N. M. Zadymova, Z. N. Markina, N. K. Evseeva, E. D. Shchukin, *Kolloidn. Zh.* **50**, 825 (1988); D. L. Dorset, *Macromolecules* **23**, 894 (1990); J. Kloubek, *Colloids Surf.* **55**, 191 (1991); D. W. Zhu, *Synthesis* **1993**, 953 (1993).
9. J. Falbe, *New Syntheses with Carbon Monoxide* (Springer-Verlag, Berlin, 1980).
10. J. D. Jamerson, R. L. Pruett, E. Billig, R. A. Fiato, *J. Organomet. Chem.* **193**, C43 (1980).
11. E. G. Kuntz, *Chemtech* **17**, 570 (1987); E. Wiebus and B. Cornils, *Chem. Ing. Tech.* **66**, 916 (1994).
12. I. T. Horváth, *Catal. Lett.* **6**, 43 (1990).
13. Separation of an *oxo* FBS system consisting of 3 ml of a mixture of 1-decene (20 percent) and 1-undecanal (80 percent) and 3 ml $C_6F_{11}CF_3$ at 40°C: upper phase, 2.7 percent $C_6F_{11}CF_3$, 19.5 percent 1-decene, and 77.8 percent 1-undecanal; lower phase, 99.3 percent $C_6F_{11}CF_3$, 0.7 percent 1-decene, and trace amounts of 1-undecanal.
14. C. Tamborski, C. E. Snyder Jr., J. B. Christian, U.S. Patent 4,454,349 (1984); H. Gopal, C. E. Snyder Jr., C. Tamborski, *J. Fluorine Chem.* **14**, 511 (1979), and references therein.
15. S. Benefice-Malouet, H. Blancou, A. Commeyras, *J. Fluorine Chem.* **30**, 171 (1985).
16. A 100-ml glass-lined autoclave was charged under N_2 with 35 g (100 mmol) 1*H*,1*H*,2*H*-perfluoro-1-octene, 0.6 g azobis(isobutyronitrile), and 0.85 g (25 mmol) PH_3 at room temperature. The mixture was stirred and heated to 100°C and kept at that temperature for 2 hours. After the reactor was cooled to room temperature, the unreacted PH_3 was vented to a scrubber containing 37 percent aqueous formaldehyde solution and 0.05 percent $RhCl_3$. Analysis by GC and ^{31}P NMR (in CF_2ClCCl_2F) showed the formation of $H_2PCH_2CH_2(CF_2)_5CF_3$ (2 percent, −139.3 ppm, t, coupling constant J_{P-H} = 189 Hz), $HP[CH_2CH_2(CF_2)_5CF_3]_2$ (4 percent, −67.1 ppm, d, J_{P-H} = 194 Hz), and $P[CH_2CH_2(CF_2)_5CF_3]_3$ (20 percent, −24.9 ppm). Addition of azobis(isobutyronitrile) (0.25 g) and heating the solution at 80°C for 8 hours resulted in the disappearance of the mono- and dialkylphosphines. The reaction mixture was diluted with 25 ml of C_6F_{14} and washed with toluene (four times with 15 ml). Distillation under vacuum (155°C at 0.3 torr) yielded 26 percent tris(1*H*,1*H*,2*H*,2*H*-perfluorooctyl)phosphine.
17. A mixture of 0.05 mmol $Rh(CO)_2(CH_3COCHCOCH_3)$ in 35 ml of toluene and 2.00 mmol $P[CH_2CH_2(CF_2)_5CF_3]_3$ in 35 ml of $C_6F_{11}CF_3$ was charged to a 300-ml autoclave under 75 psi (5 atm) CO/H_2(1:1) and heated to 100°C. A 75-ml pressure bomb was charged with 158 mmol 1-decene and attached to the autoclave. When the temperature in the autoclave reached 100°C, the 1-decene was added by using 150 psi (10 atm) CO/H_2(1:1) pressure, which was maintained during the reaction. After the reaction was complete, the reactor was cooled to room temperature. The autoclave was depressurized, and the two-phase system was separated in a separatory funnel under N_2. The ^{31}P NMR and GC analysis of the spent fluorous phase revealed that the phosphine ligand remained unchanged during the reaction. The upper phase was recharged to the cleaned and catalytically inactive autoclave. A solution of 30 ml of 1-octene in 35 ml of toluene was added under 75 psi (5 bar) CO/H_2(1:1) and heated to 100°C. The pressure was increased to 150 psi (10 bar) CO/H_2(1:1) and maintained for 24 hours. A GC analysis of the reaction mixture showed only trace amounts of conversion of 1-octene. In contrast, when the lower phase was charged to the autoclave, the hydroformylation of 1-octene proceeded to give 85 percent nonanals with *n/i* ratio of 2.9 and 8 percent octenes.
18. J. F. Liebman, A. Greenberg, W. R. Dolbier Jr., *Fluorine-Containing Molecules. Structure, Reactivity, Synthesis, and Applications* (VCH Press, New York, 1988).
19. M. Hudlicky, *Chemistry of Organic Fluorine Compounds* (Macmillan, New York, 1962).
20. Minnesota Mining and Manufacturing Co., British Patent 840,725 (1960).
21. Upper phase, 12.5 percent C_7F_{14}, 51.3 percent *n*-hexane, 36.2 percent toluene; lower phase, 58.7 percent C_7F_{14}, 27.2 percent *n*-hexane, 14.1 percent toluene.
22. A heavy-wall Pyrex tube containing a mixture of 0.5 mmol of phthalocyaninato cobalt(II) and 5 mmol of perfluorodecyl iodide under argon was heated in a heat bath at 250°C for 12 hours, and the temperature was increased to 290°C in 2 hours. After the tube was cooled to room temperature, the crude reaction product was extracted with 40 ml of perfluorohexane. The solvent was removed in vacuo at room temperature, and all of the volatile side products were removed by high vacuum at 100°C.
23. J. March, *Advanced Organic Chemistry* (Wiley, New York, 1992), pp. 956–963.
24. C. Reichardt, *Solvents and Solvent Effects in Organic Chemistry* (VCH Press, Weinheim, 1990).
25. R. S. Dickson, *Hydrogen and Carbon Compounds of Rhodium and Iridium* (Reidel, Amsterdam, 1985).
26. A solution of 0.3 mmol $P[CH_2CH_2(CF_2)_5CF_3]_3$ in 35 ml of $C_6F_{11}CF_3$ was mixed with a light yellow solution of 12.9 mg (0.05 mmol) $Rh(CO)_2$(acac) in 35 ml of toluene under argon. The resulting two-phase system contained a colorless upper phase and a slightly yellow lower phase, indicating the transfer of the rhodium from the toluene phase to the fluorous phase.
27. NMR data for $HRh(CO)\{P[CH_2CH_2(CF_2)_5CF_3]_3\}_3$ in $C_6F_{11}CF_3$: ^1H NMR: δ = −11.82 ppm (J_{P-H} = 19 Hz and J_{H-C} = 35 Hz), ^{31}P NMR: δ = 21.2 ppm (J_{P-Rh} = 148 Hz and J_{P-C} = 10 Hz).
28. The careful technical work of R. A. Cook and K. A. Eriksen is gratefully acknowledged. We are indebted to R. L. Espino, P. J. Guzi, A. Kaldor, M. G. Matturro, S. C. Mraw, P. S. Stevens, and W. Weissman for their support and encouragement.

