THE ROYAL SOCIETY

doi 10.1098/rstb.2000.0760

Folding defects in fibrillar collagens

Peter H. Byers

Departments of Pathology and Medicine, Box 357470, University of Washington, Seattle, WA 98195-7470, USA
(*pbyers@u.washington.edu*)

Fibrillar collagens have a long triple helix in which glycine is in every third position for more than 1000 amino acids. The three chains of these molecules are assembled with specificity into several different molecules that have tissue-specific distribution. Mutations that alter folding of either the carboxy-terminal globular peptides that direct chain association, or of the regions of the triple helix that are important for nucleation, or of the bulk of the triple helix, all result in identifiable genetic disorders in which the phenotype reflects the region of expression of the genes and their tissue-specific distribution. Mutations that result in changed amino-acid sequences in any of these regions have different effects on folding and may have different phenotypic outcomes. Substitution for glycine residues in the triple helical domains are among the most common effects of mutations, and the nature of the substituting residue and its location in the chain contribute to the effect on folding and also on the phenotype. More complex mutations, such as deletions or insertions of triple helix, also affect folding, probably because of alterations in helical pitch along the triple helix. These mutations all interfere with the ability of these molecules to form the characteristic fibrillar array in the extracellular matrix and many result in intracellular retention of abnormal molecules.

Keywords: collagen; procollagen; chain assembly; triple-helix formation; human genetic disorders; nucleation

1. INTRODUCTION

The collagens are a family of proteins that comprise more than 30 precursor chains. These chains combine with exquisite specificity to produce over 20 distinct trimeric molecules. Collagens have triple helices that vary in length from a few dozen to over 300 tripeptides (Gly-X-Y, in which X and Y are any amino acid other than cysteine and tryptophan in most molecules) in which glycine is in every third position. The molecules are arrayed in the extracellular matrix of tissues in which they are specifically expressed and mutations give rise to disorders characterized by structural and functional abnormalities in the tissues in which these genes are expressed (table 1).

Misfolding of collagens is the major outcome of most mutations in collagen genes. Mutations have this effect in one of two ways. They can alter the sequences of collagens that are important for initial folding of the monomer, association of chains into trimers, nucleation of the triple helix, propagation of the triple helix, secretion, processing of the amino-terminal propeptides, or orderly aggregation into supramolecular structures, the collagen fibrils. Alternatively, they alter the expression or stability of mRNA or protein chains. In some instances, this leads to homotrimer formation instead of heterotrimer production, with alterations in the structure of the mature protein. In other instances, they alter the relative amounts of mature molecules in the extracellular matrix and perturb higher order aggregation of matrix components. Although mutations have been identified in many collagen genes, this discussion will concentrate on the effects of those in the genes for fibrillar collagens in which the alterations in folding are best understood.

2. INITIAL ASSEMBLY OF THE PROCOLLAGEN MOLECULE

The proα chains of fibrillar collagen are synthesized in the rough endoplasmic reticulum (ER). Like most secreted proteins they have amino-terminal leader sequences that guide the nascent chain and associated ribosome to the outer edge of the ER and lead to the vectoral insertion of the sequence into and through the membrane. Hydroxylation of most prolyl and lysyl residues in the Y position of the triple helix begins while the chains are being synthesized and are completed when the molecule is assembled and the triple helix is stable. Hydroxylation is thought to lag about 300 residues behind synthesis. The number of hydroxylases that occupy a chain is not clear but may be more than one. Prolyl 4-hydroxylase is a tetramer that consists of two α-chains and two β-chains, the latter being protein disulphide isomerase. This complex is anchored to the inner membrane of the ER through the KDEL (Arg-Asp-Glu-Leu) sequences of the disulphide isomerase. The *PLOD1* gene encodes the major lysyl hydroxylase and its product is bound to the inner membrane of the ER, although the precise mechanism is unclear. Other lysyl hydroxylases have been identified and their genes cloned (*PLOD2* and *PLOD3*) but their specificities have not been determined. Prolyl hydroxylation is essential to folding of the triple helix at normal temperatures. In the absence of prolyl hydroxylation the melting temperature of type I procollagen is *ca.* 27 °C instead of 43 °C in the fully hydroxylated form. Type I procollagen molecules synthesized when the enzyme is inhibited by iron chelators are stored in the ER but can be hydroxylated and secreted if the

