THE ROYAL SOCIETY

doi 10.1098/rstb.2000.0761

Nuclear magnetic resonance characterization of peptide models of collagen-folding diseases

Alexei Buevich and Jean Baum[*]

Department of Chemistry, Rutgers University, Piscataway, NJ 08855-0939, USA

Misfolding of the triple helix has been shown to play a critical role in collagen diseases. The substitution of a single Gly by another amino acid breaks the characteristic repeating (Gly-X-Y)$_n$ sequence pattern and results in connective tissue disease such as osteogenesis imperfecta. Nuclear magnetic resonance (NMR) studies of normal and mutated collagen triple-helical peptides offer an opportunity to characterize folding and conformational alterations at the substitution site, as well as at positions upstream and downstream of a Gly mutation. The NMR studies suggest that the local sequences surrounding the substitution site, and the renucleation sequences N-terminal to and adjacent to the substitution site, may be critical in defining the clinical phenotype of osteogenesis imperfecta. These studies may pave the way to understanding the mechanism by which a single Gly substitution in collagen can lead to pathological conditions.

Keywords: nuclear magnetic resonance; collagen; osteogenesis imperfecta; triple helix; peptides; dynamics

1. INTRODUCTION

Nuclear magnetic resonance (NMR) spectroscopy has played a major role in elucidating protein folding by determining mechanisms by which globular proteins proceed from the unfolded state to the fully folded state (Dobson & Hore 1998; Dyson & Wright 1996, 1998). The significant advances that have been made in understanding the folding of small monomeric globular proteins can be expanded to consider the folding of non-globular proteins, such as the collagen triple-helix motif by NMR (Dill & Chan 1997; Dobson *et al*. 1998; Matthews 1993). Collagen comprises a family of extracellular matrix molecules responsible for the integrity and mechanical properties of connective tissue, including bone, tendon, skin, cartilage and cornea (Kielty *et al*. 1993; Prockop & Kivirikko 1995). Mutations in collagen are the cause of various connective tissue diseases, including osteogenesis imperfecta (OI) and hereditary aortic aneurysm (Byers 1993; Kuivaniemi *et al*. 1997). Altered triple-helix folding and collagen fibril assembly have been implicated in the aetiology of these diseases, putting collagen in the context of protein folding and aggregation diseases (Kelly *et al*. 1997; Lansbury 1997). This report describes NMR approaches for studying the folding of native and mutant collagen-like triple-helical peptides, and the use of the resulting structural and folding information to begin to understand the basis of collagen-folding diseases.

2. COLLAGEN FOLDING

The linear (Gly-X-Y)$_n$ repetitive triple-helical collagen presents a rod-like 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 Pro and Hyp, 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, and includes an unusually high concentration of Pro and Hyp, but does not include a hydrophobic collapse. This peculiarity brings into focus folding features other than those encountered in globular proteins.

In a physiological context, the folding of collagen involves molecular assembly of three chains, followed by higher-order assembly to fibrils or other associated forms (Brodsky & Shah 1995; Kielty *et al*. 1993) (figure 2). Collagen is synthesized in a procollagen form, with N- and C-terminal globular propeptides flanking the (Gly-X-Y)$_n$ central domain (Kielty *et al*. 1993). 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-X-Y)$_n$ region (Bächinger *et al*. 1978, 1980; Harrison & Stein 1990; McLaughlin & Bulleid 1997). After nucleation of the three chains, the triple helix propagates with *cis–trans* isomerization of Pro and Hyp being rate-limiting steps (Bächinger *et al*. 1978, 1980). 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 (Kielty *et al*. 1993).

[*] Author for correspondence (baum@rutchem.rutgers.edu).

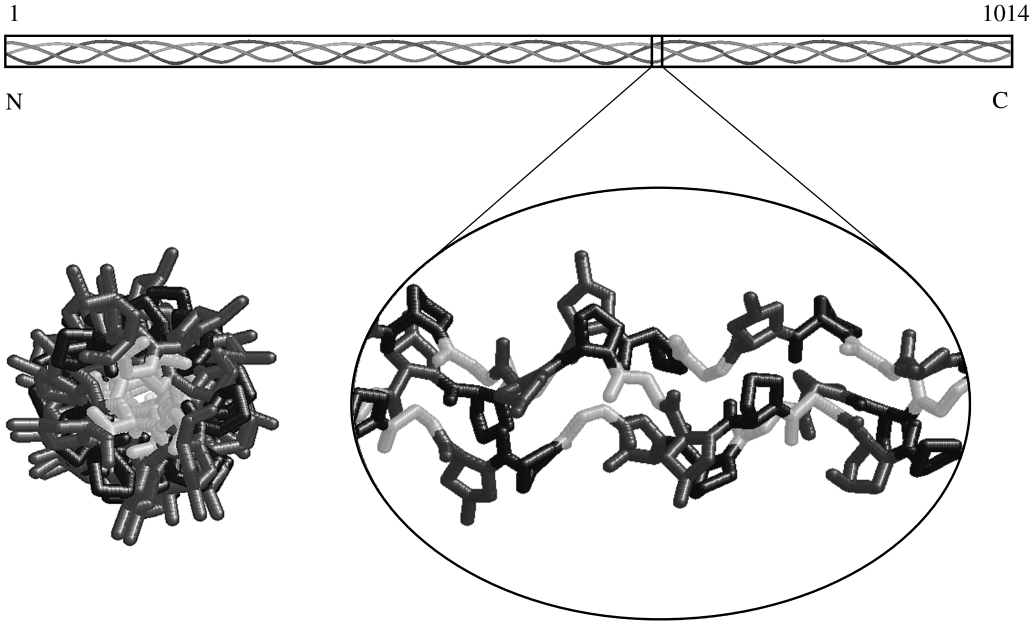

Figure 1. Schematic representation of the collagen triple-helix conformation. The triple-helical conformation consists of three polypeptide chains, each in a polyproline II-like helix, which are supercoiled about a common axis and hydrogen bonded together (Bella *et al*. 1994; Rich & Crick 1961). The 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-X-Y)$_n$. All Gly residues (light shading) are buried near a central axis, while the residues in the X and Y positions (dark shading) are largely exposed to solvent.

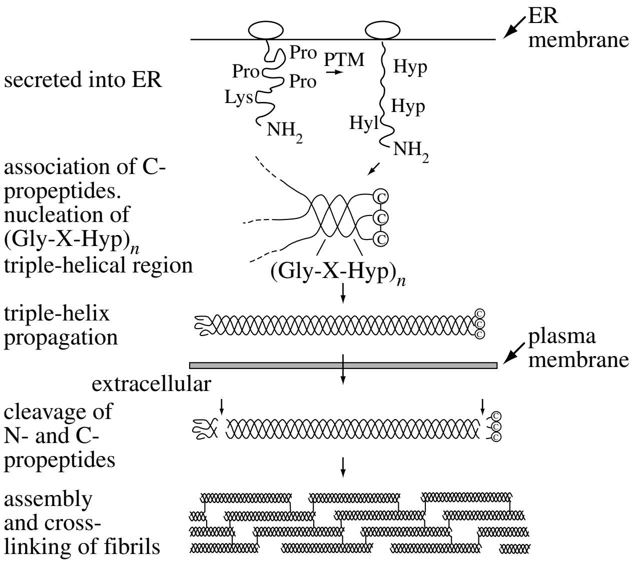
ER membrane
secreted into ER
Pro
Pro
Pro
Lys
PTM
Hyp
Hyp
Hyl
NH2
NH2
association of C-propeptides. nucleation of (Gly-X-Hyp)n triple-helical region
(Gly-X-Hyp)n
triple-helix propagation
plasma membrane
extracellular
cleavage of N- and C-propeptides
assembly and cross-linking of fibrils