21 June 1994; accepted 18 August 1994

Crystal and Molecular Structure of a Collagen-Like Peptide at 1.9 Å Resolution

Jordi Bella, Mark Eaton, Barbara Brodsky, Helen M. Berman*

The structure of a protein triple helix has been determined at 1.9 angstrom resolution by x-ray crystallographic studies of a collagen-like peptide containing a single substitution of the consensus sequence. This peptide adopts a triple-helical structure that confirms the basic features determined from fiber diffraction studies on collagen: supercoiling of polyproline II helices and interchain hydrogen bonding that follows the model II of Rich and Crick. In addition, the structure provides new information concerning the nature of this protein fold. Each triple helix is surrounded by a cylinder of hydration, with an extensive hydrogen bonding network between water molecules and peptide acceptor groups. Hydroxyproline residues have a critical role in this water network. The interaxial spacing of triple helices in the crystal is similar to that in collagen fibrils, and the water networks linking adjacent triple helices in the crystal structure are likely to be present in connective tissues. The breaking of the repeating $(X\text{-}Y\text{-}Gly)_n$ pattern by a Gly→Ala substitution results in a subtle alteration of the conformation, with a local untwisting of the triple helix. At the substitution site, direct interchain hydrogen bonds are replaced with interstitial water bridges between the peptide groups. Similar conformational changes may occur in Gly→X mutated collagens responsible for the diseases osteogenesis imperfecta, chondrodysplasias, and Ehlers-Danlos syndrome IV.

Until now, the triple helix was the only major regular protein motif that had not been elucidated by single crystal x-ray diffraction. The triple helix is characteristic of collagen proteins, and it also appears as a structural element in some proteins with host defense functions, such as the macrophage scavenger receptor (*1*) and C1q (*2*). The first models for the molecular conformation of collagen were proposed in the mid-1950s (*3*–*6*) on the basis of the unusual amino acid features of collagen and the high angle x-ray fiber diffraction pattern of tendon. The currently accepted model (*5*, *7*) consists of three polypeptide chains, each in an extended, left-handed polyproline II–like helix, which are staggered by one residue and then supercoiled about a common axis in a right-handed manner. Interchain hydrogen bonds between C=O and N–H groups stabilize the structure.

Triple helix sequence constraints are strict. Close-packing of the chains near the central axis imposes the requirement that glycine occupy every third position, generating an $(X\text{-}Y\text{-}Gly)_n$ repeating sequence. Proline and 4-hydroxyproline, which in collagens constitute about 20 percent of all residues, are found almost exclusively in the

J. Bella, M. Eaton, and H. M. Berman are in the Department of Chemistry and the Waksman Institute, Rutgers University, New Brunswick, NJ 08855, USA. B. Brodsky is in the Department of Biochemistry, UMDNJ-Robert Wood Johnson Medical School, Piscataway, NJ 08854, USA.

*To whom correspondence should be addressed.

X and Y positions, respectively. Steric constraints derived from their imino acid rings favor the polyproline II chain conformation that assembles into the triple helix. 4-Hydroxyproline (*8*) is the result of the post-translational enzymatic modification of proline in the Y position by prolyl hydroxylase. The unusual Hyp residue is always present in triple-helical domains of animal proteins although it is rarely found in other proteins. This residue provides the triple helix with greater stability than proline does, and several explanations have been offered for this effect (*7*, *9*, *10*).

Peptides that contain Gly as every third residue and have large amounts of Pro and Hyp behave as triple helices in solution and provide models for the study of more complex biological molecules (*11*). The most common triplet in collagen is Pro-Hyp-Gly, which accounts for about 10 percent of the total sequence. The peptide $(Pro\text{-}Hyp\text{-}Gly)_{10}$ forms a very stable triple helix (*9*) and models the regions of collagen with the highest content of imino acids. A peptide has been designed to model the effect of interrupting the repeating $(X\text{-}Y\text{-}Gly)_n$ pattern with a single Gly substitution (*12*, *13*). This peptide, Gly→Ala, carries one Gly to Ala substitution in the center of the 30–amino acid peptide $(Pro\text{-}Hyp\text{-}Gly)_{10}$ (*14*). At low temperatures Gly→Ala forms trimers, but shows a reduced thermal stability and a small decrease in its triple helical content when compared with the unsubstituted peptide (*12*, *13*). The Gly→Ala peptide represents a simple model for a particular collagen mutation type characterized by single Gly→X substitutions. These substitutions have been identified in several connective tissue diseases such as osteogenesis imperfecta, Ehlers-Danlos syndrome IV (*15*), and chondrodysplasia (*16*).

As illustrated by the crystal structure of the GCN4 leucine zipper (*17*), crystallographic analysis of long peptides is invaluable for characterizing some protein folds such as the α-helical coiled coil. We now describe the crystallization of the triple-helical Gly→Ala peptide, and the determination of its molecular structure at 1.9 Å resolution. Previously, single crystals were reported for one triple-helical peptide, $(Pro\text{-}Pro\text{-}Gly)_{10}$ (*18*). These crystals revealed a fiber-like structure in which the peptide molecules aggregate end-to-end, generating infinite polymers (*19*). Because of the inherent fiberlike disorder, only an average model could be obtained (*19*). In contrast, Gly→Ala crystals show no fiber-like disorder. The collagen-like triple helix that we describe allows visualization of the conformation of its individual residues and permits identification of specific interactions with bound water molecules.

Experimental data and structure determination. The synthesis and characterization of the Gly→Ala peptide have been described (*12*, *13*). A 2.0-mg sample was used for all of our crystallographic experiments. Thermal analysis experiments had shown that, at 10°C and a concentration of 2 mg/ml in 0.1 M acetic acid, this peptide is 99 percent associated as a trimer (*12*, *13*). Needle-like crystals were grown at 4°C by the hanging drop vapor diffusion method, with PEG 400 as the precipitating agent. The best crystals originated from drops containing initial concentrations of peptide from 4.0 to 6.0 mg/ml in 10 percent acetic acid, 9.5 percent PEG 400, and 0.1 percent sodium azide, equilibrated against a reservoir containing 19 percent PEG 400. Two fragments cut from a single needle were used for diffraction experiments.

