Local amino acid sequence patterns dominate the heterogeneous phenotype for the collagen connective tissue disease Osteogenesis Imperfecta resulting from Gly mutations

Jianxi Xiao,^{‡,†} Zhangfu Yang,[†] Xiuxia Sun,[†] Rayna Addabbo,^{‡,§} Jean Baum^{‡,*}

[‡]Department of Chemistry and Chemical Biology, Rutgers University, 610 Taylor Road, Piscataway, NJ 08854

[†]State Key Laboratory of Applied Organic Chemistry, Department of Chemistry, Lanzhou University, Lanzhou 730000, China

*Correspondence: jean.baum@rutgers.edu; 732/445-5666.

ABSTRACT

Osteogenesis Imperfecta (OI), a hereditary connective tissue disease in collagen that arises from a single Gly->X mutation in the collagen chain, varies widely in phenotype from perinatal lethal to mild. It is unclear why there is such a large variation in the severity of the disease considering the repeating (Gly-X-Y)_n sequence and the uniform rod-like structure of collagen. We systematically evaluate the effect of local (Gly-X-Y)_n sequence around the mutation site on OI phenotype using integrated bio-statistical approaches, including odds ratio analysis and decision tree modeling. We show that different Gly->X mutations have different local sequence patterns that are correlated with lethal and nonlethal phenotypes providing a mechanism for understanding the sensitivity of local context in defining lethal and non-lethal OI. A number of important trends about which factors are related to OI phenotypes are revealed by the bio-statistical analyses; most striking is the complementary relationship between the placement of Pro residues and small residues and their correlation to OI phenotype. When Pro is present or small flexible residues are absent nearby a mutation site, the OI case tends to be lethal; when Pro is present or small flexible residues are absent further away from the mutation site, the OI case tends to be nonlethal. The analysis also reveals the dominant role of local sequence around mutation sites in the Major Ligand Binding Regions that are primarily responsible for collagen binding to its receptors and shows that non-lethal mutations are highly predicted by local sequence considerations alone whereas lethal mutations are not as easily predicted and may be a result of more complex interactions. Understanding the sequence determinants of OI mutations will enhance genetic counseling and help establish which steps in the

collagen hierarchy to target for drug therapy.

KEY WORDS: Collagen, Osteogenesis Imperfecta, Odds ratio, Decision tree, bio-statistics

Present address: Rayna Addabbo, Department of Chemistry and Biophysics Program, University of Wisconsin-Madison, 1101 University Avenue, Madison, Wisconsin 53706.

INTRODUCTION

Understanding the relationship between collagen sequence and connective tissue disease phenotype is crucial to deciphering the molecular mechanisms of hereditary diseases as well as to developing diagnostic strategies (Bodian and Klein, 2009; Reuter et al., 2013). Osteogenesis Imperfecta (OI) is a rare heritable connective tissue disorder with characteristic clinical features including abnormal bone fragility, dentinogenesis imperfecta, blue sclera and hearing loss (Byers and Cole, 2002; Van Dijk and Sillence, 2014). The most common cause of OI is the missense mutation of a conserved Gly in the triple helical domain of Type I collagen (Shoulders and Raines, 2009). Type I collagen, the major structural protein of bone, skin, tendon and ligament, is a heterotrimer composed of two $\alpha 1$ and one $\alpha 2$ chains. Each chain has a repetitive (Gly-X-Y)_n sequence pattern, while the residues at the X and Y positions are frequently Pro and Hydroxyproline (Hyp or O), respectively (Brodsky et al., 2008). The three extended polyproline II-like chains are supercoiled around a common axis, forming the distinct triple helical structure (Ramachandran and Kartha, 1955; Rich and Crick, 1955). The close packing of the three chains requires Gly as every third residue, and any bulkier residue that replaces Gly could disrupt the triple helix (Bella et al., 1994).

The phenotypes of OI caused by mutations altering type I collagen structure vary widely and can be divided into four types according to the Sillence classification system (Sillence et al., 1979; Van Dijk and Sillence, 2014). Type I is the mildest form of OI characterized by minimal bone deformity and low impact fractures after puberty. Type III shows severe, but non-lethal features such as extremely short stature and multiple bone

