Ogawa, Y., Quagliarotti, G., Jordan, J., Taylor, C. W., Starbuck, W. C., and Busch, H. (1969), *J. Biol. Chem. 244*, 4387.
Panyim, S., and Chalkley, R. (1969), *Arch. Biochem. Biophys. 130*, 337.
Permogorov, V. I., Debabov, V. G., Sladkova, I. A., and Rebentish, B. A. (1970), *Biochim. Biophys. Acta 199*, 556.
Riordan, J. F., and Vallee, B. L. (1967), *Methods Enzymol. 11*, 565.
Shih, T., and Fasman, G. (1970), *J. Mol. Biol. 52*, 125.
Shirey, T., and Huang, R. C. (1969), *Biochemistry 8*, 4138.
Simpson, R. T. (1970), *Biochemistry 9*, 4814.
Simpson, R. T., and Sober, H. A. (1970), *Biochemistry 9*, 3103.
Walker, I. O. (1965), *J. Mol. Biol. 14*, 381.
Weber, K., and Osborn, M. (1969), *J. Biol. Chem. 244*, 4406.
Zubay, G., and Doty, P. (1959), *J. Mol. Biol. 1*, 1.

Structure of Rat Skin Collagen α1-CB8. Amino Acid Sequence of the Hydroxylamine-Produced Fragment HA1*

Gary Balian, Eva Marie Click, and Paul Bornstein†

ABSTRACT: Hydroxylamine cleavage of α1-CB8, a large CNBr-produced peptide obtained from the α1 chain of rat skin collagen, yielded two fragments. The amino acid sequence of the 99 residues in the NH_2-terminal fragment, HA1, has been determined. In keeping with other collagen sequences, glycine was found as every third amino acid and hydroxyproline was limited to the third position in the repeating triplet Gly-X-Y. However, hydroxylation of several prolyl residues in this position was incomplete, indicating that the phenomenon of partial hydroxylation extends to prolyl residues far removed from the NH_2 terminus of the α chain. A pronounced clustering of charged residues, which correlates well with the band pattern observed in electron micrographs of segment long-spacing aggregates of α1-CB8, was found in HA1. The data obtained in this study, together with other published sequences, further suggest that certain amino acids, notably leucine and phenylalanine, are distributed in a nonrandom fashion in positions X and Y of the collagen triplet.

The determination of the primary structure of α1-CB8, a large fragment obtained by CNBr cleavage of the α1 chain of rat collagen, was undertaken to provide detailed information regarding the chemistry of collagen. It is expected that the analysis of a collagen sequence of this magnitude (282 amino acids) will (1) assist in relating the structure of the molecule to its primary function as a self-aggregating unit in fiber formation; (2) augment existing knowledge of the distribution of specific amino acids in the collagen triplet, Gly-X-Y, and provide information regarding possible sequence homologies which may reflect duplication of genetic material; (3) elucidate further the pattern of hydroxylation of proline (Bornstein, 1967) and lysine (Butler, 1968; Bornstein, 1969a); (4) provide additional correlations between the distribution of charged amino acids and the band pattern in electron micrographs of SLS[1] aggregates of collagen (von der Mark *et al.*, 1970).

α1-CB8 constitutes about 25% of the length of the α1 chain of rat collagen and is located in the middle of the first half of the chain (Piez *et al.*, 1969; Rauterberg and Kühn, 1968). In the first paper in this series (Bornstein, 1970) the fractionation and amino acid compositions of the tryptic peptides of α1-CB8 were reported and the nature and mechanism of cleavage of a hydroxylamine-sensitive bond were described. The hydroxylamine-sensitive bond, thought previously to be an example of a nonpeptide bond linking intra-α-chain subunits (Gallop *et al.*, 1967), was shown to consist of a cyclic imide which formed by cyclization of an asparaginyl side chain with the subsequent (glycyl) amide group in the polypeptide chain (Bornstein, 1969b, 1970). Elucidation of the nature of the hydroxylamine-sensitive bond radically altered its significance. However, use was made of the specific cleavage of α1-CB8 by hydroxylamine to separate two large fragments, HA1 and HA2, which accounted for the amino acid composition and molecular weight of the starting material.

The availability of a point of cleavage roughly one-third of the length of α1-CB8 from the NH_2 terminus greatly simplified the determination of its amino acid sequence. The primary structure of the NH_2-terminal fragment HA1, consisting of 99 amino acids, is reported in this communication. The determination of the structure of HA2 (183 amino acids) is in progress and will be reported separately (G. Balian, E. M. Click, M. Hermodson, and P. Bornstein, in preparation).

Materials and Methods

Preparation of α1-CB8 and Cleavage with Hydroxylamine. α1-CB8 was prepared from salt-extracted lathyritic rat skin collagen (Butler *et al.*, 1967; Bornstein, 1970). Acid-extracted normal rat skin collagen was also used as a source for some

* From the Department of Biochemistry, and Research Training Unit, Department of Medicine, University of Washington, Seattle, Washington 98195. *Received July 13, 1971.* A portion of this work has appeared in preliminary form (Balian and Click, 1971). This work was supported by National Institutes of Health Contract 69-2230 and National Institutes of Health Grant AM 11248.

† Recipient of Research Career Development Award K4-AM-42582 from the U. S. Public Health Service; to whom to address correspondence.

[1] Abbreviations used are: SLS, segment long spacing; PITC, phenyl isothiocyanate; PTH, phenylthiohydantoin; DFP, diisopropyl fluorophosphate.

preparations. The procedure for hydroxylamine cleavage (Bornstein, 1970) was modified as follows. The final concentration of hydroxylamine was 2 M and incubation was performed at pH 9.0 and 45° for 2 hr. These conditions were found to increase the degree of cleavage of the primary susceptible bond in α1-CB8 to 70–80% without causing detectable nonspecific degradation.

Preparation of HA1 and HA2. The products of hydroxylamine cleavage were separated by molecular sieve chromatography on 8% agarose (Bio-Gel 1.5m, 200–400 mesh, Bio-Rad Laboratories) equilibrated with 1 M $CaCl_2$ containing 0.05 M Tris-HCl, pH 7.5 (Piez, 1968; Bornstein, 1970). Fragments HA1 and HA2 were purified by CM-cellulose column chromatography. The column was equilibrated at 40° with 0.02 M sodium acetate (pH 4.8) and the fragments eluted with a linear gradient of NaCl from 0 to 0.1 M.

Enzymatic Hydrolyses. Chromatographically purified collagenase (CLSPA, Worthington) was further purified on Sephadex G-200. The enzyme was added to peptide in a weight ratio of 1% in 0.2 M NH_4HCO_3 (pH 7.8) containing 1 mM $CaCl_2$ and 0.25 mM *N*-ethylmaleimide. The sulfhydryl reagent was added to inhibit a small amount of nonspecific protease activity which contaminated the enzyme even at this stage of purification (Peterkofsky and Diegelmann, 1971). Incubation was performed at 37° for 16 hr and terminated by lyophilization.