Cell dimensions were measured initially by precession photography at −8°C and then more accurately on an Enraf Nonius CAD4 diffractometer. The Gly→Ala peptide crystallizes in the space group C2 with a = 173.5 Å, b = 14.06 Å, c = 25.31 Å, and β = 95.82°, with one triple helix per asymmetric unit. These unit cell parameters are consistent with the dimensions of a triple-helical structure: the longest axis is twice the expected length of a polypeptide chain of 30 amino acid residues in a triple-helical conformation, while the shortest dimension is comparable to the intermolecular distances determined from equatorial spacings in fiber diffraction patterns from collagen. In addition, two strong reflections (61,1,−1) and (62,0,−1), show Bragg spacings of 2.8 Å, in agreement with the shortest axial repeat of collagen triple helices appearing in the meridian of fiber diffraction patterns.

Data were collected at −10°C on the

Table 1. Final refinement statistics for the Gly→Ala peptide.

Resolution range	8.0–1.85 Å
Number of reflections used ($F > 3\sigma$)	3393
R factor*	19.7%
Root mean square deviation	
Δbonds	0.016 Å
Δangles	2.3°
Δchiral†	1.7°

*$R = \Sigma|F_o - F_c|/F_o$. †Chirality and planarity constraints are applied via improper torsion angles.

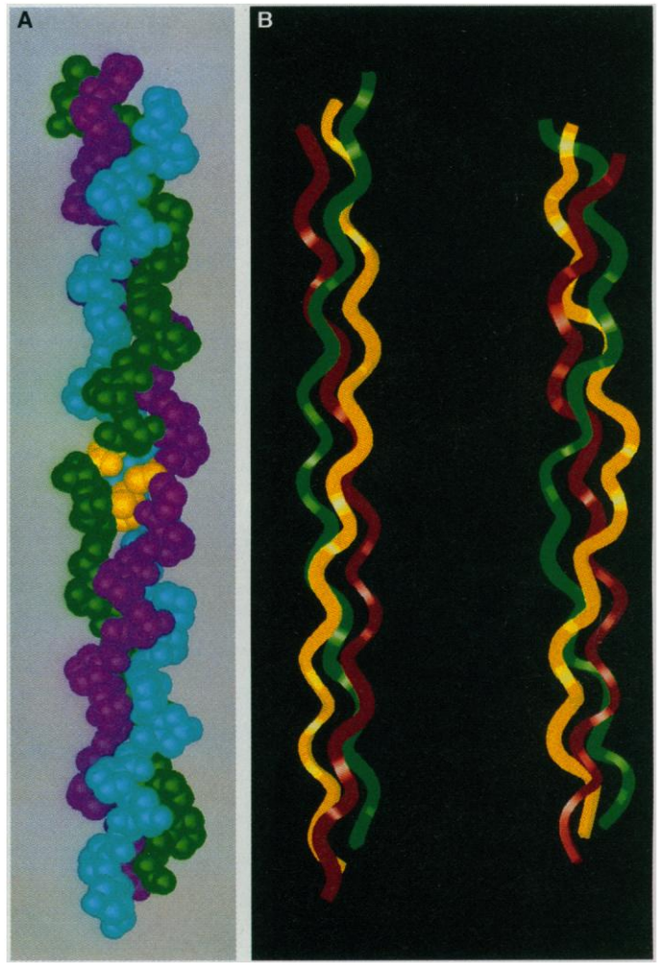

Fig. 1. (**A**) Space filling model of the Gly→Ala peptide in the crystal structure. The three alanine residues (yellow) are packed inside the triple helix. The three polypeptide chains are parallel and staggered by one residue. All residues are exposed to the solvent. (**B**) Comparison of ribbon diagrams (amino-terminal, top) of the native collagen, 10_7 helix (left) (*7*), and the Gly→Ala peptide (right; our data). The helical twist in the $(Pro\text{-}Hyp\text{-}Gly)_n$ regions of Gly→Ala is higher than that in the native collagen model. At the glycine substitution site, however, the triple helix exhibits a local unscrewing or twist relaxation. Both triple helices have essentially the same unit height, somewhat larger in the fiber model, determined from stretched collagen. The smaller length of the Gly→Ala peptide arises from the absence of the residues Gly^{30} and Gly^{90}, as well as from partial unraveling of the triple helical structure in the two terminal zones.

diffractometer with $\omega/2\theta$ scans. As the crystal did not decay appreciably, it was possible to collect weak reflections at a lower speed to improve the statistics. Intensity measurements were corrected for Lorentz-polarization and absorption, with the software package MOLEN (*20*). A total of 5595 unique reflections were measured (100 percent coverage) to a resolution limit of 1.85 Å. Beyond 1.9 Å, however, less than 50 percent of the data were observed with $I > \sigma$. Therefore this has been considered the effective resolution.

The structure was determined by molecular replacement methods with a fragment of triple-helical collagen as a probe molecule, and the symmetry and conformational angles of the fiber diffraction model from kangaroo tail tendon were used (*7*). A first fragment consisting of 27 residues, $3[(POG)_3]$, was positioned with the use of conventional rotation and translation searches in Patterson space. A search for a second fragment did not give an unambiguous solution because of the inherent symmetry of the model, and therefore a new approach was devised. The eight best solutions for this second fragment were each assigned one-eighth occupancy and used to calculate an electron density map. From this map it was possible to clarify the position of the second fragment as well as to see most of the rest of the molecule (*21*). Two software packages, MERLOT (*22*) and X-PLOR (*23*) were used in all the calculations, and the starting helical probe was built with standard bond distances and angles with the LALS program (*24*).

The first model contained 69 of 90 residues, with tripeptide units missing at both ends of the triple helix. This model was subjected to several rounds of refinement with the simulated annealing procedure (*25*) implemented in X-PLOR, manual rebuilding of the missing parts with the molecular graphics program FRODO (*26*), and least-squares refinement with X-PLOR (*27*). Our model includes 88 amino acids, 85 water molecules, and six acetic acid molecules, shows reasonable agreement with the x-ray data and stereochemical restraints (Table 1), and fits the $2F_o$-F_c map well.

Hydroxyproline is not a standard amino acid, and therefore it is absent from most of the conventional molecular modeling packages including X-PLOR. A set of parameters was derived from a statistical survey on the Cambridge Structural Database (*28*) for high resolution crystal structures of compounds containing chemical fragments matching this imino acid.

Overall structure. The Gly→Ala peptide is a rod-shaped molecule, 87 Å long and about 10 Å in diameter, with no appreciable bend (Fig. 1). The three peptide chains form a triple helix with a slight distortion at the site of the alanine substitution. A striking feature is that all residues are exposed to the solvent, either by their imino acid rings or by their carbonyl (C=O) groups. This important difference between collagen triple helices and the more common globular proteins points toward a critical participation of solvent in the triple-helical structure.

Although the Gly→Ala peptide carries a disruptive mutation in the substitution of the central Gly by Ala in every chain, the molecule still behaves as a trimer in solution (*12*, *13*). A modeling study suggested two possible conformations to overcome the steric problem caused by the addition of methyl groups to Gly sites; in one case the alanyl residues were kept in the center of the helix at the cost of losing three interchain hydrogen bonds, while a second model postulated that the region of the peptide chain containing the alanine residues would loop out of the helix (*12*). In this crystal the alanine residues are packed inside the triple helix, with no evidence of looping (Fig. 1A). However, the molecular structure shows, at the substitution site, two distinctive features that could not have been predicted, (i) a twist relaxation, and (ii) the presence of four interstitial water molecules that provide the extra hydrogen bonds necessary to maintain the triple-helical assembly.

