Site-Specific NMR Monitoring of cis−trans Isomerization in the Folding of the Proline-Rich Collagen Triple Helix†

Alexei V. Buevich,‡,§ Qing-Hong Dai,‡,§ Xiaoyan Liu,§,|| Barbara Brodsky,⊥ and Jean Baum*,§

Department of Chemistry, Rutgers University, Piscataway, New Jersey 08854, and Department of Biochemistry, Robert-Wood Johnson Medical School, Piscataway, New Jersey 08854

Received November 9, 1999; Revised Manuscript Received January 17, 2000

ABSTRACT: Understanding the folding of the proline-rich collagen triple helix requires consideration of the effects of proline cis−trans isomerization and may shed light on the misfolding of collagen in connective tissue diseases. Folding was monitored in real time by heteronuclear 2D NMR spectroscopy for the ^{15}N labeled positions in the triple-helical peptide T1−892 [GPAGPAGPVGPAGARGPAGPOGPOGPOG-POGV]. In the equilibrium unfolded monomer form, each labeled residue showed multiple peaks with interconversion rates consistent with cis−trans isomerization of Gly−Pro and Pro−Hyp bonds. Real-time NMR studies on the folding of T1−892 showed slow decay of monomer peaks and a concomitant increase in trimer peaks. Gly25 in the C-terminal rich $(Gly\text{-}Pro\text{-}Hyp)_4$ domain folds first, consistent with its being a nucleation domain. Analysis of the kinetics indicates that the folding of Gly25 is biphasic and the slower step represents cis−trans isomerization of imino acids. This illustrates that nucleation is limited by cis−trans isomerization. Monitoring Gly6, Gly10, Ala12, and Gly13 monomer and trimer peaks captures the C- to N-terminal propagation of the triple helix, which is also limited by Gly−Pro cis−trans isomerization events. The zipper-like nature of the propagation process is confirmed by the slower rate of folding of Ala6 compared to Gly13, reflecting the larger number of isomerization events encountered by the more N-terminal Ala6. The cis−trans isomerization events at multiple proline residues is a complex statistical process which can be visualized by these NMR studies.

The significant experimental and theoretical advances in understanding the folding of small monomeric globular proteins can be expanded to consider the folding of non-globular proteins, such as collagen (*1*−*4*). The linear, $(Gly\text{-}X\text{-}Y)_n$ repetitive triple-helical collagen presents a rodlike folded form that is simpler than the three-dimensional structure of globular proteins. In this conformation, the Gly residues are all buried near a central axis, while the residues in the X and Y positions are largely exposed to solvent (Figure 1). Therefore, all polar residues and hydrophobic side chains are on the exterior, and there is no hydrophobic core. The X and Y positions are frequently occupied by proline and hydroxyproline, respectively, which are important to triple-helix stability, and Gly-Pro-Hyp is the most common and most stabilizing tripeptide sequence. Therefore, the folding of the triple helix requires multichain assembly to a final folded state, includes an unusually high concentration of prolines and hydroxyprolines, but does not include a hydrophobic collapse. This brings into focus other folding features than those encountered in globular proteins.

In a physiological context, the folding of collagen involves molecular assembly of the triple-helix, followed by higher order assembly to fibrils or other associated forms (*9*, *10*). Collagen is synthesized in a procollagen form, with N- and C-terminal globular propeptides flanking the $(Gly\text{-}X\text{-}Y)_n$ central domain (*10*). Trimerization occurs through the association of three C-terminal propeptides, and then nucleation of the triple-helix takes place at the (Gly-Pro-Hyp) rich sequence found at the C-terminus of the $(Gly\text{-}X\text{-}Y)_n$ region (*11*−*14*). After nucleation of the three chains, the triple helix propagates with cis−trans isomerization of Pro and Hyp being rate-limiting steps (*12*, *13*). The completed triple-helical procollagen molecule is secreted into the extracellular matrix, and following cleavage of the propeptides, the molecules assemble into the characteristic collagen fibrils that form the supporting matrix of bone, tendon, and other connective tissues (*10*).

Connective tissue diseases have been found to result from collagen mutations leading to defective molecular or higher order folding (*15*−*17*). Misfolding has been implicated in cases of Osteogenesis Imperfecta ("brittle bone disease") where a single Gly in type I collagen has been mutated to another residue, breaking the $(Gly\text{-}X\text{-}Y)_n$ repeating pattern (*18*). Type I collagen, a heterotrimer of two α1 chains and one α2 chain, is the most common collagen and forms the major organic matrix of bone. Abnormal folding of type I collagen with Osteogenesis Imperfecta mutations have been directly observed and have been inferred from an increased degree of posttranslational modifications (*18*, *19*).

† This work was supported by NIH Grant GM-45302 (J.B.) and NIH Grant GM-60048 (B.B.).

* To whom correspondence should be addressed. E-mail: baum@rutchem.rutgers.edu.

‡ Authors contributed equally to this work.

§ Department of Chemistry, Rutgers University, Piscataway, New Jersey.

|| Present address: Colgate-Palmolive Company, 909 River Road, Piscataway, NJ 08854.

⊥ Department of Biochemistry, Robert-Wood Johnson Medical School, Piscataway, New Jersey.

10.1021/bi992584r CCC: $19.00
Published on Web 03/24/2000

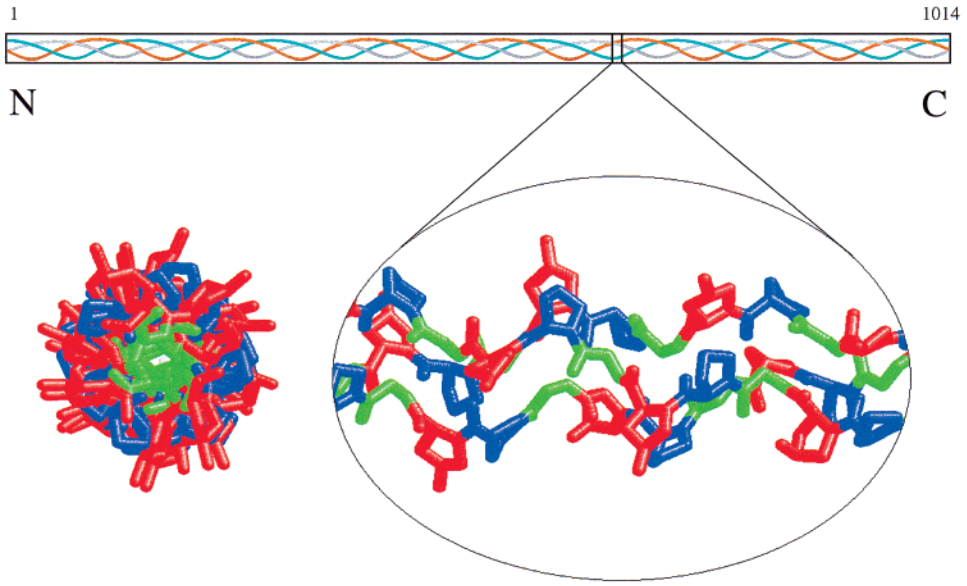

FIGURE 1: Schematic representation of the collagen triple-helix conformation. The triple-helical conformation consists of three polypeptide chains, each in a polyproline II-like helix, which are supercoiled about a common axis and hydrogen bonded together (*5*−*8*). The molecule can be a homotrimer, composed of three identical chains, or a heterotrimer, consisting of two or three distinct chain types. The conformation determines the strict sequence constraint of Gly as every third residue, so the structure can be identified by its characteristic sequence pattern of $(Gly\text{-}X\text{-}Y)_n$. All Gly (colored green) are buried near a central axis, while the residues in the X and Y positions (colored blue and red, respectively) are largely exposed to solvent. The stability of the triple-helix also requires a high content of the imino acids Pro and Hyp in the X and Y positions.

Recent progress in characterizing the conformation and folding of the triple-helix has emerged from studies on model peptides (*20*−*25*). The slow folding rate and the incorporation of ^{15}N-enriched residues allows NMR monitoring in real time of the folding of residues at specific locations (*2*, *26*). 2D heteronuclear NMR real-time folding studies applied to the triple helix are used to elucidate the residue-specific mechanism of both the nucleation and propagation steps (*24*, *27*). NMR studies of the folding of the peptide T1−892 containing an 18 residue sequence from the human αl(I) chain of type I collagen are described here. The results indicate the critical nature of imino acids. Isomerization of the high percentage of cis isomers of Pro and Hyp observed in the unfolded state were found to be the rate-limiting steps of both nucleation and propagation of the triple helix.