Figure 2. Schematic drawing of collagen folding (adapted from Baum & Brodsky (1999)). Collagen is synthesized by ribosomes bound to the cytosolic side of the endoplasmic reticulum (ER) and then is secreted into the ER. Folding requires enzymes, such as prolyl and lysyl hydroxylases, for post-translational modification (PTM) of the unfolded chains. The folding of the collagen triple helix is a multistep process and involves the association of three C-propeptide domains, followed by nucleation and propagation to form the triple helix. After triple-helix formation, the molecules are secreted from the cell, the propeptides are cleaved and the triple helix associates to form fibrils. Covalent cross-links, mediated by lysyl oxidase, are formed between molecules in the 670 Å periodic fibrils, and provide the tensile strength needed for tissue function.

Although there is a general understanding of the folding of collagen, molecular descriptions of individual steps in triple-helix folding remain elusive. Given the Gly-X-Y repeating unit of the triple helix, the identity of the X and Y residues must determine the factors that are important in folding. In the human α1(I) chain of type I collagen there are 236 Gly-Pro and Pro-Hyp bonds, indicating that Pro is an integral part of the folding of the triple helix. The distribution of triplets along the collagen triple helix is variable, with regions of low imino-acid content interleaved with regions of high imino-acid content. In § 3, we discuss elements of how the Gly-X-Y triplet distribution determines the folding of the triple helix.

3. NUCLEAR MAGNETIC RESONANCE APPROACHES TO STUDY FOLDING OF TRIPLE-HELICAL PEPTIDES

NMR is a powerful method for studying protein folding as the characterization of unfolded and partially folded equilibrium states can be combined with NMR real-time folding experiments to obtain a detailed structural and kinetic picture of folding pathways (Dobson & Hore 1998; Dobson *et al*. 1998; Dyson & Wright 1996, 1998). A number of NMR approaches have been used to understand protein folding and to characterize folding intermediates in structural and kinetic terms. NMR studies of denatured states have helped to clarify the importance of early folding, while hydrogen-exchange kinetics have been used to monitor the location of protected amides during the kinetics of protein refolding (Dyson & Wright 1998; Roder & Shastry 1999). A complementary NMR approach to hydrogen-exchange

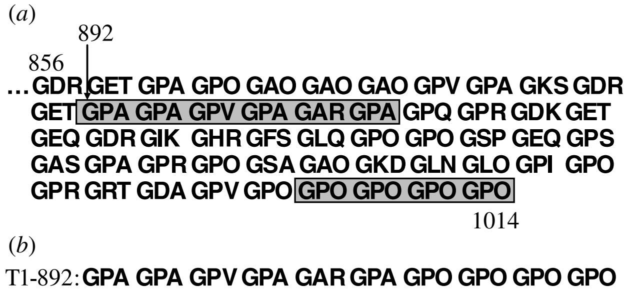
(*a*)
892
856
...GDR GET GPA GPO GAO GAO GAO GPV GPA GKS GDR
GET GPA GPA GPV GPA GAR GPA GPQ GPR GDK GET
GEQ GDR GIK GHR GFS GLQ GPO GPO GSP GEQ GPS
GAS GPA GPR GPO GSA GAO GKD GLN GLO GPI GPO
GPR GRT GDA GPV GPO GPO GPO GPO GPO
1014
(*b*)
T1-892: GPA GPA GPV GPA GAR GPA GPO GPO GPO GPO

Figure 3. Collagen-like triple-helix peptide design. (*a*) C-terminal portion of the α1 (I) collagen sequence. (*b*) T1–892 peptide sequence. The T1–892 peptide contains two segments from the α1 (I) chain of collagen. The N-terminal part of T1–892 is the 18 residue sequence starting from Gly892 of the α1 (I) chain and the C-terminal part of T1–892 is the $(GPO)_4$ segment from the C-terminal end of the α1 (I) chain. The amino-acid sequences that form the T1–892 peptide are shown in boxes in the α1 (I) collagen sequence (*a*).

and equilibrium experiments is to apply real-time kinetics to the folding of proteins. The ability to detect the loss of intensity of unfolded peaks or the increase of intensity of the folded peaks allows one to obtain a detailed kinetic picture of folding events (Baum & Brodsky 1997; Dobson & Hore 1998).

NMR studies of triple-helical model peptides offer an approach for isolating and characterizing individual events in collagen triple-helix assembly and for studying the influence of specific Gly-X-Y triplets on these steps. In contrast to collagen, which has not been crystallized and is difficult to isotopically label in bacterial expression systems, peptides have proved suitable for biophysical characterization by NMR, and by other spectroscopic methods, such as circular dichroism (CD) and X-ray crystallography (Baum & Brodsky 1997, 2000; Brodsky & Shah 1995; Mayo 1996).

Peptides that satisfy the stringent $(\text{Gly-X-Y})_n$ sequence constraint and have a high content of imino acids will form stable triple helices in aqueous solution and can be used to model different regions of a collagen chain (Fields & Prockop 1996; Goodman *et al.* 1998). Our approach is based on designing short 32–35 residue peptides, which model different properties of collagen and can be analysed by NMR spectroscopy at an atomic level. Repeating Gly-Pro-Hyp triplets are usually included at one or both ends of the sequence for stability, and other triplets denoted by Gly-X-Y are included to model collagen or understand basic principles. For example, figure 3 shows a portion of the amino-acid sequence from the human α1 (I) chain of type I collagen, starting from the C-terminal end to residue 856. The peptide design consists of combining the C-terminal $(GPO)_4$ region and an 18 residue sequence from residues 892 to 910. This peptide, denoted T1–892, is proposed to be a good model of the collagen chain as it incorporates the C-terminal (Gly-Pro-Hyp)-rich environment and a $(\text{Gly-X-Y})_n$ sequence that includes the 901 site of a Gly to Ser OI mutation (Yang *et al.* 1997). The folding of T1–892 is a very slow process, taking place in the order of minutes or even hours (Engel 1987; Liu *et al.* 1996). This slow folding of the triple helix allows direct experimental monitoring by NMR, which can define details of the process and lead to proposed mechanisms (Baum & Brodsky 1997, 1999).

(a) *Triple-helix folding: nucleation*

From the earliest studies on collagen folding, it was proposed that regions rich in Pro and Hyp and with *trans* peptide bonds would act as nucleation sites, because the Φ and Ψ angles of the rigid imino-acid ring are very close to those of polyproline II and those of the final collagen triple helix (Harrington & Von Hippel 1961). 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 that of a non-imino acid. Second, an imino acid in the X and Y position restricts the conformational space of the preceding residue such that only Ψ angles of 120° are populated (Brant *et al.* 1967; Schimmel & Flory 1968). Similar Ψ angles ($140° \pm 20°$) 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 with other Gly-X-Y sequences. 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. Examination of the amino-acid sequences of fibril-forming collagens shows that three to five Gly-Pro-Hyp sequences, or strings of Gly-X-Hyp triplets, are often seen at the C-terminus of the $(\text{Gly-X-Y})_n$ region (Buevich *et al.* 2000). Studies on recombinant type III collagen have shown that triple-helix folding does not take place when the $(\text{Gly-X-Hyp})_6$ region at its C-terminus is deleted, although a smaller number of Gly-X-Hyp repeats may be sufficient (Bulleid *et al.* 1997).

