicals, twice crystallized), and incubating at 37 °C for 30 min. The reaction was stopped by the addition of 1 drop of glacial acetic acid. Peptide α1(II)-CB8-T5 (1 μmol) was treated with 0.02 μmol of chymotrypsin in 1 mL of 0.05 M Tris-HCl buffer (pH 7.4), 0.001 M $CaCl_2$ for 4 h at 37 °C. The method for cleavage of peptides with bacterial collagenase has been described (Butler, 1970).

Purification of Proteolytic Cleavage Products. Separation by gel filtration on Sephadex G-50s was performed with a 1.5 × 140 cm column eluted with 0.2 M acetic acid. Purifications employing phosphocellulose chromatography were performed as described (Butler et al., 1977). Smaller peptides resulting from collagenase or trypsin cleavage were purified on a 0.9 × 150 cm column of Chromobeads A eluted with a gradient formed from pyridine acetate buffers (Schroeder, 1967) as described (Butler et al., 1974b).

Edman Degradation. Automated Edman degradation was performed with a Model 890C Beckman automatic sequencer operated at 56 °C. Larger fragments were degraded by the Slow Protein-Quadrol program (No. 042772, described in the Beckman Sequencer Manual) and smaller ones by the newer Slow Peptide-DMAA program (No. 102974, described in the Beckman publication *In Sequence*, April, 1975). More recently we have used the 0.1 M Quadrol procedure as described by Brauer et al. (1975) and modified by Beckman (Program No. 030176, *In Sequence*, May, 1976).

Pth-amino acids were initially identified by gas–liquid chromatography (Pisano et al., 1972) with 4-ft U-shaped glass columns of 10% DC-560. Selected residues were also analyzed after trimethylsilylation. In addition, the Pth-amino acids from every cycle were analyzed by thin-layer chromatography (Inagami and Murakami, 1972). Pth-arginine was identified by the phenanthrenequinone spot test.

In order to improve the retention of small peptides in the reaction cup of the sequencer, they were reacted with Ans in the presence of EDC essentially as described by Foster et al. (1973). A freshly prepared aqueous solution of EDC and Ans (each 10 mM) was adjusted to pH 4.0 with 0.2 N NaOH. The peptide to be derivatized was dissolved in 0.5 mL of water and placed in the spinning cup of the sequencer; 0.1 mL of the Ans-EDC solution was added and the reaction was allowed to proceed at 56 °C for 30 min under a stream of nitrogen. The mixture was dried and the regular program for automated Edman degradation begun.

Subtractive Edman degradation was performed as described by Balian et al. (1971).

Amino Acid Analysis. Samples were hydrolyzed with constant-boiling HCl at 108 °C for 18–24 h in a nitrogen atmosphere. After drying by rotary evaporation, samples were analyzed either on a Beckman 120C amino acid analyzer modified for single-column analysis (Miller and Piez, 1966) or on a Beckman 121M analyzer by methods already described (Butler et al., 1977).

Analysis of Hydroxylysine Glycosides. The levels of Glc-Gal-Hyl and Gal-Hyl in a peptide were determined on a Beckman 119 automatic amino acid analyzer following hydrolysis of peptides in 2 N NaOH (0.1–1.0 mL) at 108 °C for 24 h in sealed, alkali-resistant tubes. After cooling to 24 °C, the pH of the hydrolysates was adjusted to 5.3 and diluted tenfold by the addition of 0.5 N HCl and distilled water. Aliquots of the diluted hydrolysates (0.25 mL) were then applied to the column of the amino acid analyzer. Elution of amino acids and resolution of the hydroxylysine glycosides was achieved by employing 0.35 M (Na^+) sodium citrate buffer, pH 5.3, throughout the initial 240 min of each run, followed by elution with buffer D (Miller, 1972) for an additional 100 min. Under these conditions Glc-Gal-Hyl is eluted at 54 min and is well-resolved from the mixture of acidic and most of the neutral amino acids which are not retained by the analyzer column. Glc-Gal-Hyl is followed by elution of tyrosine at 60 min, phenylalanine at 68 min, Gal-Hyl at 88 min, and free hydroxylysine at 160 min.

Results

The COOH-Terminal Chymotryptic Peptide from α1(II)-CB11. Treatment of α1(II)-CB11 with chymotrypsin in a substrate:enzyme molar ratio of 200:1 gave rise to a minimum of six peptides in varying sizes and in differing quantities which were initially separated by gel filtration on Sephadex G-50s (not shown). The peptide that eluted last from this column (C6) was readily purified by rechromatography on phosphocellulose; it gave a composition (Table I) indicative of its origin from the COOH terminus of α1(II)-CB11 since it contained homoserine. Peptide C6 contained 40 residues and was always present in substantially higher quantities than the other chymotryptic fragments. We assumed that peptide C6 contained

[2] Abbreviations used are: CNBr, cyanogen bromide; Pth, phenylthiohydantoin; Ans, 2-amino-1,5-naphthalenedisulfonic acid; EDC, *N*-ethyl-*N'*-(dimethylaminopropyl)carbodiimide; Hse, homoserine; Glc-Gal-Hyl, 2-*O*-α-D-glucopyranosyl-*O*-β-D-galactopyranosylhydroxylysine; Gal-Hyl, *O*-β-D-galactopyranosylhydroxylysine.

TABLE I: Composition[a] of α(II)-CB11-C6 and the Products Obtained after Digestion of This Peptide with Trypsin.

Amino acid	Peptide[b]			
	C6	T1	T2	T3
4-Hyp	4.3	1.0	2.4	1.0
Asp	2.1	1.1	1.1	
Thr	0.9	0.8		
Ser	0.9	0.2	1.0	
Hse	1.0			0.8
Glu	4.1	1.1	2.0	1.1
Pro	4.5	1.2	3.4	
Gly	13.1	4.0	6.6	2.4
Ala	3.4	1.0	2.1	
Val	1.4		0.8	0.7
Leu	0.3		0.1	
Lys	0.9	0.9		
Arg	3.2	1.0	1.9	
Total	40	12	22	6

[a] Values are expressed as residues per peptide. A blank indicates that a value was less than 0.1 residue per peptide. [b] See the text and Figure 1 for clarification of the nomenclature.

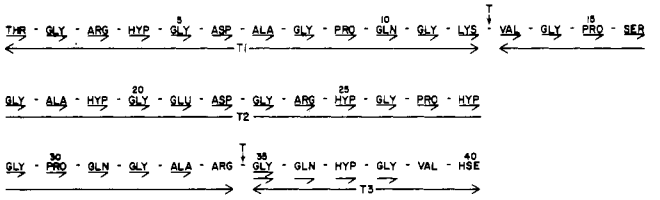

FIGURE 1: The amino acid sequence of α1(II)-CB11-C6. The short half arrows (—) indicate the identification of a residue by Edman degradation. The tryptic (T) peptides are indicated by long arrows (↔).

