[4] Mahalanobis, P. C., *Proc. Nat. Inst. Sci. India*, **12**, 49 (1936).

[5] Bryan, J. G., *Harvard Educat. Rev.*, **21**, 90 (1951).

[6] Siegel, F. L., M. K. Roach, and L. R. Pomeroy, these PROCEEDINGS, **51**, 605 (1964).

[7] Cooley, W. W., and P. R. Lohnes, in *Multivariate Procedures for the Behavioral Sciences* (New York: John Wiley, 1962), pp. 124–132.

[8] Williams, R. J., and F. L. Siegel, *Am. J. Med.*, **31**, 325 (1961).

A SUBUNIT MODEL FOR THE TROPOCOLLAGEN MACROMOLECULE[*, †]

BY JOHN A. PETRUSKA AND ALAN J. HODGE

DIVISION OF BIOLOGY, CALIFORNIA INSTITUTE OF TECHNOLOGY

Communicated by M. Delbrück, March 26, 1964

In this article is presented a novel subunit model for the tropocollagen (TC) macromolecule, the monomeric unit of soluble collagen.[1] The TC macromolecule is a rigid rod of definite length composed of three helical polypeptide strands that are equal (or very nearly equal) in length.[2–4] The strands are of two types chemically, designated α_1 and α_2.[4] Each macromolecule contains two α_1 strands and one α_2 strand.[4] Our thesis is that each α_1 strand is a repeating sequence of a subunit σ_1, and that the α_2 strand is a repeating sequence of a *shorter* subunit σ_2. It is proposed specifically, after close examination of experimental data, that the subunits σ_1 and σ_2 have lengths in the ratio 7:5. This proposal leads to the "self-limiting" model for the TC macromolecule illustrated in Figure 1. In this model the strands terminate in register when each α_1 strand comprises *five* σ_1 subunits and the α_2 strand comprises *seven* σ_2 subunits. The over-all molecular length L is the "beat period" of the two subunit lengths,

$$L = 5l_1 = 7l_2 = 35l_0,$$

where l_1 and l_2 are the lengths of σ_1 and σ_2, respectively, along the axis of the macromolecule, and $l_0 = l_1/7 = l_2/5$ is their highest common denominator.

Physical-Chemical and Analytical Data.—The TC macromolecules extracted from vertebrate skin or tendon are seen by physical-chemical methods to be rigid rods of dimensions *ca.* 3000 × 15 Å and molecular weight *ca.* 300,000.[2] They exhibit a large negative optical rotation in aqueous solution approaching that of poly-L-proline.[5] Their optical rotation is consistent with the triple-helical structure for collagen indicated by X-ray diffraction studies of tendon.[6, 7] This is a structure in which the three constituent polypeptide strands, each containing 33 per cent glycine, are individually left-handed helices resembling poly-L-proline in its normal (poly-L-proline II) configuration in water. The three left-handed helices are held together by sequential hydrogen bonding made possible by their 33 per cent glycine content, and are wound around each other in the right-handed sense. Rich and Crick[6] have presented a model for the collagen triple helix based on the standard Pauling-Corey values for bond distances and bond angles. In this model one complete right-handed turn of the three strands around a common axis is made every 30 residues per strand.

Upon denaturation, the TC macromolecules show in the ultracentrifuge a fundamental α component of molecular weight *ca.* 100,000, plus varying amounts of β

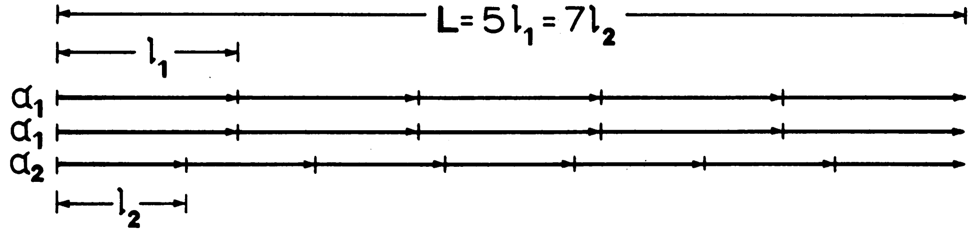

FIG. 1.—Schematic representation of the "self-limiting" 5:7 subunit model proposed here for the TC macromolecule. This model is based on two subunits whose lengths l_1 and l_2 are in the ratio 7:5. Each α_1 strand comprises five identical subunits of length l_1; the α_2 strand, seven identical subunits of length l_2. The over-all molecular length L is the "beat period" of the two subunit lengths, $L = 5l_1 = 7l_2$.