NMR spectroscopy provides a method for examining the nucleation process in detail. The slow rate of folding of the triple helix, together with the NMR assignments of both monomer and trimer peaks, sets the stage for monitoring the kinetics of folding directly (figure 4). The NMR assignments of peptide T1–892 are seen in the heteronuclear single quantum correlation (HSQC) spectrum with ^{15}N-labelled residues at positions Gly25 (at the C-terminal end), Ala13 (near the centre) and Ala6 (at the N-terminal end) (figure 4*a*). The spectrum indicates an equilibrium between monomer and trimer forms that are in slow conformational exchange on the NMR timescale. In addition, multiple monomer peaks are observed for each labelled residue arising from slow interconversion between *cis* and *trans* forms in the unfolded state. During the folding process of the model T1–892 peptide, the kinetics of folding were monitored by following the decrease in the intensity of the monomer peaks and/or the increase in the intensity of the trimer peaks (figure 4*b*) (Buevich *et al.* 2000). The observation that Gly25 folds faster than the central and N-terminal residues indicates that folding begins at the C-terminus for this peptide and that the C-terminal $(GPO)_4$ region is indeed a good nucleation domain. The kinetics of folding of residues Gly25 were biphasic, with a fast initial phase that cannot be measured by NMR, and a slower phase,

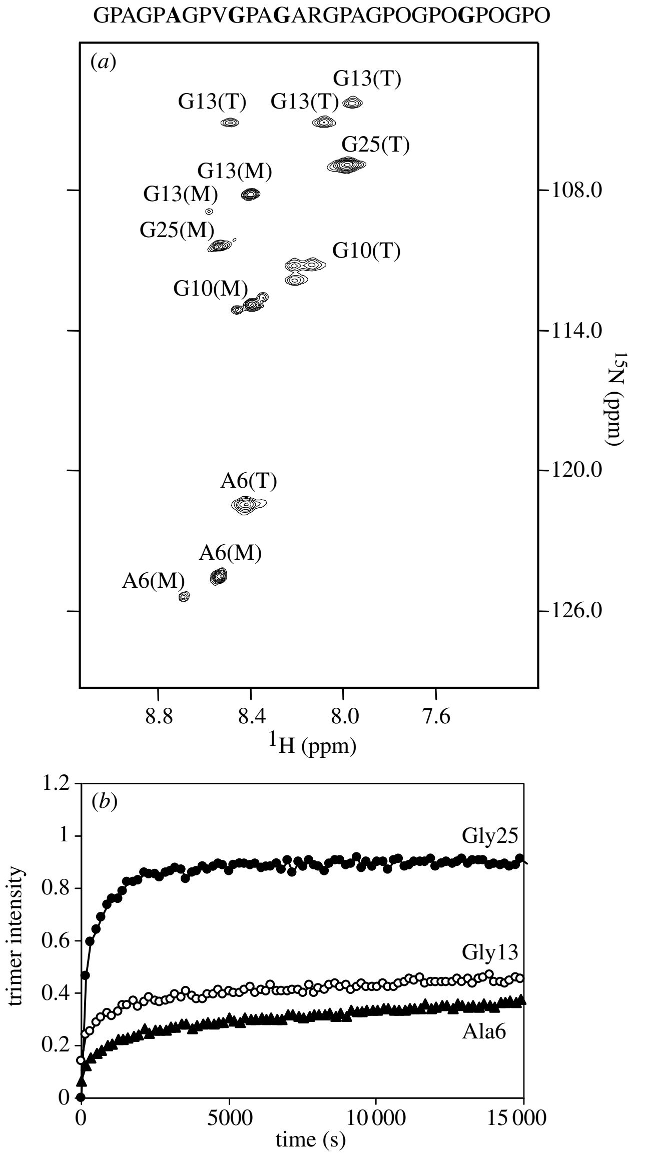

Figure 4. (*a*) ^{1}H–^{15}N HSQC spectrum of T1–892 (9 mM in H_2O:D_2O = 9:1, pH 2.3, 10 °C). The sequence for the peptide is shown on the top with the ^{15}N-labelled positions denoted in bold. In the spectrum, monomer and trimer peaks are denoted as M and T, respectively. (*b*) Real-time folding kinetics of T1–892 depicted by the time-dependent change in intensity of the trimer peaks for Gly25, Gly13 and Ala6 residues. Resonance intensities were measured as volume integrals of the signals in the ^{1}H–^{15}N HSQC spectra recorded every 3 min after initiation of refolding by a rapid temperature quench from 50 to 10 °C (the melting point of peptide T1–892 is 27 °C; Buevich *et al.* 2000). The rate of Gly25 folding exceeds the rates of Gly13 and Ala6 folding, suggesting that the triple-helix folding of T1–892 is initiated at the C-terminal GPO-rich region.

with a first-order rate constant of $10^{-3}\,\mathrm{s}^{-1}$ for both monomer decay and trimer formation. The slow folding rate was 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.

Statistical calculations of the *cis*–imino-acid bonds present in the unfolded state for the C-terminal-rich $(\mathrm{GPO})_4$ nucleation region predict that 42% of the monomer chains contain all-*trans* peptide bonds, while 58% have one or more *cis*–imino-acid peptide bonds (Buevich *et al.* 2000). The experimental data show that the fast phase in the kinetics of folding of Gly25 has *ca.* 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. A model for nucleation that is consistent with the biphasic kinetics of Gly25 was proposed and is illustrated in the following scheme:

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

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

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

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

where M, $\mathrm{M^T}$, $\mathrm{M^C}$, $\mathrm{M^{CC}}$, $\mathrm{M^{CCC}}$ 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 parenthesis. Thus, our NMR studies show that the nucleation process is inherently fast for the *trans* form, but it cannot be completed until *cis–trans* isomerization leads to three chains with all-*trans* residues.

(b) *Triple-helix folding: propagation*

After a stable triple-helix nucleus is formed, how does it propagate to the rest of the molecule? A series of elegant experiments performed in the late 1970s by Engel and co-workers suggest that the propagation of the collagen triple helix proceeds from the C-terminal nucleation site in a C- to N-terminal direction, via a zipper-like mechanism (Bächinger *et al.* 1978, 1980; Bruckner *et al.* 1978; Engel 1987). But it is not clear if the triple helix zips up one triplet at a time, or whether there are larger cooperative folding units, as suggested by calorimetry (Privalov *et al.* 1979). It is also not known whether the rate of propagation is relatively uniform, if there are kinetic intermediates, or if it slows down and speeds up, depending on the local amino-acid sequence. The rate-limiting step in triple-helix propagation has been shown to be *cis–trans* isomerization of the peptide bonds of imino acids (Bächinger *et al.* 1978, 1980). *In vivo* there is a *cis–trans* prolyl isomerase that catalyses this step (Bächinger 1987).