Table 1. *Collagen genes and their disorders*

collagen type	gene	chromosomal location	protein	disorders
I	*COL1A1*	17q21.31-q22.05	proα1(I)	osteogenesis imperfecta Ehlers–Danlos syndrome type VIIA
	COL1A2	7q22.1	proα2(I)	osteogenesis imperfecta Ehlers–Danlos syndrome type VIIB Ehlers–Danlos syndrome type II
II	*COL2A1*	12q13.11-q13.2	proα1(II)	Stickler syndrome type I Wagner syndrome type II spondylepiphyseal dysplasia congenita Kniest dysplasia hypochondrogenesis achondrogenesis type II spondylo-metaphyseal-epiphyseal dysplasia (SMED), Strudwick type
III	*COL3A1*	2q31	proα1(III)	Ehlers–Danlos syndrome type IV Ehlers–Danlos syndrome type III (?)
IV	*COL4A1*	13q34	proα1(IV)	
	COL4A2	13q34	proα2(IV)	
	COL4A3	2q36-q37	proα3(IV)	Alport syndrome, recessive
	COL4A4	2q36-q37	pro4(IV)	Alport syndrome, recessive
	COL4A5	Xq22	pro5(IV)	Alport syndrome, X linked
	COL4A6	Xq22	pro6(IV)	Alport syndrome, X linked with leiomyomatosis
V	*COL5A1*	9q34.2-q34.3	proα1(V)	Ehlers–Danlos syndrome type I Ehlers–Danlos syndrome type II
	COL5A2	2q31	proα2(V)	Ehlers–Danlos syndrome type I
	COL5A3	not mapped	proα3(V)	
VI	*COL6A1*	21q22.3	proα1(VI)	Bethlem myopathy
	COL6A2	21q22.3	proα2(VI)	Bethlem myopathy
	COL6A3	2q37	proα3(VI)	Bethlem myopathy
VII	*COL7A1*	3p21.3	proα1(VII)	epidermolysis bullosa, recessive dystrophic epidermolysis bullosa, dominant dystrophic epidermolysis bullosa, pretibial
VIII	*COL8A1*	3q12-q13.1	proα1(VIII)	
	COL8A2	1p34.4-p32.3	proα2(VIII)	
IX	*COL9A1*	6q13	proα1(IX)	multiple epiphyseal dysplasia
	COL9A2	1p33-p32.2	proα2(IX)	multiple epiphyseal dysplasia, type II
	COL9A3	20q13.3	proα3(IX)	multiple epiphyseal dysplasia, type III
X	*COL10A1*	6q21-q22.3	proα1(X)	metaphyseal chondrodysplasia, Schmid type spondylometaphysealdysplasia, Japanese type
XI	*COL11A1*	1p21	proα1(XI)	Stickler syndrome type III Marshall syndrome
	COL11A2	6p21.3	proα2(XI)	Stickler syndrome type II otospondylomegaepiphyseal dysplasia (OSMED) Weissenbacher–Zweymuller syndrome non-syndromic deafness (DFNA13)
XII	*COL12A1*	6	proα1(XII)	
XIII	*COL13A1*	10q22	proα1(XIII)	
XIV	*COL14A1*	8q23	proα1(XIV)	
XV	*COL15A1*	9q21-q22	proα1(XV)	
XVI	*COL16A1*	1p34	proα1(XVI)	
XVII	*COL17A1*	10q24.3	proα1(XVII)	epidermolysis bullosa, generalized atrophic benign
XVIII	*COL18A1*	21q22.3	proα1(XVIII)	
XIX	*COL19A1*	6q12-q14	proα1(XIX)	

block to hydroxylation is removed, or if the cells are incubated at a lower temperature to permit helix formation. No naturally occurring mutations in the components of the prolyl hydroxylase complex are known. Lysyl hydroxylation leads to the formation of more stable cross-links in the mature extracellular molecule and its absence produces the recessively inherited disorder Ehlers–Danlos syndrome type VI in which skin, ligament and joint tensile properties are altered. HSP47 (also known as colligin), an ER-resident protein anchored by a KDEL receptor, appears to bind non-hydroxylated chains more avidly than hydroxylated chains (Asada *et al.* 1999; Koide *et al.* 1999). If correct, this provides one part of the solution to the problem of maintaining the long proα chains in a favourable conformation until the carboxy-terminal propeptide, the locus of chain–chain interaction, is synthesized and released.