MATERIALS AND METHODS

Sample Preparation. Three peptides with analogous sequences containing different ^{15}N-labeled positions (in boldface, Scheme 1) were purchased from Synpep Corporation (Dublin, CA).

Scheme 1

	1 5 10 15 20 25 30
T1-892a:	GPAGPAGPVGP**AG**ARGPAGPOGPO**G**POGPOGV
T1-892b:	GPAGP**A**GPV**G**PAGARGPAGPOGPOGPOGPOGV
T1-892c:	GPAGP**A**GPV**G**PA**G**ARGPAGPOGPO**G**POGPOGV

The peptide solutions used in NMR experiments were prepared with 10% D_2O/90% H_2O and pH 2.5 was adjusted by titration with 5 mM HCl. HPLC purity of the samples was 95−99%, and their identities were confirmed by mass spectrometry.

NMR Spectroscopy. NMR experiments were performed on Varian Unity Plus 500 MHz or Varian Inova 600 MHz spectrometers equipped with a triple resonance probe and pulsed field gradient. Two-dimensional data sets were processed on a Silicon Graphics workstation using FELIX 97 (Molecular Simulation Inc.). Heteronuclear single quantum coherence spectroscopy (HSQC) spectra were recorded at 0−40 °C. All the pulse sequences employed enhanced sensitivity pulsed field gradient techniques (*28*). To determine the exchange rates between interconverting monomer peaks, HSQC-NOESY experiments (*29*) were performed at 40 °C with a series of mixing times: 30, 60, 80, 100, 200, 300, 400, 500, and 750 ms. The ^{15}N NOE relaxation measurements were performed in the presence and absence of ^{1}H saturation (*30*).

Folding Experiments. The NMR folding experiments were performed as previously described (*27*). The sample was denatured outside of the NMR spectrometer by heating for 15 min at 50 °C and then immersed in a highly concentrated KCl solution at about −9 °C (cooled by dry ice and ethanol mixture) or ice−water bath at 0 °C for 20 s to quickly reduce the temperature to 10 °C. A series of 2D ^{1}H-^{15}N HSQC spectra were acquired every 3 min at 10 °C. The dead time of the first spectrum was 60 s. Kinetics of folding were monitored by measuring cross-peak volumes as a function of time. The folding experiment was repeated three times, and average intensities were used in the analysis. Due to the very different relaxation rates of the monomer and trimer and intermediate species, the intensity in HSQC spectra of fast relaxing trimer and intermediate resonances are decreased compared to the slowly relaxing monomer resonances. This was due to the fact that the efficiency of magnetization transfer in the HSQC experiment is, among other factors, also dependent on the relaxation rates (*31*). Therefore, scaling of intensities was required for the kinetic data analysis, which was done assuming that the loss of monomer in the folding process was equal to the accumulation of the sum of the

trimer and intermediate peak intensity (mass conservation law). On the basis of this assumption, the scaling coefficients were calculated for each residue separately and each scaling coefficient, η, was calculated as an average value of four coefficients corresponding to four different time intervals between first and second, first and third, ..., and first and fifth time points: $\eta = [\sum(M_1 - M_i)/(T_i - T_1)]/4$, where $i = 2, 3, ..., 5$, and M_i and T_i are the intensity of monomer and trimer (or trimer plus intermediate) peaks at the ith-time point. Scaling coefficients η were 2.72 ± 0.46, 4.52 ± 0.41, 3.49 ± 0.51, and 3.13 ± 0.53 for Ala6, Gly10, Gly13, and Gly25, respectively. Scaled data were normalized assuming that the initial intensity of monomer was equal to one.

Curve Fitting Procedure. Kinetic data were analyzed using KinFit program (D. V. Dearden, Brigham Young University, Provo, UT). KinFit is a program written in Visual Basic as a Microsoft Excel 5.0 application macro. This program solves a set of coupled ordinary differential equations (*32*) and determines the best fit to the raw data by adjusting the rate constants using the Marquardt minimization algorithm (*33*).

Exchange Rate Calculations. To determine the exchange rates of two interconverting species in the monomer state, spectra were recorded for a series of mixing times using an HSQC-NOESY experiment at 40 °C. The exchange rates were extracted from the peak volumes of both auto and exchange peaks through a least-squares fitting procedure as described (*34*). The intensity of the auto peaks for two interconverting states are given by

$$I_{11}(T) = \frac{I_1(0)[-(\lambda_2 - a_{11})e^{-\lambda_1 T} + (\lambda_1 - a_{11})e^{-\lambda_2 T}]}{(\lambda_1 - \lambda_2)} \quad (1)$$

$$I_{22}(T) = \frac{I_2(0)[-(\lambda_2 - a_{22})e^{-\lambda_1 T} + (\lambda_1 - a_{22})e^{-\lambda_2 T}]}{(\lambda_1 - \lambda_2)} \quad (2)$$

while the intensities of the exchange peaks corresponding to the transfer of the magnetization from 1 to 2 $I_{12}(T)$ and from 2 to 1 $I_{21}(T)$ are given by

$$I_{12}(T) = \frac{I_1(0)(-a_{21}e^{-\lambda_1 T} - a_{21}e^{-\lambda_2 T})}{(\lambda_1 - \lambda_2)} \quad (3)$$

$$I_{21}(T) = \frac{I_2(0)(-a_{12}e^{-\lambda_1 T} - a_{12}e^{-\lambda_2 T})}{(\lambda_1 - \lambda_2)} \quad (4)$$

where $\lambda_{1,2} = 1/2\{(a_{11} + a_{22}) \pm [(a_{11} - a_{22})^2 + 4k_{12}k_{21}]^{1/2}\}$, $a_{11} = R_1 + k_{12}$, $a_{12} = -k_{21}$, $a_{21} = -k_{12}$, $a_{22} = R_2 + k_{21}$, R_1 and R_2 are the longitudinal relaxation rates of magnetization in sites 1 and 2, $I_1(0)$ and $I_2(0)$ denote the amount of longitudinal nitrogen magnetization associated with states 1 and 2 at the start of the mixing period T, and k_{12} and k_{21} are the exchange rates for the magnetization converting from site 1 to 2 and 2 to 1, respectively. A least-squares fitting procedure was employed to extract the longitudinal decay and chemical exchange rates by fitting the expressions given in eqs 1−4 to the measured intensities of auto and exchange peaks. $K_{eq} = I_2(0)/I_1(0)$, $K_{eq} = k_{12}/k_{21}$, were used to constrain the relative values of k_{12} and k_{21}.

RESULTS

A triple-helical peptide, T1−892, was designed to contain an 18 residue amino acid sequence from the α1 chain of type I collagen, residues 892−910, including the 901 site of a Gly to Ser Osteogenesis Imperfecta mutation. A C-terminal sequence of four repeating Gly-Pro-Hyp units was added to stabilize the triple helix and to model the C-terminal nucleation site found in collagen. For NMR studies, peptides were made with ^{15}N-enriched residues at specific positions, including one with labels at Ala6, Gly10, Gly13, and Gly25, a second one with labels at Ala12, Gly13, and Gly25, and a third one with labels at Ala6 and Gly10. Peptide T1−892 adopts a triple-helical conformation with a melting temperature of 26 °C (*35*). NMR assignments and dynamics of the ^{15}N-enriched residues in the triple helix indicated that all positions were in a highly ordered, triple-helical form (*36*). In addition to defining the final state of the triple helix, characterization of the folding process requires elucidation of the monomer form and monitoring of the kinetics of assembly of the triple helix in real time.

Monomer conformation of T1−892. To examine the unfolded monomer form, 2D ^{1}H-^{15}N HSQC spectra of T1−892 were recorded at 40 °C, a temperature above the triple-helix melting point ($T_m = 26$ °C) (*35*). More than one peak was observed for each residue in the monomer form of T1−892 at 40 °C. For residues Ala12, Gly13, and Ala6, there is one major monomer peak (M_1) with 86−88% of the total intensity and one well-defined minor peak (M_2) with 12−14% of the total intensity (Figure 2a and Table 1). To ascertain whether the major and minor peaks are interconverting conformers, HSQC-NOESY experiments were performed at 40 °C on the T1−892 peptide with labeled Ala12, Gly13, and Gly25 residues (Figure 2b). These spectra show two exchange cross-peaks between the major and minor forms (Figure 2b), indicating the existence of an equilibrium and slow exchange between the two forms. The rates of exchange between the major and minor peaks were determined using a series of HSQC-NOESY experiments with variable mixing times (Figure 2c). For Ala12, an exchange rate of 0.066 s^{-1} was determined for the major to minor peak direction and 0.40 s^{-1} for the reverse reaction (Table 2). Similar rates were seen for Gly13. The free energy of activation, $\Delta G^{\neq}$, for the interconversion between monomers is 84 kJ/mol for Ala12 and 85 kJ/mol for Gly13, respectively.