The propagation step can be monitored by NMR using specifically labelled triple-helical peptides. In T1–892, the propagation step was examined by monitoring the rate of folding of the labelled residues in the central and N-terminal region, Gly13 and Ala6, respectively. Similar to nucleation, propagation kinetics represented by Ala6 and Gly13 were biphasic. The fast phase cannot be monitored and the slow phase of monomer decay has a first-order rate constant of 10^{-3} to $10^{-4}\,\mathrm{s}^{-1}$, which is somewhat slower than expected for

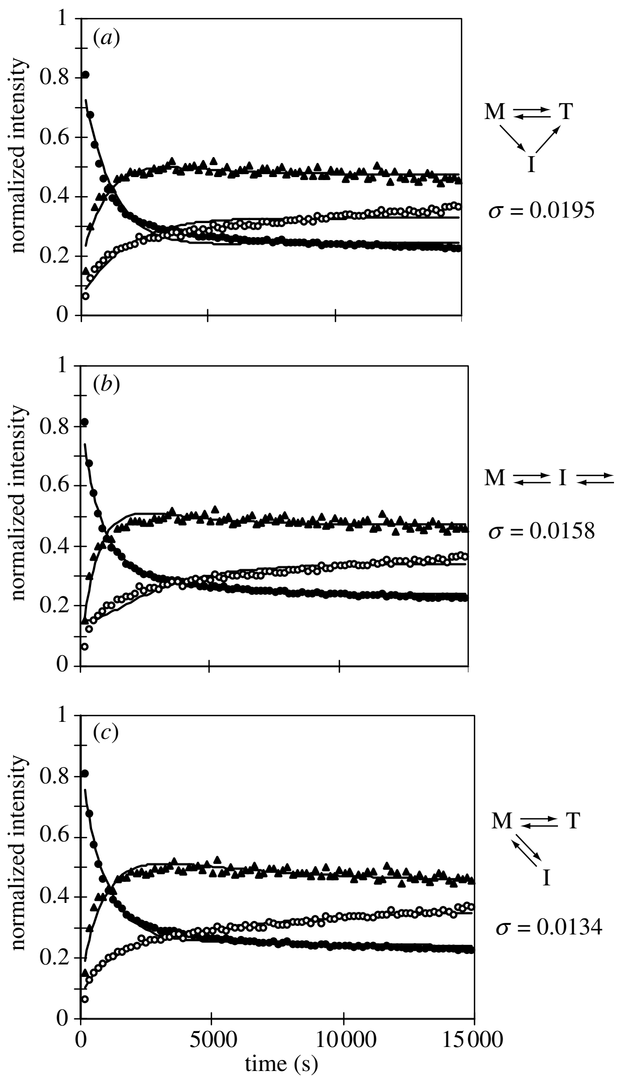

Figure 5. Analysis of the refolding kinetics of the T1–892 peptide at the Ala6 position by three kinetic models: (*a*) parallel pathways, (*b*) on-pathway intermediate and (*c*) off-pathway intermediate. Intensities of monomer (solid circles), intermediate (solid triangles) and trimer (open circles) peaks of Ala6 were measured from ^{1}H–^{15}N HSQC spectra recorded every 3 min after initiation of refolding by fast temperature quench from 50 to 10 °C. Normalization of intensities was performed as described earlier (Buevich *et al.* 2000). Three different kinetics models (shown on the right of the data) were fitted to the monomer (M), intermediate (I) and trimer (T) species. First-order rate constants were assumed in all models and the root mean square deviations of the analysis are indicated. The best fit (solid line) was achieved with the third model (*c*), indicating that the intermediate is best described as an off-pathway intermediate in the kinetics of trimer formation.

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. The zipper-like nature of the propagation is confirmed by the slower rate of folding of Ala6 versus Gly13 (figure 4*b*). Folding intermediates were observed for the residues in the propagation domain. While analysing the kinetics of monomer, trimer and intermediate peaks of T1–892, several kinetic mechanisms were examined (figure 5). These mechanisms include parallel pathway, on-pathway and off-pathway intermediate species. The best fit was achieved with the model that includes off-pathway intermediate species, suggesting the formation of misfolded non-native forms of T1–892 in the process of triple-helix assembly.

To establish the dynamics of the observed kinetic intermediate, ^{15}N nuclear Overhauser effect (NOE) relaxation experiments were performed on the protein during the refolding process. These experiments showed that the dynamics of the intermediate are very similar to that of the native trimer. The apparent line width of the trimer and intermediate peaks are nearly identical, suggesting a trimer-like conformation for the intermediate. NMR studies on peptides with ^{15}N-enriched amino acids at specific sites is an important method for defining the relationship between folding and amino-acid sequence (Baum & Brodsky 1997, 1999).

We have designed a model peptide of the triple helix with a C-terminal (Gly-Pro-Hyp)-rich region and six Gly-X-Y triplets from natural collagen at the amino end to characterize individual events in collagen triple-helix folding. We have found that folding is initiated in the C-terminal region and propagates in a zipper-like fashion to the N-terminal end suggesting that the peptide is a good model for collagen folding (Buevich *et al.* 2000). Now, with a good model system for folding and optimized NMR experiments, we are placed to understand the effect of mutations resulting in disease on the conformation, dynamics and folding of the triple helix.

4. MISFOLDING OF THE COLLAGEN TRIPLE HELIX IN DISEASE

A number of hereditary connective tissue diseases have been associated with mutations in the collagen triple helix (Byers 1993; Kuivaniemi *et al.* 1997). The most common mutation is a single base change that leads to the replacement of a Gly by another residue anywhere along the (Gly-X-Y)$_n$ sequence, breaking the repeating tripeptide pattern. The best-characterized case of Gly substitutions is found in OI, or brittle bone disease, in which there is defective mineralization of bones (Byers 1993; Kuivaniemi *et al.* 1997). The severity of the disease varies widely, ranging from mild cases with multiple fractures to perinatal lethal cases. Type I collagen is the major collagen in bone, and over 250 distinct Gly substitution mutations have been identified as causes of individual OI cases. The residues substituted for Gly are most frequently Ser and Cys, followed by significant numbers of Arg, Asp, Val, Ala and Glu, all of which are generated by a single base change (Byers 1993; Kuivaniemi *et al.* 1997) (figure 6). Mutations are found all along the collagen chain, suggesting that the loss of a Gly at any site in the triple helix has pathological consequences. It is interesting to note that certain mutations result in mild mutations whereas others tend to be more lethal. For example, Asp mutations are all lethal whereas Ala mutations are both non-lethal and lethal.