Digestion of HA1 with thermolysin (crystalline, Daiwa Kasei, K. K. Osaka, Japan) was performed at 37° using 1 μmole of peptide dissolved in 2 ml of H_2O. The solution was adjusted to pH 7.5 with NaOH, 10 μl of a 0.5-mg/ml solution of thermolysin in 1 mM calcium acetate was added and the pH maintained at 7.5 in a pH-Stat with 0.02 M NaOH. After 1 hr another 10-μl aliquot of enzyme was added, the reaction continued for 30 min and terminated by reduction of the pH to 2.5. HA1-T13 was dissolved in 0.2 M NH_4HCO_3 buffer (pH 7.8) and digested with thermolysin for 3 hr at 37° using a 1:100 weight ratio of enzyme to substrate.

Papain digestion was carried out as previously described (Bornstein, 1967). A 2–5% weight ratio of enzyme to substrate was used and digestion was limited to 24 hr. Tryptic digestion was performed as described (Bornstein, 1970).

For carboxypeptidase digestion a known amount of peptide was dissolved in 1 ml of 1 M Tris-HCl (pH 8.0) containing 2 mM $CaCl_2$ and digested for 16 hr at room temperature with 10 μl of carboxypeptidase B (1 mg/ml, Worthington, DFP treated). The temperature of the sample was raised to 37°, 15 μl of carboxypeptidase A (1 mg/ml) was added, and aliquots containing 15 nmoles of peptide were removed at 0.5-, 1-, 3-, 8-, and 14-hr intervals. The peptide was then added to an equal volume of 50% acetic acid and the solution applied directly to an amino acid analyzer column.

Ion-exchange chromatography and automated analysis of peptides were performed largely as described previously (Bornstein, 1970). In order to achieve better resolution of peptides eluting from cation-exchange columns, 0.05 M pyridine acetate (pH 2.5) was included as the initial buffer in the gradient and an additional buffer, 0.5 M pyridine acetate (pH 3.75), was used (Bradshaw *et al.*, 1969).

Paper chromatography and electrophoresis were performed as described (Bornstein, 1970). Descending paper chromatography was also performed in butanol–acetic acid–water (12:3:5, v/v) and butanol–pyridine–water (1:1:1, v/v).

Determination of Amide Groups. The presence of side-chain amides in peptides found to contain one glutamic or aspartic acid after acid hydrolysis was determined by the electrophoretic mobility of the peptide at pH 6.5 in comparison with amino acid and peptide standards. In several instances the change in electrophoretic mobility occurring after Edman degradation could be used to identify side-chain amides or acids in peptides containing more than one such residue. In one case (residue 23) the nature of the side chain was determined by identification of the PTH derivative by gas–liquid chromatography (Pisano and Bronzert, 1969).

Amino Acid Sequence Determinations. The amino acid sequences of two short tryptic peptides (T9 and T6) were determined by sequential Edman degradation and identification of newly formed NH_2-terminal amino acids as their dansyl derivatives (Bornstein, 1969a). These sequences were confirmed as part of the NH_2-terminal sequence of HA1, determined by an automatic amino acid sequencer. All other sequence determinations were made by the subtractive Edman procedure (Konigsberg, 1967). Reagents were obtained from Pierce Chemical Co. and were Sequanal grade. A known amount of peptide was dissolved in a mixture of 0.5 ml of pyridine and 0.5 ml of 5% *N*-ethylmorpholine acetate (adjusted to pH 9.2 with acetic acid) to give a final pH of 9.5. PITC (20 μl) was added and the reaction performed in a nitrogen atmosphere at 50° for 30 min. The solution was lyophilized and cleaved with 200 μl of trifluoroacetic acid under nitrogen for 30 min at 40°. The dried reactants were suspended in 150 μl of H_2O and extracted three times with 1.5 ml of benzene. The aqueous layer was lyophilized, dissolved in pyridine–morpholine buffer, and an aliquot containing 5–15 nmoles of peptide hydrolyzed for amino acid analysis.

The NH_2-terminal sequence of HA1 was determined with a Beckman Model 890A automatic amino acid sequencer using a modification of the procedure of Edman and Begg (1967). The PTH derivatives of amino acids were identified directly by gas–liquid chromatography (Pisano and Bronzert, 1969).

Trifluoroacetylation. HA1 (2 μmoles) in 0.2 M Na_2CO_3 (pH 10.0) was treated with ethyl thioltrifluoroacetate at 25° for 1 hr as described by Goldberger (1967). The pH was maintained at 10.0 with 1 M NaOH. The peptide was desalted on Bio-Gel P2 (100–200 mesh, Bio-Rad Laboratories) equilibrated with 0.2 M NH_4HCO_3 (pH 7.8) and digested with trypsin. The trypsin was then inactivated with 5 μl of 1 M DFP and trifluoroacetyl groups removed by treatment with 1 M piperidine at 0° for 2 hr. After desalting on P2 (200–400 mesh), the tryptic peptides were separated by chromatography on ion-exchange columns.

Maleylation. Reaction of HA1 with maleic anhydride was performed as described by Butler *et al.* (1969). Following tryptic digestion maleyl groups were removed by incubation in 0.2 M pyridine acetate (pH 3.2) at 60° for 6 hr and the tryptic peptides separated by ion-exchange column chromatography.

Amino acid analyses were performed on a Beckman 120C analyzer modified for accelerated single-column gradient elution (Miller and Piez, 1966). Corrections for hydrolytic losses and incomplete release of valine have been reported (Bornstein, 1970). The use of a range card for high sensitivity made it possible to obtain accurate analyses on quantities of peptide as low as 2–3 nmoles.

Results

Preparation of α1-CB8-HA1. The agarose elution pattern of a hydroxylamine digest of α1-CB8 is illustrated in Figure 1.

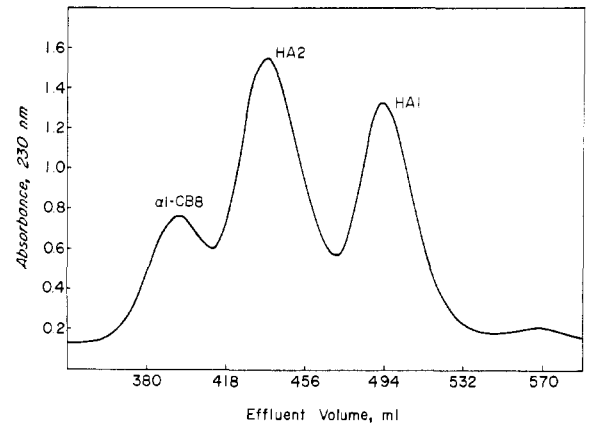


FIGURE 1: Agarose elution pattern of fragments obtained by cleavage of α1-CB8 with hydroxylamine.