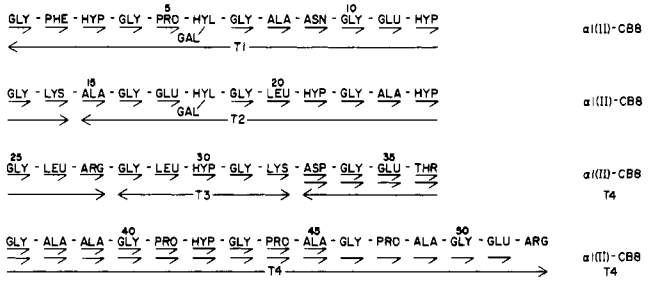

FIGURE 2: The covalent structure of the first 51 residues of α1(II)-CB8. Amino acids identified by Edman degradation are indicated by half arrows (—). Cycles 6 and 18 yielded no identifiable Pth-amino acid in any significant quantity and are thus "blanks." The positions of the first four tryptic (T) peptides of α1(II)-CB18 are indicated by the long arrows (↔).

2 mol of valine which was incompletely released by acid hydrolysis or was present in a site partially occupied by leucine (see later).

About 1.1 μmol of C6 was reacted with Ans in the presence of EDC and subjected to automated Edman degradation utilizing the Slow Peptide-DMAA program. In this way the sequence of 34 of the first 35 residues of C6 was established. The results are detailed in Table II and summarized in Figure 1. To further substantiate this sequence and to obtain the structure at the COOH terminus, 2 μmol of C6 was treated with trypsin and the products were separated by gel filtration on Sephadex G-50s. Three tryptic peptides were obtained with the amino acid compositions (Table I) predicted from the sequence data and from the original composition of C6. The composition of tryptic peptide T2 (Table I) indicated that residue 34 was arginine. Next, about 0.5 μmol of peptide C6-T3 was reacted with Ans and degraded four cycles in the automated sequencer, giving the sequence Gly-Gln-Hyp-Gly (Figure 1).

Peptides C6 and C6-T2 consistently contained subintegral amounts of valine and leucine (Table I) with the sum of the two closer to integral values. The data suggest that sequence heterogeneity exists at residue 13 of C6 and involves valine in α1(II)Major and leucine in α1(II)Minor. However, no Pth-leucine was observed at cycle 13 of the Edman degradation (Table II); therefore, it is unclear at this time whether heterogeneity occurs in this site of the α1(II) chain. The low value for valine may be due to incomplete release on acid hydrolysis. Although peptide C6-T3 also gave subintegral values for valine, it contained no leucine; therefore, heterogeneity at position 39 (Figure 1) can be ruled out. Apparently the Val-Hse bond in this sequence is only partially cleaved by acid hydrolysis.

The residue at position 19 of α1(II)-CB11-C6 is a partially hydroxylated proline[3] since both proline and hydroxyproline were found in this site. This conclusion is consistent with the composition of peptides C6 and of C6-T2 (Table I); nonintegral values were found for these two amino acids while the sum of the values was closer to an integral.

Edman Degradation of α1(II)-CB8. The peptide was subjected to automated Edman degradation using the Slow Protein-Quadrol program with 0.5 M Quadrol. In the first attempt, 0.6 μmol of peptide was utilized and the Pth-amino acids from 43 of the first 45 cycles were identified (Figure 2). Utilizing the yields of a number of the Pth-glycines, the repetitive yield was determined to be 95%. In a second experiment 0.45 μmol of α1(II)-CB8 was degraded and the Pth-amino acids for 29 cycles identified with results identical with those of the first experiment. The repetitive yield for this second run was 94% when calculated from the levels of Pth-glycine at cycles 15 and 23. As indicated in Figure 2, cycles 6 and 18 gave "blanks"; that is, no Pth-amino acids were observed either on thin-layer or gas–liquid chromatograms.

In addition to Pth-hydroxyproline, Pth-proline was observed at cycles 12, 21, and 24 of the Edman degradation of α1(II)-CB8, indicating partial hydroxylation of prolines in these positions. This phenomenon is also reflected in the proline and hydroxyproline contents of tryptic peptides, T1 and T2 (see later).

The Tryptic Peptides of α1(II)-CB8. Nine major peptide fragments, liberated by digestion of α1(II)-CB8 with trypsin, were purified as follows. Two large peptides, T5 and T9 with 45 and 32 amino acids, respectively, were separated from the smaller ones (containing 3 to 19 residues) by gel chromatography on Sephadex G-50s (chromatogram not shown). Peptides T5 and T9 were separated from each other by phosphocellulose chromatography (Figure 3). The seven smaller tryptic peptides from the included volume of Sephadex G-50s were

[3] Hydroxyproline residues are formed by the enzymatic hydroxylation of proline residues after formation of procollagen chains in the endoplasmic reticulum (see Prockop et al., 1976). This posttranslational event has been shown to be incomplete by sequence analysis of collagen chains (Bornstein, 1967); thus, a given site in the chains may contain only hydroxyproline, or may contain hydroxyproline and proline. In the presentation of the sequences, these partially hydroxylated prolines are given as hydroxyproline. They are detected by identification of both Pth-proline and Pth-hydroxyproline at the same cycle in Edman degradation or by amino acid analysis (i.e., by noting subintegral values for hydroxyproline while the proline values are elevated). In determining the number of residues in a peptide from amino acid analysis, the nearest integral of the total of the hydroxyproline and proline values is used, instead of considering them separately.

TABLE II: Summary of Identification[a] and Recoveries[b] of Pth-Amino Acids in the Various Cycles of Automated Edman Degradation of α1(II)-CB11-C6.