The population of the minor peak and the free energy of activation for Ala12 and Gly13 are very similar to those previously reported for the cis form of Pro in short oligopeptides (*38*, *40*−*42*). For example, Gly−Pro bonds in model peptides show 14% cis form and 86% trans (*38*). And normally, the activation energy for cis−trans isomerization about X−Pro bonds is ∼85 kJ/mol (*42*). This close correspondence indicates that Gly-Pro cis−trans isomerization is the source of the major and minor peaks observed for Ala12 and Gly13. However, in the sequence of T1−892 [Gly10-Pro11-Ala12-Gly13-Ala14-Arg15], neither Ala12 nor Gly13 is directly involved in an X−Pro bond. It is likely that their interconverting peaks represent cis−trans conformers at the preceding Gly10−Pro11 bond. Ala6 has a cis−trans ratio very similar to Ala12 and Gly13, suggesting that it is related to cis−trans conformers at one Gly-Pro position. Given the sequence environment of Ala6 [Gly4-Pro5-Ala6-Gly7-Pro8-Val9], it is not possible to conclude whether the conformers are reflecting isomerization at the Gly4−Pro5 or Gly7−Pro8 bonds.

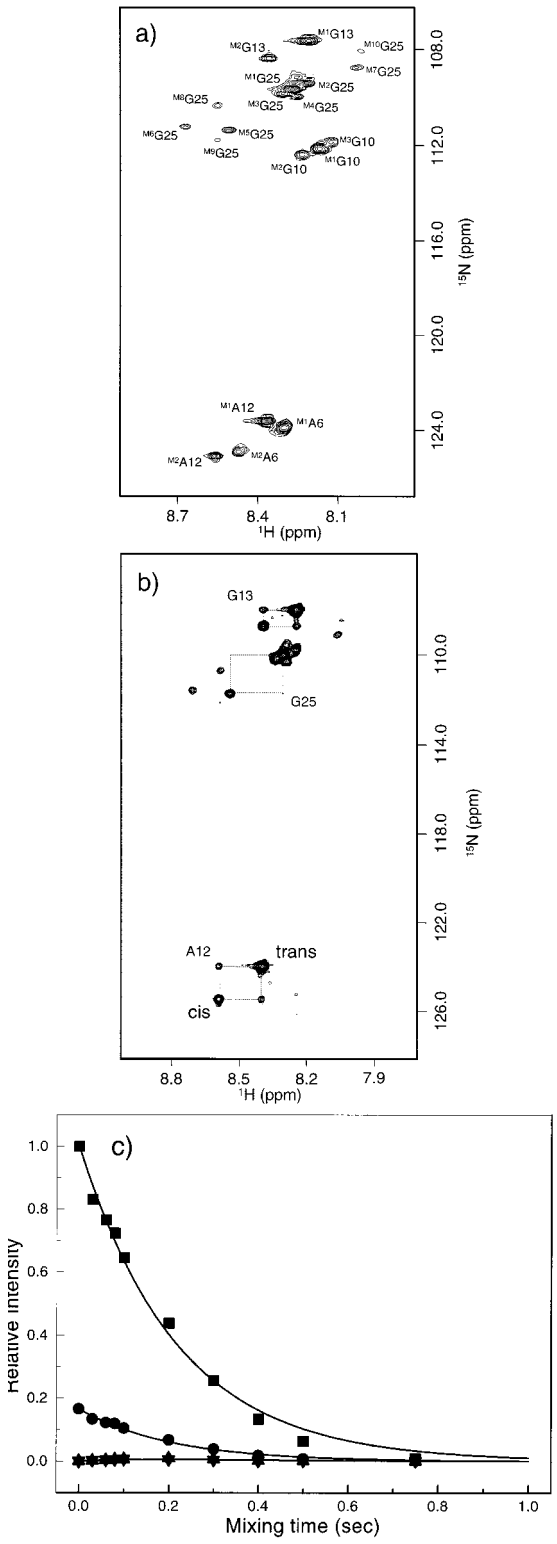

FIGURE 2: Charaterization of monomer states of T1−892 in 90% H_2O/10% D_2O (pH 2.3) at 40 °C. (a) ^1H-^{15}N HSQC spectrum of thermally denatured T1−892 at 40 °C. The peaks corresponding to the monomer states of individual residues are denoted with a superscript M. (b) HSQC-NOESY spectrum of T1−892 at 40 °C with mixing time 500 ms. Exchange cross-peaks of major and minor peaks for Ala12, Gly13, and Gly25 are shown with the dotted lines. The cross-peaks for Gly25 can only be observed at a lower contour level. (c) Mixing time dependence of chemical exchange between the two monomer states of Ala12 in T1−892 at 40 °C. The upper two curves describe the decay of auto peaks, while the lower curves show the buildup/decay of the exchange peaks. The curves are the best least-squares fit to the experimental data.

Table 1: Experimental and Theoretical Percentages of Different Monomeric Peaks for Individual Residue in Peptide T1−892 at 40 °C (concentration 9.2 mM, solvent 90% H_2O/10% D_2O, pH 2.3)[a]

residue	experimental value (%)	theoretical value (%)	possible configuration state
G25M1	62.3	61.8	ttttt
G25M2	10.8	9.88	tcttt/tttct
G25M3	11.2	9.88	tcttt/tttct
G25M4	4.2	3.95	ctttt/ttctt/ttttc
G25M5	3.9	3.95	ctttt/ttctt/ttttc
G25M6	2.4	3.95	ctttt/ttctt/ttttc
G25M7	1.7	(1.56	(tctct/ccttt
G25M8	1.6	0.63	tcctt/tcttc
G25M9	1.3	0.63	cttct/ttcct
G25M10	0.65	0.63)	tttcc)
G13M1	84.2	86.3	t
G13M2	15.8	13.7	c
A12M1	85.8	86.3	t
A12M2	14.2	13.7	c
G10M1	74.4	74.5	tt
G10M2	11.4	11.8	ct/tc
G10M3	13.9	11.8	ct/tc
A6M1	85.5	86.3	t
A6M2	14.5	13.7	c

[a] Theoretical values for G25 are calculated based on multiple X-P bonds in G-P-O-G-P-O-G(25)-P-O (2GP/3PO); while for G13, A12, theoretical values are calculated based on single G-P bond by in G(10)-P-A(12)-G(13)-A-R; For G10, double G−P bonds in G-P-V-G(10)-P-A are used; For A6, single G-P bond is G-P in G-P-A(6)-G-P-V is used. Percentages are G-P, 86.3% trans, 13.7% cis; P-O, 94% trans, 6% cis. The assignment of the possible configuration state for individual peak is based on its percentage and also chemical shifts (*41*, *42*).

Table 2: Rate Constants (k_{ct}, k_{tc}) and Free Energy of Activation ($\Delta G^{\neq}_{ct}$, $\Delta G^{\neq}_{tc}$) for cis−trans Isomerization of the X-Pro Peptide Bond in T1−892 at 40 °C

	A12	G13	G25
k_{ct} (s^{-1})	0.066	0.052	0.030
k_{tc} (s^{-1})	0.40	0.32	0.39
$\Delta G^{\neq}_{ct}$ (kJ/mol)[a]	83.9	84.5	85.9
$\Delta G^{\neq}_{tc}$ (kJ/mol)[a]	79.2	79.7	79.2

[a] $\Delta G^{\neq}$ was calculated as described in ref *37*.

The other two labeled residues, Gly10 and Gly25, also show major and minor peaks in their monomer forms. However, there is more than one minor peak observed for these residues and the major/minor peak ratio of intensities is not consistent with isomerization at a single Gly−Pro bond. Gly10 shows one major peak (74% intensity), together with two minor peaks (11 and 14%). This is consistent with conformers reflecting cis−trans isomerization at two Gly−Pro bonds. The sequence environment of Gly10 [Gly7-Pro8-Val9-Gly10-Pro11-Ala12] suggests that Gly10 is sensitive to the preceding Gly7−Pro8 isomerization as well as Gly10−Pro11.