HA1 was recovered in stoichiometric amounts when provision was made for incomplete cleavage (as evidenced by residual α1-CB8). The recovery of HA2 was less complete, a finding consistent with the observed absorbance of the HA2 peak in Figure 1. Preliminary evidence indicates that an additional point of cleavage by hydroxylamine exists some 40–50 amino acids from the COOH terminus of HA2 (G. Balian and P. Bornstein, unpublished observations). HA2 is therefore degraded further to a fragment intermediate in size between HA1 and HA2. Evidence for this fragment exists in acrylamide gel electrophoresis patterns of hydroxylamine digests of α1-CB8 (see Figure 3, Bornstein, 1970).

Order of Tryptic Peptides in HA1 (Figure 2). The order of the first three tryptic peptides was established by the determination of the sequence of the first 12 amino acids in HA1 using the automatic sequencer, *viz.*, Gly-Pro-Arg-Gly-Leu-Hyp-Gly-Glu-Arg-Gly-Arg-Hyp. This sequence was confirmed by determination of the internal sequences of the three tryptic peptides T9, T6, and T8 (see below) and by analyses of thermolysin digests of HA1 and is consistent only with the order T9-T6-T8.

Blockage of ϵ-amino groups by either trifluoroacetylation or maleylation followed by tryptic digestion resulted in the isolation of the overlapping peptide T13-T4b. These procedures together with Edman degradation of the T4 fraction demonstrated conclusively that the sequence thought initially to represent a single octadecapeptide T4 (Bornstein, 1970) was in reality two very similar nonapeptides T4a and T4b. For reasons which are not clear the second lysyl residue in HA1, at the COOH terminus of T14, failed to react

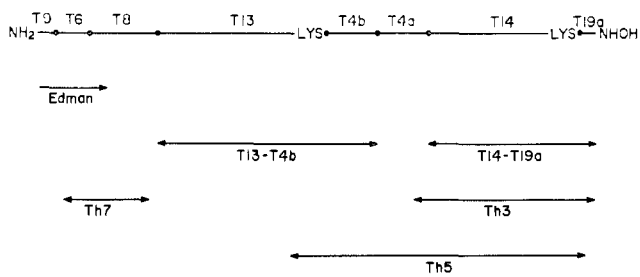
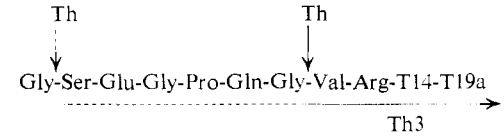


FIGURE 2: The order of the tryptic peptides in HA1. The overlapping sequences T13-T4b and T14-T19a were obtained by trifluoroacetylation and incomplete tryptic digestion, respectively. Th3, Th5, and Th7 represent thermolysin-produced peptides.

TABLE I: Amino Acid Composition of Th3 Compared with Its Component Tryptic Peptides.[a]

	Th3[b]	T4a(Val-Arg)-T14-T19a
Hydroxyproline	3.5	3.6
Aspartic acid	2.8	3.1
Threonine	(0.4)	
Serine	0.6	
Glutamic acid	2.7	1.8
Proline	4.2	3.7
Glycine	11.9	10.3
Alanine	6.9	7.2
Valine	1.0	1.0
Hydroxylysine	0.1	0.1
Lysine	1.0	0.9
Arginine	1.4	1.0

[a] Values are expressed as residues per peptide. [b] The additional content of serine, glutamic acid, proline, and glycine in Th3 may be explained by partial thermolytic cleavage at the Gly–Ser bond in T4a as well as at the Gly–Val bond, *viz.*



The additional larger peptide was incompletely separated from Th3.

with either ethyl thioltrifluoroacetate or maleic anhydride. However, T19a was incompletely released by trypsin from T14, presumably due to the proximity of this lysine to the COOH terminus of the chain. The overlapping peptide, T14-T19a, was thereby isolated.

Digestion of HA1 with thermolysin provided the additional overlaps necessary to complete the order of the tryptic peptides. Thermolysin-produced peptides were separated by chromatography on cation-exchange resins and on Sephadex G-50. Despite the mildness of the conditions used for digestion, elution patterns were complex indicating a considerable number of thermolysin-susceptible bonds in HA1.

The amino acid composition of peptide Th3 indicates that it includes the terminal dipeptide Val-Arg from T4a, plus T14 and T19a (Table I). Since T19a contains hydroxamate (Bornstein, 1970) it must be the COOH-terminal tryptic peptide in HA1. The order of the tryptic peptides in Th3 must therefore be Val-Arg(T4a)-T14-T19a. Additional peptides were isolated from thermolysin digests (data not reported) with compositions consistent with the orders T6-T8 and T13-T4b-T4a-T14. These findings unambiguously define the order of the tryptic peptides in HA1 as: T9-T6-T8-T13-T4b-T4a-T14-T19a (Figure 2).

Internal Sequences of the Tryptic Peptides

T9: Gly-Pro-Arg (Residues 1–3). The sequence of this tripeptide was determined by Edman degradation and dansylation and confirmed with the automatic amino acid sequencer.

T6: Gly-Leu-Hyp-Gly-Glu-Arg (Residues 4–9). The sequence of this hexapeptide was determined by Edman degradation and dansylation and confirmed with the automatic se-

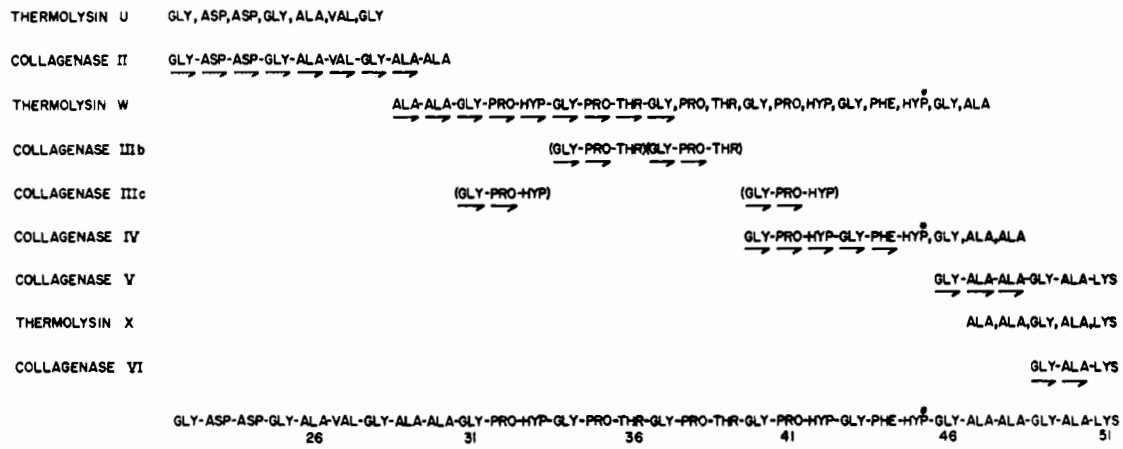


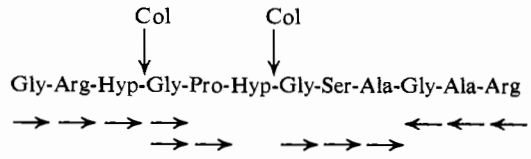
FIGURE 3: Amino acid sequence of T13. Collagenase- and thermolysin-produced peptides are arranged in order and indicate the extent of sequence overlap obtained. Only the major thermolysin peptides are listed. Horizontal arrows indicate the extent of Edman degradation. In position 45, proline is incompletely hydroxylated (Hyp*).