It is only once the chain is synthesized in full that molecular assembly begins. At this point it is not clear what relationships exist between the chain and modifying enzymes or chaperones. It is likely that chains remain associated with both the hydroxylases, perhaps with HSP47 and possibly with other modifying enzymes. The carboxy-terminal propeptide folds to form two internal disulphide bonds—a process no doubt aided by disulphide isomerase—to form a major loop and a smaller, internal, minor loop. The globular domains of the proα chains are substrates for protein disulphide isomerase, the β-subunit of prolyl 4-hydroxylase, but the interaction is independent of association with the α-subunits of the protein (Wilson *et al.* 1998). Guided by association domains near the beginning of the major loop, correct chain association then occurs. This is no mean feat as a cell may synthesize as many as six chains (for example, the two chains of type I procollagen, the chains of type III procollagen, and the three chains of type V procollagen) at the same time in the same region of the ER. These association domains are thought, on the basis of domain swapping experiments, to be a discontinuous region of about 18 amino acids that permit the unique associations to occur (Lees *et al.* 1997). There is flexibility in chain association, no doubt to provide different functional facets of the resultant molecules. For example, proα1(I) chains can form homotrimers or, much more often, heterotrimers in which one proα1(I) is replaced by proα2(I). These molecules do not contain any of the other chain types. With procollagens made in cartilage and the vitreous body there is more flexibility allowed. For example, the proα1(II) chain of type II collagen also interacts with the proα1(XI) and proα2(XI) to form type XI collagen (Eyre & Wu 1987). In the vitreous body, the proα2(V) chain substitutes for the proα2(XI) chain (Mayne *et al.* 1993).

In type I procollagen, two proα1(I) chains or one proα1(I) chain and a single proα2(I) chain associate; the order of initial association appears to be random but determines the identity of the third chain brought into the molecule. Triple-helix formation is nucleated at the carboxy-terminal end of that domain, probably through the interaction of sequences very rich in hydroxyproline. From experimental studies it appears that helix nucleation requires two or more tripeptides in which hydroxyproline is in the Y position. Helix formation is propagated toward the N-terminal end of the molecule, probably in a punctuated fashion. In the triple helical domain the proline amide bonds are randomly distributed between the *cis* and the *trans* conformation, but in the final triple helix all are *trans* (Bachinger *et al.* 1978). The propagation of the triple-helical structure requires isomerization to the *trans* form, which is accomplished, in part, by the action of the enzyme peptidyl–prolyl *cis–trans*-isomerase (Bachinger 1987). It has been proposed that presence of *cis* bonds slows propagation of triple-helix formation. The triple helix has regions of high density of hydroxyproline and others of low density. The high-density regions may represent areas of secondary nucleation that are required to permit the ordered propagation of the triple helix.

Triple-helix formation removes proα chains from their interactions with the modifying enzymes and is the signal for release from the ER. These molecules are stacked in the Golgi bodies in a laterally aggregated form. The mechanisms of aggregation are not clear but the structures are reminiscent of collagen forms that can be induced by incubation with small charged molecules like ATP or large charged sulphated glycosaminoglycans. Whether similar molecules play such a role in the Golgi bodies is not clear, although it is the site of synthesis of glycosaminoglycan side-chains of proteoglycans. Release from the cell leads to further rearrangement and a different set of aggregate forms in which end-to-end overlaps generate the familiar collagen fibrils of the extracellular matrix. Some mutations in type I procollagen genes, which result in aberrant modification and defective secretion, do not appear to result in abnormal fibrils in soft tissues but may have dramatic effects on mineralization in bone. One explanation is that the fibrils in soft tissues are more plastic and do not need to accommodate the crystalline arrays of matrix mineral that forms bone.

3. MOLECULAR DEFECTS

The organization of the fibrillar collagen genes is one determinant of the effects of mutations on proteins and, as a result, on protein folding. The type I collagen genes have more than 50 exons. In *COL1A1* they occupy *ca.* 20 kb of genomic DNA and in the *COL1A2* gene they are distributed over *ca.* 40 kb. The *COL3A1* gene is similar to the *COL1A2* gene in size. The amino-terminal propeptide is encoded within the first six exons, the triple helical domain is encoded within parts of 44 exons (6–49), and the carboxy-terminal propeptides in the remaining exons. Each exon of the triple helix starts with a glycine codon (GGN) and ends with a codon for a Y-position amino acid so that skipping of a 'cassette' leaves the triplet structure intact, although the placement of charged and hydrophobic residues is disrupted relative to the normal sequences (Prockop & Kivirikko 1995).

Although most mutations that lead to osteogenesis imperfecta fall within the triple helical domain (probably a reflection both of the density of phenotypically active sites and its length), the carboxy-terminal extensions of both chains are functionally important and mutations in those domains have phenotypic effects (see §§(a)–(d)). Mutations that alter sequences in the carboxy-terminal propeptide have very different effects on molecular behaviour than those that affect sequences in the triple helix.

(a) *Substitutions for glycine residues in the triple helix*

Single nucleotide substitutions within glycine codons are the most common type of mutation. Three characteristics determine the phenotypic effects of mutations that result in substitutions for glycine residues within the triple helical coding region of type I collagen genes: the position of the substitution; the nature of the substituting amino acid; and the chain in which the substitution occurs (Byers 1993; Marini *et al.* 1993; Prockop & Kivirikko 1995).