It has been suggested that the degree of OI phenotype is related to the degree to which folding is perturbed by the mutation in collagen (Bonadio & Byers 1985; Byers

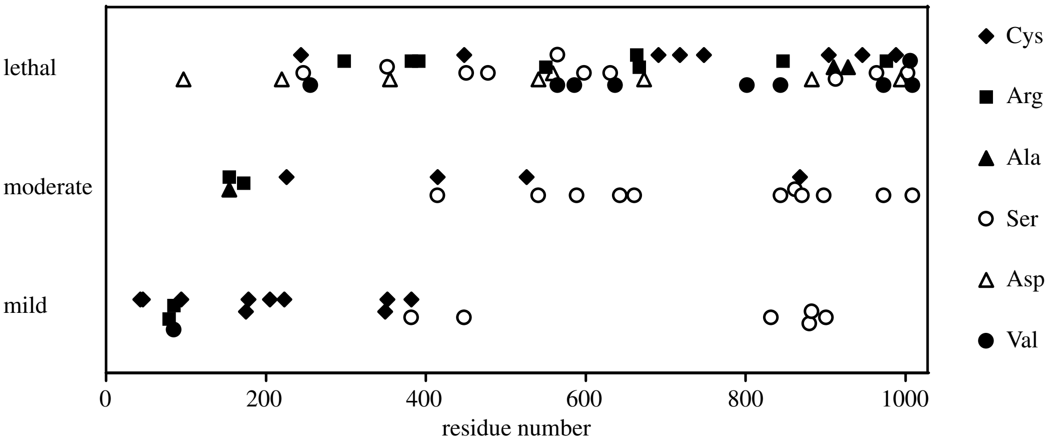

Figure 6. Positions along the α1 chain of type I collagen of Gly→X mutations found in cases of OI (Beck *et al.* 2000). Residue-specific Gly→X mutations are noted with different symbols. Lethal, moderate and mild phenotypes of OI disease are shown separately.

POGPOGPOGPOGPO**A**POGPOGPO**G**POGPOG GPAGPAGPV**S**P**AG**ARGPAGPOGPO**G**POGPO

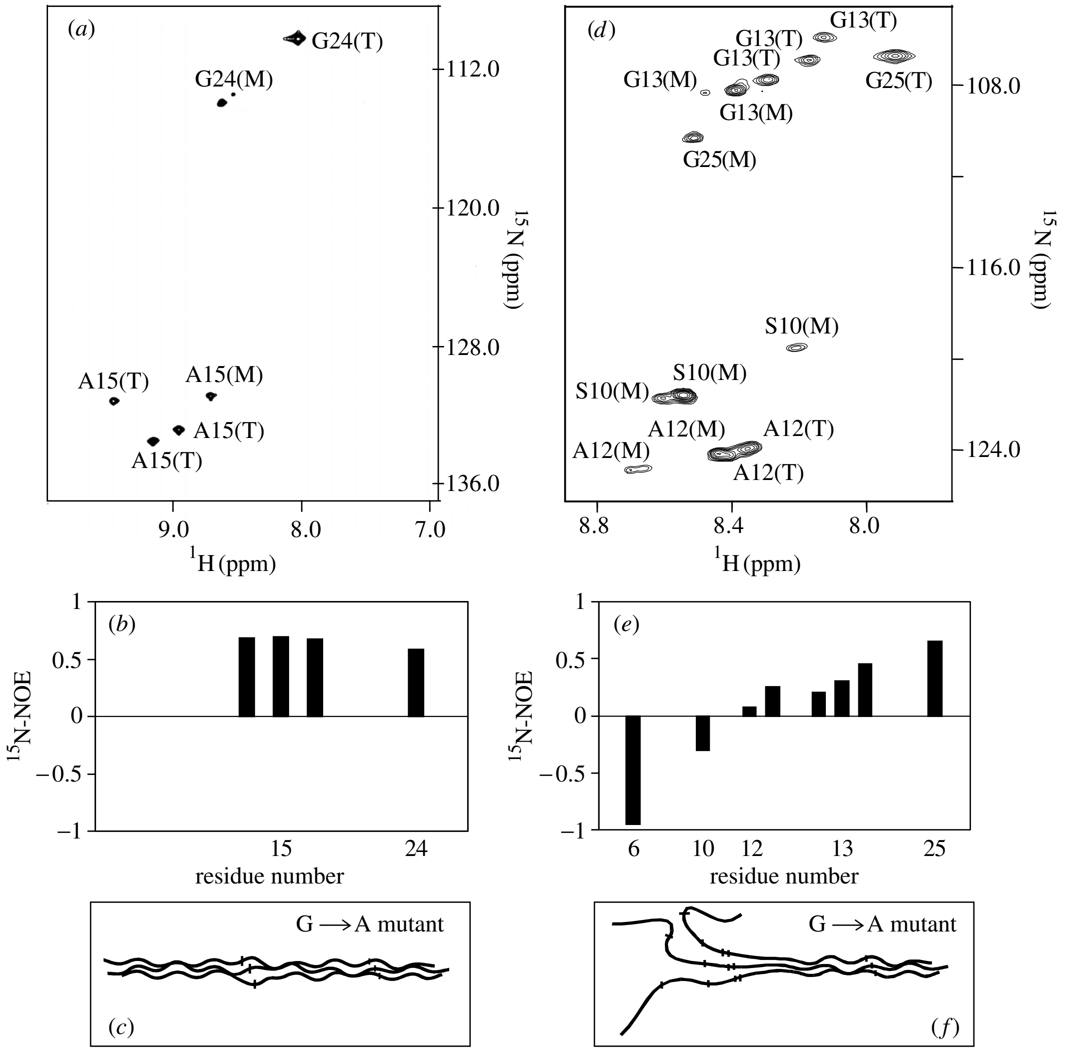

Figure 7. NMR characterization of two model OI peptides: Gly→Ala (*a–c*) and T1–892[G901S] (*d–f*). The sequences for each peptide are shown above with the ^{15}N-labelled positions denoted in bold. (*a*,*d*) ^{1}H–^{15}N HSQC spectra recorded under equilibrium conditions at 10 °C for the Gly→Ala and T1–892[G901S] peptides, respectively. Resonances of monomer and trimer peaks are assigned as M and T, respectively. (*b*,*e*) ^{1}H–^{15}N heteronuclear NOE dynamics experiments indicating a rigid triple helix along the length of the Gly→Ala peptide including at the mutation site, whereas the T1–892[G901S] peptide shows a gradient of mobility from the C- to N-terminal end with essentially monomer-like dynamics at the Gly→Ser mutation site. (*c*,*f*) Schematic diagrams of the triple-helix conformation of the Gly→Ala and T1–892[G901S] peptides suggested by the X-ray crystal structure and NMR data (Bella *et al.* 1994; Liu 1997; Liu *et al.* 1998).

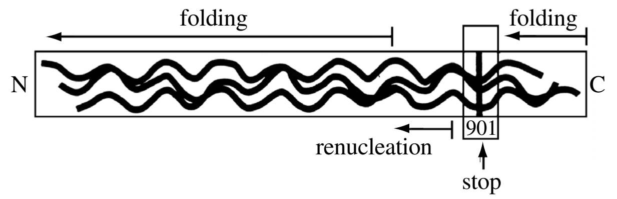

Figure 8. Schematic diagram of the folding of OI collagen containing a Gly→Ser mutation at position 901. After C-terminal nucleation of collagen, the folding propagates as far as the mutation site, propagation is then terminated, and folding stops. For folding to continue in collagen, renucleation must occur after the mutation site.

1993; Engel & Prockop 1991; Raghunath *et al.* 1994). All OI collagens show increased levels of lysyl hydroxylation and glycosylation N-terminal to the mutation site. This increase has been suggested to result from abnormal folding, as post-translational enzymatic modifications can occur only on the unfolded chain. Thus, the appearance of increased amounts N-terminal to the Gly substitution site suggests that the mutation delayed triple-helix formation at the site, extending the time during which collagens can be modified by enzymes which may act only on the unfolded state. Direct monitoring of the folding of several OI collagens using enzymatic digestion has confirmed a slower folding rate (Raghunath *et al.* 1994). Peptides have been designed to investigate the role of the local sequences surrounding the substitution site and NMR approaches have been used in order to define the nature of the folding defect at the site and upstream and downstream of the site.