TABLE II: The Amino Acid Sequence of T8.

1. Subtractive Edman Degradation[a]

	Step				
	0	1	2	3	4
Hydroxyproline	2.20	2.04	2.16	*1.21*	1.18
Serine	1.03	1.04	1.07	1.08	1.00
Proline	0.92	1.13	1.22	1.05	0.98
Glycine	4.00	*3.40*	2.72	2.73	*2.35*
Alanine	1.90	1.96	1.67	1.84	1.82
Arginine	1.98	2.00	*1.10*	0.92	1.20
Sequence		Gly	Arg	Hyp	Gly

2. Collagenase and Carboxypeptidase Digestion[b]



[a] Values are expressed as residue per peptide. Residues in italic type are those lost at each step. [b] The horizontal arrows indicate the extent of Edman degradation (⟶) or carboxypeptidase digestion (⟵). The vertical arrows indicate points of cleavage by collagenase (Col).

TABLE III: Amino Acids Released as a Function of Time by Sequential Digestion of Peptides T8, T4b, and T4a with Carboxypeptidases B and A.[a]

		Reaction Time (hr)				
		0.5	1	3	8	14
T8	Arginine	16.9	16.8	17.7		16.7
	Alanine	4.6	6.4	6.4		9.2
	Glycine	1.4	1.3	1.8		2.4
T4b	Arginine	15.3	12.1	12.5	12.6	
	Alanine	5.5	5.8	7.3	9.8	
	Glycine	1.8	1.8	2.4	4.1	
	Glutamine	1.2	1.0	1.1	1.6	
T4a	Arginine	16.9	15.5	17.2	17.6	
	Valine	14.2	13.4	14.5	14.5	
	Glycine	6.3	7.1	8.0	11.6	
	Glutamine	1.2	1.4	1.7	2.7	

[a] Expressed as nanomoles per unit of peptide.

quencer. In addition, the sequence of the first four residues was determined by the subtractive Edman procedure. The presence of only 0.25 equiv of NH_3 after acid hydrolysis and the neutral charge of T6 at pH 6.5 indicate that glutamic acid rather than glutamine exists in position 8.

T8: Gly-Arg-Hyp-Gly-Pro-Hyp-Gly-Ser-Ala-Gly-Ala-Arg (*Residues 10–21*). The sequence of the first four amino acids in T8 was determined by subtractive Edman degradation (Table II). The COOH-terminal sequence Gly-Ala-Arg was determined by carboxypeptidase B and A digestion (Table III). Three collagenase-produced peptides were separated by high-voltage electrophoresis at pH 6.5. Their order in T8 was apparent from knowledge of the sequence at the NH_2 and COOH termini of the tryptic peptide. Subtractive Edman degradations on the second and third collagenase peptides, as indicated in Table II, served to complete the sequence of T8.

T13: Gly-Asp-Asp-Gly-Ala-Val-Gly-Ala-Ala-Gly-Pro-Hyp-Gly-Pro-Thr-Gly-Pro-Thr-Gly-Pro-Hyp-Gly-Phe-Hyp-Gly-Ala-Ala-Gly-Ala-Lys* (*Residues 22–51*).[2] Subtractive Edman degradations indicated that the initial sequence of T13 was Gly-Asx-Asx-Gly. Collagenase digestion produced six major peptides (Table IV and Figure 3) which were resolved by electrophoresis at pH 6.5, by paper chromatography in butanol–pyridine–water, and by ion-exchange chromatography. The sequences of the collagenase-produced peptides were obtained by subtractive Edman degradation and their relative positions in T13 were established as follows. The sequence of peptide II, Gly-Asx-Asx-Gly-Ala-Val-Gly-Ala-Ala, corresponds to the NH_2-terminal

[2] The symbol Hyp* is used to indicate an incompletely hydroxylated prolyl residue.

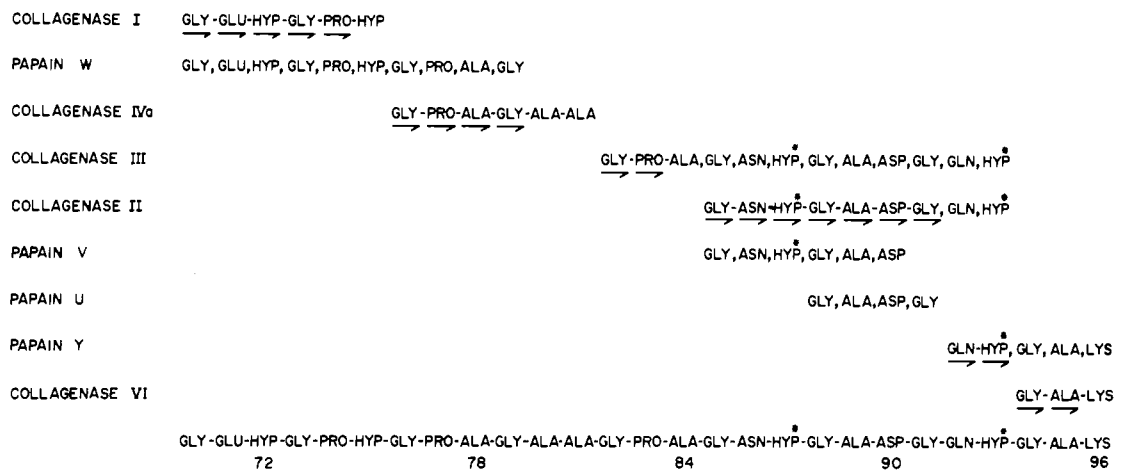


FIGURE 4: Amino acid sequence of T14. Collagenase- and thermolysin-produced peptides are arranged in order and indicate the extent of sequence overlap obtained. Only the major papain peptides are listed. Horizontal arrows indicate the extent of Edman degradation. Prolyl residues in positions 87 and 93 are incompletely hydroxylated (Hyp*).