The twist relaxation can be best understood by examination of Fig. 1B, where a ribbon diagram of Gly→Ala is compared to that of the native collagen fiber model (*7*). The conformation of Gly→Ala in the $(POG)_n$ zones shows tighter winding compared to the fiber model. Similar behavior has been observed for $(PPG)_{10}$, in which triple helices follow a 7_5 screw symmetry (*19*) instead of the 10_7 screw symmetry observed in fiber collagen (*7*). However, Gly→Ala shows a twist relaxation in the zone where the substitution from Gly to Ala occurs. The three alanine residues accommodate themselves inside the triple helix through the unscrewing of the molecule in that region.

The helical parameters of the individual polypeptide chains show a broad distribution from residue to residue, a twist angle of roughly 60°, and a unit height of about 8.4 Å per tripeptide (Fig. 2). The unscrewing disruption at the substitution site is obvious in all three chains but more pronounced in

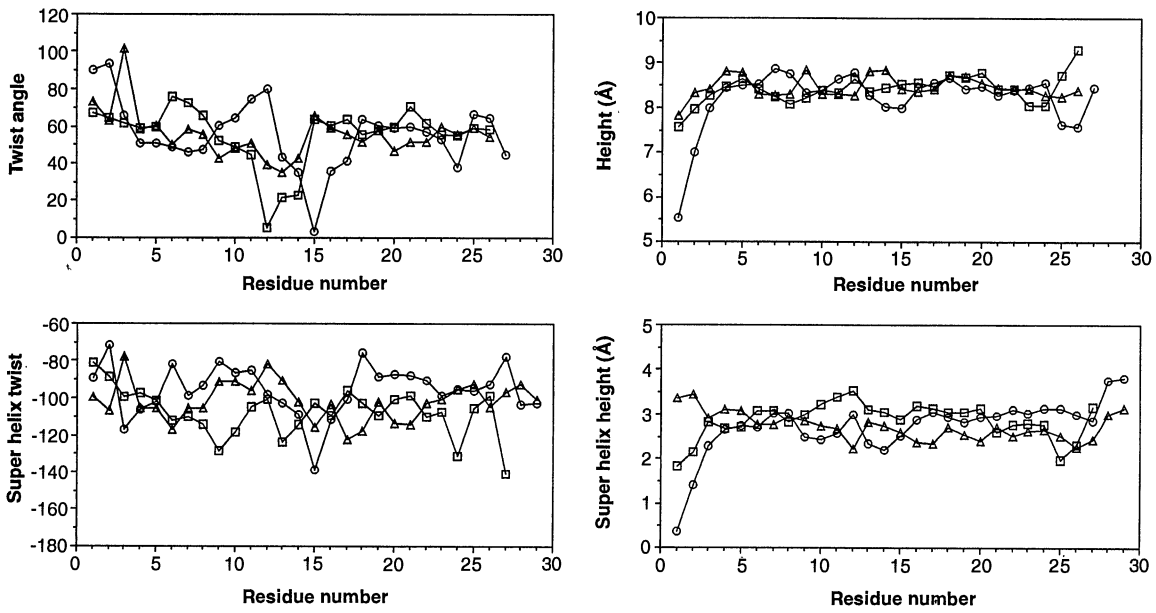

Fig. 2. Variation of helical twist and height and superhelical twist and height for Gly→Ala in the crystal structure (□ chain 1, ○ chain 2, and △ chain 3). For selection of the best axis the whole molecule was reoriented to minimize the discrepancy in cylindrical radii between atoms that are pseudo-equivalent by screw symmetry. For the basic helix, unit height and twist were measured between two contiguous tripeptides in sequence; for the super helix, one tripeptide was related to the next screw-related tripeptide in the clockwise neighboring chain. Parameters of the superhelix clearly have a broader distribution. Average twist rates are about 60° for the basic helix and −100° for the superhelix. Unit heights are, on average, 8.4 Å for the basic helix and 2.8 Å for the superhelix.

two of them (chains 1 and 2). The twist angle in the $(POG)_n$ zones of Gly→Ala differs from that of native collagen (36°, tenfold symmetry) and $(PPG)_{10}$ (51.4°, sevenfold symmetry). In contrast, the tripeptide unit height remains quite constant along the three chains and its variation in the substitution zone is not significantly different from the overall variation. The superhelical twist shows an even broader distribution, that centers around −100°, and arises from the variation of twist along the three individual chains staggered by 2.82 Å on average. The average unit height over the $(POG)_n$ zones is 2.86 Å, which is consistent with the observed height in unstretched collagen or model polypeptides (*5*, *6*) but shorter than that measured for stretched collagen (*7*).

Chain conformation and hydrogen bonding. Three zones become apparent in this peptide: the termini zone, the collagen zone, and the substitution zone. The boundary limits of these zones are determined by the interchain hydrogen bonding (Fig. 3). All φ and ψ torsion angles are scattered around average values closer to the 7_5 model for $(PPG)_{10}$ than to the 10_7 model for native collagen in fibers (Table 2), a finding consistent with the more tightly wound character of Gly→Ala. Examination of the variation of φ and ψ along the peptide sequence shows that torsion angles of all three chains are similar in the collagen zones. In the substitution zone, where the triple helix unwinds locally, most of the torsion angles remain similar to those in the collagen zone. Only five residues have one torsion angle that significantly deviates from the average values (Table 2). That small degree of torsional variability is responsible for the major changes observed in the helical twist (Fig. 2A).

In the termini zone, the triple helical conformation is somewhat unraveled. At acidic pH, the three NH_2-terminal groups have positive charges and cause the three chains to pull apart from one another. The COOH-terminal groups become partially charged because the final pH falls into their expected pK range in triple helical peptides (*29*, *30*). The packing of the triple helices in the crystal places terminal charged groups of symmetry-related molecules close in space, and as a result nonstereospecific electrostatic interactions introduce appreciable conformational disorder in that zone. The best fit of the end groups of chains 2 and 3 into the electronic density has Hyp^{32} and Hyp^{62} in a cis conformation. In this way the prolyl terminal groups of these two chains become effectively separated from each other as well as from the prolyl ring of chain 1.

The puckering of the imino acids shows predominantly the Pro(down)–Hyp(up) pattern reported for native collagen (*7*). Similar observations for $(POG)_{10}$ in solution are supported by two-dimensional nuclear magnetic resonance studies (*31*).