(a) *NMR studies of mutated peptides*

The effect of collagen mutations on the triple-helix conformation, dynamics and folding have been investigated by NMR approaches applied to model triple-helical peptides with specific amino-acid labels (Baum & Brodsky 1997, 1999, 2000; Brodsky & Ramshaw 1997). Gly substitutions were introduced first into Gly-Pro-Hyp repeating peptides. For example, the peptide (Pro-Hyp-Gly)$_4$-Pro-Hyp-Ala-(Pro-Hyp-Gly)$_5$ with a single Gly→Ala change (denoted Gly→Ala peptide) was shown to have reduced stability by CD spectroscopy relative to (Pro-Hyp-Gly)$_{10}$ (Long *et al.* 1992, 1993). The X-ray crystal structure showed that the peptide adopts a triple-helical conformation with local untwisting and breaking of hydrogen bonds at the Ala substitution site (Bella *et al.* 1994). NMR studies of the Gly→Ala peptide (figure 7*a*,*b*) were conducted to examine the conformation, dynamics and folding in solution (Liu 1997). Assignments for the Gly→Ala peptide indicate that the Ala15 residue gives rise to three trimer peaks, while the Gly24 residue gives rise to one trimer peak (figure 7*a*). The three peaks for Ala15 arise because the three residues are not in a symmetrical environment due to the staggering of the three chains. ^{15}N-relaxation experiments were performed in order to examine the mobility of the ^{15}N-labelled residues on the picosecond to nanosecond time-scale (figure 7*b*). The ^{15}N-NOE data indicate that the Ala residues are rigid on these time-scales and that they are similar to the NOEs at the terminal Gly-Pro-Hyp region of the Gly→Ala peptide and to the NOEs observed for the fully triple-helical and rigid (Pro-Hyp-Gly)$_{10}$ (Liu 1997). Real-time folding experiments have shown that the C-terminal Gly folds faster than the central Ala mutation site, indicating that the peptide may be nucleating at the C-terminal end of the peptide and folding more slowly in the central region. Two-dimensional NOE experiments performed on this peptide identified NOE cross-peaks at the Ala mutation site that were consistent with the bulge seen in the X-ray crystal structure at the Ala position (Bella *et al.* 1994). In particular, two interchain NOEs between ^{1}Ala(NH) and ^{2}Ala(C$_\beta$H), and ^{2}Ala(NH) and ^{3}Ala(C$_\beta$H) gave solution distances of 2.55 and 2.72 Å as compared with 2.58 and 2.76 Å observed in the X-ray crystal structure. Therefore, the NMR conformational and dynamics data in solution suggest that the Ala15 position is incorporated into the triple helix in a manner similar to the X-ray structure (figure 7*c*).

More realistic model peptides were synthesized containing a sequence of the α1(I) chain of type I collagen in which a non-lethal OI mutation is found (Yang *et al.* 1997). A comparison of the NMR features of the peptides Gly→Ala and T1–892[G901S] are shown in figure 7. The sequences of both peptides are given and the labelled positions are shown in bold. Assignments have been made for the five labelled positions in T1–892[G901S], allowing the monitoring of the dynamics and folding at these sites (figure 7*d*). As opposed to the Gly→Ala peptide, for which the Ala mutation site is both incorporated into the triple helix and retains its rigid nature, NMR studies on T1–892[G901S] have indicated that C- to N-terminal propagation is stopped by a Gly→Ser mutation (Liu *et al.* 1996). The magnitude of the ^{15}N-NOE peaks decreases in a gradient-like manner from the C- to N-terminal direction. ^{15}N-relaxation studies of T1–892[G901S] mutant showed that while the C-terminal end of the peptide retains the triple helix, the Ser substitution site and residues N-terminal to it exhibit the mobility of a random coil (figure 7*e*). The introduction of a Gly→Ser substitution into the T1–892 peptide induced an asymmetrical disruption of the uniform triple helix (figure 7*f*).

The NMR data on the two mutant peptides suggest that the context of the amino-acid residues that surround the mutation site may be important in defining the folding defect. The presence of the rigid Gly-Pro-Hyp triplets around the Gly→Ala mutation appears to force the incorporation of the Ala into the triple helix, whereas the less rigid Gly-Pro-Y residues that surround the Gly→Ser mutation in peptide T1–892 do not provide an environment that readily promotes triple-helix formation (figure 7*c*,*f*).

(b) *Renucleation in collagen*

The termination of triple-helix propagation at the mutation site in model peptide T1–892[G901S] is consistent with the increased post-translational modification level that is N-terminal to mutation sites seen in OI collagens, suggesting that the model peptides are beginning to define the nature of the folding defect. In collagens, however, full-length triple-helical molecules can be formed, indicating that folding can continue past the substitution site. The peptide results suggest that

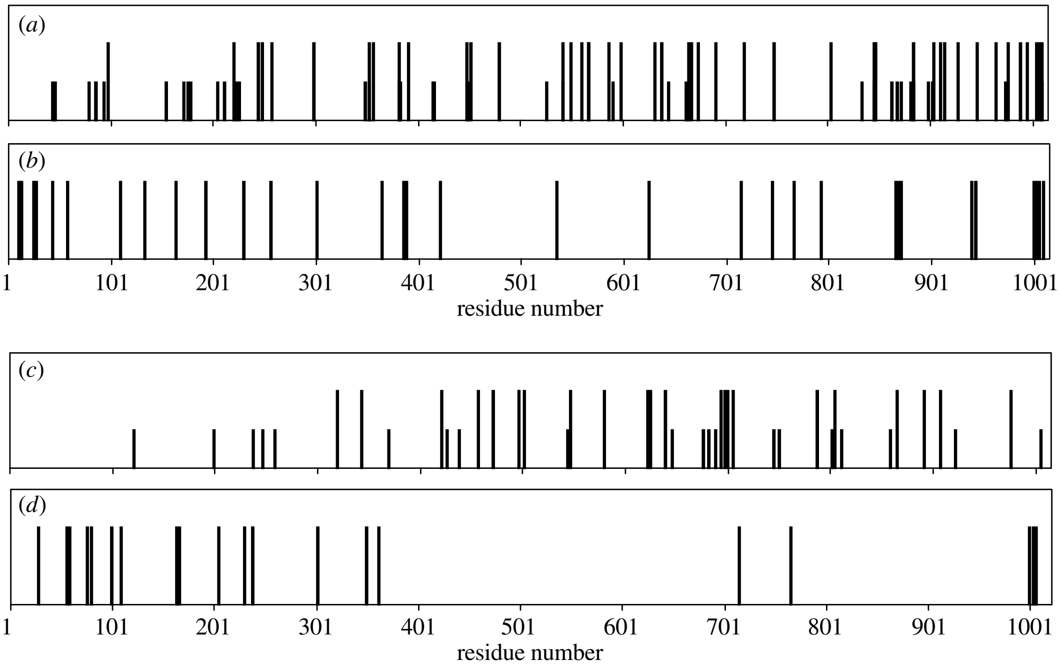

Figure 9. Bar representation of OI mutations and renucleation (GXO-GXO) domains in α1(I) and α2(I) chains of collagen. Long and short bars in (*a*,*c*) represent lethal and non-lethal (OI) mutations in α1(I) and α2(I) collagen chains, respectively. Distribution of renucleation domains in α1(I) and α2(I) chains are shown in (*b*,*d*). The majority of the mutations in OI collagen tend to occur in regions of the chain that do not have many renucleation sequences.

renucleation must occur successfully in collagen N-terminal to the substitution site (figure 8).