TABLE IV: Amino Acid Compositions of T13 and Its Collagenase- and Thermolysin-Produced Peptides.[a]

		Collagenase Peptides						Thermolysin Peptides		
	T13	II	IIIb	IIIc	IV	V	VI	U	W	X
Hydroxyproline	2.7			0.9	1.8				2.4	
Aspartic acid	2.1	1.9						1.7	(0.2)	
Threonine	1.8		0.9						1.9	
Glutamic acid	(0.3)									
Proline	4.4		1.0	1.0	1.2				4.7	
Glycine	10.0	3.2	1.1	1.1	3.3	2.1	1.0	3.1	7.4	1.1
Alanine	6.2	3.0			2.0	3.0	1.0	1.2	2.9	3.1
Valine	1.1	1.0						0.9	(0.3)	
Phenylalanine	1.0				1.0				1.0	
Hydroxylysine	0.2									0.1
Lysine	0.8					1.1	1.0			0.8

[a] Values are expressed as residues per peptide. A space indicates less than 0.2 residue. Residues in parentheses are fractional residues thought to be impurities.

sequence of T13. The lysine-containing tripeptide (VI) is COOH terminal and had the sequence Gly-Ala-Lys. The existence of the overlapping peptide (V), Gly-Ala-Ala-Gly-Ala-Lys, indicated that peptide IV, Gly-Pro-Hyp-Gly-Phe-(Hyp*, Gly,Ala,Ala), precedes peptide VI. The hexapeptide (Gly,Pro,Hyp,Gly,Phe,Hyp*) was also isolated from a separate collagenase digest confirming the sequence of residues 40–45 in T13. Peptides IIIb and IIIc were found to have the sequences Gly-Pro-Thr and Gly-Pro-Hyp, respectively. However, since T13 contains two residues of threonine and between two and three residues of hydroxyproline, the above tripeptides must originate from repeating sequences in T13 (Figure 3).

Digestion of T13 with thermolysin produced three major peptides (Table IV and Figure 3). Thermolysin U with the composition (Gly,Asx,Asx,Gly,Ala,Val,Gly) originates from the NH_2 terminus of T13, and thermolysin X with the composition (Ala,Ala,Gly,Ala,Lys) originates from the COOH terminus. The remainder of the amino acids in T13 are included in thermolysin W. The amino acid composition of thermolysin W (Table IV) suggests heterogeneity resulting from minor cleavage at the Ala–Val bond (residues 26–27). Subtractive Edman degradation indicated the sequence of the first 9 amino acids in thermolysin W to be Ala-Ala-Gly-Pro-Hyp-Gly-Pro-Thr-Gly. Consequently, the four tripeptides in collagenase fractions IIIb and IIIc must be positioned as shown Figure 3. The incomplete hydroxylation of the prolyl residue in position 45, as indicated by amino acid analysis of collagenase peptide IV, was confirmed by the isolation of a peptide from a papain digest of T13 (data not shown) which contained phenylalanine and proline but no hydroxyproline.

Aspartic acid was assigned to residue 23 by indentification of the PTH amino acid by gas–liquid chromatography. Aspartic acid was also assigned to residue 24 by the negative charge, at pH 6.5, of collagenase peptide II after two subtractive Edman degradations.

T4b: Gly-Glu-Ala-Gly-Pro-Gln-Gly-Ala-Arg (Residues 52–60). Subtractive Edman degradation on a mixture of T4a and T4b (thought to be a single peptide T4) clearly showed the presence of two similar peptides (Table V). The initial sequence Gly-Glx-Ala-Gly-Pro was assigned to T4b after

this peptide was separated as the overlapping sequence T13-T4b, the trifluoroacetyl group removed, and the peptides T13 and T4b generated by tryptic digestion. The COOH-terminal sequence of T4b was found to be Gln-Gly-Ala-Arg by sequential carboxypeptidase B and A digestion of the purified peptide (Table III). The presence of glutamic acid in position 53 is indicated by the neutrality of T4b at pH 6.5. After two Edman degradations the peptide became basic at pH 6.5 confirming that residue 53 is glutamic acid and residue 57 glutamine.

T4a: Gly-Ser-Glu-Gly-Pro-Gln-Gly-Val-Arg (Residues 61–69). The initial sequence of T4a, Gly-Ser-Glx-Gly-Pro-, was obtained by subtractive Edman degradation on a mixture of T4a and T4b and confirmed by subsequent amino acid analysis and collagenase digestion of the purified peptide (Table V). T4a was obtained free from T4b by ion-exchange chromatography of a tryptic digest of trifluoroacetylated HA1. The COOH-terminal sequence of T4a was found to be Gln-Gly-Val-Arg by carboxypeptidase digestion of the purified peptide (Table III). T4a was neutral at pH 6.5 and became basic after three Edman degradations confirming that residue 63 is glutamic acid and residue 66 is glutamine.

T14: Gly-Glu-Hyp-Gly-Pro-Hyp-Gly-Pro-Ala-Gly-Ala-Ala-Gly-Pro-Ala-Gly-Asn-Hyp-Gly-Ala-Asp-Gly-Gln-Hyp*-Gly-Ala-Lys (Residues 70–96).* The NH_2-terminal sequence obtained by subtractive Edman degradation was found to be Gly-Glx-Hyp-Gly-Pro-Hyp-Gly-Pro. Cleavage with collagenase yielded five peptides which were separated by a combination of electrophoresis at pH 6.5 and chromatography in butanol–acetic acid–water (Table VI and Figure 4). The amino acid sequence of collagenase peptide I was Gly-Glx-Hyp-Gly-Pro-Hyp which localized its position as the NH_2-terminal hexapeptide. This peptide was acidic at pH 6.5 indicating that residue 71 is glutamic acid. Collagenase peptide II had the sequence Gly-Asx-Hyp*-Gly-Ala-Asx-Gly-(Glx,Hyp*). The composition of III was equal to that of II plus three residues (Gly,Pro,Ala). Since the NH_2-terminal sequence of III was found to be Gly-Pro the sequence of III was established as Gly-Pro-Ala-Gly-Asx-Hyp-Gly-Ala-Asx-Gly-(Glx,Hyp*). The two remaining collagenase-produced peptides IVa and VI had the sequences Gly-Pro-Ala-Gly-Ala-Ala and Gly-Ala-Lys, respectively. Peptide VI, by virtue of the presence of lysine, constitutes the COOH-terminal sequence of the tryptic peptide T14.

Digestion of T14 with papain produced four major peptides (Table VI and Figure 4) which were separated by electrophoresis at pH 6.5. The composition of papain peptide Y overlaps the two COOH-terminal residues of collagenase peptides II and III, and the COOH-terminal tripeptide VI, thus restricting the position of collagenase peptide IVa to residues 76–81 between collagenase peptides I and III (Figure 4). The amino acid compositions of papain peptides V and Y reflect directly the degree of underhydroxylation of proline in positions 87 and 93.