Cycle	Conclusion	Identification method: Gas–liquid chromatography, Non-Silylated	Identification method: Gas–liquid chromatography, Silylated	Identification method: Thin-layer chromatography, solvent XM[c]	Recovery[b] (%)
1	Thr	–	–	Thr	
2	Gly	Gly		Gly	51
3	Arg[d]	–	–	–	
4	Hyp	Hyp		Hyp	
5	Gly	Gly		Gly	
6	Asp	–	Asp	Asp	
7	Ala	Ala		Ala	37
8	Gly	Gly		Gly	55
9	Pro	Pro		Pro	24
10	Gln	–	Glu/Gln	Glu/Gln	
11	Gly	Gly		Gly	45
12	Lys	–	Lys	Lys	
13	Val	Val		Val	41
14	Gly	Gly		Gly	27
15	Pro	Pro		Pro	21
16	Ser	–	–	Ser	
17	Gly	Gly		Gly	30
18	Ala	Ala		Ala	29
19	Hyp	Pro/Hyp		Pro/Hyp	
20	Gly	Gly		Gly	21
21	Glu	–	Glu	Glu	
22	Asp	–	Asp	Asp	
23	Gly	Gly		Gly	22
24	Arg[d]	–	–	–	
25	Hyp	Hyp		Hyp	
26	Gly	Gly		Gly	15
27	Pro	Pro		Pro	8
28	Hyp	Hyp		Hyp	
29	Gly	Gly		Gly	15
30	Pro	Pro		Pro	
31	Gln	–	Glu/Gln	Glu/Gln	–
32	Gly	Gly		Gly	
33	Ala	Ala		Ala	
34	Arg	–	–	–	
35	Gly	Gly		Gly	

[a] The positive identification of a Pth-amino acid is denoted by the name of the parent amino acid under the identification method. A dash indicates that, though the method was utilized, identification of the Pth-amino acid was not achieved. [b] Recoveries are given as percent of the starting peptide material. The amounts of Pth-amino acids were calculated from peak heights on gas–liquid chromatograms compared with those of standard Pth-amino acids. [c] Solvent XM is xylene and methanol in an 8:1 proportion. Identities of the Pth-amino acids were aided by spraying dried thin-layer chromatograms with ninhydrin (Inagami and Murakami, 1972). [d] Identified by phenanthrenequinone spot test.

TABLE III: Amino Acid Composition[a] of the Tryptic Peptides of α1(II)-CB8.

Amino acid	T1	T2	T3	T4	T5	T6	T7	T8	T9	Summation	α1(II)-CB8[b]
4Hyp	1.5	0.8	1.0	1.1	5.5		1.1	1.0	3.8	16	14
Asp	0.9			1.0	1.1				1.1	4	5
Thr				0.9					1.9	3	3
Ser				0.2	1.0			1.9		3	3
Hse									1.0	1	1
Glu	1.0	1.0		2.3	6.8	1.1	1.1	2.1	2.3	17	16
Pro	1.6	1.5		3.1	5.4			0.2	4.1	15	18
Gly	4.8	3.9	2.0	6.2	15.3	1.0	2.3	4.3	10.8	50	50
Ala	1.0	2.0		3.9	3.3			1.1	4.1	15	14
Val					1.9					2	2
Leu		1.8	1.0		2.0			0.9	1.9	8	8
Phe	1.0			0.1	0.9		0.8			3	3
Hyl	1.0	1.1							1.0	3	3
Lys	1.1		0.9		0.9					3	3
Arg		1.1		1.0	0.9	0.9	0.8	0.9		6	6
Total	14	13	5	19	45	3	6	12	32	149	149

[a] Residues per peptide. A blank indicates that the level was less than 0.1 residue per peptide. [b] From Miller and Lunde (1973).

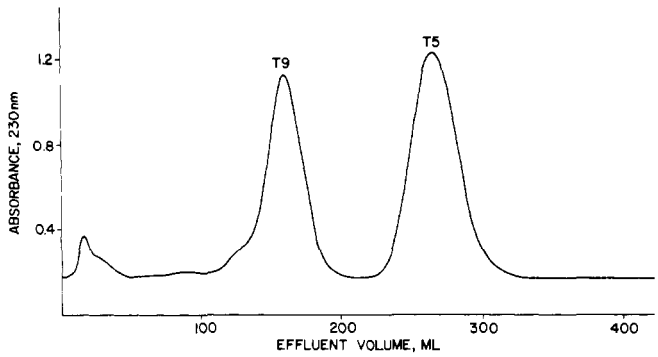

FIGURE 3: Phosphocellulose chromatography of two large tryptic peptides (T5 and T9) derived from α1(II)-CB8. The 1.5 × 15 cm column of 140-200 mesh Whatman p1 resin was eluted at 24 °C with a linear gradient of 0 to 0.15 M NaCl in 0.001 M sodium formate buffer, pH 3.6, over a total volume of 1000 mL.

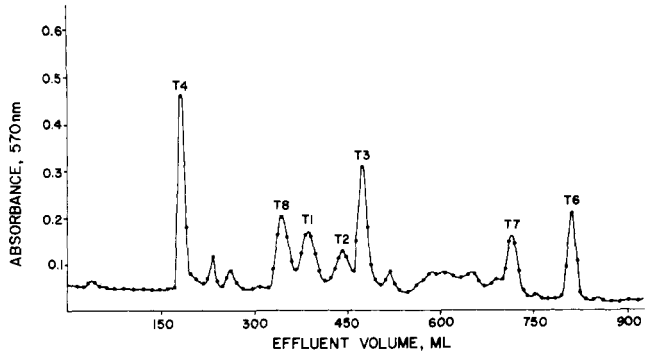

FIGURE 4: Ion-exchange chromatography of the smaller tryptic peptides of α1(II)-CB8, separated from peptides T5 and T9 by gel filtration on Sephadex G-50s. The Chromobeads A column (0.9 × 150 cm) was equilibrated with 0.2 M pyridine acetate buffer (pH 3.1) at 40 °C (Schroeder, 1967) by pumping it at 50 mL per h. After application of the sample, the column was eluted with a linear gradient formed from 750 mL each of 0.2 M pyridine acetate, pH 3.1 (starting buffer), and 2.0 M pyridine acetate, pH 5.0, as the limiting buffer. Fractions of 5 mL were collected and analyzed for ninhydrin-positive material with a Technicon autoanalyzer.

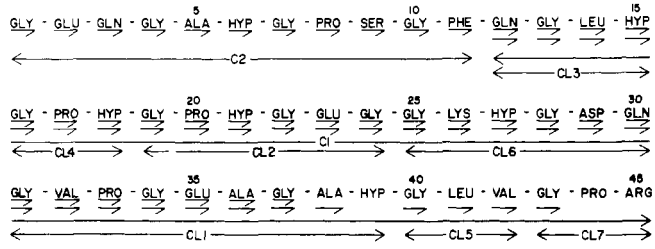

FIGURE 5: The amino acid sequence of α1(II)-CB5-T5. Edman degradations are marked with half arrows (→). The chymotryptic (C) and collagenase (CL) peptides are indicated by long arrows (↔).

purified by ion-exchange chromatography on a Chromobeads A column as illustrated in Figure 4. The ninhydrin-positive material eluting between 550 and 700 ml on this chromatogram was not consistently observed; amino acid analysis indicated that it contained only small amounts of peptide material relative to the seven designated peptide peaks. The peaks marked T1, T2, T8, and T4 contained small amounts of contaminating material and were each further purified by passage over a column of CM-cellulose in 0.02 M sodium acetate buffer, pH 4.8, as described by Butler et al. (1974b). Peptides T3, T7, and T6 were not further purified. The compositions of the nine tryptic peptides, given in Table III, indicated that they were highly pure, since the values for the amino acids were near integral values. The summation of the amino acids found in these peptides agreed well with the composition reported for α1(II)-CB8 by Miller and Lunde (1973) and thus represents the complete sequence.