The nature of a 'renucleation' sequence in collagen has not been defined. A number of different triplet sequences may act as renucleation sequences; however, it is proposed that a good 'renucleation' sequence, that allows folding to be resumed after an interruption, may be similar to a good nucleation sequence that initiates folding. Studies by Bulleid and Brodsky have begun to define the requirements for the size and amino-acid composition of nucleation sequences in collagen. Bulleid *et al.* (1997) have shown that the simplest nucleation domain in collagen is a minimum of two Gly-Pro-Hyp triplets at the C-terminal end of the protein. Brodsky and co-workers (Ackerman *et al.* 1999) have established the propensity of different Gly-X-Y triplets to promote nucleation in model triple-helical peptides. They have shown that folding is fastest with Hyp in the Y position and that the identity of the residue in the X position for Gly-X-Hyp triplets does not dramatically affect the rate. Based on these data, it is possible that Gly-X-Hyp-Gly-X-Hyp sequences could function as renucleation sequences in collagen (B. Brodsky and J. Baum, unpublished data).

Applying this definition of renucleation to the α1(I) and α2(I) chains of collagen, we have identified 32 and 20 renucleation sites in the α1(I) and α2(I) chains, respectively (figure 9*b*,*d*). It is interesting that the renucleation sites in both chains have a tendency to be closer to the N- and C-terminal ends of the chains, with some predominance at the N-terminus for the α2(I) chain. In accord with this observation, the distribution of renucleation domains within the sequence is somewhat anticorrelated with the distribution of mutation positions (figure 9*a*,*c*). Thus, the middle region of the α1(I) chain, residues 300–700, contains nearly 50% of all lethal mutations, whereas only seven out of the 32 nucleation sites (22%) are located within the same region. This tendency is even more striking in the α2(I) sequence, in which all lethal mutations are present in the region that does not contain renucleation domains (figure 9*c*,*d*). These data are suggestive of a relationship between the positions of the mutation sites and the renucleation sites in collagen.

The NMR approaches allow a detailed understanding of the folding by providing a picture of the effect of Gly→X mutations on the unfolded form, during the folding process, and on the conformation of the mutant peptide. The results indicate that a number of factors may be important in determining the degree of abnormal folding. The amino-acid context in which the mutation occurs, the type of amino acid that is substituted, and the distribution of the renucleation sequences adjacent to and N-terminal to the mutation site may play an important role in defining the folding at the mutation site (Beck *et al.* 2000). NMR studies on triple-helical peptides which model OI collagen may open a new dimension to defining the molecular basis of this biomedical problem and may provide a definition of the conformation and precise folding defect which can serve as a starting point for therapeutic treatment of OI.

REFERENCES

Ackerman, M. S., Bhate, M., Shenoy, N., Beck, K., Ramshaw, A. M. & Brodsky, B. 1999 Sequence dependence of the folding of collagen-like peptides—single amino acids affect the rate of triple-helix nucleation. *J. Biol. Chem.* **274**, 7668–7673.

Bächinger, H. P. 1987 The influence of peptidyl-prolyl *cis–trans* isomerase on the *in vitro* folding of type III collagen. *J. Biol. Chem.* **262**, 17 144–17 148.

Bächinger, H. P., Bruckner, P., Timpl, R. & Engel, J. 1978 The role of *cis–trans* isomerization of peptide bonds in the coil/triple helix conversion of collagen. *Eur. J. Biochem.* **90**, 605–613.

Bächinger, H. P., Bruckner, P., Timpl, R., Prockop, D. J. & Engel, J. 1980 Folding mechanism of the triple helix in type-III collagen and type-II pN-collagen. Role of disulfide bridges and peptide bond isomerization. *Eur. J. Biochem.* **106**, 619–632.

Baum, J. & Brodsky, B. 1997 Real-time NMR investigations of triple helix folding and collagen folding diseases. *Fold. Design* **2**, R53–R60.

Baum, J. & Brodsky, B. 1999 Folding of peptide models of collagen and misfolding in disease. *Curr. Opin. Struct. Biol.* **9**, 122–128.

Baum, J. & Brodsky, B. 2000 Folding of the collagen triple helix and its naturally occurring mutants (ed. R. H. Pain), pp. 330–347. Oxford University Press.

Beck, K., Chan, V. C., Shenoy, N., Kirkpatrick, A., Ramshaw, J. A. M. & Brodsky, B. 2000 Destabilization of osteogenesis imperfecta collagen-like model peptides correlates with the identity of the residue replacing glycine. *Proc. Natl Acad. Sci. USA* **97**, 4273–4278.

Bella, J., Eaton, M., Brodsky, B. & Berman, H. M. 1994 Crystal and molecular structure of a collagen-like peptide at 1.9 Å resolution. *Science* **266**, 75–81.

Bonadio, J. & Byers, P. 1985 Subtle structural alterations in the chains of type I procollagen produce osteogenesis imperfecta type II. *Nature* **316**, 363–366.

Brant, D. A., Miller, W. G. & Flory, P. J. 1967 Conformational energy estimates for statistically coiling polypeptide chains. *J. Mol. Biol.* **23**, 47–65.

Brodsky, B. & Ramshaw, J. A. 1997 The collagen triple-helix structure. *Matrix Biol.* **15**, 545–554.

Brodsky, B. & Shah, N. K. 1995 The triple-helix motif in proteins. *FASEB J.* **9**, 1537–1546.

Bruckner, P., Bächinger, H. P., Timpl, R. & Engel, J. 1978 Conformationally distinct domains in the amino-terminal segment of type III procollagen and its rapid helix-coil transition. *Eur. J. Biochem.* **90**, 595–603.

Buevich, A. V., Dai, Q.-H., Liu, X., Brodsky, B. & Baum, J. 2000 Site-specific NMR monitoring of *cis–trans* isomerization in the folding of the proline-rich collagen triple helix. *Biochemistry* **39**, 4299–4308.

Bulleid, N. J., Dalley, J. A. & Lees, J. F. 1997 The C-propeptide domain of procollagen can be replaced with a transmembrane domain without affecting trimer formation or collagen triple helix folding during biosynthesis. *EMBO J.* **16**, 6994–6701.

Byers, P. H. 1993 Osteogenesis imperfecta. In *Connective tissue and its heritable disorders: molecular, genetic, and medical aspects* (ed. P. M. Royce & B. Steinmann), pp. 317–350. New York: Wiley Liss.

Dill, K. A. & Chan, H. S. 1997 From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **4**, 10–19.

Dobson, C. M. & Hore, P. J. 1998 Kinetics studies of protein folding using NMR spectroscopy. *Nat. Struct. Biol.* **5**(Suppl.), 504–507.

Dobson, C. M., Sali, A. & Karplus, M. 1998 Protein folding: a perspective from theory and experiment. *Angewante Chem. Int. Ed.* **37**, 868–893.

Dyson, H. J. & Wright, P. E. 1996 Insight into protein folding from NMR. *A. Rev. Phys. Chem.* **47**, 369–395.