Gly25 shows at least 10 peaks of varying intensities, in addition to the major peak (Table 1). This suggests that Gly25, in the Gly19-Pro20-Hyp21-Gly22-Pro23-Hyp24-Gly25-Pro26-Hyp27-Gly28-Pro29-Hyp30 C-terminal sequence, is sensitive to isomerization at multiple Gly−Pro and/or Pro−Hyp sites (Figure 3). Theoretical intensities for different combinations of cis−trans isomers were calculated based on the fact that imino acid−imino acid bonds have about 6% cis isomers while X−imino acid bonds have about

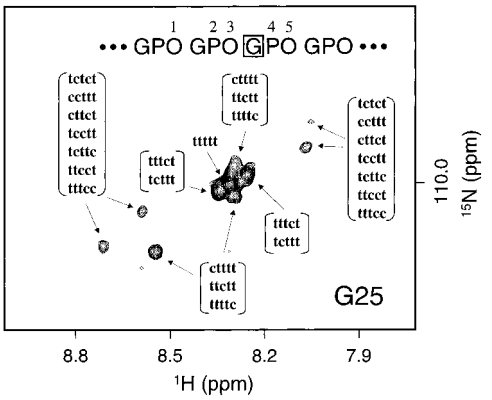

FIGURE 3: ^{1}H-^{15}N HSQC of monomer resonances of Gly25 at 40 °C. A set of possible conformational state assignment (cis/trans) are indicated next to the corresponding resonances. The tentative assignments were done based on the comparison of the experimental peak intensities with the theoretical intensities expected for different sets of cis and trans conformers in the five X-imino acid bonds: P20-O21-G22-P23-O24-G25-P26−O27. The experimental and theoretical intensities are listed in Table 1. Each of these bonds was labeled from N- to C-terminal by letters c and t, where letter c stands for cis and t stands for trans bond conformation. For example, the conformational configuration with four trans-bonds and one cis bond between G22 and P23 will be labeled as tcttt.

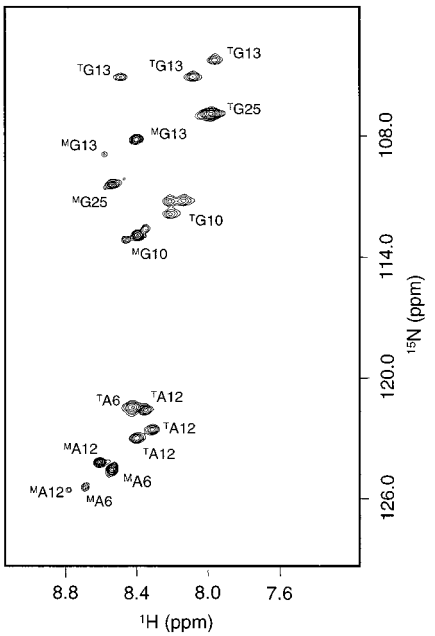

FIGURE 4: ^{1}H-^{15}N HSQC spectrum of T1−892 at 10 °C. The peaks corresponding to the monomer states of T1−892 are denoted with a superscript M, while those peaks corresponding to the triple helical states are denoted with a superscript T.

14% (*38*) (Table 1). The relative intensities of the major and minor peaks were most consistent with isomerization at five individual X−imino acid sites (2GP and 3PO): Pro20−Hyp21, Gly22−Pro23, Pro23−Hyp24, Gly25−Pro26, and Pro26−Hyp27 (Table 1). Tentative assignments for Gly25 monomer peaks were made by comparing experimental intensities to the theoretical ones in Table 1. On the basis of intensities alone, it was not possible to assign cross-peaks without ambiguity. For example, a tttct configuration cannot be distinguished from a tcttt one and the cross-peak was labeled as being either one or the other. However, one cross-peak could be assigned directly to ttttt configuration.

HSQC-NOESY exchange experiments for Gly25 showed one exchange cross-peak between the major peak and one of the minor forms (Figure 2b), which corresponds to a cis bond at a Pro−Hyp site based on the fact that the relative intensity of the minor form was 7.1%. The major (trans) to minor (cis) peak exchange rate at 40 °C is 0.030 s^{-1}, a value slower than observed for Ala12 and Gly13, while the reverse rate was 0.39 s^{-1}, similar to that seen for Ala12 and Gly13 (Table 2). Other expected exchange peaks for Gly25 were difficult to analyze due to the overlap with the major peak and the weak intensities of the minor forms with more than one cis bond (Table 1). At low temperature, the monomer peaks seen at high temperature are still present but are now accompanied by peaks assigned to the triple-helix (*36*) (Figure 4).

Kinetics of Folding of Peptide T1−892. The folding of T1−892 can be followed in real time by heteronuclear NMR by monitoring the appearance of the trimer peaks of the five labeled residues and the disappearance of their corresponding monomer peaks as the temperature is dropped from 50 to 10 °C (Figure 5). The kinetics of the monomer decay of Ala6, Gly10, Ala12, Gly13, and Gly25 could not be fit to any single kinetic order using the entire data set. Therefore, analysis of the data was performed assuming biphasic kinetics. The shape of the monomer decay suggests a fast phase between the initiation of folding and the first data point, and then a slow phase describing monomer to trimer folding, M ⇄ T.

The slow phases of the monomer decay curves for Ala6, Gly10, Gly13, and Gly25 were all fit to first-order kinetics with rate constants on the order of 10^{-3} s^{-1} (Table 3). Results indicate that the fast phase of Gly25 disappears more rapidly than the other residues, and the slow phase of Gly25 is faster than Ala6 by a factor of 1.6, suggesting that the folding process is initiated at the C-terminal end.

The triple-helical peak for Gly25 increases in parallel with the monomer decay; therefore, the data for Gly25 can be fit simultaneously to monomer decay and trimer formation assuming biphasic kinetics (Table 3, Figure 5). As expected, the rate constants for the slow phase obtained from this fit are similar to those obtained by fitting to the monomer decay only. The appearance of triple-helix peaks for the other residues, Ala6, Gly10, and Gly13, is delayed relative to Gly25. The triple helical and monomer peaks for Gly10 were fit similarly to Gly25.

Folding intermediates were observed for three of the residues. For Ala6, Ala12, and Gly13, new nonmonomer peaks arise as the trimer peaks are formed (Figure 6). For example, in the native triple-helical peptide, Ala6, has two small monomer peaks and one peak in the trimer region. In the early stages of folding, Ala6 shows one peak in the trimer region, at least four nonnative peaks near the trimer peak and one nonnative peak near the monomer peak. With time, the nonnative peaks decay and the trimer peak grows (Figure 6). The nonnative peaks of Ala6 may be considered as an intermediate form in the folding pathway. Similar nonnative peaks suggesting intermediates were observed for Gly13. However, Ala6 showed a higher proportion of intermediate relative to Gly13.

The kinetics of folding for Ala6 and Gly13 are biphasic, similarly to those of Gly25 and Gly10. To fit the slow phase for the Ala6 and Gly13 trimer and intermediate folding,

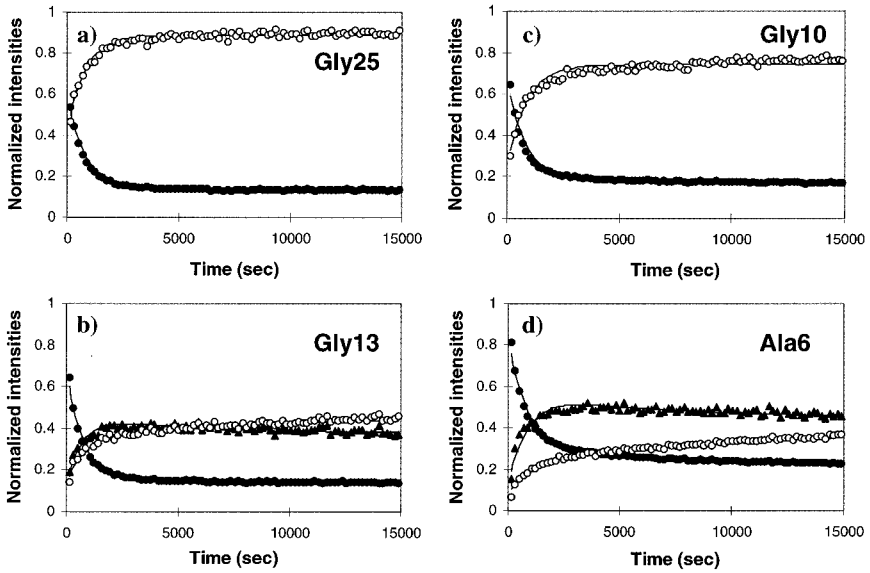

FIGURE 5: NMR real-time folding profiles of Gly25 (a), Gly13 (b), Gly10 (c), and Ala6 (d) residues in T1−892c at 10 °C: monomer (●), trimer (○), and intermediate (▲). Data were collected and resonance volumes were determined as described under Materials and Methods. All curves (—) represent best fit of the second slow phase of the biphasic folding kinetics. In all cases, the slow phase was fit starting from the second time interval. Rate constants for Gly25, Gly13, Gly10, and Ala6 obtained from these fits are shown in Table 3.