Of 60 possible Pro carbonyl and Gly amide groups, 46 participate in interchain hydrogen bonds which follow the pattern known as model II of Rich and Crick (*5*) (Fig. 3). Of the 14 groups that do not participate in this hydrogen bonding, six are located in the termini zone where the molecule is more disordered; either their partners are missing or the whole residue has not been located satisfactorily in the density maps (Gly^{30} and Gly^{90}). Far more interesting is the region around the Gly→Ala substitution site where the normal interchain hydrogen bonds are precluded because of the steric hindrance introduced by the methyl groups of the Ala residues. The orientation of these methyl groups tends to pull the three chains apart so that the increased Pro:C=O---H–N:Ala distances prevent hydrogen bond formation (measured distances between Pro:O and Ala:N range from 4.3 Å in the Pro^{13}-Ala^{75} pair to 5.0 Å in Ala^{45}-Pro^{73}). An additional hydrogen bond between the flanking Gly^{48} residue and Pro^{76} is also affected as a consequence of the overall disruption (the N---O distance is 4.7 Å). Water molecules are seen to establish bridges between the same groups that otherwise would participate in direct interchain hydrogen bonds (Fig. 3). The presence of these interstitial waters was totally unexpected; their inclusion during the refinement of the structure significantly im-

Table 2. Main chain conformational angle statistics in the Gly→Ala peptide. Average values are compared with those of fiber models for $(Pro\text{-}Pro\text{-}Gly)_{10}$ and native collagen; Ala residues are classified with Gly residues. Torsion angles in the substitution zone that differ significantly from their respective average values are underlined. Residues in the termini zones are excluded from this analysis.

Average main chain conformational angles

Torsion angle	Gly→Ala	$(PPG)_{10}$ 7_5 helix (*19*)	Collagen 10_7 helix (*7*)
ω Pro	179.9 ± 1.8	178.2	180.0
φ Pro	−72.6 ± 7.6	−75.5	−72.1
ψ Pro	163.8 ± 8.8	152.0	164.3
ω Hyp	178.5 ± 1.5	−176.8	180.0
φ Hyp	−59.6 ± 7.3	−62.6	−75.0
ψ Hyp	149.8 ± 8.8	147.2	155.8
ω Gly (Ala)	177.3 ± 3.1	178.2	180.0
φ Gly (Ala)	−71.9 ± 9.6	−70.2	−67.6
ψ Gly (Ala)	174.1 ± 11.9	175.4	151.4

Main chain conformational angles in the substitution zone

Residue	φ	ψ
Pro^{13}	−61	145
Hyp^{14}	−52	131
Ala^{15}	−61	131
Ala^{45}	−81	158
Pro^{46}	−60	179
Hyp^{47}	−56	127
Gly^{48}	−62	159
Pro^{73}	−73	165
Hyp^{74}	−61	149
Ala^{75}	−104	169
Pro^{76}	−78	159

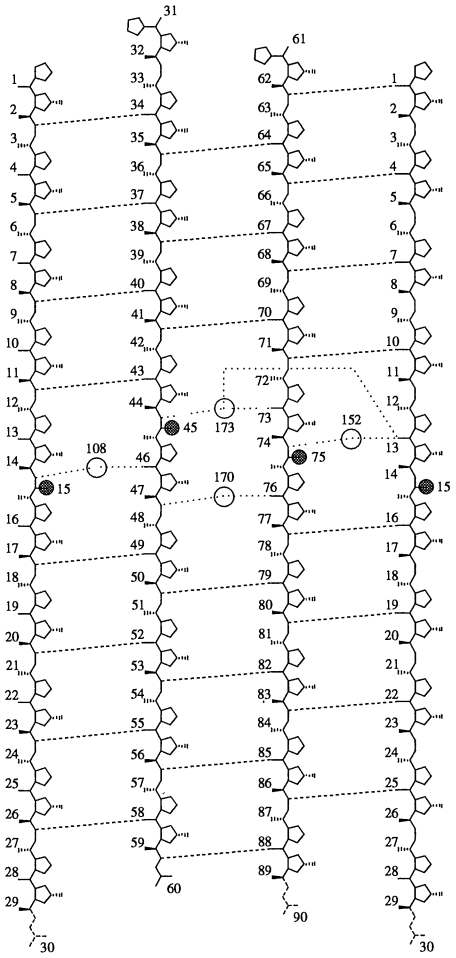

Fig. 3. Numbering scheme for Gly→Ala and schematics of interchain hydrogen bonding. The three chains are staggered by one residue and go clockwise from 1 to 3. This is a cylindrical projection so that chain 1 is repeated at the right side to provide a clearer description of the chain 3 → chain 1 hydrogen bonds. The mutual vertical displacement between the chains leaves some groups at the ends without their partner to establish an interchain hydrogen bond. The termini zones are outside the limits defined by the interchain hydrogen bonds. Collagen zones comprise residues 2–12, 34–44, 64–72, 16–27, 49–59, and 77–87, whereas the substitution zone contains residues 13–15, 45–48 and 73–76. Residues Hyp^{32} and Hyp^{62} have cis peptide bonds and Gly^{30} and Gly^{90} are missing from the current model. At the substitution site the basic hydrogen bonding is broken but the three chains still interconnect through four interstitial water molecules. Water numbering starts at residue 101. Only interstitial waters are shown here.

proved the quality of the electron density maps (Fig. 4).

Interstitial waters, cylinder of hydration, and hydroxyprolines. The presence of interstitial waters in the Gly→Ala crystal structure is consistent with the thermodynamic results in solution. A single change of one amino acid causes a decrease of 33°C in the melting temperature of this peptide (T_m = 29°C), when compared with the reference $(POG)_{10}$ (T_m = 62°C). Most of this destabilization is entropic in origin (*13*) and cannot be explained through loss of hydrogen bonds. A model for Gly→Ala in solution with interstitial waters can account for these experimental findings. The number of hydrogen bonds is the same as in $(POG)_{10}$, and therefore no large differences in ΔH are expected between the two peptides. In contrast, there is an entropy penalty if solvent molecules are mediators in interchain hydrogen bonding; the free energy of triple-helical Gly→Ala in solution increases with respect to that of $(POG)_{10}$, and its T_m decreases accordingly.

The Gly→Ala peptide is coated with a layer of water molecules that are hydrogen bonded to the Gly and Hyp carbonyl groups as well as to the 4-OH groups of Hyp residues. Thus, water molecules in intimate contact with the peptide acceptor groups create a first cylinder of hydration, which ensures that the peptide is not exposed to the bulk solvent. Most water molecules in this cylinder are networked to one another and form the hydration core to which other molecules are bonded. There are several types of water bridges. Intrachain bridges include ones in which a carbonyl in one chain hydrogen bonds via one or more water molecules to a hydroxyl or another carbonyl group in the same chain. Interchain bridges include ones in which water molecules interconnect acceptors from different peptide chains. There are also water bridges between symmetry related molecules as well as some water molecules that bond to only one atom in the peptide (Fig. 5).