First position substitutions in glycine codons lead to substitution of glycine by arginine, serine, cysteine and tryptophan—in approximately that order of frequency—due to the number of synonymous codons and the relative frequency of nucleotides in the third position. Second position substitutions give rise to alanine, valine, glutamic acid and aspartic acid codons. The relative frequency of substitutions by alanine, valine, glutamic acid and aspartic acid do not correspond to codon frequencies and the reasons for the disparities are not clear. In the proα1(I) chain, the product of the *COL1A1* gene, substitutions of valine, glutamic acid, aspartic acid and arginine produce severe, usually lethal, phenotypes from the carboxy-terminal end to the region of about residue 200 of the triple helix (Byers 1993; Kuivaniemi *et al.* 1997). The severity of similar substitutions within the proα2(I) chain seems somewhat milder, perhaps because there is a single proα2(I) chain in a molecule and there are two proα1(I) chains (Byers 1993; Kuivaniemi *et al.* 1997). Alternatively, the helix may be more tolerant of irregularity in the triple helix in the proα2(I) chain. Tryptophan substitutions are extremely rare, probably because of the rarity of GGG glycine codons in both genes.

In contrast to the effects of the large residue substitutions, substitutions by serine and cysteine for glycine have a startlingly punctuated effect on phenotype. That is, there are alternating regions of mild and severe phenotypes with mutations along the triple helix. These two substitutions might be anticipated to have different effects within the molecule. There are no cysteine residues within the triple helical domain of proα1(I) or proα2(I) chains. When the mutation is heterozygous (as almost all are), then half the chains synthesized have a cysteine residue in the triple helix. If it is in proα1(I) chains, then one-quarter of all trimers will contain two such chains, one-half contain one chain in which a glycine is substituted, and the remaining one-quarter are normal. In the molecules that contain two abnormal chains intramolecular disulphide bonds form. When these molecules are examined by rotary shadowing, a 'kink' can be observed in the triple helix at the apparent site of the substitution (Vogel *et al.* 1988; Lightfoot *et al.* 1992). Because such kinks are found only in molecules with substitutions of cysteine for glycine (Lightfoot *et al.* 1992), it is likely that the disulphide bonds stabilize an interaction between chains in some molecules that may not be the most stable conformation or, in order to stabilize, introduce a bend in the helix. The 50% of molecules that have a single abnormal chain no doubt have altered thermal stability but may also expose a free sulphhydryl group for interactions with nearby molecules in the matrix. Substitution of cysteine for glycine results in slowing of the folding of the triple helix, at least as recognized by the appearance of protease-resistant disulphide-bonded dimers. The extent of slowing of triple-helix formation correlated with a decrease in thermal stability of the molecules that contained the dimers. No clear relationship between the position of the mutation and phenotype was observed but those mutations that lowered thermal stability or slowed helix propagation most were associated with the worst phenotypes. Similar studies have not been done with chains that contain other substitutions for glycine because no readily visible markers are present that allow assessment of folding of chains that contain altered sequences. Substitutions by these two residues in the triple helical domain of the proα2(I) chain also have a discontinuous effect on phenotype.

Substitutions for glycine residues alter the triple helical structure (Bella *et al.* 1994; Baum & Brodsky 1999; Klein & Huang 1999) and produce delay in triple-helix formation. The helix appears to propagate normally from the region of nucleation at the carboxy-terminal end to the domain of the mutation, where it either ceases or slows at 37 °C (Raghunath *et al.* 1994). The effect on helical structure has been suggested by both modelling studies and by analysis of peptides with substitutions (Vogel *et al.* 1988; Bella *et al.* 1994; Baum & Brodsky 1999; Klein & Huang 1999). In one model, using space-filling blocks, a bubble is formed in the domain of the mutation and chain order appears to differ. In the others, derived from crystal structure, NMR studies (Bella *et al.* 1994; Baum & Brodsky 1999) and modelling using molecular dynamics (Klein & Huang 1999), there is a change in the pitch of the helix in the region of the substitution that propagates an altered relationship among the chains to their amino-terminal ends.

These studies appear to provide a ready explanation for one confusing finding—the overmodification of lysyl residues by both increased hydroxylation and glycosylation that occurs in an asymmetrical fashion in molecules that contain abnormal chains (Bonadio & Byers 1985). Both the delay in helix formation and the change in the relationship among residues in the three chains help to explain why overmodification of lysyl residues occurs largely amino-terminal to the site of mutations. In the normal molecule, post-translational modification continues until the triple helix is formed and stable. In the presence of mutations in the triple helix the disruption of the helical motif does not allow the chains to come into proximity. A new site of nucleation must be found amino-terminal to the disruption and, while the search occurs, post-translational modification continues. The association of molecules that contain abnormal chains with the modifying enzymes is also a factor that accounts for the delay in secretion.