Table 3: First Order Rate Constants for the Slow Step of T1−892c Folding for Ala6, Gly10, Gly13, and Gly25[a]

	monomer data (s^{-1}) $M \underset{k_2}{\overset{k_1}{\rightleftharpoons}} T$		monomer and trimer data (s^{-1}) $M \underset{k_2}{\overset{k_1}{\rightleftharpoons}} T$		monomer, intermediate, and trimer data (s^{-1}) $M \underset{k_2}{\overset{k_1}{\rightleftharpoons}} T$, $M \underset{k_4}{\overset{k_3}{\rightleftharpoons}} I$			
residue	$k_1 \times 10^3$	$k_2 \times 10^3$	$k_1 \times 10^3$	$k_2 \times 10^3$	$k_1 \times 10^3$	$k_2 \times 10^3$	$k_3 \times 10^3$	$k_4 \times 10^3$
Ala6	0.8	0.26			0.21	0.14	0.63	0.34
Gly10	1.2	0.27	0.9	0.22				
Gly13	1.3	0.22			0.49	0.16	0.71	0.27
Gly25	1.3	0.20	1.1	0.17				

[a] Fits are performed to the monomer decay only, to monomer and trimer data and to monomer, intermediate and trimer data.

different kinetic models including intermediate on-pathway, parallel, and off-pathway were examined. The best fit was found for the off-pathway intermediate folding model, and the slow second phase for monomer to trimer folding was fit with first-order rate constants that were on the order of 10^{-4} s^{-1} (Table 3). The rate constants of trimer formation, k_1, indicate that Gly25 folds faster than Gly13 and Ala6. In addition, Ala6 trimer folds 2.5 times slower than Gly13 trimer, whereas the rates constants of intermediate folding, k_3, for Ala6 and Gly13 are similar and on the order of 10^{-4} s^{-1}.

Characterization of the Kinetic Intermediate. To establish the nature of the observed kinetic intermediate, ^{15}N NOE relaxation experiments were performed on the protein during the refolding process (Figure 7). Low temperature conditions were chosen such that the observed nonnative species was highly populated during the experiment and changed only by approximately 20−30% for the duration of the 18 h experiment. ^{15}N NOE relaxation data was obtained at the same time on the monomer form and the native trimer form as a control. Previous relaxation data, including R_1, R_2, and NOE have indicated that the triple-helix form of T1−892 is a rigid uniform rod in solution. From NOE data in Figure 7, it is clear that the dynamics of a large percent of nonnative species are very similar to those of the native trimer form rather than the monomer form, indicating that most of the nonnative form is quite rigid. The apparent line widths of the trimer and intermediate peaks are nearly identical, suggesting that the intermediate aggregation state is trimer-like. In addition, the chemical shifts of the nonnative forms of Ala12, Gly13, and the major parts of Ala6 are very close to those of the native triple helix, suggesting that the nonnative form that is observed in the kinetics experiments is similar in packing and dynamics to the native triple-helix form.

DISCUSSION

NMR characterization of the cis−trans isomerization role in the folding of peptide T1−892 allows integration of collagen triple-helix folding into general concepts derived from globular protein studies. For globular proteins with one proline residue, 85−95% of the population contains only trans residue and will undergo fast folding, while 5−15% of the population will fold slowly, with the rate-limiting step being the cis−trans interconversion of an X−Pro bond (*43*). In contrast, for proline-rich collagen molecules, e.g., type α1(I) collagen chain with 236 Gly−Pro and Pro−Hyp bonds, only one molecule in 10^{11} of the monomer population contains all trans residues. In the imino acid rich triple-helix peptide with nine Gly−Pro and four Pro−Hyp bonds studied here, only 20% of the equilibrium monomer unfolded state has all trans residues. In addition, three monomer chains must

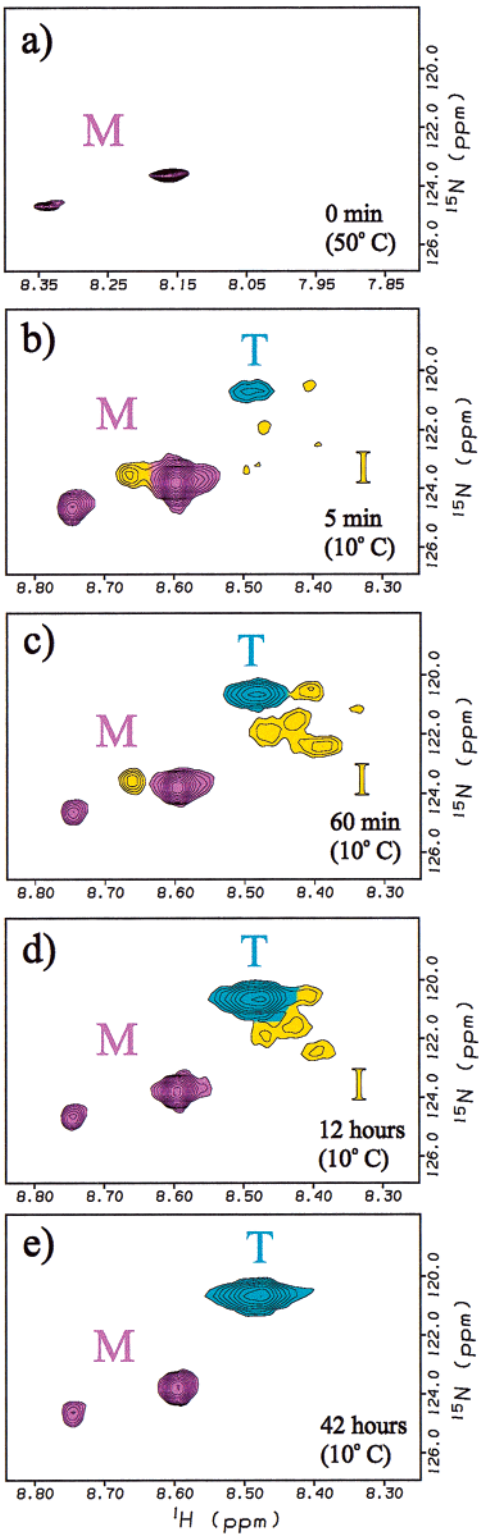

FIGURE 6: ^{1}H-^{15}N HSQC spectra of ^{15}N-labeled Ala6 in T1−892c obtained during a real-time NMR folding experiment. Monomers are shown in cyan, intermediates shown in yellow, and trimers shown in blue. (a) Well-resolved spectrum of thermally denatured monomer state at 50 °C recorded before the initiation of folding, (b) spectrum recorded at 10 °C after 5 min, (c) spectrum recorded at 10 °C after 60 min, (d) spectrum recorded at 10 °C after 12 h, and (e) equilibrium spectrum at 10 °C obtained after 42 h of folding. The large difference of resonance line widths in the ^{15}N dimension in the spectra recorded at 50 and 10 °C was mainly due to the experimental conditions and postacquisition transformation. Since only 32 complex data points in the ^{15}N dimension were acquired in spectra at 10 °C, they required much stronger postacquisition line broadening compared with the spectrum at 50 °C recorded with 128 complex points the ^{15}N dimension.

come together to form a triple helix, requiring all trans bonds in three chains. Thus, imino acid cis−trans isomerization is a dominant and integral feature of the folding pathway of the triple helix.

cis−trans Isomerization of Monomer Form. The equilibrium unfolded state of peptide T1−892 is a heterogeneous mixture of interconverting monomer species undergoing cis−trans isomerization at X−imino acid bonds, similar to that seen for other collagen-like monomer peptides (*37*). The percentages observed at each of the five sites in the peptide are consistent with values reported for Gly−Pro (14%) and Pro−Hyp (7%) in small model peptides, as well as averages reported for collagen (*38*, *44*). These observations indicate that the cis−trans isomerization at each site is independent of its neighboring sequence. Given the equilibrium values, 20% of the total unfolded population of T1−892 is calculated to be in the all trans form (Figure 8a), while 80% of the unfolded state will have one or more cis-imino acid sites. Monitoring the kinetics of folding of individual residues in different regions along the peptide gives information about the folding pathway.