and γ components of molecular weight *ca.* 200,000 and 300,000.[4, 8] By column chromatography on carboxymethyl cellulose,[4] the α component is found to consist of two chemically distinct polypeptide strands of molecular weight *ca.* 100,000, designated α_1 and α_2. These have the same glycine content (33%) but significantly different contents of most of the other amino acids. The β component, on the other hand, is a mixture of α_1—α_1 and α_1—α_2, while the γ component appears to be α_1—α_1—α_2, where the dashes represent covalent nonpeptide *crosslinks* between the strands. These crosslinks are observed to form over a period of time *after* the macromolecules are assembled.[9] The molar ratio of α_1 to α_2 in the total pool is 2, and there is no appreciable amount of α_2—α_2 present.[4] These findings lead to the conclusion that each macromolecule contains two α_1 strands and one α_2 strand that initially are not crosslinked.

Besides the variable number of interstrand crosslinks formed with time, there always seem to be 10–20 *intra*strand nonpeptide links present in the TC macromolecule.[10] These are cleaved by hydroxylamine under conditions that leave the peptide bonds intact, and are tentatively identified as ester or imide bonds involving aspartic acid.[10] Their cleavage results in fragmentation of the strands. The resultant fragments have yet to be properly characterized. However, the hypothesis that they are produced by the breaking of bonds between true subunits in the strands looks promising.[10] The finding of fragments averaging in molecular weight around 25,000 by Gallop and co-workers[10] suggests there are at least four subunits per strand. Preliminary experiments in this laboratory indicate that on carboxymethyl cellulose columns the fragments are separable into two major components whose total weight ratio is approximately 2:1.[11] The molecular weights of the components are currently being investigated by sedimentation equilibrium.

From the compositional data published for the α_1 and α_2 strands[4] one can get an upper limit for the number of possible *identical* subunits in these strands of 1,000 or so amino acids. If there are n identical subunits, then the numbers of each kind of residue should approach integral multiples of n. The smallest residue number (apart from 0) should be n. For the α_1 strand, the compositional data[4] indicate there may be as many as five subunits per strand of 1,000 residues; for the α_2 strand, as many as eight subunits.

Electron Microscopic Data.—The ability of the TC macromolecules to form *segment-long-spacing* (SLS) paracrystallites allows unusually precise electron microscopic observations of molecular details.[3, 12, 13] In the SLS crystallites (readily formed by adding adenosine triphosphate to acidic TC solutions) the macromolecules lie side by side with their equivalent features aligned transversely in register.[12] Molecular features appear in bands normal to the axis of the macromolecule. By "positive staining" with phosphotungstic acid (PTA), one can locate the bands containing the basic amino acids (principally arginine).[12] By allowing excess PTA to remain on the grid to give a "negative contrast" outline of the SLS crystallites, one can also locate the macromolecular ends.[3, 13]

Some 50 distinct bands are observed in the SLS crystallites of the calf skin TC macromolecule on "positive staining" with PTA. How these are related to the bands in the *native-type* collagen fibril has been established both by growing SLS crystallites on the fibrils and by "optical synthesis" of the native-type band pattern from the SLS band pattern.[12] It is seen that the native-type fibril is a staggered array of parallel TC macromolecules, in which there is an axial displacement of macromolecules relative to nearest neighbors by a distance D equal to the observed (690 Å) axial repeat period. Each band in the native-type fibril corresponds to a summation of certain SLS bands spaced D apart. Thus the SLS bands can be indexed in terms of sets of bands, each having a spacing of D, with each set corresponding to a particular band in the native-type fibril.[12]

The apparent repeat at intervals of D, however, is not a true subunit repeat within the TC macromolecule. It applies (approximately) to the positions of the bands, but not to their intensities.[12] In a given set of bands spaced D apart, the members usually have quite different staining intensities, i.e., they represent loci that are chemically different.

