11. J. Kreyenberg, *Z. Angew. Entomol.* **14**, 140 (1929).
12. W. M. Hoskins and R. Craig, *Physiol. Rev.* **15**, 525 (1935).
13. L. Seamans and L. C. Woodruff, *Kansas Entomol. Soc.* **12**, 73 (1939).
14. D. B. Long, *Trans. Roy. Entomol. Soc.* **104**, 543 (1953).
15. Supported in part by NSF grant GB 3150.
* Present address: Department of Laser Biology, Pasadena Foundation for Medical Research, Pasadena, California.

7 May 1968

Collagen: Relatively Invariant (Helical) and Variable (Nonhelical) Regions

Abstract. *The structural identity of certain helical regions of collagen from human and rat skin equals or exceeds that of other homologous proteins. In contrast, the short nonhelical sequences in the two proteins, although homologous, differ appreciably in structure. The requirements of the collagen helix and the numerous intermolecular interactions characteristic of collagen may restrict the number of functionally acceptable amino acid replacements occurring during evolution.*

Studies of the structural changes which occur during evolution indicate that the number of amino acid replacements in homologous proteins roughly parallels interspecial phyletic distances, as estimated by morphological taxonomic criteria (*1*). The relatively invariant regions which exist are frequently involved in a specialized function of the protein, but intramolecular side-chain interactions associated with a specific tertiary structure also serve to select against amino acid substitutions. However, as suggested by the comparative biochemistry of globin chains (*2*), selective pressures at a functional level appear to be exerted toward preserving the overall conformation of a protein rather than a unique amino acid sequence.

In keeping with the structure proposed for collagen (*3*), the interior of the highly asymmetric cylindrical molecule is occupied by the helical backbone of the three polypeptide chains, whereas all amino acid side chains are directed exteriorly. The formation of the triple-stranded collagen helix requires that every third amino acid residue be glycine, and, in order for the structure to be stable in vivo, a minimum pyrrolidine (proline plus hydroxyproline) content is necessary (*4*). The side chains of the remaining 45 percent of the amino acids do not appear to participate significantly in intramolecular interactions. If there are no mutational "cold spots" in the DNA coding for collagen, amino acid residues which are not critical to the function of the protein may be subject to change by mutation, and the positions occupied by these amino acids would be expected to show interspecial variation. By extension of this reasoning, intraspecial polymorphism could result from the formation of allotypic collagen chains produced by numerous allelic genes.

To investigate this matter, the amino acid composition and partial sequence of portions of the $\alpha 1$ and $\alpha 2$ chains of human skin collagen (HSC) were determined in three subjects. There were no differences in composition in the isologous proteins. In addition, with the exception of a short sequence near the amino terminus the composition and sequence (insofar as determined) of HSC and rat skin collagen (RSC) were very similar.

Specimens of skin were obtained at autopsy from three infants who had died a respiratory death shortly after delivery. Pathologic examination revealed no abnormalities in connective tissues. Rat skin was obtained from male Sprague-Dawley rats weighing 100 to 150 g. Collagen was successively extracted from skin with 1*M* NaCl, 0.5*M* acetic acid, and 5*M* guanidine, and purified as described (*5*). Single-chain (α) and double-chain (β) components were separated by carboxymethyl (CM) cellulose chromatography (*5, 6*). Isolated $\alpha 1$ and β_{12} components were cleaved nonenzymatically at methionyl residues with cyanogen bromide (CNBr) (*5, 7*), and the resulting fragments were separated by chromatography on CM-cellulose at *p*H 4.8 (*7*) and on phosphocellulose (*5*). Individual CNBr-produced fragments were cleaved with trypsin or chymotrypsin, and the enzymatic digests were separated by column chromatography on Bio-Gel P-2 or phosphocellulose. Peptide maps of tryptic digests were obtained by two-dimensional chromatography and electrophoresis (*8*). Amino acid analyses were performed with a Beckman 120C analyzer modified for high-speed single-column gradient elution (*9*).

The CM-cellulose and phosphocellulose elution patterns of CNBr digests of the $\alpha 1$ chain of HSC and RSC were very similar. The position of elution of only one peptide, $\alpha 1$-CB1, was clearly different in digests of the two proteins. Three peptides, $\alpha 1$-CB1, $\alpha 1$-CB2, and $\alpha 1$-CB3 were further compared largely because of their relative ease of purification. The phosphocellulose elution patterns of CNBr digests of β_{12} (the $\alpha 1$-$\alpha 2$ dimer) from HSC and RSC differed only in the position of elution of the cross-linked peptide, β_{12}-CB1. Peptide $\alpha 2$-CB2 was purified and analyzed.