Nucleation. The observation that Gly25-folds faster than the central and N-terminal residues indicates that folding begins at the C-terminus for this peptide. The Gly25 residue is located in the $(GPO)_4$ domain, and previous data suggest that the Gly-Pro-Hyp rich repeats can act as a nucleation site for triple-helix formation (*45*, *46*). The kinetics of folding of residues Gly25 are biphasic, with a fast initial phase that cannot be measured by NMR and a slower phase, with a first-order rate constant of 10^{-3} s^{-1} for both monomer decay and trimer formation. The slow folding rate is in good agreement with previously reported rates for cis−trans isomerization at Gly−Pro bonds, consistent with cis−trans isomerization as the rate-limiting step. Calculation of the cis−imino acid bonds present in the unfolded state for $(GPO)_4$ predicts that, for a statistical distribution of unfolded conformations (Figure 8b), 42% of the monomer chains contain all trans peptide bonds, while 58% have one or more cis-imino acid peptide bonds. The experimental data shows that the fast phase in the kinetics of folding of Gly25 has approximately 47% of the population folded, consistent with the theoretical 42% trans form which could form a stable nucleus quickly and represent the fast-folding population. The slow phase of Gly25 represents molecules that nucleate slowly following mostly one cis−trans isomerization event, since about 40% of the molecules contain a single cis imino acid bond. (Figure 8b). A model for nucleation that is consistent with the biphasic kinetics of Gly25 is proposed and illustrated in Scheme 2, where M, M^T, M^C, M^{CC}, M^{CCC},

Scheme 2

$$M^T \xrightarrow{fast} N \qquad (M^T = 0.42\ M)$$

$$M^C \xrightarrow{slow} M^T \xrightarrow{fast} N \qquad (M^C = 0.39\ M)$$

$$M^{CC} \xrightarrow{slow} M^C \xrightarrow{slow} M^T \xrightarrow{fast} N \qquad (M^{CC} = 0.16\ M)$$

$$M^{CCC} \xrightarrow{slow} M^{CC} \xrightarrow{slow} M^C \xrightarrow{slow} M^T \xrightarrow{fast} N \qquad (M^{CCC} = 0.05\ M)$$

and N denote the whole ensemble of monomers, monomers with all trans bonds, with one, two, and three cis bonds, and nucleated species, respectively. The initial proportion of these species are given in parentheses.

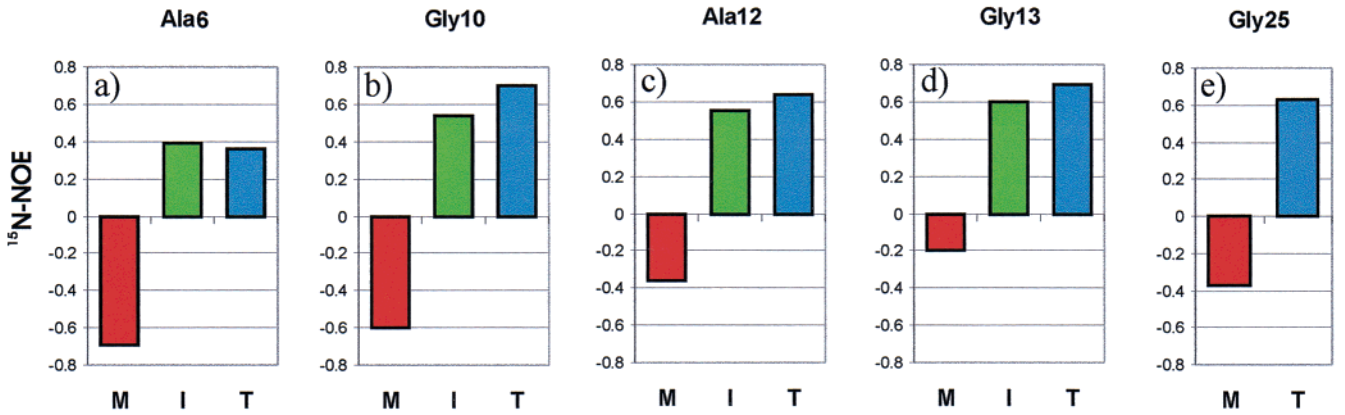

FIGURE 7: ^{1}H-^{15}N heteronuclear NOE of labeled residues in T1−892a and T1−892b peptides at 10 °C: Ala6 (a), Gly10 (b), Ala12 (c), Gly13 (d), and Gly25 (e). Note NOE for monomer (M) peaks are negative, whereas NOE for the trimer (T) and the majority of intermediate (I) peaks are positive.

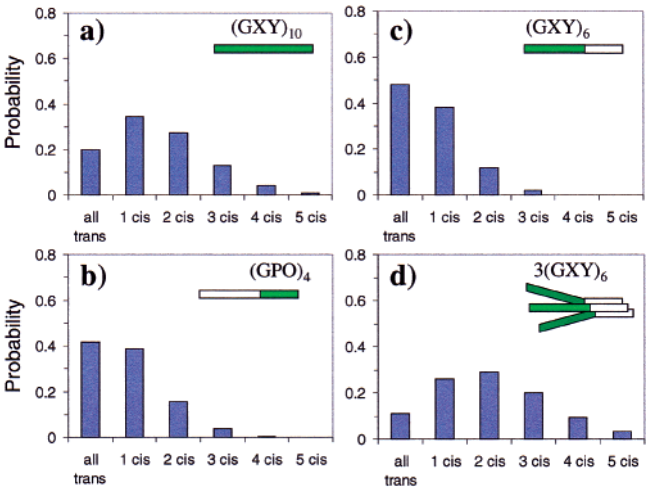

FIGURE 8: Statistical probabilities of states with different number of cis-X-Pro bonds for the whole T1−892 monomer sequence (a), for the C-terminal $(GPO)_4$ region of T1−892 (b), for the N-terminal $(GXY)_6$ region of monomer T1−892 (d), and for the three chains of T1−892, which nucleated into the trimer form at $(GPO)_4$ region (c). Calculations were carried out based on the assumption that each Gly-Pro and Pro-Hyp bond has 14% cis-Pro and 7% cis-Hyp, respectively.

Folding of T1−892 is initiated from the C-terminal $(GPO)_4$ domain, even though this domain has only 40% probability of all trans peptide bonds, and three of these must participate in nucleation. The N-terminal end of T1−892 has a lower imino acid content and slightly higher all trans population than the C-terminus, but nucleation does not take place at this end. This suggests that the highly restricted ϕ, ψ conformational space of imino acids in the Gly-Pro-Hyp repeating C-terminal region is critical to initiating the triple-helix nucleation step. The presence of imino acids in the X or Y position of the Gly-X-Y sequence has two effects. First, the ϕ space of the X or Y position is more restricted relative to a nonimino acid. Second, an imino acid in the X and Y position restrict the conformational space of the preceding residue such that only ψ angles of 120° are populated (*47*, *48*). Similar ψ angles (140 ± 20°) are found in the collagen triple helix, so Gly-Pro-Hyp sequences in the monomer form may be correctly preformed for triple helix formation. This makes them better nucleation sites compared to the more conformationally divergent Gly-Pro-Y sequences at the N-terminal end. In contrast to folding of globular proteins, where removal of Pro residues eliminates the slow folding step, elimination of imino acid residues from a collagen triple helix would prevent the nucleation process and make the structure unstable. This confirms early proposals that regions rich in Pro and Hyp would act as nucleation sites, because the ϕ,ψ angles of the rigid imino acid ring are very close to those of polyproline II and those of the final collagen triple helix (*49*) and is consistent with the strings of Gly-Pro-Hyp or Gly-X-Hyp sequences often seen at the C-terminus of the collagen $(Gly\text{-}X\text{-}Y)_n$ region (*11*).