Some abnormal molecules are incorporated into the matrix, although not into the most highly cross-linked, stable portion of tissues like bone (Bateman *et al.* 1987). In *in vitro* studies, inclusion of even a small proportion of molecules that harbour abnormal chains readily leads to failure of fibril formation or formation of abnormal fibrils (Torre-Blanco *et al.* 1992).

(b) *Mutations that affect triple-helix nucleation*

Triple-helix nucleation requires at least two triplets that contain hydroxyproline in the Y position (Bulleid *et al.* 1997). In type I procollagen mutations in the last

five tripeptides of the triple helix are not always lethal—particularly if they occur in the proα2(I) chain—and do permit nucleation, helix propagation and secretion. In contrast, mutations between the glycine at positions 994 and 1021 of the triple helix in the proα1(III) chain interfere with nucleation to the extent that many molecules that contain these abnormal chains fail to fold and few are secreted (U. Schwarze and P. H. Byers, unpublished data). Although these molecules can achieve triple helix at lower temperatures, at 37 °C most do not. The unfolded molecules accumulate in the ER, bound largely to the hydroxylases.

(c) *Splice-site mutations and deletion or insertion of Gly-X-Y triplets*

Splice-site mutations are the second most common mutation in type I collagen genes that result in the osteogenesis imperfecta phenotypes (Kuivaniemi *et al.* 1997). These mutations result either in exon-skipping or in the use of alternative or cryptic sites. The former result in retention of the canonical triple-helical triplet structure while the latter may or (more usually) may not. The effects of the use of cryptic splice sites depend on whether the spliced product can be resolved by the spliceosome and if it is in translational reading frame. Those that change reading frame and lead to a premature termination codon in the mRNA are generally unstable.

Within the *COL1A1* gene, exon-skipping mutations are generally very severe or lethal if they are located 3′ to exon 14 (Byers 1993). For example, mutations in splice sites that result in skipping of exons 14, 20, 22 (P. H. Byers, unpublished data), 27 (Pepin *et al.* 1997), 30 (P. H. Byers, unpublished), 44 (Byers 1990) and 47 (P. H. Byers, unpublished data) are associated with the OI type II phenotype. In the *COL1A2* gene, the chain seems more tolerant of deletions that affect sequences closer to the carboxy-terminal end of the protein as there are no lethal exon-skipping mutations 5′ of exon 27. In both genes when exon skipping is found in more 3′ locations, it is often associated with the simultaneous use of cryptic sites in the same allele, some of which lead to unstable mRNAs and thus, presumably, mitigate the severity of the expected phenotype. Small insertions or deletions that result in introduction or deletion or Gly-X-Y tripeptide units have similar effects and may be mild or lethal depending on their location.

The effects of these mutations seem, at first glance, more difficult to explain. Like substitutions for glycine, alteration from the native sequence, even with retention of triplet integrity, results in overmodification of chains amino-terminal to the site of the sequence alteration. Furthermore, the thermal stability of these molecules is lower than normal. The best explanation for these effects probably lies in the non-homogeneity of the triple helix along its length (Bachinger & Davis 1991; Bachinger *et al.* 1993). The sequence variations are associated with regions of slightly different pitch to the helix. Any large-scale alteration in sequence thus changes the relationships among chains so that in heterotrimers two chains at most maintain the correct register. If this is the case, then folding is impaired and the chains remain accessible while the molecules fail to be transported from the ER. Like molecules with substitutions for glycine in the triple helix, these stay associated with modifying enzymes, thus accounting in part for transport delay or failure.

(d) *Mutations in the carboxy-terminal propeptide*

The carboxy-terminal propeptides assure correct chain–chain interactions in the ER where chains of several different collagen types coexist. The failure to find interspecies chain association, in the absence of deliberate mutations, is a testament to the specificity of these binding domains (Lees *et al.* 1997). No mutations have been identified in the chain association domains but mutations have been identified in other regions that either eliminate chain assembly completely or alter the efficiency of chain association.

Mutations in the *COL1A1* gene that completely abolish chain recognition result in the mild osteogenesis imperfecta type I phenotype (Willing *et al.* 1990) in the heterozygote because mature molecules must have at least two proα1(I) chains. Without sufficient proα1(I) chains, excess proα2(I) chains are not incorporated and are rapidly degraded. The end result is similar to the failure to synthesize proα1(I) chains and the phenotype of osteogenesis imperfecta type I results. The abnormal proα1(I) chains may be degraded through the cytoplasmic proteosome system (Fitzgerald *et al.* 1999); the fate of the excess proα2(I) chains is less clear.