Dyson, H. J. & Wright, P. E. 1998 Equilibrium NMR studies of unfolded and partially folded proteins. *Nat. Struct. Biol.* **5**(Suppl.) 499–503.

Engel, J. 1987 Folding and unfolding of collagen triple helices. In *Advances in meat research*, vol. 4 (ed. A. M. Pearson, T. R. Dutson & A. J. Bailey), pp. 145–161. New York: Van Nostrand Reinhold.

Engel, J. & Prockop, D. J. 1991 The zipper-like folding of collagen triple-helices and the effects of mutations that disrupt the zipper. *A. Rev. Biophys. Biophys. Chem.* **20**, 137–152.

Fields, G. B. & Prockop, D. J. 1996 Perspectives on synthesis and applications of triple-helical collagen model peptides. *Biopolymers* **40**, 345–357.

Goodman, M., Bhumralkar, M., Jefferson, E. A., Kwak, J. & Locardi, E. 1998 Collagen mimetics. *Biopolymers* (*Pept. Sci.*) **47**, 127–142.

Harrington, W. & Von Hippel, P. H. 1961 Formation and stabilization of the collagen-fold. *Arch. Biochem. Biophys.* **92**, 110–113.

Harrison, R. K. & Stein, R. L. 1990 Substrate specificities of the peptidyl prolyl *cis–trans* isomerase activities of cyclophilin and FK-506 binding protein: evidence for the existence of a family of distinct enzymes. *Biochemistry* **29**, 3813–3816.

Kelly, J. W., Colon, W., Lai, Z., Lashuel, H. A., McCulloch, J., McCutchen, S. L., Miroy, G. J. & Peterson, S. A. 1997 Transthyretin quaternary and tertiary structural changes facilitate misassembly into amyloid. *Adv. Protein Chem.* **60**, 161–181.

Kielty, C. M., Hopkinson, I. & Grant, M. E. 1993 Connective tissue and its hereditable disorders, molecular, genetic and medical aspects. In *The collagen family: structure, assembly, and organization in the extracellular matrix* (ed. P. M. Royes & B. U. Steinmann), pp. 103–147. New York: Wiley Liss.

Kuivaniemi, H., Tromp, G. & Prockop, D. 1997 Mutations in fibrillar collagens (types I, II, II, and IX), fibril-associated collagen (type IX), and network-forming collagen (type X) cause a spectrum of diseases of bone, cartilage, and blood vessels. *Hum. Mutat.* **9**, 300–315.

Lansbury Jr, P. T. 1997 Structural neurology: are seeds at the root of neuronal degeneration? *Neuron* **19**, 1151–1154.

Liu, X. 1997 *Real-time NMR investigation of triple helical peptide folding and collagen folding diseases*. New Brunswick, NJ: Rutgers University.

Liu, X., Siegel, D. L., Fan, P., Brodsky, B. & Baum, J. 1996 Direct NMR measurement of the folding kinetics of a trimeric peptide. *Biochemistry* **35**, 4306–4313.

Liu, X., Kim, S., Dai, Q.-H., Brodsky, B. & Baum, J. 1998 NMR shows asymmetric loss of triple helix in peptides modeling a collagen mutation in brittle bone disease. *Biochemistry* **33**, 15 528–15 533.

Long, C. G., Li, M.-H., Baum, J. & Brodsky, B. 1992 Nuclear magnetic resonance and circular dichroism studies of a triple-helical peptide with a glycine substitution. *J. Mol. Biol.* **225**, 1–4.

Long, C. G., Braswell, E., Zhu, D., Apigo, J., Baum, J. & Brodsky, B. 1993 Characterization of collagen-like peptides containing interruptions in the repeating Gly-X-Y sequence. *Biochemistry* **32**, 11 688–11 695.

McLaughlin, S. H. & Bulleid, N. J. 1997 Molecular recognition in procollagen chain assembly. *Matrix Biol.* **16**, 369–377.

Matthews, R. C. 1993 Pathways of protein folding. *A. Rev. Biochem.* **62**, 653–683.

Mayo, K. H. 1996 NMR and X-ray studies of collagen model peptides. *Biopolymers* **40**, 359–370.

Privalov, P. L., Tiktopulo, E. I. & Tischenko, V. M. 1979 Stability and mobility of the collagen structure. *J. Mol. Biol.* **127**, 203–216.

Prockop, D. J. & Kivirikko, K. I. 1995 Collagens: molecular biology, diseases, and potentials for therapy. *A. Rev. Biochem.* **64**, 403–434.

Raghunath, M., Bruckner, P. & Steinmann, B. 1994 Delayed triple helix formation of mutant collagen from patients with osteogenesis imperfecta. *J. Mol. Biol.* **236**, 940–949.

Rich, A. & Crick, F. H. C. 1961 The molecular structure of collagen. *J. Mol. Biol.* **3**, 483–506.

Roder, H. & Shastry, M. R. 1999 Methods for exploring early events in protein folding. *Curr. Opin. Struct. Biol.* **9**, 620–626.

Schimmel, P. R. & Flory, P. J. 1968 Conformational energies and configurational statistics of copolypeptides containing L-proline. *J. Mol. Biol.* **34**, 105–120.

Yang, W., Battinenni, M. & Brodsky, B. 1997 Amino acid sequence environment modulates the disruption by osteogenesis imperfecta glycine substitutions in collagen-like peptides. *Biochemistry* **36**, 6930–6935.

Discussion

S. Lindquist (*Howard Hughes Medical Institute, University of Chicago, IL, USA*). I would expect that prolyl isomerases would play a role *in vivo*, so have you explored the effect of adding such enzymes?

J. Baum. We have added prolyl isomerase to our control peptide, and it does increase the rate of folding of the nucleation and the propagation, but the intermediate is not affected by prolyl isomerase.

J. W. Kelly (*Department of Chemistry, Scripps Research Institute, La Jolla, CA, USA*). Have you tried solid-state NMR on these samples to see if they anneal with time to produce one homogeneous structure?

J. Baum. No. These peptides are highly soluble, and we have not yet managed to get them to form fibrils.

P. H. Byers (*Departments of Pathology and Medicine, University of Washington, Seattle, WA, USA*). Have you observed any slippage in the helices that form? To form a triple helix, the chains have to rotate about one another. What provides the anchor needed for this rotation to occur?

J. Baum. Based on the NOE data, in the cases we have looked at, there is no mis-staggering. I do not know how the supercoil forms.

R. Seckler (*Physik. Biochemie., University of Potsdam, Germany*). Collagen triple helices can be stabilized by increasing the number of Gly-Pro-Hyp triplets, and one can think of several reasons why nature would not have done that during the evolution of collagen sequences. Are there any known disease-linked mutations that would increase triple-helix stability?

J. Baum. I am not aware of any.

P. H. Byers. Such mutations occur in the vicinity of the proteolytic processing site. They impede the proteolytic processing necessary for fibril formation.

A. Helenius (*Swiss Federal Institute of Technology (ETH), Institute of Biochemistry, Zurich, Switzerland*). In the natural situation *in vivo* the chains are put in register by the C-terminal domain. Why do you not need that in your experiments? What concentrations do you need to get the triple helix to form?

J. Baum. We believe that the high concentrations we use simulate the effect of the globular C-propeptide. By NMR the lowest concentration we can look at is a few millimolar, lower with CD. So we think the diffusion problem is circumvented by the high concentrations we use.