Propagation. Nucleation of collagen is followed by propagation from the C- to N-terminal direction (*12*, *13*). Similarly to nucleation, propagation kinetics represented by Ala6 and Gly13 are biphasic. The fast phase cannot be monitored and the slow phase has a first-order rate constant of 10^{-4} s^{-1}, which is somewhat slower than expected for cis−trans isomerization. The slower rate may arise because the propagation rate is limited first by the slow cis−trans step during nucleation and then by sequential cis−trans events as the triple helix folds from the C-terminus to the N-terminus. A propagation mechanism for the triple helix folding is proposed and encompasses fast propagation trans conformations that represent the fast phase and slow interconverting cis conformations that represent the slow phase (Figure 9). To describe propagation events, one must consider the statistics of cis−trans isomerization for three chains. There are five Gly−Pro bonds in each chain after the nucleation step; therefore, only 10% percent of the molecules are in the all trans conformation and can fold fast to the final all trans form. The slow phase represents 90% of the molecules, which must undergo one or more cis−trans isomerization events (Figure 8d). The closer a residue is to the N-terminus end point, the more Gly−Pro residues must be traversed by the triple helix before that residue is folded and the more possibilities of different cis−trans branch points and complicated pathways. For example, there are three Gly−Pro sites C-terminal to Ala6 (outside the nucleation domain), and the triple helix may propagate to Ala6 through zero (26%), one (44%), two (25%), or three (5%) cis isomerization events, whereas there is only one Gly-Pro site C-terminal to Gly13, which therefore undergoes only zero (64%) or one (31%) isomerization event (forms with occupancy less than 5% were neglected). The rates of Ala6 and Gly13 trimer formation are related to the number of

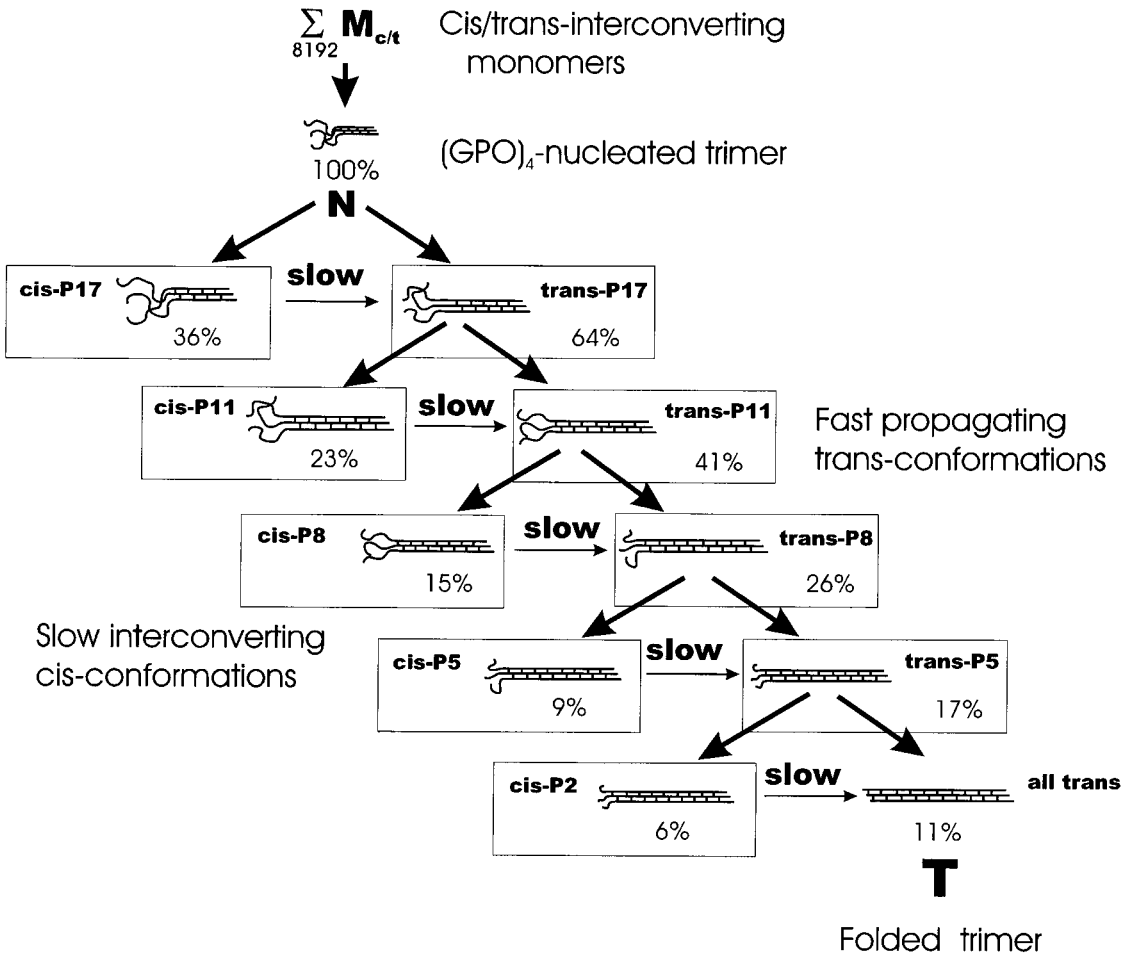

FIGURE 9: Schematic illustration of the proposed folding mechanism for proline-rich T1−892 peptide. The monomer state is represented by the ensemble of interconverting conformations, where the number of conformations is equal to 8192 (2^{13}, 13 is the number of different X-Pro and Pro-Hyp bonds in T1−892). After the nucleation of three chains at the C-terminal $(GPO)_4$ region, the folding propagates in a zipper-like manner from the C- to N-terminus. The trans forms undergo a fast folding propagation, whereas forms with a cis Pro at either Pro5, Pro8, Pro11, or Pro17 are incapable of forming trimer and are engaged in a slow interconvertion to the trans form. The percentages of the cis forms were calculated based on the assumption that there are 100% nucleated species and that each Gly−Pro bond in the nucleated trimer has a 14% probability of a cis-Pro.

respective cis−trans interconversion events. Gly13 has approximately 2.5 times more cis interconversion events than Ala6. The experimental observation of a 2.5 times slower rate of folding of trimer for Ala6 relative to Gly13 is consistent with the cis form statistics of the possible pathways for reaching Gly13 versus Ala6. The NMR results are consistent with the molecule undergoing a zipper-like propagation process from the C- to N-terminus limited by statistical progression of cis−trans isomerization events. In addition, the first data point in the trimer formation which represents the all trans burst phase shows a relative difference of 2.3 between Ala6 and Gly13 consistent with the ratio of 64/26 expected for the all trans form.

Intermediate. Transient intermediates have been seen in real-time folding studies of globular proteins (*50*, *51*), but have not been observed to date in the folding of triple-helical peptides (*27*). Therefore, it was surprising to find indications of an intermediate in the folding of the T1−892 triple-helical peptide. The chemical shift dispersion and the ^{15}N relaxation experiments both support a rigid triple-helical nature for the intermediate, similar to that seen in the native state. The folding and unfolding rates from monomer to intermediate are similar to those seen for monomer to trimer, indicating the involvement of cis−trans isomerization in the formation of both these species. The best fit to the kinetics of Ala6 and Gly13 were an off-pathway model for the intermediate, but the limitation of this approach consists of its inability to incorporate our knowledge about monomer and intermediate as an ensemble of different conformations. Therefore, it is quite possible that the nature of the intermediate is more complex and heterogeneous. Possibilities for the intermediate conformation may include misstraggering or altered supercoiling. Future analysis will help determine the character of the intermediate.