In contrast mutations that do permit chain association usually have a highly deleterious outcome, perhaps because they entrap molecules, leading to marked overmodification of the unfolded portions of the triple helix that remain associated with the chaperones that assist chain association. These mutations lead to activation of the unfolded protein response pathway. In cells that carry many of these mutations the stress–response proteins BiP (GRP78) and GRP94 are synthesized at high levels and appear to interact with the abnormal proteins, perhaps directing them to sites of degradation as they are often short lived (Chessler & Byers 1993; Lamande *et al.* 1995). These abnormal molecules may also be subject to degradation through the proteosome (Fitzgerald *et al.* 1999).

(e) *Effects of mutations on extracellular fibril formation and mineralization*

Once abnormal molecules are secreted, their effects on phenotype are probably modulated through several factors—the influence of the mutation on ability to process the molecule, the effect of the processed or unprocessed procollagen on fibril formation, and the effects of altered fibril structure on mineralization.

Even at considerable distance from the amino-terminal end of the triple helix, many mutations alter the efficiency of the N-protease, which cleaves the amino-terminal propeptide (Vogel *et al.* 1987; Lightfoot *et al.* 1992). The propeptide is normally folded back on the helical portion. It is clear that deletion of the sequences of exon six of either the proα1(I) or proα2(I) chain, which contain the N-protease cleavage site, alters the structure of that fold. There are no equivalent experiments with molecules known to harbour point mutations but some of these molecules have slowed cleavage rates. Bulky substitutions and deletions of sequence within the triple helix are more apt to alter rates than small residue substitutions. No doubt this influences fibrillogenesis.

Studies done with purified molecules indicate that not only is the rate of fibrillogenesis slowed but also the structure of the fibrils formed is altered (Kadler *et al.* 1991, 1996; Parkinson *et al.* 1997). It is not clear how these findings can be reconciled readily with studies of tissues of infants and adults with osteogenesis imperfecta that demonstrate normal fibrillar arrangements. Presumably the complexity of the biological system provides a buffer that facilitates fibril formation, perhaps even in the presence of abnormal fibrils.

Even when fibrils appear normal, mineralization is almost certainly disturbed. Mineral crystal size is often smaller than normal and it is possible to capture unmineralized fibrils from these tissues (Traub *et al.* 1994; Culbert *et al.* 1995). The fragility of bone in the presence of abnormal fibrils probably reflects the altered crystal structure as well as the change in the fibrillar matrix.

4. TRANSLATION OF GENOTYPE TO PHENOTYPE

Perhaps the most complex task in the study of disorders that result from mutations in fibrillar collagen is to understand the relationship between genotype (the specific mutation in a single host) and the phenotype in that person. The context of this difficulty is best perceived if we remember that, for example, the osteogenesis imperfecta phenotype is the physiological response to a mutation in a single collagen allele (most of the time). This means that the entire genetic background, the product of millions of years of evolution in the context of normal collagen alleles, influences the expression of the single change. It is not surprising that it is not straightforward to find molecular correlates of clinical severity. Several points about these attempts are clear. First, there is not a saturation set of mutations. As a result domains are unspecified in part because the data are incomplete. Second, it may require much more complete analysis of the efficiency of secretion and the effects of mutations on fibrillogenesis and mineralization before these mutations are understood at the phenotypic level. Finally, the role of alterations in several other genes needs to be understood.

Thus while there is considerable hope that phenotype–genotype relationships ultimately will be understood, hope for any near-term comprehension needs to be tempered by the reality of the number of factors that will need to be understood.

Original research described here was supported in part by grants from the National Institutes of Health (AR21557, AR41223).

REFERENCES

Asada, S., Koide, T., Yasui, H. & Nagata, K. 1999 Effect of HSP47 on prolyl 4-hydroxylation of collagen model peptides. *Cell Struct. Funct.* **24**, 187–196.

Bachinger, 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.

Bachinger, H. P. & Davis, J. M. 1991 Sequence specific thermal stability of the collagen triple helix. *Int. J. Biol. Macromol.* **13**, 152–156.

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

Bachinger, H. P., Morris, N. P. & Davis, J. M. 1993 Thermal stability and folding of the collagen triple helix and the effects of mutations in osteogenesis imperfecta on the triple helix of type I collagen. *Am. J. Med. Genet.* **45**, 152–162.

Bateman, J. F., Chan, D., Walker, I. D., Rogers, J. G. & Cole, W. G. 1987 Lethal perinatal osteogenesis imperfecta due to the substitution of arginine for glycine at residue 391 of the α.1(I) chain of type I collagen. *J. Biol. Chem.* **262**, 7021–7027.

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

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. H. 1985 Subtle structural alterations in the chains of type I procollagen produce osteogenesis imperfecta type II. *Nature* **316**, 363–366.

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**, 6694–6701.