The distribution of amides and acidic amino acids in T14 was determined as follows. Residues 71 and 90 were identified as glutamic and aspartic acids, respectively, by the negative charges at pH 6.5, of collagenase I and papain U. Residues 86 and 92 were shown to be asparagine and glutamine, respectively, by the electrophoretic mobilities at pH 6.5, of papain peptides V and Y.

T19a: Gly-Ala-Asn (Residues 97–99). This tripeptide was known to be COOH terminal in HA1 by its hydroxamate content. The sequence was established by dansylation and sequential Edman degradation. In a previous report (Bornstein, 1970) it had been noted that this tryptic peptide could not be located in elution patterns from ion-exchange columns. This may result from the existence of the peptide in three forms with different charges, *i.e.*, Gly-Ala-Asp, Gly-Ala-Asp-NHOH, and Gly-Ala-Asp-$(NHOH)_2$. In addition, the tripeptide is incompletely released by trypsin. Although found as the hydroxamate of aspartic acid, the COOH-terminal residue of HA1 is believed to exist as asparagine in α1-CB8 (see Bornstein, 1970, and Discussion).

Complete amino acid sequence of HA1 is summarized in Figure 5. The basis for the assignment of acid and amide groups in the sequence is summarized in Table VII. As in other amino acid sequences from the helical region of collagen (Bornstein, 1967; Butler, 1970; Highberger *et al.*, 1971; Butler and Ponds, 1971; von der Mark *et al.*, 1970), glycine is present as every third residue and hydroxyproline is limited to position Y in the collagen triplet Gly-X-Y. In three instances (residues 45, 87, and 93) hydroxylation of prolyl residues in position Y was grossly incomplete. This aspect and other features of the amino acid sequence of HA1 are considered in the Discussion.

TABLE V: Amino Acid Composition and Sequence of T4a and T4b.

1. Composition of T4a, T4b, and Their Derivatives[a]

	T4a				T4b	
	Residues 64–69[b]	Collagenase Peptides I	Collagenase Peptides II		Residues 54–60[c]	
Serine	0.9			0.9		
Glutamic acid	2.2	1.0		2.3	1.8	1.1
Proline	1.1	1.1		1.0	1.2	1.0
Glycine	3.2	2.0	1.1	2.1	3.1	2.1
Alanine	(0.3)				2.0	1.9
Valine	0.9	1.0	1.0			
Arginine	1.0	1.0	1.0		0.9	1.0

2. Subtractive Edman Degradation of a Mixture of T4a and T4b[d]

	Step					
	0	1	2	3	4	5
Serine	1.03	1.17	*0.43*	0.38	0.29	0.37
Glutamic acid	4.12	3.98	*3.52*	*2.75*	2.56	2.20
Proline	2.00	1.86	1.93	1.73	1.83	*0.12*
Glycine	5.69	*4.05*	4.35	4.45	*3.49*	2.80
Alanine	1.94	1.89	1.89	*1.29*	1.09	0.87
Valine	1.03	1.08	1.01	1.10	1.10	0.96
Arginine	1.88	1.79	2.18	2.08	1.98	2.07
T4a		Gly	Ser	Glu	Gly	Pro
T4b		Gly	Glu	Ala	Gly	Pro

[a] Values are expressed as residues per peptide. A space indicates less than 0.2 residue. Residues in parentheses are fractional residues thought to be impurities. [b] Obtained after three Edman degradations of T4a. [c] Obtained after two Edman degradations of T4b. [d] Values are expressed as residues per peptide. Residues in italic type are those lost at each step.

TABLE VI: Amino Acid Compositions of T14 and Its Collagenase- and Papain-Produced Peptides.[a]

		Collagenase Peptides					Papain Peptides			
	T14	I	II	III	IVa	VI	U	V	W	Y
Hydroxyproline	3.1	1.7	1.2	1.5				0.6	1.9	
Aspartic acid	2.1		1.9	2.0			1.1	2.1		0.4
Glutamic acid	2.0	1.0	1.1	1.1					1.1	1.0
Proline	3.7	1.1	0.8	1.5	1.0	1.0		0.2	1.9	0.6
Glycine	9.1	2.0	3.0	3.8	2.0	1.0	2.0	2.3	4.0	1.1
Alanine	6.1	(0.4)	1.0	1.9	3.1		1.0	1.1	1.1	1.0
Hydroxylysine	0.1									
Lysine	0.9					0.9				1.0

[a] Values are expressed as residues per peptide. A space indicates less than 0.2 residue. Residues in parentheses are fractional residues thought to be unpurities.

Discussion

The determination of the amino acid sequence of α1-CB8-HA1 extends the known structure of the α1 chain of rat skin collagen to a total of 234 residues from the NH_2 terminus (Bornstein, 1967; Butler, 1969a; Butler and Ponds, 1971; Kaug *et al.*, 1967). When added to other known sequences in the helical regions of rat collagen (Highberger *et al.*, 1971) and in calf skin collagen (von der Mark *et al.*, 1970), these sequences provide a basis for the investigation of patterns in the distribution of amino acids in collagen. Although the majority of amino acids appear to be equally distributed between positions X and Y in the triplet sequence Gly-X-Y, the data suggest a nonrandom distribution for certain amino acids in addition to glycine and the imino acids. Leucine and phenylalanine have thus far been found exclusively in position X. When the known positions of two leucyl and two phenylalanyl residues in α1-CB8-HA2 are included (G. Balian *et al.*, in preparation) a total of nine leucyl and five phenylalanyl residues are so located. Earlier studies using Edman degradation of mixtures of collagenase peptides of a fish collagen (Greenberg *et al.*, 1964) also indicated a marked preponderance of valine, leucine, and phenylalanine in position X. However, amino acids with bulky hydrophobic side chains are not excluded from position Y, since in the sequences referred to above one of three isoleucyl, three of five valyl, and three of four methionyl residues were located in position Y. The majority of arginyl and lysyl (including hydroxylysyl) residues occurred in position Y (26 of 35), whereas 11 of 14 glutamyl residues were found in position X. On the other hand, the 10 aspartyl residues so far encountered were distributed equally between positions X and Y.