Homologous peptides in RSC and HSC were identified by similarities in elution from phosphocellulose and by amino acid composition. The amino acid compositions of $\alpha 1$-CB2 from HSC and RSC were the same with the exception of the degree of hydroxylation of proline (Table 1). The chromatographic properties and amino acid compositions of the chymotryptic fragments of the two peptides were also the same. The compositions of $\alpha 2$-CB2 (Table 1) and $\alpha 1$-CB3 (Table 2) were extremely similar in the two proteins.

Table 1. Amino acid compositions of some CNBr-produced fragments from rat and human skin collagen. Results are given as residues per peptide. Values are the average of three or more determinations. Lack of a number indicates that the amino acid was either entirely absent or present as less than 0.1 residue. Residues in parentheses are fractional residues thought to be impurities.

Amino acid	$\alpha 1$-CB1		$\alpha 1$-CB2		$\alpha 2$-CB2	
	HSC	RSC*	HSC	RSC*	HSC	RSC
4-Hydroxyproline			5.5	4.9	2.9	2.8
Aspartic acid	1.1	1.0			1.9	2.8
Threonine	1.0					1.0
Serine	2.9	2.0	1.8	2.2	1.8	1.1
Homoserine†	0.9	0.9	1.0	1.0	0.8	0.9
Glutamic acid	2.1	1.0	3.9	4.0	1.3	1.1
Proline	2.1	1.8	6.0	6.9	2.9	3.1
Glycine	4.2	3.2	11.4	12.1	9.7	9.4
Alanine	(0.2)	1.1	2.1	2.1	3.2	2.1
Valine	1.2	1.8			1.1	1.0
Isoleucine	1.1					
Leucine	1.0		1.0	1.0	1.1	1.0
Phenylalanine			1.0	0.9		
Tyrosine	1.8	1.2				
Lysine	0.9	1.0				
Arginine			1.1	1.0	2.8	3.0

* Data from Bornstein and Piez (*5*). † Includes homoserine lactone.

In addition to the difference in ratio of proline to hydroxyproline in α1-CB3 from the two species, differences occurred in aspartic acid (or asparagine), threonine, and alanine in α1-CB3, and aspartic acid (or asparagine), threonine, serine, and alanine in α2-CB2. Peptide maps of the tryptic peptides of RSC and HSC α2-CB2 and α1-CB3 were virtually identical, indicating very similar amino acid sequences. Since hydroxylation of proline occurs after peptide-bond synthesis and is under the control of a cellular enzyme and a number of cofactors (*10*), differences in the degree of hydroxylation of proline do not indicate a change in the genetic material coding for collagen. Incomplete hydroxylation of individual prolyl residues has been demonstrated in α1-CB2 from collagen of normal rat skin and tail tendon (*11*).

The similarity in chromatographic properties of the above-mentioned three CNBr-derived fragments of HSC and RSC and in their enzymatic fragments indicates that amino-acid substitutions were probably limited to those detected by differences in composition and that replacements involved asparagine rather than aspartic acid. On the basis of amino acid composition, approximately five substitutions occur in a total sequence of 215 amino acids. All five substitutions are very conservative or fairly conservative (*2*), and four of the five may involve changes of only one base.

In contrast, the composition of α1-CB1 differed considerably in HSC and RSC (Table 1). Partly because of its unusual amino acid composition (for collagen), the chemistry of RSC α1-CB1 has been extensively investigated (*5*). This peptide constitutes the amino terminal sequence of the α1 chain and participates in the formation of covalent cross-links between chains by transformation of its lysyl side chain to a reactive lysyl-derived aldehyde (*5, 13*). The homology of HSC α1-CB1 to RSC α1-CB1 was indicated by its similar and uniquely atypical amino acid composition and by the isolation of a related aldehyde-containing peptide which differed from HSC α1-CB1 in lacking only the single lysyl residue.

According to present concepts, the amino acid sequences of HSC and RSC α1-CB1 do not permit the formation of the triple helical conformation which is characteristic of collagen. The different structure assumed by the amino terminus of the molecule may in part lead to its participation in the formation of interchain covalent cross-links. This function is preserved despite the changes in amino acid composition and sequence which are observed in the case of α1-CB1 from RSC and HSC (*14*). In view of the presence of tyrosine in the nonhelical region (*5, 13, 15*), differences in antigenicity may result from this species-specific variation in structure.

In contrast, the helical body of the collagen molecule appears to be unusually constant in amino acid composition and sequence. This study reveals a structural identity in certain helical regions of RSC and HSC which may be as high as 97 percent (exclusive of differences in hydroxylation of proline and lysine). This identity exceeds that of homologous proteins such as human and mouse α and β hemoglobin chains (88 and 84 percent, respectively) and human and rabbit cytochrome c (92 percent) (*16*).

Table 2. Amino acid composition of α1-CB3 from rat and human skin collagen. Values are the average of three or more determinations. Lack of a number indicates that the amino acid was either entirely absent or present as less than 0.1 residue per peptide. Residues in parentheses are fractional residues thought to be impurities.