The alignment of tryptic peptides T1, T2, T3, and T4 was deduced from the sequence analysis of α1(II)-CB8 (see Figure 2). Peptides T5, T6, and T7 were aligned by comparison of their compositions and sequences (see later) to those of three similar tryptic peptides from bovine, chick and rat α1(I)-CB3 (Fietzek et al., 1972; Butler et al., 1974b; Dixit et al., 1975). The homoserine of peptide T9 shows it to be from the COOH terminus of α1(II)-CB8. And finally, peptide T8 was aligned by deduction. Though identical in size to a tryptic peptide from α1(I)-CB3, peptide T8 is dissimilar in composition and sequence and could not be placed by homology.

The hydroxylysine in peptides T1 and T2 (Table III) suggested that the "blanks" encountered on automated Edman degradation (see above) at positions 6 and 18 were due to the presence of glycosylated hydroxylysines. The insensitivity of the hydroxylysine peptide bonds in these two peptides to proteolysis by trypsin also indicated that these amino acids bore carbohydrates. To test this possibility, samples of T1 and T2 were subjected to alkaline hydrolysis and the hydrolysates were chromatographed on the amino acid analyzer. Peptide T1 yielded Gal-Hyl but no Glc-Gal-Hyl. A small amount of hydroxylysine (9% of the level of Gal-Hyl) was also detected on the chromatogram. The hydrolysate of peptide T2 also yielded Gal-Hyl, a small amount of hydroxylysine (6% of the Gal-Hyl level) and no Glc-Gal-Hyl. These data show that positions 6 and 18 of α1(II)-CB8 are Gal-Hyl moieties (Figure 2).

Sequence Analysis of Peptide T4 (Residues 33–51). Approximately 0.7 μmol of T4 was derivatized with Ans in the presence of EDC and subjected to 18 cycles of Edman degradation with the Slow Peptide-DMAA sequence program. From this determination and from the knowledge that arginine was the COOH-terminal amino acid, the amino acid sequence of peptide T4 was shown to be: Asp-Gly-Glu-Thr-Gly-Ala-Ala-Gly-Pro-Hyp-Gly-Pro-Ala-Gly-Pro-Ala-Gly-Glu-Arg. Utilizing the yields of Pth-glycine at cycles 5, 8, 11, and 14, the repetitive yield during the Edman degradation of T4 was determined to be 94%.

Amino Acid Sequence of Peptide T5 (Residues 52–96). The complete sequence of this tryptic peptide is indicated in Figure 5 along with diagrammatic representations of the methodology employed. Peptide T5 (1.18 μmol) was coupled to Ans in the presence of EDC and subjected to automated Edman degradation with the 0.1 M Quadrol procedure of Brauer et al. (1975). By this procedure the first 37 residues of T5 were identified (Figure 5). Using several criteria, the repetitive yield for the first 25 steps was determined to be 94–95%.

Next the peptide was cleaved with chymotrypsin and the products were separated by gel filtration on Sephadex G-50s; two major peptide fragments (C1 and C2) were obtained containing 34 and 11 residues, respectively (Table IV). The composition of C2, when compared with the sequence of T5 (Figure 5), indicated that it originated from the NH_2 terminus and that chymotrypsin had cleaved the peptide at the Phe-Gln bond (residues 11–12).

In one experiment chymotrypsin also partially cleaved the Leu-Val bond (residues 41–42) of T5. The evidence for this conclusion was lowered levels of arginine and valine in peptide C1 and the presence of the valine and arginine-containing peptide as a contaminant of C2.

The chymotryptic fragment, C1 (1.2 μmol), was subjected to 32 cycles of automated Edman degradation after coupling to Ans. The Slow Peptide-DMAA program of the automated

TABLE IV: Composition[a] of Peptides Isolated after Cleavage of Peptide T5 with Chymotrypsin[b] and of C1 with Collagenase.[c]

Amino acid	C2	C1	CL3	CL4	CL2	CL6	CL1	CL5	CL7
Hyp	1.0	4.2	0.9	1.0	0.9	0.7	1.1		
Asp		1.1				1.0			
Ser	1.0								
Glu	2.0	4.2	0.8		1.1	1.1	1.0		
Pro	1.0	4.7		1.0	1.0	0.3	0.8		1.0
Gly	4.1	11.8	1.2	1.0	2.8	1.8	2.9	1.1	1.0
Ala	1.0	2.3			0.2		1.9		
Val		1.9					1.0	0.9	
Leu		2.0	1.0					1.0	
Phe	1.0								
Lys		0.9				0.9			
Arg		0.9							1.0
Total	11	34	4	3	6	6	9	3	3
Recovery (%)	*d*	*d*	53	80	90	51	51	41	58

[a] Residues per peptide. A blank indicates that the level was below 0.1 residue. [b] See the text for nomenclature. [c] See the text and Figures 5 and 6 for nomenclature. [d] Not determined.

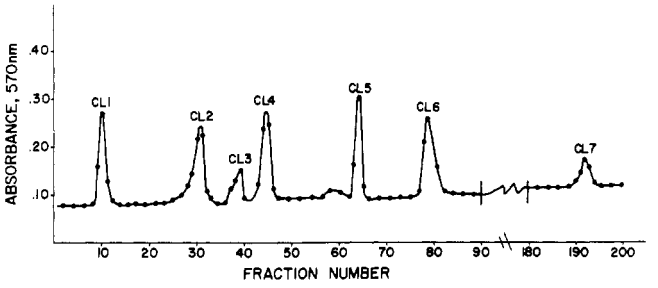

FIGURE 6: Separation of the peptides derived from T5-C1 after cleavage with collagenase by chromatography on the Chromobeads A column. The details of the method are the same as in Figure 4.

sequencer was employed. By this procedure the Pth-amino acids at 31 of the first 34 residues of C1 were identified (Figure 5). These results verified those obtained for cycles 12 to 37 in the Edman degradation of peptide T5. The repetitive yield for the Edman degradation of C1, determined from a log-regression curve constructed from the levels of Pth derivatives of glycine, leucine, alanine, and valine at the various cycles, was 91.4%. Both Pth-proline and Pth-hydroxyproline were detected at cycle 16 of C1, indicating partial hydroxylation of the proline residue. This observation is consistent with the compositions of peptide C1 and of C1-CL6 (Table IV).