While Hyp residues may cause a subtle conformational effect in triple-helical structures like Gly→Ala, it is clear that they play a role intimately related to the cylinder of hydration that wraps around the triple helix. As seen in Fig. 5, Hyp residues are the keystones supporting the water network. Their hydroxyl groups act as anchoring points for several multispan water bridges that can go back to the same chain, to another chain or to a symmetry-related molecule. The predominant upward puckering observed for Hyp residues may provide a favorable orientation of the 4-OH groups for water bridging to other acceptor groups on the peptide. When Hyp puckers upward, the 4-OH group occupies an axial position in the imino acid ring and the Cγ-Oδ bond is oriented tangentially to the hypothetical surface of a cylinder wrapped around the triple helix. In contrast, in a Hyp residue puckered downward, the 4-OH group occupies an equatorial position, and the Cγ-Oδ bond points out radially from the triple helix, approximately perpendicular to the cylindrical surface.

Gly→Ala triple helices are arranged in layers as a consequence of the C-centered unit cell. Within every layer the molecules show hexagonal antiparallel packing (Fig. 6), with an axis-to-axis distance of 14 Å. The layer thickness is 86.7 Å (half of the longest unit cell axis) which corresponds to the length of a complete peptide. Because of the

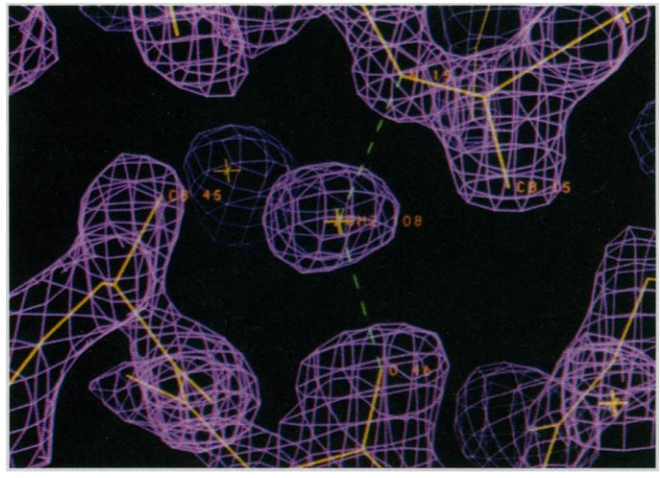

Fig. 4. Electron density map $2F_o - F_c$ in the interstitial water region. Water 108 bridges the gap between O:Pro^{46} and N:Ala^{15}. The two methyl groups Cβ:Ala^{15} and Cβ:Ala^{45} preclude the formation of a regular hydrogen bond between those residues. Map is contoured at 1.5σ.

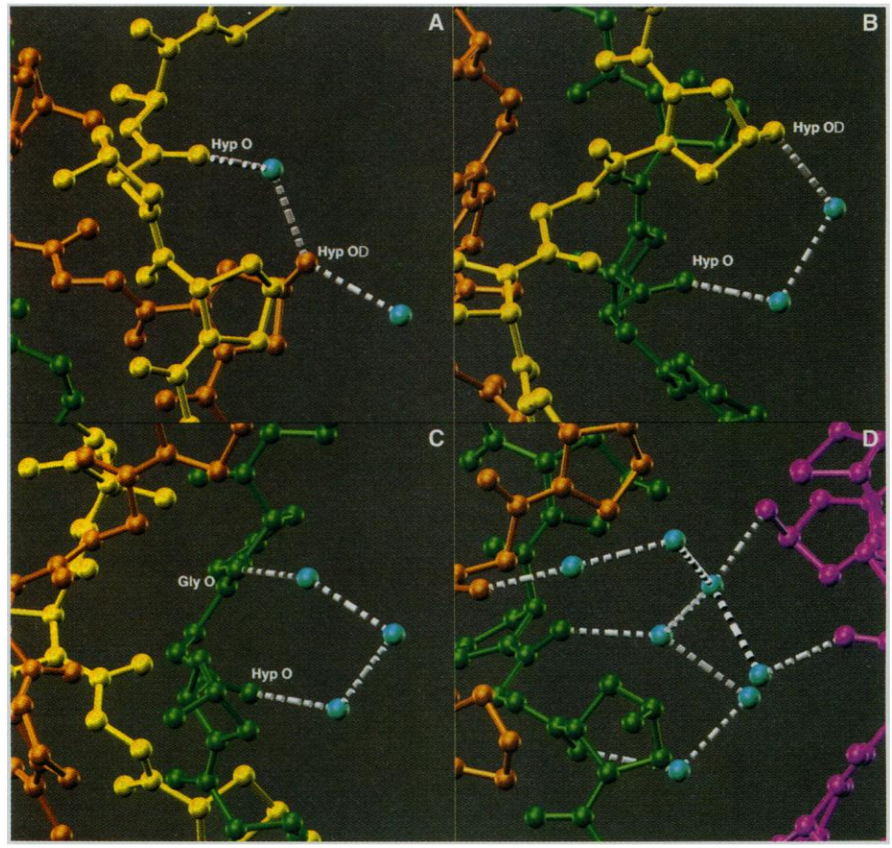

Fig. 5. Several examples of peptide-solvent hydrogen bonding. (**A**) Interchain water bridge involving one water molecule. The double water binding at the 4-OH group in Hyp residues (OD in the figure), is a repetitive motif along this crystal structure. (**B**) Interchain water bridge involving two water molecules. (**C**) Intrachain water bridge involving three water molecules and carbonyl oxygens from the peptide. (**D**) A network of water molecules interconnecting neighboring triple helices through hydrogen bonding. Carbonyl and hydroxyl peptide groups provide anchoring points for the bridge system.

disorder at the terminal regions, our model cannot provide a detailed picture of the end-to-end interactions present at the layer interfaces.

Lateral packing is highly defined in this crystal structure. The 14 Å spacing between Gly→Ala triple helices is too long for direct contact to occur. Instead, highly ordered water molecules form multispan hydrogen-bonded bridges that connect neighboring molecules (Fig. 5D). This represents an important difference from most globular protein crystals, in which only the first spheres of hydration contain well-defined water molecules, and a large amount of disordered solvent fills the gap between protein molecules. In this crystal a highly structured network of water molecules interconnects neighboring triple helices. This type of network may be related to the attractive hydration forces that have been suggested to function in collagen assemblies (*32*).

Implications for collagen. The 1.9 Å resolution structure of the Gly→Ala peptide confirms the two key features determined from fiber diffraction studies on collagen structure (*5*, *7*): the supercoiling of three individual polyproline II helices into a triple helical structure, and the formation of interchain C=O---H–N hydrogen bonds that keep the three helices in a specific register. In addition, it provides new information relevant to all triple helices and also specifically to those bearing a glycine substitution.