Furthermore, after locating the ends of the macromolecule by the "negative contrast" technique, it is seen that the over-all length L of the TC macromolecule is a *nonintegral* multiple of D.[3, 13, 14] For the calfskin TC macromolecule in particular, upon measuring the length in terms of the distance between bands known to be spaced at integral multiples of D, it is found that $L = 4.40 \pm 0.02\ D$.[3, 13]

The coordinates and intensities of the bands in a number of PTA-stained SLS crystallites of the calfskin TC macromolecule have been measured in this laboratory. These data enable us to explore what kinds of subunit models are possible for the TC macromolecule. Since there are two kinds of strands in the TC macromolecule (α_1 and α_2), it is necessary to have at least two kinds of subunits. Our data indicate that, provided the strands do indeed have subunits (i.e., more than one apiece), a model based on two types of subunits of *equal* length can be ruled out. We cannot get such a model to comply with the data without assigning more than 10 subunits to each strand and without invoking complicated subunit arrangements that make all three strands different. On the other hand, by allowing the two subunit types to be of *unequal* length, we obtain a satisfactory fit upon assigning *five* identical subunits of length $L/5$ to each α_1 strand and *seven* identical subunits of length $L/7$ to the α_2 strand. The resultant 5:7 subunit model illustrated in Figure 1, when fitted to one of our best sets of data, gives the results shown in Figure 2. Other models of this nature based on two unequal-length subunits, with subunit lengths in the ratio of two whole numbers up to 8 (e.g., 4:7, 5:8), have also been examined. None of these fit the data as well as the 5:7 subunit model.

Properties of the 5:7 Subunit Model.—In the 5:7 subunit model proposed here for the TC macromolecule (Fig. 1), the subunit lengths along the triple helical axis are

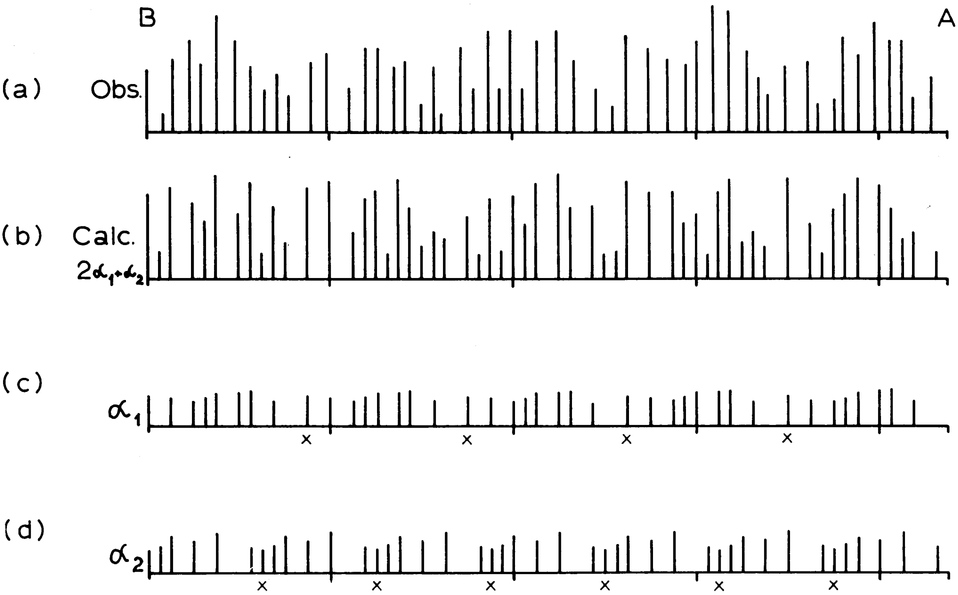

FIG. 2.—The results obtained upon fitting the 5:7 subunit model to the band pattern of PTA-stained calfskin SLS. (*a*) The observed SLS pattern mapped on a horizontal line that represents the over-all molecular length of 4.375 native-type periods (left to right, *B* end to *A* end). The vertical lines indicate the peak centers and relative peak heights (staining intensities) found from a densitometer trace across one particularly well-ordered SLS crystallite of the calfskin TC macromolecule. (*b*) The calculated pattern obtained after analyzing the observed data in terms of the 5:7 subunit model. (*c*) The contribution assigned to *each* of the two α_1 strands. (*d*) The contribution assigned to the single α_2 strand. The α_1 and α_2 strands are given five and seven subunits, respectively. The links between the subunits in each case (*c* and *d*) are indicated by X's.