In order to verify these results and to identify the remaining amino acids in the sequence of T5, 0.9 μmol of peptide C1 was cleaved with collagenase and the products were separated on a Chromobeads A column (Figure 6). Seven peptides ranging in size from 4 to 9 residues were obtained in excellent yield. The compositions (Table IV) indicated that these collagenase peptides were highly pure and were consistent with the sequence data obtained from the automated Edman degradations of both T5 and C1. The composition of CL1 showed that residue 39 of T5 was hydroxyproline while that of CL7 indicated residue 44 to be proline. Arginine was known to be the COOH-terminal amino acid from trypsin specificity.

Peptide T5 contained a residue each of lysine and arginine (see Table III) though it was derived from α1(II)-CB8 after trypsin cleavage. This observation is explained by the occurrence of Lys-Hyp and Lys-Pro sequences (residues 26–27) known to be resistant to cleavage by trypsin (Butler and Ponds, 1971; Grimm and Grassmann, 1964; Hannig and Nordwig, 1967).

TABLE V: Subtractive Edman Degradation of the Tryptic Peptides α1(II)-CB8-T6 and α1(II)-CB8-T7.[a]

	α1(II)-CB8-T6		
Cycle	Gly	Glu	Arg
0	1.00	1.12	0.94
1	*0.15*	1.05	0.95

	α1(II)-CB8-T7				
Cycle	Gly	Phe	Hyp	Glu	Arg
0	2.09	0.91	0.95	1.11	0.98
1	*1.2*	0.92	0.82	1.08	0.98
2	1.15	*0.08*	0.85	1.06	0.93
3	1.08	0.00	*0.00*	1.00	0.92
4	*0.68*	0.00	0.00	1.07	0.93

[a] Compositions for the Edman degradation cycles are given as residues per peptide. The value for the residue removed by Edman degradation at each cycle is italicized.

Sequence of Peptides T6 and T7 (Residues 97–105). These small tryptic peptides have compositions (Table III) identical with two obtained from rat α1(I)-CB3 (Butler et al., 1974b). The sequence of T6 was determined to be Gly-Glx-Arg and that of T7, Gly-Phe-Hyp-Gly-Glx-Arg by subtractive Edman degradation (Table V).

To identify the Glx residues in positions 2 and 5 of T6 and T7, respectively, each was subjected to automated Edman degradation with the Slow Peptide-DMAA program after coupling to Ans. In addition to verifying the sequence data given above, the experiments showed that residues 2 and 5 of T6 and T7 were each glutamic acid and not glutamine.

Structure of Peptide T8 (Residues 106–117). About 0.45 μmol of T8 was coupled to Ans in the presence of EDC and subjected to automated sequence analysis with the Slow Peptide-DMAA program. The following sequence was obtained: Gly-Ser-Hyp-Gly-X-Gln-Gly-Leu-Gln-Gly-Ala. Utilizing the yields of Pth-glycine the repetitive yield for this run was calculated to be 86%. Arginine was known to be the COOH-terminal amino acid from trypsin specificity. Since the composition of T8 (Table III) showed it to contain 12 amino acids, including two serines, the unidentified residue at position 5 must be serine. The data show that the complete sequence of T8 is: Gly-Ser-Hyp-Gly-Ser-Gln-Gly-Leu-Gln-Gly-Ala-Arg.

TABLE VI: Summary of Identification and Recovery of Pth-Amino Acids from the Automated Edman Degradation of α1(II)-CB8-T9.[a]

Cycle	Conclusion	Identification method: Gas-liquid chromatography, Non-Silylated	Identification method: Gas-liquid chromatography, Silylated	Identification method: Thin-layer chromatography, solvent XM	Recovery (%)
1	Gly	Gly		Gly	52
2	Leu	Leu		Leu	52
3	Hyp	Hyp		Hyp	
4	Gly	Gly		Gly	43
5	Thr	Thr		Thr	
6	Hyp	Hyp		Hyp	
7	Gly	Gly		Gly	38
8	Thr	Thr		Thr	
9	Asp	-	Asp	Asp	
10	Gly	Gly		Gly	25
11	Pro	Pro		Pro	18
12	Gly-Gal-Hyl[b]	-	-	-	
13	Gly	Gly		Gly	14
14	Ala	Ala		Ala	13
15	Ala	Ala		Ala	20
16	Gly	Gly		Gly	12
17	Pro	Pro		Pro	10
18	Ala	Ala		Ala	9
19	Gly	Gly		Gly	11
20	Pro	Pro		Pro	9
21	Hyp	Hyp		Hyp	
22	Gly	Gly		Gly	7
23	Ala	Ala		Ala	4
24	Gln	-	Glu/Gln	Glu/Gln	
25	Gly	Gly		Gly	5
26	Pro	Pro		Pro	
27	Hyp	Hyp		Hyp	
28	Gly	Gly		Gly	3
29	Leu	Leu		Leu	3
30	Gln	-		Gln	

[a] See Table II for details of this reporting procedure. [b] See the text for the assignment of Glc-Gal-Hyl to this position.

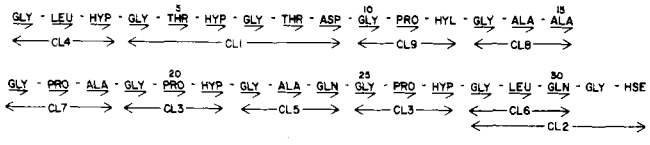

FIGURE 7: The covalent structure of α1(II)-CB8-T9. Collagenase (CL) peptides are indicated by long arrows (↔) and Edman degradations by half arrows (—). Residue 12, not identified by Edman degradation, was shown to yield hydroxylysine after acid hydrolysis of CL9 (Table VII). It was identified as a mixture of Glc-Gal-Hyl and Gal-Hyl by analysis of an alkaline hydrolysate of α1(II)-CB8-T9 (see the text).

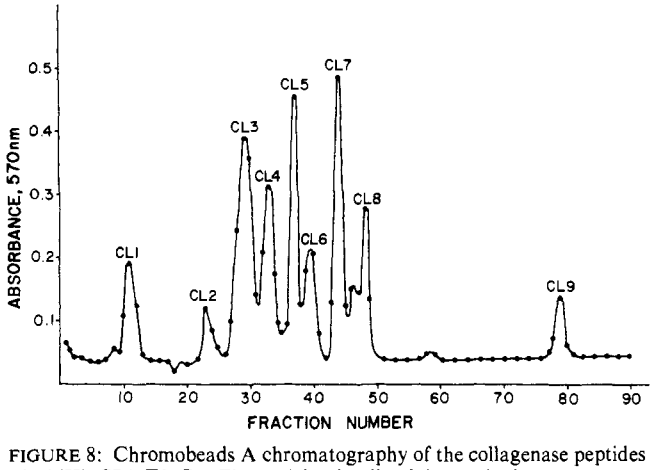

FIGURE 8: Chromobeads A chromatography of the collagenase peptides of α1(II)-CB8-T9. See Figure 4 for details of the method.