1) Triple helices can change their twist through small variations of main chain torsion angles, without appreciable changes in their unit height or interchain hydrogen bonding pattern. These variations can even accommodate point disruptions of the optimal sequence. Both this peptide in its collagen zone and the fiber-like model for $(Pro\text{-}Pro\text{-}Gly)_{10}$ show a smaller twist than that observed on average for native collagen. It is unlikely that this variation in twist arises from crystal packing effects, because the molecular packing in Gly→Ala crystals is somewhat reminiscent of that of native collagen triple helices and is clearly different from that in $(PPG)_{10}$ crystals (*19*). The difference in twist could relate to amino acid sequence differences, since both the Gly→Ala peptide and $(PPG)_{10}$ have imino acid contents (66 percent) much higher than found in collagen (about 20 percent), and imino acids impose a more restrained conformation with restricted torsion angles (*33*). The potential for torsional flexibility exhibited by the Gly→Ala triple helix is likely to be present in collagen in the form of local twist variations along the chain sequence. These variations can result either from the occurrence of triplet types other than POG, like PYG, XOG, and XYG, or more specifically from the appearance of particular residues in X and Y positions. This sequence dependence of the triple-helical conformation will be clarified as crystal structures of other peptides containing non-POG triplets are determined. Such conformational variability may be involved in binding and recognition phenomena.

2) Water molecules can effectively mediate interchain hydrogen bonds whenever a point disruption of the optimal sequence exists. This behavior can be more general in triple helices: nonimino acids located in the X and Y positions may participate in interchain hydrogen bonding bridges mediated by water molecules in an analogous way to the interstitial waters in Gly→Ala.

3) Triple helices are surrounded by a highly structured cylinder of hydration that determines their lateral separation in macromolecular assemblies. Although the packing of the Gly→Ala molecules in this crystal is pseudo-hexagonal and antiparallel, it shows the same interaxial spacing as that observed in semi-crystalline rat tail tendon, also pseudo-hexagonal but parallel (*34*). This similarity suggests that the 13 to 14 Å lateral spacing observed in collagen assemblies is sequence independent and becomes dictated by the effective diameter of the cylinder of hydration coating the triple helices. In contrast, the axial relation observed in the staggering of collagen molecules leading to fibrils must be the result of direct interaction between nonimino acid residues with long side chains.

4) Hydroxyproline residues interact with the surrounding cylinder of hydration, rather than with the peptide chain, providing more hydrogen bonding loci for water molecules and effectively enhancing the solvation of the triple helix. Other residues with polar side chains could accomplish the same hydration effect, but would introduce more conformational freedom into the individual chains and therefore destabilize their extended form. From this point of view, Hyp is the optimal choice and its use by nature is widespread, not only in all types of collagen but also in triple helical domains of noncollagenous proteins (*1*, *2*).

5) The subtle effect of the glycine substitution in the homotrimeric Gly→Ala peptide may clarify the structural consequences of some collagen mutations. This

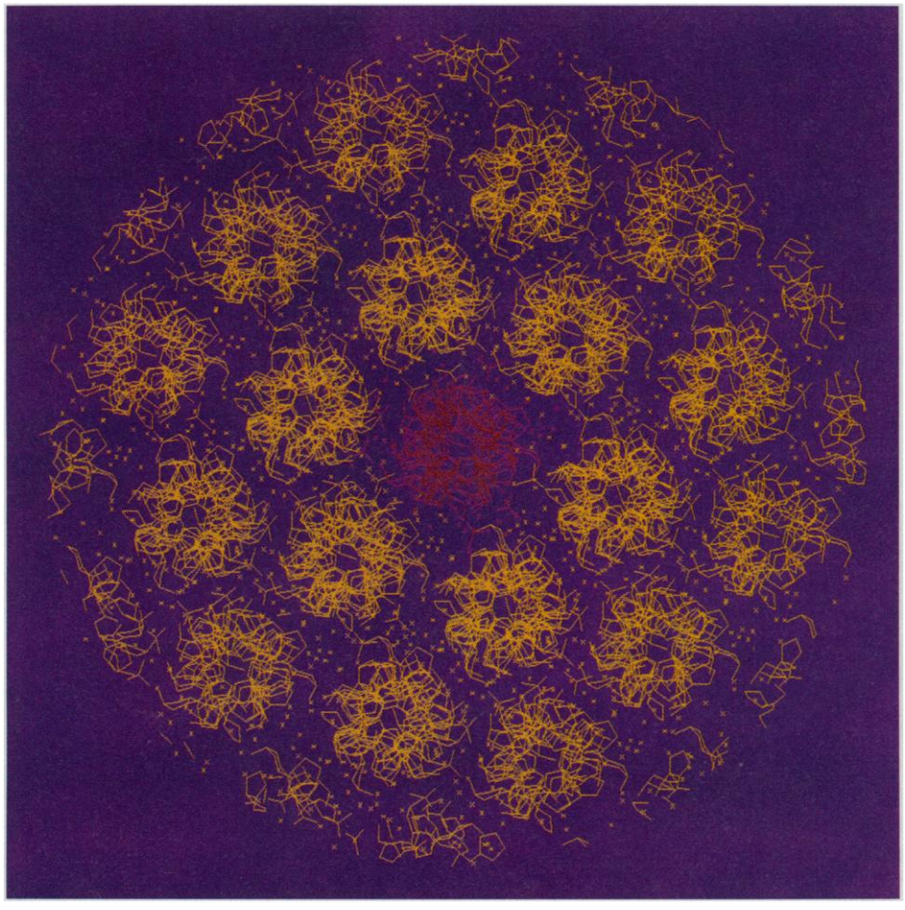

Fig. 6. A view of the packing of Gly→Ala looking down the helix axis, which is approximately parallel to the crystallographic axis *a*. Only one layer of molecules is shown. Reference and symmetry molecules are shown in different colors.

crystal structure provides structural information on the effect of a glycine substitution in a triple helix, an alteration which usually leads to pathological states in fibrillar collagens (*15*, *16*). A rare dominant connective tissue disease, osteogenesis imperfecta, is in many cases due to a single Gly→X substitution in one allele of either the α1(I) or α2(I) chains of the heterotrimeric type I collagen, $[\alpha1(I)]_2$ α2(I) (*15*). In homotrimeric collagens, a single Gly→X substitution in one allele of the α1(III) chain in type III collagen, $[\alpha1(III)]_3$, has been identified in some cases of Ehlers-Danlos syndrome type IV, a dominant disorder in which arterial rupture may occur (*15*); also, a Gly→Glu substitution in the α1(II) chain in type II collagen, $[\alpha1(II)]_3$, has been identified in a case of chondrodysplasia, a disorder characterized by abnormal formation and growth of cartilage (*16*). In all cases, the resulting collagen molecules assemble from a mixture of chains generated by one mutant and one normal allele, so that type I molecules may have 0, 1, or 2 altered chains whereas type II or type III ones may have 0, 1, 2, or 3 altered chains. These mutant collagen molecules can exhibit decreased thermal stability, decreased collagen secretion, and increased proteolytic sensitivity, and it has been suggested that this kind of mutation represents a defect in the folding of the triple helix (*15*). In contrast, all three chains of the Gly→Ala peptide have the Gly substitution. Here, triple helix torsional flexibility has allowed the accommodation of the unnatural Ala residues with minor conformational changes at the substitution site. However, although the local disruption is small, the untwisting of the triple helix leads to an alteration of the spatial relation between the triple helical segments at both sides of the substitution. This loss of spatial coherence could contribute to the long range effects seen for glycine substitutions (*35*). The nonequivalence of the three chains could be more disruptive, leading to a bent or kinked molecule such as that seen for the triple helix in C1q (*36*) or proposed for some mutant collagens (*35*). Studies on heterotrimeric triple helices with additional collagen-like sequences are needed to clarify the sequence dependent nature of biologically important features of collagen and their alterations in diseased states.