Byers, P. H. 1990 Brittle bones–fragile molecules: disorders of collagen gene structure and expression. *Trends Genet.* **6**, 293–300.

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.

Chessler, S. D. & Byers, P. H. 1993 BiP binds type I procollagen proα chains with mutations in the carboxy-terminal propeptide synthesized by cells from patients with osteogenesis imperfecta. *J. Biol. Chem.* **268**, 18 226–18 233.

Culbert, A. A., Lowe, M. P., Atkinson, M., Byers, P. H., Wallis, G. A. & Kadler, K. E. 1995 Substitutions of aspartic acid for glycine-220 and of arginine for glycine-664 in the triple helix of the proα1(I) chain of type I procollagen produce lethal osteogenesis imperfecta and disrupt the ability of collagen fibrils to incorporate crystalline hydroxyapatite. *Biochem. J.* **311**, 815–820.

Eyre, D. & Wu, J.-J. 1987 Type XI or 1a2a3a collagen. In *Structure and function of collagen types* (ed. R. Mayne & R. E. Buregeson), pp. 261–282. Orlando, FL: Academic Press.

Fitzgerald, J., Lamande, S. R. & Bateman, J. F. 1999 Proteosomal degradation of unassembled mutant type I collagen pro-α1(I) chains. *J. Biol. Chem.* **274**, 27 392–27 398.

Kadler, K. E., Torre-Blanco, A., Adachi, E., Vogel, B. E., Hojima, Y. & Prockop, D. J. 1991 A type I collagen with substitution of a cysteine for glycine-748 in the α1(I) chain copolymerizes with normal type I collagen and can generate fractal-like structures. *Biochemistry* **30**, 5081–5088.

Kadler, K. E., Holmes, D. F., Trotter, J. A. & Chapman, J. A. 1996 Collagen fibril formation. *Biochem. J.* **316**, 1–11.

Klein, T. E. & Huang, C. C. 1999 Computational investigations of structural changes resulting from point mutations in a collagen-like peptide. *Biopolymers* **49**, 167–183.

Koide, T., Asada, S. & Nagata, K. 1999 Substrate recognition of collagen-specific molecular chaperone HSP47. Structural requirements and binding regulation. *J. Biol. Chem.* **274**, 34 523–34 526.

Kuivaniemi, H., Tromp, G. & Prockop, D. J. 1997 Mutations in fibrillar collagens (types I, II, III, and XI), 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.

Lamande, S. R., Chessler, S. D., Golub, S. B., Byers, P. H., Chan, D., Cole, W. G., Sillence, D. O. & Bateman, J. F. 1995 Endoplasmic reticulum-mediated quality control of type I collagen production by cells from osteogenesis imperfecta patients with mutations in the proα1(I) chain carboxy-terminal propeptide which impair subunit assembly. *J. Biol. Chem.* **270**, 8642–8649.

Lees, J. F., Tasab, M. & Bulleid, N. J. 1997 Identification of the molecular recognition sequence which determines the type-specific assembly of procollagen. *EMBO J.* **16**, 908–916.

Lightfoot, S. J., Holmes, D. F., Brass, A., Grant, M. E., Byers, P. H. & Kadler, K. E. 1992 Type I procollagens containing substitutions of aspartate, arginine, and cysteine for glycine in the proα1(I) chain are cleaved slowly by N-proteinase, but only the cysteine substitution introduces a kink in the molecule. *J. Biol. Chem.* **267**, 25 521–25 528.

Marini, J. C., Lewis, M. B., Wang, Q., Chen, K. J. & Orrison, B. M. 1993 Serine for glycine substitutions in type I collagen in two cases of type IV osteogenesis imperfecta (OI). Additional evidence for a regional model of OI pathophysiology. *J. Biol. Chem.* **268**, 2667–2673.

Mayne, R., Brewton, R. G., Mayne, P. M. & Baker, J. R. 1993 Isolation and characterization of the chains of type V/type XI collagen present in bovine vitreous. *J. Biol. Chem.* **268**, 9381–9386.

Parkinson, J., Brass, A., Canova, G. & Brechet, Y. 1997 The mechanical properties of simulated collagen fibrils. *J. Biomech.* **30**, 549–554.

Pepin, M., Atkinson, M., Starman, B. J. & Byers, P. H. 1997 Strategies and outcomes of prenatal diagnosis for osteogenesis imperfecta: a review of biochemical and molecular studies completed in 129 pregnancies. *Prenat. Diag.* **17**, 559–570.

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.

Torre-Blanco, A., Adachi, E., Romanic, A. M. & Prockop, D. J. 1992 Copolymerization of normal type I collagen with three mutated type I collagens containing substitutions of cysteine at different glycine positions in the α1(I) chain. *J. Biol. Chem.* **267**, 4968–4973.