The Covalent Structure of Peptide T9. Approximately 1.4 μmol of T9 was coupled to Ans and degraded with the Slow Peptide-DMAA program of the automatic sequencer. Twenty-nine residues of the first 30 of T9 were identified as detailed in Table VI and summarized in Figure 7. To strengthen these conclusions and to identify the unknown amino acid at residue 12, peptide T9 was cleaved with collagenase and the resultant small peptides were separated by ion-exchange chromatography on Chromobeads A (Figure 8). The analyses of these nine peptides (Table VII) were compatible with the data from automated sequence analysis (see Figure 7). The compositions indicated that all but one (CL7) of the nine peptides were of high purity. Even though peptide CL7 is slightly contaminated by CL6, the major component in this fraction was obviously derived from residues 16–18 of T9. The composition of CL9 indicated that hydroxylysine was present in position 12 of peptide T9; the resistance to proteolysis by trypsin suggested that it was a glycosylated hydroxylysine. In order to test this possibility, an alkaline hydrolysate of peptide T9 was analyzed on the amino acid analyzer. Glc-Gal-Hyl and Gal-Hyl were found in a ratio of 1.54:1; free hydroxylysine was present in small quantities (<1% of the total of Glc-Gal-Hyl and Gal-Hyl). Thus, position 12 of T9 is occupied by both Glc-Gal-Hyl and Gal-Hyl; it is shown in Figure 9 as the disaccharide since that is the predominant form.

Discussion

The covalent structure of residues 363–551 of bovine α1(II) chains, as determined by the studies presented here, is depicted

TABLE VII: Composition[a] of Peptides Isolated after Cleavage of T9 with Collagenase.[b]

Amino acid	CL4	CL1	CL9	CL8	CL7	CL3	CL5	CL6	CL2
Hyp	0.8	0.8				1.0			
Asp		1.1							
Thr		1.9							
Hse									1.0
Glu					0.3		0.9	1.0	1.0
Pro	0.2	0.3	1.0		0.9	1.0			
Gly	1.1	1.9	1.0	1.1	1.3	1.0	1.1	1.0	2.1
Ala				1.9	1.1		1.0		
Leu	0.9				0.3			0.9	0.9
Hyl			1.0						
Total	3	6	3	3	3	3	3	3	5
Recovery (%)	39	32	24	14	30	57[c]	40	22	21

[a] Residues per peptide. A blank indicates levels below 0.1 residue. [b] See Figure 7 and the text for clarifications of the nomenclature. [c] The actual yield of this peptide was 114%, but is given as half this value because a Gly-Pro-Hyp sequence occurs twice (residues 19–21 and 25–27).

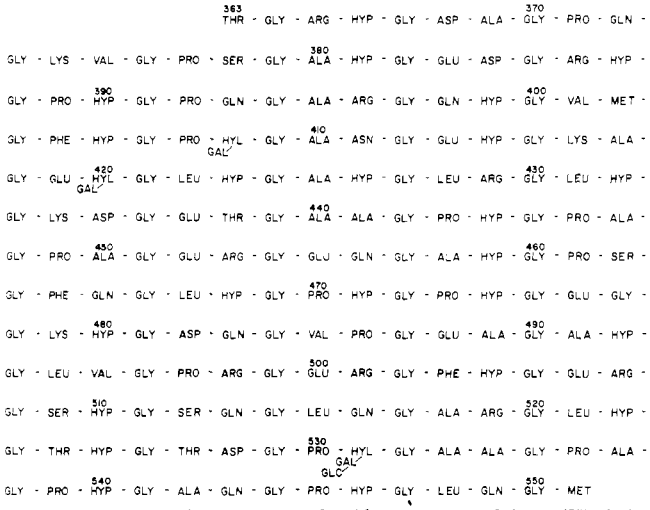

FIGURE 9: The covalent structure of residues 363–551 of the α1(II) chain from bovine type II collagen.

in Figure 9. The data show that this area of α1(II), like that from residues 1–162 (Butler et al., 1976), displays a repeating Gly-X-Y amino acid sequence. Except for short segments near both ends, this type of repeating structure is found in all collagen α chains studied to date (for reviews see Fietzek and Kühn, 1976; Miller, 1976). The occurrence of glycine as every third residue is necessary for the formation of the collagen triple helix.

Another general feature of the sequence of collagen chains is the unequal distribution of certain amino acids in the X and Y positions of the recurring collagen triplet (Fietzek and Kühn, 1976). In the α1(II) sequence reported here, all eight of the leucines and the three phenylalanines occur in the X position. A similar situation exists for these residues in α1(I) and α2 (Fietzek and Kühn, 1976; Traub and Fietzek, 1976; Dixit et al., 1977a,b). This phenomenon has been explained in terms of steric hindrance for certain bulky residues when in the Y positions (Traub, 1974).

In the sequence of 189 amino acids presented in this paper, 51 are different from those of α1(I).[4] Thus, about 73% of the residues of this portion of α1(II) are identical with those of α1(I) chains. If the invariant glycines are not considered in these calculations, the level of homology of the X and Y positions of α1(I) and α1(II) chains in this region is about 60%. These numbers are lower than similar calculations for residues 1–162 from the NH_2-terminal helical portions of the chains, where the overall level of sequence homology was 81% and that of the X and Y positions was 73%, for the α1(I) and α1(II) chains (Butler et al., 1976).

The amino acid sequences of α1(I) and α2 chains for residues 361 to 660 are available (Fietzek and Kühn, 1976; Dixit et al., 1977a,b); in fact, for much of this span, information is available for three species (bovine, rat, and chick). The experiments on α1(II) presented here, along with those on the NH_3-terminal segment of α1(II)-CB10 (G. Francis, W. T. Butler, and J. E. Finch, Jr., unpublished data) disclose the sequence of this same segment of α1(II). In addition, some data are available on the sequence of α1(III) in this region (Fietzek and Kühn, 1976). By comparing this sequence data for α1(I), α2, and α1(II) chains, and a portion of α1(III), one can find certain amino acids that are the same for these chains and could be considered to be "invariant." Table VIII lists the results of such a comparison giving the amino acid and position of these invariant residues in the X and Y positions of the repeating collagen triplet. It should be emphasized that, of the amino acids listed in Table VIII, some will prove *not* to be invariant, since the conclusions were arrived at from relatively few (4–7) comparisons. Nevertheless, the fact that three, and sometimes four, divergent collagen chains are involved strengthens the analysis.