REFERENCES AND NOTES

1. T. Kodama *et al.*, *Nature* **343**, 531 (1990).
2. B. Brodsky-Doyle, K. R. Leonard, K. B. M. Reid, *Biochem. J.* **159**, 279 (1976).
3. G. N. Ramachandran and G. Kartha, *Nature* **176**, 593 (1955).
4. A. Rich and F. H. C. Crick, *ibid.*, p. 915.
5. ———, *J. Mol. Biol.* **3**, 483 (1961).
6. For a review on early work on collagen see R. D. B. Fraser and T. P. MacRae, [*Conformation in Fibrous Proteins* (Academic Press, New York, 1973)].
7. R. D. B. Fraser, T. P. MacRae, E. Suzuki, *J. Mol. Biol.* **129**, 463 (1979).
8. Standard one- and three-letter abbreviations are used for proline, Pro, P; glycine, Gly, G; and alanine, Ala, A. For the nonstandard imino acid 4-hydroxyproline, Hyp, O are used. Thus, PPG means a Pro-Pro-Gly triplet, and POG means a Pro-Hyp-Gly triplet.
9. S. Sakakibara *et al.*, *Biochim. Biophys. Acta* **303**, 198 (1973).
10. J. Engel, H. Chen, D. J. Prockop, H. Klump, *Biopolymers* **16**, 601 (1977).
11. E. Heidemann and W. Roth, *Adv. Polym. Sci.* **43**, 143 (1982).
12. C. G. Long, M. H. Li, J. Baum, B. Brodsky, *J. Mol. Biol.* **225**, 1 (1992).
13. C. G. Long *et al.*, *Biochemistry* **32**, 11688 (1993).
14. The complete sequence of the Gly→Ala peptide is $(POG)_4POA(POG)_5$.
15. H. Kuivaniemi, G. Tromp, D. J. Prockop, *FASEB J.* **5**, 2052 (1991).
16. R. Bogaert et al., *J. Biol. Chem.* **267**, 22522 (1992).
17. E. K. O'Shea, J. D. Klemm, P. S. Kim, T. Alber, *Science* **254**, 539 (1991).
18. S. Sakakibara *et al.*, *J. Mol. Biol.* **65**, 371 (1972).
19. K. Okuyama, K. Okuyama, S. Arnott, M. Takayanagi, M. Kakudo, *ibid.* **152**, 427 (1981).
20. C. K. Fair, MOLEN (Enraf-Nonius, Delft, Netherlands, 1990).
21. J. Bella *et al.*, in preparation; the starting model was inaccurate in its twist: in the $(POG)_n$ regions the rate of twist of the model was slightly smaller than the final twist observed, whereas around the Gly→Ala substitution the rate of twist was much larger. The use of a small probe molecule and the combination of multiple quasi-degenerate solutions from the molecular replacement search provided a rather good picture of the main features of the molecule. The model allowed us to locate the two $(POG)_n$ domains but did not reproduce the central region. Fourier synthesis with several solutions from the molecular replacement search produced electron density maps that allowed the central region to be built.
22. P. M. D. Fitzgerald, *J. Appl. Crystallogr.* **21**, 273 (1988).
23. A. T. Brünger, X-PLOR, version 3.1 (Yale Univ. Press, New Haven, CT, 1992).
24. P. J. Campbell-Smith and S. Arnott, *Acta Crystallogr.* **A34**, 3 (1978).
25. A. T. Brünger, J. Kuriyan, M. Karplus, *Science* **235**, 458 (1987).
26. T. A. Jones, *J. Appl. Crystallogr.* **11**, 268 (1978).
27. Placing the sequence in register with the electron density is not a trivial problem in a repeating structure. A "POG-only" model was kept through much of the refinement, and therefore we did not impose which glycine residues would eventually become alanines. The major concern was not to misplace the three chains by one tripeptide in the direction of the triple helix axis. We repeatedly checked the model at both ends to identify extra electron density, as well as at the glycine α carbons, looking for density suggesting methyl groups. Difference Fourier synthesis maps calculated with a model of 84 amino acids revealed many possible water sites with distances and orientations compatible with hydrogen bond formation to the peptide chain. After 45 molecules of water were fitted in this first cylinder of solvation, the overall quality of the electron density maps improved, and it became possible to locate acetic acid molecules. The methyl groups of the alanine residues were identified at the last stages of refinement and were reconfirmed by difference electron density maps with these residues omitted. The final positioning of the alanines is consistent with the untwisting of the triple helix at the central region and with the location of the interstitial waters bridging the three chains in that region. Gly^{30} and Gly^{90} are not included in the current model because they were not located satisfactorily in the electronic density map.
28. F. H. Allen, O. Kennard, R. Taylor, *Acc. Chem. Res.* **16**, 146 (1983); for Pro, Gly, and Ala residues we used the parameters derived also from the CSD: R. A. Engh and R. Huber, *Acta Crystallogr.* **A47**, 392 (1991).
29. R. A. Berg, B. R. Olsen, D. J. Prockop, *J. Biol. Chem.* **245**, 5759 (1970).
30. M. G. Venugopal, J. A. M. Ramshaw, E. Braswell, D. Zhu, B. Brodsky, *Biochemistry* **33**, 7948 (1994).
31. M.-H. Li, P. Fan, B. Brodsky, J. Baum, *ibid.* **32**, 7377 (1993).
32. S. Leikin, D. C. Rau, V. A. Parsegian, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 276 (1994).
33. Molecular modeling studies suggest that imino acid–rich triplets induce a more twisted conformation. See for example M. H. Miller and H. A. Scheraga, *J. Polymer Sci. Symp.* **54**, 171 (1976).
34. R. D. B. Fraser, T. P. MacRae, A. Miller, E. Suzuki, *J. Mol. Biol.* **167**, 497 (1983).
35. B. E. Vogel *et al.*, *J. Biol. Chem.* **263**, 19249 (1988); S. J. Lightfoot *et al.*, *ibid.* **267**, 25521 (1992).
36. E. Kilchherr, H. Hofmann, W. Steigemann, J. Engel, *J. Mol. Biol.* **186**, 403 (1985).

37. We thank C. G. Long for providing the Gly→Ala peptide and for helpful discussions. Supported by NIH grant GM 21589 (H.M.B.), NIH grant AR 19626 (B.B.), a postdoctoral fellowship from the Ministerio de Educación y Ciencia of Spain (J.B.), and in part by a grant from the Pittsburgh Supercomputing Center funded through the NIH and NSF. Coordinates have been deposited in the Brookhaven Protein Data Bank with entry code number 1CAG.

30 March 1994; accepted 16 August 1994