$l_1 = 7l_0$ and $l_2 = 5l_0$, where $l_0 = L/35$. The observed molecular length L in terms of the native-type period D is $L = 4.40 \pm 0.02\ D$.[3, 13] Substituting $L = 35l_0$, we find $D = 7.95 \pm 0.04\ l_0$ or $8l_0$ within experimental error. Accordingly, the length parameters in the model can be expressed as integral multiples of $D/8$:

$$\begin{aligned} l_0 &= D/8 = 0.125\ D \\ l_1 &= 7\ D/8 = 0.875\ D \\ l_2 &= 5\ D/8 = 0.625\ D \\ L &= 35\ D/8 = 4.375\ D \end{aligned}$$

The value of D indicated by low-angle X-ray diffraction studies of wet tendon is 690 ± 10 Å.[15] This yields

$$\begin{aligned} l_0 &= 86 \pm 1\ \text{Å} \\ l_1 &= 604 \pm 9\ \text{Å} \\ l_2 &= 431 \pm 6\ \text{Å} \\ L &= 3020 \pm 45\ \text{Å} \end{aligned}$$

The corresponding numbers of amino acid residues can be obtained by dividing by 2.90 ± 0.05 Å, the residue-residue distance along the triple helical axis indicated by wide-angle X-ray diffraction studies of wet tendon.[6, 7]

We then find the following: (*a*) The element of subunit length, l_0, spans 30 ± 1 residues per strand. (*b*) The subunit σ_1 (in each of the two α_1 strands) contains

210 ± 7 residues. (*c*) The subunit σ_2 (in the α_2 strand) contains 150 ± 5 residues. (*d*) Each of the three strands in the TC macromolecule contains 1050 ± 35 residues.

The value of 30 ± 1 residues per strand obtained for l_0 is seen to equal one complete right-handed turn of the three strands about a common axis in the Rich-Crick model[6] for the collagen triple helix. Thus our subunit model predicts that in the fully coiled TC macromolecule all the length parameters match *integral* multiples of the major, right-handed turn of the triple helix. It predicts that the subunit lengths l_1 and l_2 match seven and five turns, respectively, the native-type period D matches eight turns, and the over-all molecular length L matches 35 turns.

Two other important parameters of the native-type fibril, besides D, also are predicted to match integral multiples of the major turn of the triple helix. One of these is the "gap" or "hole" between nearest *nonoverlapping* macromolecules in the fibril axis.[3] Since $L = 4.375\ D$, it follows that the nearest macromolecules lying in a straight line must be displaced axially relative to one another by 5 D and be separated by a gap of $5\ D - L = 0.625\ D$. This gap of 0.625 D equals $5l_0$ or five major turns of the triple helix. Interestingly, this is exactly the length of the *shorter* subunit, σ_2. The other parameter of the native-type fibril is the *end overlap* between nearest macromolecules displaced axially relative to one another by 4 D.[3, 13] It is this overlap that is responsible for the *end junction* in both native-type fibrils and the ordered fibrous aggregates of period 4 D known as F-SLS.[3, 13] Its value is $L - 4\ D = 0.375\ D = 3l_0$ or three major turns of the triple helix.

Concluding Remarks.—A more detailed discussion of the structural implications of the 5:7 subunit model, and further experimental details, will be presented in forthcoming publications. Before closing, however, we wish to mention two other bits of supporting information that will be discussed later. We have found that the compositional data for the α_1 strand published for rat skin and tendon and carp swim bladder,[4] when normalized to 1,050 total residues, do indeed correlate best statistically with multiples of 5. Also, we have found that the compositions of the homogeneous crosslinked triple-peptides isolated from tryptic digests of denatured calfskin collagen[16] yield favorable results when analyzed mathematically on the assumption that each triple-peptide consists of two of a particular σ_1 peptide and one σ_2 peptide. They lead to 10 distinct (hypothetical) tryptic peptides for σ_1 and five for σ_2, with residue totals close to 210 and 150, respectively, and over-all compositions like those expected for σ_1 and σ_2 from known α_1 and α_2 compositions.

It is obvious that a model of the type proposed here has inherent biological advantages. It provides a simple method of forming a large macromolecular species from a relatively small amount of genetic information. In the case of the TC macromolecule, only the polypeptide subunits of 210 and 150 amino acid residues need to be genetically determined. The model suggests the macromolecule is formed by progressive addition of subunits to form a three-stranded structure until the three strands terminate in register, i.e., reach the "beat period" of the two subunit lengths, and no further overlap is possible. Such a mechanism of formation could ensure production of a population of macromolecules homogeneous with respect to length and composition without requiring complicated ribosomal machinery. It is reasonable to expect that other multistranded fibrous protein macromolecules, for example, α-proteins like myosin and paramyosin, might also be formed in a similar manner, i.e., from two or more *unequal-length* subunits, by progressive subunit addition until the "beat period" is reached.