In an earlier publication (Butler et al., 1974a), we speculated that invariant amino acids of collagen α chains may occur in clusters which would reflect the evolutionary preservation of common structural features of the collagen molecules in *regions,* rather than in scattered *individual* amino acids. An examination of the comparative sequences of residues 361–660 revealed several regions where the invariant residues were apparently clustered. For example, in the span from residues 401–420, half the amino acids (7/14) in the X and Y positions are invariant, while of the next 20 positions only one invariant amino acid occurs. For the area from residues 461–480, 64% (9/14) of the amino acids in X and Y are the same in all chains while in similar spans before (residues 441–460) and after (residues 481–500) this segment only 13% and 23% of the X and Y amino acids are invariant. These data add credibility

[4] This comparison involves rat α1(I) chain for residues 363–402 and bovine α1(I) for 403–551. The former sequence for the bovine species has not been published.

TABLE VIII: Amino Acid Residues in the X and Y Positions Which Are Identical in α1(I), α2 and α1(II) Chains and in Portions of the α1(III) Chain.[a]

Residue No.[b]	Amino acid	No. of comparisons[c]	Residue No.[b]	Amino acid	No. of comparisons[c]	Residue No.[b]	Amino acid	No. of comparisons[c]
366	Hyp	5	476	Glu	6	594	Hyp	4
371	Pro	5	479	Lys	5	596	Pro	4
374	Lys	5	480	Hyp	5	597	Ala	4
377	Pro	5	492	Hyp	6	605	Glu	4
384	Asp	5	498	Arg	6	608	Pro	4
392	Pro	5	500	Glu	6	611	Pro	4
396	Arg	4	501	Arg	6	612	Ala	4
403	Phe	7	504	Hyp	6	618	Arg	4
404	Hyp	7	506	Glu	6	621	Hyp	4
406	Pro	7	519	Arg	6	624	Arg	4
408	Lys	7	525	Hyp	6	626	Glu	4
414	Hyp	6	528	Asp	5	629	Pro	4
419	Glu	6	531	Lys	5	632	Pro	4
440	Ala	6	552	Hyp	5	635	Phe	4
444	Hyp	6	554	Glu	5	636	Ala	5
455	Glu	6	555	Arg	5	638	Pro	5
464	Phe	6	564	Lys	5	641	Ala	5
465	Gln	6	578	Ala	5	645	Hyp	5
467	Leu	6	582	Asp	5	647	Ala	5
469	Hyp	6	585	Arg	4	648	Lys	5
470	Pro	6	587	Leu	4	650	Glu	5
473	Pro	6	591	Ile	4			

[a] Determined by comparing published sequence data for α1(I), α1(III), and α2 chains with that of the α1(II) chain contained in the present paper and that of Francis, Butler, and Finch (unpublished). See the text for more details. [b] Numbering begins with the first glycine of the repeating Gly-X-Y sequence (Hulmes et al., 1973; Fietzek and Kühn, 1976). A disparity of numbers occurs after residue 614 because of the finding of an extra triplet of amino acids in α2 (Dixit et al., 1977b) and in α1(II) (G. Francis, W. T. Butler, and J. E. Finch, Jr., unpublished). [c] Comparisons include data from α1(I) and α2 chains of rat, chick, and bovine species, from the bovine α1(III) chain and from residues 403–438 and 551–582 of the bovine α1(II) chain. The numbers indicate how many sequences were available for comparison in a given area. In all cases data for at least one α1(I), α2, or α1(II) chain were utilized.

to the hypothesis of clustering of invariant residues because of the added number of sequences available for comparison.

Some amino acids appear to be present as invariant residues in a higher frequency[5] than expected from their quantity in collagen. Phenylalanine is invariant more than two times as often as expected; of four phenylalanines in residues 363–661 of α1(II), three are also present in all α1(I) and α2 chains examined. The frequency of invariance of the charged amino acids, arginine, lysine (or hydroxylysine) and glutamic acid, is almost twice as high as expected in this span. Thus, the importance of these charged residues in collagen structure, pointed out by several authors (see Fietzek and Kühn, 1976), seems to be verified by the present comparative results.

Acknowledgments

We gratefully acknowledge the technical assistance of Mr. Glen Bridges and Ms. Virginia Wright. We thank Dr. Tadashi Inagami, Vanderbilt University, for helpful discussions and aid in the early phases of this research, and Dr. Sayru Dixit, the University of Tennessee Medical School, for the copy of a manuscript on the sequence of α2-CB3 prior to its publication.

[5] The expected frequency of occurrence of an amino acid as an invariant residue, if this were on a random basis, was calculated from the overall content of an amino acid in mammalian collagens. For example, both type I and type II collagen chains contain approximately 14 residues of phenylalanine per 1000 amino acids. The 300 residue span considered here would thus contain, on the average, $14 \times 0.3 = 4.2$ phenylalanine residues. Indeed, 4 residues of phenylalanine do occur in α1(I), α2, and α1(II) chains in this region. Since only 65 of the 200 residues in the X and Y positions were judged to be invariant (Table VII), the number of phenylalanine residues which would occur *by chance* as invariant in residues 361–660 would be $4.2 \times 65/200 = 1.36$.

References

Balian, G., Click, E. M., and Bornstein, P. (1971), *Biochemistry 10,* 4470.

Bornstein, P. (1967), *Biochemistry 6,* 3032.

Brauer, A. W., Margolies, M. N., and Haber, E. (1975), *Biochemistry 14,* 3029.

Butler, W. T. (1970), *Biochemistry 9,* 44.

Butler, W. T., Finch, J. E., Jr., and Miller, E. J. (1977), *J. Biol. Chem. 252,* 630.

Butler, W. T., Miller, E. J., and Finch, J. E., Jr. (1976), *Biochemistry 15,* 3000.

Butler, W. T., Miller, E. J., Finch, J. E., Jr., and Inagami, T. (1974a), *Biochem. Biophys. Res. Commun. 57,* 190.

Butler, W. T., Piez, K. A., and Bornstein, P. (1967), *Biochemistry 6,* 3771.

Butler, W. T., and Ponds, S. L. (1971), *Biochemistry 10,* 2076.

Butler, W. T., Underwood, S. P., and Finch, J. E., Jr. (1974b), *Biochemistry 13,* 2946.

Dixit, S. N., Kang, A. H., and Gross, J. (1975), *Biochemistry 14,* 1929.

Dixit, S. N., Seyer, J. M., and Kang, A. H. (1977a), *Eur. J. Biochem. 73,* 213.

Dixit, S. N., Seyer, J. M., and Kang, A. H. (1977b), *Eur. J. Biochem.* (in press).

Fietzek, P. P., and Kühn, K. (1976), *Int. Rev. Connect. Tissue Res. 7,* 1.