Relation to Collagen Folding. The folding properties deduced from NMR studies on the peptide T1−892 are relevant to collagen folding. Although the studies reported here were carried out on a self-associating model peptide, the high concentrations necessary for NMR studies overcome the diffusion problem. These results relate to the folding of collagen in vivo where three associated globular C-propeptides provide a tethered end that constrains the Gly-X-Y sequences such that nucleation sequences of the form Gly-Pro-Hyp can initiate the triple-helix. At the C-terminus of the $(Gly\text{-}X\text{-}Y)_n$ domain of collagen α1(I) chain, there are five sequential Gly-Pro-Hyp triplets which act as the nucleation domain (*45*). The nucleation domain in peptide T1−892 is the C-terminal $(Gly\text{-}Pro\text{-}Hyp)_4$ region. Our NMR

studies show the nucleation process is inherently fast but that this process cannot take place until cis−trans isomerization leads to three chains with all trans residues. Again, propagation of the triple-helix in T1−892 is rapid, but limited by a statistical progression as it encounters multiple Gly−Pro bonds. About one-third of all X residues is Pro and one-third of all Y residues is Hyp for the $(Gly\text{-}X\text{-}Y)_{338}$ in type I collagen, suggesting that a step by step progression, one or two triplets at a time, is an intrinsic feature dictated by the frequent encounters with cis-imino acids. In collagen, the process of cis−trans isomerization is accelerated by the enzyme prolyl cis−trans isomerase, but it is still slow (on the order of 5−10 min) (*13*, *52*). We suggest that a Gly substitution, such as found in Osteogenesis Imperfecta, slows down the fast step of folding without affecting cis−trans isomerization, and the NMR techniques described here can be applied to characterize the abnormal folding resulting from such mutations.

ACKNOWLEDGMENT

We would like to acknowledge helpful discussions with Konrad Beck, John Ramshaw, and Manjiri Bhate.

REFERENCES

1. Dyson, H. J., and Wright, P. E. (1998) *Nat. Struct. Biol. 5* (Suppl.), 499−503.
2. Dobson, C. M., and Hore, P. J. (1998) *Nat. Struct. Biol. 5* (Suppl.), 504−507.
3. Dill, K. A., and Chan, H. S. (1997) *Nat. Struct. Biol. 4*, 10−19.
4. Onuchic, J. N. (1997) *Proc. Natl. Acad. Sci. U.S.A. 94*, 7129−7131.
5. Rich, A., and Crick, F. H. C. (1961) *J. Mol. Biol. 3*, 483−506.
6. Nemethy, G., and Scheraga, H. A. (1982) *Biopolymers 28*, 1573−1584.
7. Bella, J., Eaton, M., Brodsky, B., and Berman, H. M. (1994) *Science 266*, 75−81.
8. Bella, J., Brodsky, B., and Berman, H. M. (1995) *Structure 3*, 893−906.
9. Brodsky, B. (1990) *Folding and higher order structure in fibrous proteins*, AAAS, Washington.
10. Kielty, C. M., Hopkinson, I., and Grant, M. E. (1993) *The Collagen Family: Structure, Assembly, and Organization in the Extracellular Matrix*, Wiley Liss, New York.
11. McLaughlin, S. H., and Bulleid, N. J. (1997) *Matrix Biol. 16*, 369−377.
12. Bächinger, H. P., Bruckner, P., Timpl, R., Prockop, D. J., and Engel, J. (1980) *Eur. J. Biochem. 106*, 619−632.
13. Bächinger, H. P., Bruckner, P., Timpl, R., and Engel, J. (1978) *Eur. J. Biochem. 90*, 605−613.
14. Harrison, R. K., and Stein, R. L. (1990) *Biochemistry 29*, 3813−3816.
15. Kuivaniemi, H., Tromp, G., and Prockop, D. (1997) *Hum. Mutat. 9*, 300−315.
16. Prockop, D. J., and Kivirikko, K. I. (1995) *Annu. Rev. Biochem. 64*, 403−434.
17. Byers, P. H. (1993) in *Connective Tissue and Its Heritable Disorders: Molecular, Genetic, and Medical Aspects* (Royce, P. M., and Steinmann, B., Eds.) pp 317−350, 351−407, Wiley Liss, New York.
18. Bächinger, H. P., Morris, N. P., and Davis, J. M. (1993) *Am. J. Med. Gen. 45*, 152−162.
19. Raghunath, M., Bruckner, P., and Steinmann, B. (1994) *J. Mol. Biol. 236*, 940−949.
20. Goodman, M., Feng, Y., Melacini, G., and Taulane, J. P. (1996) *J. Am. Chem. Soc. 118*, 5156−5157.
21. Mayo, K. H. (1996) *Biopolymers 40*, 359−370.
22. Feng, Y., Melacini, G., Taulane, J. P., and Goodman, M. (1996) *J. Am. Chem. Soc. 118*, 10351−10358.
23. Fields, G. B., and Prockop, D. J. (1996) *Biopolymers 40*, 345−357.
24. Baum, J., and Brodsky, B. (1999) *Curr. Opin. Struct. Biol. 9*, 122−128.
25. Baum, J., and Brodsky, B. (1999) in *Protein Folding: Frontiers in Molecular Biology* (Pain, R. H., Ed.) Oxford University Press (in press).
26. Baum, J., and Brodsky, B. (1997) *Folding Des. 2*, R53−R60.
27. Liu, X., Siegel, D. L., Fan, P., Brodsky, B., and Baum, J. (1996) *Biochemistry 35*, 4306−4313.
28. Kay, L. E., Keifer, P., and Saarinen, T. (1992) *J. Am. Chem. Soc. 114*, 10663−10665.
29. Fesik, S. W., and Zuiderweg, E. R. P. (1988) *J. Magn. Reson. 78*, 588−593.
30. Farrow, N. A., Muhandiran, R., Singer, A., Pascal, S. M., Kay, C. M., Gish, G., Shoelson, S. E., Pawson, T., Forman-Kay, J. D., and Kay, L. E. (1994) *Biochemistry 33*, 5984−6003.
31. Cavanagh, J., Fairbrother, W. J., Palmer, A. G., III, and Skelton, N. J. (1996) *Protein NMR Spectroscopy: Principles and Practice*, Academic Press, Inc, New York.
32. Shampine, L. F., and Gordon, M. K. (1975) *Computer Solutions of Ordinary Differential Equaiotns: The Initial Value Problem*, W. H. Freeman, San Francisco.
33. Bevington, P. R., and Robinson, D. K. (1992) *Data Reduction and Error Analysis for the Physical Sciences*, McGraw-Hill, New York.
34. Farrow, N. A., Zhang, O., Forman-Kay, J. D., and Kay, L. E. (1994) *J. Biomol. NMR 4*, 727−734.
35. Yang, W., Battinenni, M., and Brodsky, B. (1997) *Biochemistry 36*, 6930−6935.
36. Liu, X., Kim, S., Dai, Q.-H., Brodsky, B., and Baum, J. (1998) *Biochemistry 33*, 15528−15533.
37. Günther, H. (1995) *NMR spectroscopy*, John Wiley & Sons, New York.
38. Grathwohl, C., and Wüthrich, K. (1976) *Biopolymers 15*, 2025−2041.
39. Grathwohl, C., and Wüthrich, K. (1981) *Biopolymers 20*, 2623−2633.
40. Schmid, F. X. (1993) *Annu. Rev. Biophys. Biomol. Struct. 22*, 123−143.
41. Reimer, U., Scherer, G., Drewello, M., Kruber, S., Schutkowski, M., and Fischer, G. (1998) *J. Mol. Biol. 279*, 449−460.
42. Mayo, K. H., Parra-Diaz, D., McCarthy, J. B., and Chelberg, M. (1991) *Biochemistry 30*, 8251−8267.
43. Fischer, H., and Schmid F. X. (1990) *Biochemistry 29*, 2205−2212.
44. Sarkar, S. K., Young, P. E., Sullivan, C. E., and Torchia, D. A. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 4800−4803.
45. Bulleid, N. J., Wilson, R., and Lees, J. F. (1996) *Biochem. J. 317*, 195−202.
46. Bansal, M., and Ananthanarayanan, V. S. (1988) *Biopolymers 27*, 299−312.
47. Brant, D. A., Miller, W. G., and Flory, P. J. (1967) *J. Mol. Biol. 23*, 47−65.
48. Schimmel, P. R., and Flory, P. J. (1968) *J. Mol. Biol. 34*, 105−120.
49. Harrington, W., and von Hippel, P. H. (1961) *Arch. Biochem. Biophys. 92*, 110.
50. Balbach, J., Forge, V., vanNuland, N. A. J., Winder, S. L., Hore, P. J., and Dobson, C. M. (1995) *Nat. Struct. Biol. 2*, 865−870.
51. Balbach, J., Steegborn, C., Schindler, T., and Schmid, F. X. (1999) *J. Mol. Biol. 285*, 829−842.
52. Bächinger, H. P. (1987) *J. Biol. Chem. 262*, 17144−17148.

BI992584R