Summary.—A subunit model is presented for the triple-stranded tropocollagen macromolecule, based on two kinds of polypeptide subunits having lengths in the ratio 7:5. The three constituent strands of molecular weight *ca.* 100,000 are postulated to terminate in register at a length equal to the "beat period" of the two subunit lengths. Each of the two α_1 strands is described as a repeating sequence of five identical subunits σ_1; the single α_2 strand, as a repeating sequence of seven identical subunits σ_2. The assignment of five and seven subunits to the two kinds of strands is consistent with various experimental data but rests primarily on an analysis of the band pattern in SLS crystallites of the calfskin TC macromolecule observed in the EM after phosphotungstic-acid staining. It is found that both the intensities and positions of the bands can be fitted satisfactorily to a 5:7 subunit model of the type proposed. When normalized to the observed length of the TC macromolecule, the model indicates that the subunit lengths l_1 and l_2, the period D in the native-type fibril, and the over-all molecular length L are all *integral* multiples of a quantity $l_0 = 86 \pm 1$ Å spanning 30 ± 1 residues per strand. It yields $l_1 = 7l_0$, $l_2 = 5l_0$, $D = 8l_0$, $L = 35l_0$. The quantity l_0 equals one complete, right-handed turn of the three strands around a common axis in the Rich-Crick model for the collagen triple helix.

* Supported by a grant GM-06965 of the National Institutes of Health, U.S. Public Health Service.

† The substance of this article was presented at the Biophysical Society Meeting held in Chicago, Ill., Feb. 26–28, 1964.

[1] Gross, J., J. H. Highberger, and F. O. Schmitt, these PROCEEDINGS, **40**, 679 (1954).

[2] A comprehensive review of physical-chemical data for the TC macromolecule up to 1961 is given by W. F. Harrington and P. H. von Hippel, *Advan. Protein Chem.*, **16**, 1 (1961). Early molecular weight measurements, e.g., by H. Boedtker and P. Doty, *J. Am. Chem. Soc.*, **78**, 4267 (1956), indicated a value around 350,000. However, recent measurements, e.g., by K. A. Piez, E. A. Eigner, and M. S. Lewis, *Biochemistry*, **2**, 58 (1963), and by R. V. Rice, E. K. Casassa, R. E. Kerwin, and M. D. Maser, *Arch. Biochem. Biophys.* (in press), indicate the molecular weight of the TC macromolecule is close to 300,000.

[3] Hodge, A. J., and J. A. Petruska, in *Aspects of Protein Structure*, ed. G. N. Ramachandran (New York: Academic Press, 1963), p. 289.

[4] Piez, K. A., E. A. Eigner, and M. S. Lewis, *Biochemistry*, **2**, 58 (1963); Piez, K. A., M. S. Lewis, G. Martin, and J. Gross, *Biochim. Biophys. Acta*, **53**, 596 (1961).

[5] Harrington, W. F., and M. Sela, *Biochim. Biophys. Acta*, **27**, 24 (1958).

[6] Crick, F. H. C., and A. Rich, *Nature*, **176**, 780 (1955); Rich, A., and F. H. C. Crick, *J. Mol. Biol.*, **3**, 483 (1961).

[7] Ramachandran, G. N., and G. Kartha, *Nature*, **174**, 269 (1954); Ramachandran, G. N., in *Aspects of Protein Structure*, ed. G. N. Ramachandran (New York: Academic Press, 1963), p. 39.

[8] Altgelt, K., A. J. Hodge, and F. O. Schmitt, these PROCEEDINGS, **47**, 1914 (1961).

[9] Martin, G. R., K. A. Piez, and M. S. Lewis, *Biochim. Biophys. Acta*, **69**, 472 (1963).

[10] Gallop, P. M., S. Seifter, and E. Meilman, *Nature*, **183**, 1659 (1959); Gallop, P. M., *Biophys. J.*, **4**, 79 (1964).

[11] Bailey, A. J., and A. J. Hodge, unpublished data.

[12] Hodge, A. J., and F. O. Schmitt, these PROCEEDINGS, **46**, 186 (1960).

[13] Hodge, A. J., and J. A. Petruska, in *Electron Microscopy*, Fifth International Congress for Electron Microscopy, Philadelphia, Aug. 29–Sept. 5, 1962, ed. S. S. Breese, Jr. (New York: Academic Press, 1962), vol. 1, paper QQ-1.

[14] Olsen, B. R., *Z. Zellforsch. Mikroskop. Anat.*, **59**, 184, 199 (1963).

[15] Bear, R. S., *Advan. Protein Chem.*, **7**, 69 (1952).

[16] Grassmann, W., K. Hannig, and M. Schleyer, *Z. Physiol. Chem.*, **322**, 71 (1960).