The possibility that a preferred distribution exists for a number of amino acids in addition to glycine and the imino acids is of interest although additional data are required to fully substantiate this point. The presently accepted model of the collagen helix (Traub *et al.*, 1969) places no restrictions on the distribution of amino acids in positions X and Y, but interactions between collagen molecules in the collagen fiber may dictate a preferential segregation to enhance the aggregative properties of the protein. Should a nonrandom distribution of amino acids exist, the appearance of short repetitive sequences in collagen as seen in the sequence of von der

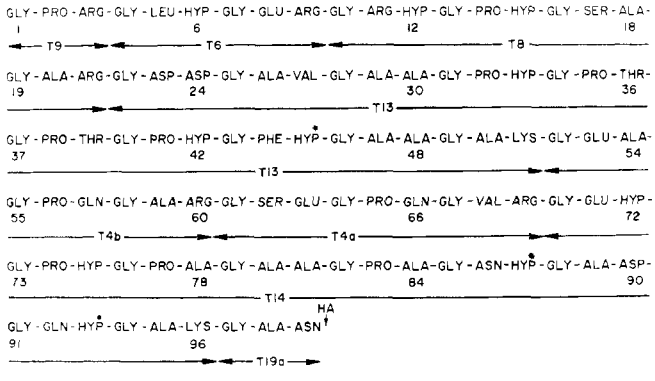


FIGURE 5: The complete amino acid sequence of α1-CB8-HA1. Only the tryptic peptides are indicated. Hydroxylation of prolyl residues at positions 45, 87, and 93 is incomplete as indicated by Hyp*. Hydroxylamine (HA) cleaves α1-CB8 at the asparaginyl bond in position 99 to form HA1 and HA2.

TABLE VII: Assignment of Acid and Amide Groups in α1-CB8-HA1.

Residue	Assignment	Method of Determination
8	Glu	HVE,[a] T6; NH_3 content of T6
23	Asp	PTH derivative[b]
24	Asp	HVE, T13-collagenase II after two Edman degradations
53	Glu	HVE, T4b
57	Gln	HVE, T4b after two Edman degradations
63	Glu	HVE, T4a
66	Gln	HVE, T4a after three Edman degradations
71	Glu	HVE, T14-collagenase I
86	Asn	HVE, T14-papain V
90	Asp	HVE, T14-papain U
92	Gln	HVE, T14-papain Y
99	Asn	See Bornstein (1970)

[a] High-voltage electrophoresis at pH 6.5. [b] Identification of PTH derivative of the amino acid by gas–liquid chromatography.

TABLE VIII: Substrate Specificity of Bacterial Collagenase in α1-CB8-HA1.[a]

Susceptible	Resistant
↓	
Gly-Pro-Hyp-Gly-X-Y	Gly-Glu-Hyp-Gly-Pro-Hyp
Gly-Pro-Hyp-Gly-Pro-Y	Gly-Ser-Glu-Gly-Pro-Gln
Gly-Pro-Y-Gly-X-Hyp	Gly-Ala-Asp-Gly-Gln-Hyp
Gly-X-Hyp-Gly-X-Y	Gly-Asn-Hyp-Gly-Ala-Asp
	Gly-Asp-Asp-Gly-Ala-Val
	Gly-Ser-Ala-Gly-Ala-Arg
	Gly-Ala-Val-Gly-Ala-Ala
Partially Susceptible	
⇣	
Gly-Ala-Ala-Gly-Ala-Lys	

[a] The symbols X and Y indicate any amino acid other than glycine, proline, or hydroxyproline.

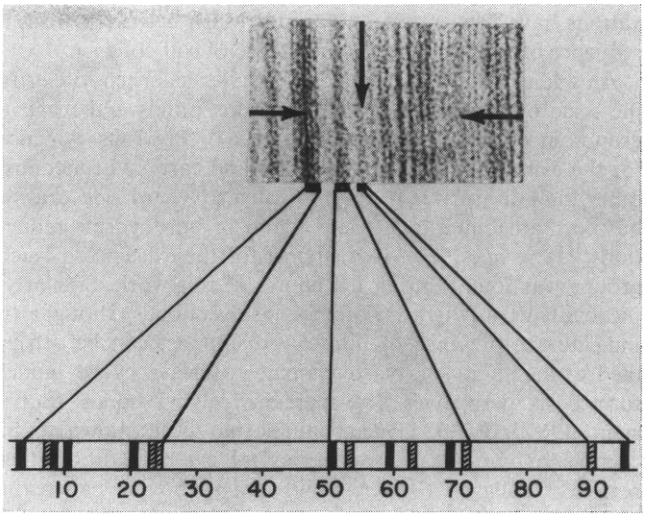


FIGURE 6: The band pattern in an electron micrograph of SLS aggregates of α1-CB8 from calf collagen (stained with phosphotungstic acid and uranyl acetate) is correlated with the distribution of basic and acidic amino acids in HA1. α1-CB8 extends between the two horizontal arrows with the NH_2-terminal end of the sequence to the left. The vertical arrow indicates the point of cleavage by hydroxylamine. The solid and hatched vertical bars represent the positions of basic and acidic amino acids, respectively.

Mark *et al.* (1970) could not be taken as evidence for duplication of genetic material.

In keeping with other sequence analyses of collagen, as well as with enzymatic studies using both synthetic and collagen peptides as substrates (see Rhoads *et al.*, 1971, for a recent discussion), hydroxyproline in HA1 is limited to positions immediately preceding glycine, *i.e.*, in position Y. Although no evidence was found for incomplete hydroxylation in the majority of positions designated as occupied by hydroxyproline in Figure 5, underhydroxylation to the extent of 10% or less would not have been detected by the procedures routinely used during the course of this work. In the case of positions 45, 87, and 93 incomplete hydroxylation was readily apparent. The hydroxyproline content in these positions was of the order of 40–80%. A similar degree of hydroxylation of certain prolyl residues was observed in sequence analyses of α1-CB2 (Bornstein, 1967) and hydroxylation to the extent of about 90% was observed in several positions in α1-CB4 (Butler and Ponds, 1971). In addition, a prolyl residue in position Y (preceded by lysine in position X) was not hydroxylated in α1-CB4. The widespread distribution of incompletely hydroxylated prolyl residues is supported by the finding that an increase of 5–15% in hydroxyproline content occurs after *in vitro* hydroxylation of collagen from various sources with collagen proline hydroxylase (Rhoads *et al.*, 1971).

The available data do not indicate a structural basis for this pattern of partial hydroxylation. Triplets in which proline in position Y is incompletely hydroxylated have now been found containing another prolyl residue or an acidic, basic, or hydrophobic amino acid. The finding that incomplete hydroxylation occurs in α1-CB8 makes it less likely that the phenomenon results from a lowered affinity of collagen proline hydroxylase for the NH_2-terminal region of a nascent α chain. Incomplete hydroxylation therefore probably does not result from an altered conformation of α chains near the NH_2 terminus (Bornstein, 1967) or from the possible requirement for an unusually extensive binding site for the hydroxylating enzyme (Prockop, 1970).