fractures, and Type IV has the broadest phenotype which are characterized by moderate bone deformation (Sillence et al., 1979; Cohn and Byers, 1990). Type II is the most severe form and is mostly perinatal lethal, while some type II-A babies can survive into childhood (Van Dijk and Sillence, 2014). It is observed that mutations in the $\alpha 1(I)$ chain lead more often to a lethal phenotype than mutations in the $\alpha 2(I)$ chain (Marini et al., 2007). Since type I collagen heterotrimers are composed of two $\alpha 1(I)$ chains and only one $\alpha 2(I)$ chain, mutations in the $\alpha 1(I)$ chain have a larger opportunity resulting in abnormal heterotrimers than those in the $\alpha 2(I)$ chain (Sykes, 1985; Prockop et al., 1989). In more than 700 missense mutations in Type I collagen reported to result in OI, Gly residues are most frequently replaced by Ser or Cys, but Arg, Val, Asp, Glu, and Ala are also well represented (Dagleish, 1997, 1998; Marini et al., 2007). It has been shown that the order of severity of OI phenotype in collagen is Ala, Ser < Cys < Arg < Val < Glu, Asp (Beck et al., 2000). To understand and be able to predict OI pathogenesis it is critical to determine the sequence determinants of OI mutations. This will enhance genetic counseling, as there will be a clearer relationship between the type and the position of the mutation within the collagen sequence and the phenotype.

Despite the uniform nature of the collagen sequence and structure, the same type of Gly mutation at different positions along the collagen chain frequently results in very different OI phenotypes. It is still unclear why there is such a large variation in the severity of diseases resulting from seemingly similar mutations along a rather uniform rod-like structure. A number of ideas have been proposed to explain how a G→X mutation affects OI and include: 1) the identity of the substituted amino acid (Byers, 1989; Marini et al., 2007); 2) the local sequence environment (Bryan et al., 2011; Xiao et al., 2011b); 3) the position of the mutation relative to the C-terminus (Bodian et al., 2008); and 4) the position relative to collagen interaction sites (Xu et al., 2008b; Parkin et al., 2011). Due to the complexity of collagen self-association and ligand binding, multiple variables need to be considered to predict and understand the basis for the clinical phenotypes of OI: these include understanding the relationship between local sequence and OI phenotype and defining the role of higher order function, such as ligand binding or self association in OI phenotype.

The heterogeneous distribution of OI mutations along the collagen chain has attracted extensive attention to build genotype-phenotype models (Marini et al., 2007; Bodian et al., 2008). A gradient model underscores the importance of the positions of Gly mutations relative to the C-terminus of collagen in the $\alpha 1(I)$ chain, while a regional model highlights the importance of potential interference of the mutations with the interaction of collagen with other biomolecules in both the $\alpha 1(I)$ and $\alpha 2(I)$ chains (Sztrolovics et al., 1993; Wang et al., 1993; Marini et al., 2007). The gradient model is supported by the observation that the mutations within 200 residues of the N-terminal in the $\alpha 1(I)$ chain are essentially non-lethal (Bodian et al., 2008). However, this model does not represent the data well beyond the N-terminal region, particularly for Ser and Cys mutations (Marini et al., 2007; Bodian et al., 2008). Experimental biophysical structural studies have shown that the Gly-X-Y triplets in fibrillar collagens fold into triple helices but that the conformational details such as hydrogen bonding, dihedral angles, side-chain interactions and bending, are dependent on specific sequences (Kramer et al., 1999; Persikov et al., 2000; Emsley et al., 2004; Gauba and Hartgerink, 2007; Hohenester et al., 2008; Xu et al., 2008a; Fallas et al., 2009; Xiao and Baum, 2009). Specific local amino acid sequence patterns have been found to play critical roles in the interaction of collagen with integrin and MMP (Emsley et al., 2000; Lauer-Fields and Fields, 2002; Stultz, 2002; Raynal et al., 2006; Xiao et al., 2010; Bertini et al., 2012; Lu and Stultz, 2013). It has also been shown that similar mutations located in different local sequence environments display very diverse effects on triple helix stability, structure, dynamics and folding (Bella et al., 1994; Li et al., 2005; Tsai et al., 2005; Bryan et al., 2007; Li et al., 2009; Xiao et al., 2011a; Xiao et al., 2011b). Despite significant progress in understanding certain aspects of the biophysics of OI, general rules or guidelines that correlate primary sequence and OI phenotype are still not defined.