Amino acid	RSC α1-CB3		HSC α1-CB3	
	Res/1000	Res/Peptide	Res/1000	Res/Peptide
4-Hydroxyproline	102	15.2	95.7	14.3
Aspartic acid	47.0	7.0	42.1	6.3
Threonine	14.1	2.1	(1.3)	(0.2)
Serine	19.8	3.0	21.7	3.2
Homoserine*	6.1	0.9	6.0	0.9
Glutamic acid	108	16.1	107	15.9
Proline	92.4	13.8	96.5	14.4
Glycine	336	50.0	333	49.6
Alanine	135	20.1	150	22.4
Valine	28.0	4.2	29.5	4.4
Isoleucine				
Leucine	20.0	3.0	20.9	3.1
Tyrosine				
Phenylalanine	18.9	2.8	19.9	3.0
Hydroxylysine	0.8	0.1	1.6	0.2
Lysine	31.1	4.6	31.6	4.7
Histidine				
Arginine	39.9	5.9	42.2	6.3

* Includes homoserine lactone.

A rather striking similarity in the electron micrographs of segment-long-spacing collagen aggregates from calf skin and the mesoglea of the sea anemone has recently been demonstrated (*17*), indicating a relative constancy in composition. It seems likely that in addition to the requirements of the collagen helix, the numerous intermolecular interactions involving both the aggregation of collagen and its association with other connective tissue macromolecules, severely limit the number of amino-acid substitutions which the function of the protein can accommodate (*18*).

PAUL BORNSTEIN
Research Training Unit, Department of Medicine, and *Department of Biochemistry, University of Washington School of Medicine, Seattle*

References and Notes

1. E. Margoliash and A. Schejter, *Advan. Protein Chem.* **21**, 113 (1966); W. M. Fitch and E. Margoliash, *Science* **155**, 279 (1967).
2. E. Zuckerkandl and L. Pauling, in *Evolving Genes and Proteins*, V. Bryson and H. J. Vogel, Eds. (Academic Press, New York, 1965), p. 97.
3. G. N. Ramachandran, in *Treatise on Collagen*, G. N. Ramachandran, Ed. (Academic Press, New York, 1967), vol. 1, p. 103.
4. J. Josse and W. F. Harrington, *J. Mol. Biol.* **9**, 269 (1964).
5. P. Bornstein and K. A. Piez, *Biochemistry* **5**, 3460 (1966).
6. K. A. Piez, E. A. Eigner, M. S. Lewis, *ibid.* **2**, 58 (1963).
7. P. Bornstein and K. A. Piez, *Science* **148**, 1353 (1965).
8. W. T. Butler, K. A. Piez, P. Bornstein, *Biochemistry* **6**, 3771 (1967).
9. E. J. Miller and K. A. Piez, *Anal. Biochem.* **16**, 320 (1966).
10. S. Udenfriend, *Science* **152**, 1335 (1966); J. J. Hutton, Jr., A. L. Tappel, S. Udenfriend, *Arch. Biochem. Biophys.* **118**, 231 (1967); K. I. Kivirikko and D. J. Prockop, *Proc. Nat. Acad. Sci. U.S.* **57**, 782 (1967).
11. P. Bornstein, *J. Biol. Chem.* **242**, 2572 (1967); *Biochemistry* **6**, 3082 (1967). The presence of less than one residue of hydroxylysine in α1-CB3 suggests that incomplete hydroxylation of lysine also occurs. Incomplete hydroxylation may account for the consistent and pronounced trailing of collagen peptides observed during paper chromatography.
12. A. H. Kang, P. Bornstein, K. A. Piez, *Biochemistry* **6**, 788 (1967).
13. P. Bornstein, A. H. Kang, K. A. Piez, *Proc. Nat. Acad. Sci. U.S.* **55**, 417 (1966); *Biochemistry* **5**, 3803 (1966).
14. The length of the amino-acid sequence prior to the first methionyl residue differs in the α1 chains of HSC and RSC. A short segment of the α1 chain of RSC may be removed either in vivo or during purification of the protein; however, preliminary sequence studies of HSC α1-CB1 indicate that this could not account for all of the differences in composition observed in the amino terminal peptides.
15. M. P. Drake, P. F. Davison, S. Bump, F. O. Schmitt, *Biochemistry* **5**, 301 (1966); P. F. Davison, L. Levine, M. P. Drake, A. Rubin, S. Bump, *J. Exp. Med.* **126**, 331 (1967).
16. R. V. Eck and M. O. Dayhoff, *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Silver Spring, Maryland, 1966).
17. A. Nordwig and U. Hayduk, *J. Mol. Biol.* **26**, 351 (1967).
18. Supported in part by NIH grant AM-11248 and NSF grant GB-6079. The assistance of F. Arguelles is acknowledged.

6 May 1968