Fietzek, P. P., Wendt, P., Kell, I., and Kühn, K. (1972), *FEBS Lett. 26,* 74.

Foster, J. A., Bruenger, C. L., Hu, C. L., Albertson, K., and Franzblau, C. (1973), *Biochem. Biophys. Res. Commun. 53,* 70.

Grimm, L., and Grassmann, W. (1964), *Hoppe-Seyler's Z. Physiol. Chem. 337,* 161.
Hannig, K., and Nordwig, A. (1967), in Treatise on Collagen, Vol. 1, Ramachandran, G. N., Ed., London, Academic Press, p 73.
Hulmes, D. J. S., Miller, A., Parry, D. A. D., Piez, K. A., and Woodhead-Galloway, J. (1973), *J. Mol. Biol. 79,* 137.
Inagami, T., and Murakami, K. (1972), *Anal. Biochem. 47,* 501.
Miller, E. J. (1971), *Biochemistry 10,* 1652.
Miller, E. J. (1972), *Biochemistry 11,* 4903.
Miller, E. J. (1976), *Mol. Cell. Biochem. 13,* 165.
Miller, E. J., and Lunde, L. G. (1973), *Biochemistry 12,* 3153.
Miller, E. J., and Piez, K. A. (1966), *Anal. Biochem. 16,* 320.
Piez, K. A. (1968), *Anal. Biochem. 26,* 305.
Pisano, J. J., Bronzert, T. J., and Brewer, H. G., Jr. (1972), *Anal. Biochem. 45,* 43.
Prockop, D. J., Berg, R. A., Kivirikko, K. I., and Uitto, J. (1976), in Biochemistry of Collagen, Ramachandran, G. N., and Reddi, A. H., Ed., New York, N.Y., Plenum Press, p 163.
Schroeder, W. A. (1967), *Methods Enzymol. 11,* 351.
Traub, W. (1974), *Isr. J. Chem. 12,* 435.
Traub, W., and Fietzek, P. P. (1976), *FEBS Lett. 68,* 245.
Trelstad, R. L., Kang, A. H., Igarashi, S., and Gross, J. (1970), *Biochemistry 9,* 4993.

RNA Primers in SV40 DNA Replication: Identification of Transient RNA–DNA Covalent Linkages in Replicating DNA†

Stephen Anderson, Gabriel Kaufmann,‡ and Melvin L. DePamphilis*

ABSTRACT: SV40 DNA, replicating in isolated nuclei, contains RNA–DNA covalent linkages which were quantitated by measuring the release of [2′(3′)-^{32}P]rNMPs from [^{32}P]-DNA incubated in KOH (^{32}P-label transfer assay). More than 96% of the ^{32}P label released during this incubation was shown to be in [2′(3′)-^{32}P]rNMPs by chemically converting it into cyclic [2′:3′-^{32}P]rNMPs and then enzymatically cleaving the cyclic nucleotides to produce [3′-^{32}P]rNMPs. [α-^{32}P]dNTP, incorporated into DNA, was identified as the ^{32}P donor because the amount of ^{32}P-label transferred was proportional to the specific radioactivity of the labeled substrate. All 16 possible rN–dN linkages were found in SV40 replicating DNA at frequencies that suggested a near-random distribution on the genome. These RNA–DNA covalent linkages behaved as transient intermediates in DNA synthesis; they disappeared at the same rate that nascent 4S DNA chains ("Okazaki pieces") were joined to the growing daughter strands. Therefore, these linkages exhibited kinetic properties consistent with the proposed role of RNA as a primer for discontinuous DNA synthesis. When 4S DNA joining was inhibited by the absence of cytosol, the disappearance of RNA–DNA covalent linkages was not prevented. Inhibition of DNA synthesis with either *ara*-CTP or *ara*-ATP also failed to block the removal of RNA–DNA covalent linkages. Thus, the excision of these putative RNA primers does not appear to require either the concomitant joining of 4S DNA chains or DNA synthesis.

DNA synthesis generally occurs contemporaneously on both sides of a replication fork during semiconservative DNA replication. This requires one of the daughter strands to grow in the 3′ to 5′ direction, despite the fact that all known DNA polymerases synthesize DNA only in the 5′ to 3′ direction. Okazaki and his co-workers (1968) reconciled these observations by postulating a mechanism of discontinuous DNA synthesis, whereby short pieces of nascent DNA are repeatedly initiated. This allows DNA synthesis to proceed simultaneously away from as well as toward the replication fork. However, none of the known DNA polymerases are able to initiate DNA synthesis de novo; synthesis always requires a 3′-OH terminated polynucleotide "primer" hydrogen-bonded to the template strand. The primer may be provided by the synthesis of an oligoribonucleotide for each nascent DNA chain. Such putative RNA primers must be transient since mature nonreplicating DNA does not contain RNA–DNA covalent linkages.

Recent work in eukaryotic DNA replication has been directed toward identifying and isolating oligoribonucleotides covalently linked to the 5′ termini of newly synthesized DNA chains (Pigiet et al., 1974; Hunter and Francke, 1974a; Reichard et al., 1974; Tseng and Goulian, 1975a; Waqar and Huberman, 1975a,b; Kaufmann et al., 1977; general review of RNA primers, Kornberg, 1976). The RNA priming hypothesis predicts that such putative RNA primers will be excised at a rate equal to or faster than the rate that short nascent pieces of DNA are joined to growing daughter strands. To test this notion, we have undertaken a study of the removal of RNA primers using nuclei isolated from SV40[1]-infected CV-1 cells.

† From the Department of Biological Chemistry, Harvard Medical School, Boston, Massachusetts 02115. *Received June 23, 1977.* This work was supported by the National Institutes of Health Grant CA 15579-03. S. A. was supported by National Service Award CA 09031. G. K. was supported in part by a fellowship from the European Molecular Biology Organization. M. L. D. is an Established Investigator of the American Heart Association.

‡ Present address: Department of Biochemistry, The Weizmann Institute of Science, Rehovot, Israel.

[1] Abbreviations used: SV40, Simian virus 40; SV40(I) DNA, covalently closed superhelical viral DNA; SV40(II) DNA, duplex circular viral DNA containing at least one single-strand interruption; SV40(RI) DNA, replicative intermediates of SV40 DNA; Hepes-Na, sodium *N*-2-hydroxyethylpiperazine-*N*′-2-ethanesulfonate; EDTA, sodium ethylenediaminetetraacetate; Tris, tris(hydroxymethyl)aminomethane; NTP and dNTP, ribo- and deoxyribonucleoside triphosphates, respectively; *ara*-CTP, cytosine 1-β-D-arabinoside-5′-triphosphate; *ara*-ATP, adenine 9-β-D-arabinoside 5′-triphosphate.