The use of bacterial collagenase in the determination of the amino acid sequence of α1-CB8-HA1 provided an additional opportunity to investigate the substrate specificity of this enzyme. In these studies the enzymatic activity used undoubtedly represents a mixture of collagenases A and B (Harper and Kang, 1970) with the activity of the less specific enzyme, collagenase B, predominating. Table VIII summarizes both the sequences cleaved and those resistant to the enzyme. All cleavages resulted in an NH_2-terminal glycyl residue. Sequences containing an imino acid in the triplet on either or both sides of the potentially susceptible bond were readily cleaved. Resistant Y–Gly bonds were characterized by the absence of an imino acid or by the presence of an acidic amino acid on adjacent triplets. The absence of an imino acid on adjacent triplets did not confer complete resistance but no evidence has been found, in this or other published work, for the ability of a purified collagenase to cleave a collagen chain in close proximity to a negatively charged amino acid.

It was of interest to compare the band pattern of SLS aggregates of renatured α1-CB8 observed in the electron microscope (Hodge and Schmitt, 1960; Rauterberg and Kühn, 1968) with the amino acid sequence of the fragment. In Figure 6 the extent of α1-CB8 is delineated by the two horizontal arrows and the vertical arrow marks the position of cleavage by hydroxylamine, as determined by the relative number of amino acids in HA1 and HA2. The distribution of charged amino acids in HA1 correlates extremely well with the observed band pattern obtained by staining with phosphotungstic acid and uranyl acetate. This correspondence between SLS band pattern and amino acid sequence was sufficiently exact to permit the former to be used as an initial indication of the position of certain tryptic peptides in both HA1 and HA2. Although the electron micrograph in Figure 6 is of calf α1-CB8, no detectable differences in band patterns exist between rat and calf α1-CB8 (K. Kühn, personal communication). Clustering of acidic and basic amino acids is seen, with the regions noted as light bands occupied predominantly or exclusively by apolar amino acids. Similar obser-

vations have been made by von der Mark *et al.* (1970) in a sequence of 112 amino acids in α1-CB6 of calf collagen.

An additional objective of these studies has been to clarify the issue of the presence of unorthodox bonds and reactive groups in collagen (Gallop *et al.*, 1967). Previous evidence for the existence of amino aldehydes and carbonyl functions, other than derivatives of lysyl and hydroxylysyl side chains, has been attributed to a side reaction of borohydride reduction (Paz *et al.*, 1970). No support for the existence of such groups was found during the course of this work. Similarly, the existence of hydroxylamine-sensitive bonds, thought to indicate the presence of intrachain subunits, can be attributed to the ability of hydroxylamine to cleave cyclic imides formed as derivatives of asparaginyl–glycyl bonds (Bornstein, 1969b, 1970). Hydroxylamine has been shown to be capable of cleaving an asparaginyl–glycyl bond in another region of collagen (Butler, 1969) and in bovine pancreatic ribonuclease (Bornstein and Balian, 1970). Support for the suggestion that asparaginyl–glycyl bonds are cleaved in preference to aspartyl–glycyl bonds by hydroxylamine (Bornstein, 1970) comes from the finding that in HA1 aspartyl–glycyl bonds at residues 24–25 and 90–91 (Figure 5) are resistant to cleavage. The only other asparaginyl residue in HA1 (residue 86) precedes hydroxyproline and therefore cannot form a cyclic imide.

Acknowledgments

We are indebted to Dr. M. Hermodson for performance of the automatic sequencer and gas chromatographic determinations, to Dr. K. Kühn for permission to use the electron micrograph in Figure 6, to Dr. Koiti Titani for helpful advice and discussions, and to Felicia Arguelles for skillful technical assistance.

References

Balian, G., and Click, E. M. (1971), *Fed. Proc., Fed. Amer. Soc. Exp. Biol. 30*, 1196.

Bornstein, P. (1967), *Biochemistry 6*, 3082.

Bornstein, P. (1969a), *Biochemistry 8*, 63.

Bornstein, P. (1969b), *Biochem. Biophys. Res. Commun. 36*, 957.

Bornstein, P. (1970), *Biochemistry 9*, 2408.

Bornstein, P., and Balian, G. (1970), *J. Biol. Chem. 245*, 4854.

Bradshaw, R. A., Garner, W. H., and Gurd, F. R. N. (1969), *J. Biol. Chem. 244*, 2149.

Butler, P. J. G., Harris, J. I., Hartley, B. S., and Leberman, R. (1969), *Biochem. J. 112*, 679.

Butler, W. T. (1968), *Science 161*, 796.

Butler, W. T. (1969), *J. Biol. Chem. 244*, 3415.

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

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

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

Edman, P., and Begg, G. (1967), *Eur. J. Biochem. 1*, 80.

Gallop, P. M., Blumenfeld, O. O., and Seifter, S. (1967), *in* Treatise on Collagen, Vol. 1, Ramachandran, G. N., Ed., New York, N. Y., Academic Press, p 339.

Goldberger, R. F. (1967), *Methods Enzymol. 11*, 317.

Greenberg, J., Fishman, L., and Levy, M. (1964), *Biochemistry 3*, 1826.

Harper, E., and Kang, A. H. (1970), *Biochem. Biophys. Res. Commun. 41*, 482.

Highberger, J. H., Kang, A. H., and Gross, J. (1971), *Biochemistry 10*, 610.

Hodge, A. J., and Schmitt, F. O. (1960), *Proc. Nat. Acad. Sci. U. S. 46*, 186.

Kaug, A. H., Bornstein, P., and Piez, K. A. (1967), *Biochemistry 6*, 788.

Konigsberg, W. (1967), *Methods Enzymol. 11*, 461.

Miller, E. J., and Piez, K. A. (1966), *Anal. Biochem. 16*, 320.

Paz, M. A., Henson, E., Rombauer, R., Abrash, L., Blumenfeld, O. O., and Gallop, P. M. (1970), *Biochemistry 9*, 2123.

Peterkofsky, B., and Diegelmann, R. (1971), *Biochemistry 10*, 988.

Piez, K. A. (1968), *Anal. Biochem. 26*, 305.

Piez, K. A., Miller, E. J., Lane, J. M., and Butler, W. T. (1969), *Biochem. Biophys. Res. Commun. 37*, 801.

Pisano, J. J., and Bronzert, T. J. (1969), *J. Biol. Chem. 244*, 5597.

Prockop, D. J. (1970), *in* Chemistry and Molecular Biology of the Intercellular Matrix, Vol. 1, Balazs, E. A., Ed., New York, N. Y., Academic Press, p 339.

Rauterberg, J., and Kühn, K. (1968), *FEBS (Fed. Eur. Biochem. Soc.) Lett. 1*, 230.

Rhoads, R. E., Udenfriend, S., and Bornstein, P. (1971), *J. Biol. Chem. 246*, 4138.

Traub, W., Yonath, A., and Segal, D. M. (1969), *Nature (London) 221*, 914.

von der Mark, K., Wendt, P., Rexrodt, F., and Kühn, K. (1970), *FEBS (Fed. Eur. Biochem. Soc.) Lett. 11*, 105.