Traub, W., Arad, T., Vetter, U. & Weiner, S. 1994 Ultrastructural studies of bones from patients with osteogenesis imperfecta. *Matrix Biol.* **14**, 337–345.

Vogel, B. E., Minor, R. R., Freund, M. & Prockop, D. J. 1987 A point mutation in a type I procollagen gene converts glycine 748 of the α1 chain to cysteine and destabilizes the triple helix in a lethal variant of osteogenesis imperfecta. *J. Biol. Chem.* **262**, 14 737–14 744.

Vogel, B. E., Doelz, R., Kadler, K. E., Hojima, Y., Engel, J. & Prockop, D. J. 1988 A substitution of cysteine for glycine 748 of the α1 chain produces a kink at this site in the procollagen I molecule and an altered N-proteinase cleavage site over 225 nm away. *J. Biol. Chem.* **263**, 19 249–19 255.

Willing, M. C., Cohn, D. H. & Byers, P. H. 1990 Frameshift mutation near the 3′ end of the COL1A1 gene of type I collagen predicts an elongated proα1(I) chain and results in osteogenesis imperfecta type I. *J. Clin. Invest.* **85**, 282–290. [Published erratum appears in *J. Clin. Invest.* **85** following p. 1338 (1990).]

Wilson, R., Lees, J. F. & Bulleid, N. J. 1998 Protein disulfide isomerase acts as a molecular chaperone during the assembly of procollagen. *J. Biol. Chem.* **273**, 9637–9643.

Discussion

J. W. Kelly (*Department of Chemistry, Scripps Research Institute, La Jolla, CA, USA*). With regard to structure, is the register precise enough so that you could grow crystals?

P. H. Byers. Do you mean the registration in the carboxy-propeptide or in the triple helix?

J. W. Kelly. The triple helix.

P. H. Byers. Very short peptides can be crystallized, but if you try to crystallize much longer ones they usually form fibres. The structure of the triple helix is relatively well known. The packing dimensions and the three-dimensional structure are known of that region. We have no idea about the structure of the carboxy-propeptide and how that directs folding or how mutations in this region interfere with the initial chain aggregation; this is the structure we need to understand.

A. Helenius (*Swiss Federal Institute of Technology (ETH) Institute of Biochemistry, Zurich, Switzerland*). Some of the mutations result in products that are rapidly degraded, while others result in products that accumulate in the ER. Is there any rule that you can see to explain why all these misfolded proteins are not degraded rapidly?

P. H. Byers. The ones that accumulate in the rough ER all have mutations in the triple helix. They all seem to remain associated with the prolyl hydroxylase. So there is a different association with prolyl hydroxylase once the chain begins to trimerize. They are all partly folded; the ones in the nucleation domain at least have a well-folded carboxy-terminal propeptide. The ones that have mutations in the carboxy-propeptide appear to be the ones that are degraded, probably through a proteosome mechanism, but we do not know what the targeting signals are.

A. Helenius. Does this indicate that the ones that are stable look like they are folded to the quality control system?

P. H. Byers. Yes and no. Up to a certain point in the molecule, probably about midway through the triple helix, something that interferes with folding leads to some secretion of even the poorly folded molecules, but also a lot of these are retained. The retention is related in part to where in the chain the mutations occur. Carboxy-terminal mutations tend to be retained more than those that are N-terminal. It depends on how many prolyl hydroxylase molecules are bound to the chain that determines how well it is retained.

Anonymous. Do you know of any mutations that affect the chain recognition sequence?

P. H. Byers. There are none that we know about.

B. E. P. Swoboda (*Department of Biological Sciences, University of Warwick, UK*). Collagen is a very strong signal for platelet aggregation. Are there any small changes, equivalent to polymorphisms, that cause major effects on platelet aggregation, such as lead to thrombosis?

P. H. Byers. I know of no firm evidence on this point. People who have Ehlers–Danlos syndrome type IV and mutations in type III collagen do not have increased

thrombogenicity, but do have very fragile blood vessel walls, and bruise easily. These walls are fragile in part because they are not built correctly; the thickness of the vessel wall is less than normal. In some of these patients, the vessels are translucent, to the extent that blood flow can be seen through major arterial vessels. It is this fragility that is the problem, not that the vessels present a thrombogenic surface.

I. Helenius. How common are such patients?

P. H. Byers. Brittle bone disease, or *osteogenesis imperfecta*, occurs in about one in every 10 000 people. Some of the forms of osteoarthritis that are familial are relatively common, and some are due to mutations in type II collagen. These mutations do not have anywhere near the frequency on an individual basis of those that cause cystic fibrosis, but in aggregate, mutations in collagen genes probably sum to those found in cystic fibrosis.