Here, we introduce bio-statistical approaches to systematically evaluate the effect of local sequence around the mutation site on OI lethality. Odds ratio analysis has been widely used in medical studies to measure the association of potential risk factors to a dichotomous outcome (Bosanquet et al., 2014). We show, using odds ratio analysis, that different types of Gly->X mutations have different neighboring local consensus sequence patterns that are correlated to OI lethality. We build decision tree models, using information from the odds ratio

analysis, that successfully predict the OI lethality from local amino acid sequences for the dominant Ser, Cys and Arg mutations in the $\alpha 1(I)$ and Ser mutations in the $\alpha 2(I)$ chains. The development of mutation specific local consensus sequences allows us to understand why the same type of substitution, such as Ser, results in a large variation in the severity of disease along the collagen chain. Our analysis, for the first time, reveals that the identity of the residues that surround the Gly mutations in the Major Ligand Binding Regions (MLBR) of collagen play a determinant role for the non-lethal OI mutations in these interaction regions. Our data supports the view that non-lethal OI phenotype in the MLBR can be predicted primarily from the local surrounding sequence around the mutation site whereas lethal phenotypes may arise from more complex interactions.

MATERIALS and METHODS

Assembly of data The mutations are collected from the Database of Collagen Mutations (www.le.ac.uk/genetics/collagen) at the University of Leicester (Dagleish, 1997). Sequences of the native proteins $\alpha 1(I)$ and $\alpha 2(I)$ are obtained from Genbank entries NP_000079.2 and NP_000080.2, respectively (Benson et al., 2013). Amino acids are numbered according to the nomenclature guidelines set by the Human Genome Variation Society (den Dunnen and Antonarakis, 2001). The lethality of OI cases is referred to the OI consortium (Marini et al., 2007) or the original publication. Lethal cases include patients who did not survive the immediate postnatal period as well as prenatal lethal cases. Due to the complexity of the clinical characteristics of OI, it is difficult to determine the exact phenotypes in some cases (Van Dijk and Sillence, 2014). In this bioinformatics analysis, only Gly missense mutations in the triple helix domain of type I collagen that have known OI lethality are included to elucidate the genotype-phenotype relationships. 615 cases of Gly missense mutations with known OI lethality in the $\alpha 1(I)$ chain and 488 cases in the $\alpha 2(I)$ chain were collected (Table S1, S2).

Odds ratio analysis The odds ratio is defined as the ratio of the odds of the presence versus the absence of a particular type of residue at a certain position leading to lethal OI. To study the effects of local amino acid sequences on OI phenotypes, the amino acids are classified into five types: Pro (Pro and Hyp), Small (Gly, Ala,

Cys, Ser and Thr), Charged (His, Lys, Arg, Glu and Asp), Hydrophobic (Val, Ile, Leu, Met, Phe, Tyr and Trp) and Polar (Asn and Gln). The position of a certain local amino acid around a Gly mutation is numbered as P_i if it is the i th residue prior to the mutation or P_i' if it is the i th residue following the mutation (Figure 1). For example, if at a certain position P_i , the Pro residues occur 'a' times in all the lethal OI cases and 'b' times in nonlethal cases, while non-Pro residues occur 'c' and 'd' times in lethal and nonlethal cases, respectively; then the odds of the presence of Pro at position P_i leading to lethal phenotype is a/b ; and the odds for the absence of Pro is c/d ; therefore, the odds ratio for Pro at position P_i leading to lethal phenotype is ad/bc (Figure 1). The number of lethal or nonlethal cases includes all the diagnosed patients, and each patient is counted once. For Gly substitutions with both lethal and nonlethal outcomes, the number of lethal and nonlethal cases is counted separately. Since the residues substituting Gly (Ala, Ser, Cys, Glu, Arg, Asp, Val) show very different odds leading to lethal OI, the odds ratio is calculated for each type of substitution separately. The P_2' position for the Ser mutation is illustrated as an example (Figure 1). Pro occurs 25 and 41 times in lethal and nonlethal OI cases, respectively; while non-Pro residues occur 31 and 156 times in lethal and nonlethal OI cases, respectively. Therefore, the odds ratio of Pro leading to lethal OI at the P_2' position for the Ser mutation is calculated to be 3.1. Its 95% confidence interval is (1.6, 5.8).

	$P_{i+1} P_i$	$P_2 P_1$	$P_1' P_2'$	$P_i' P_{i+1}'$
	---X-Y---Gly-X-Y-Gly-X-Y-Gly---X-Y---			
Pro	a/b		25/41	
non-Pro	c/d		31/156	
odds ratio	ad/bc		3.1	

Figure 1. Illustration of calculation of the odds ratio for Pro at position P_i leading to lethal phenotype. The position of a certain local amino acid around a Gly mutation (colored in red) is numbered as P_i if it is the i th residue prior to the mutation or P_i' if it is the i th residue following the mutation. At a certain position P_i , the Pro residues occur a times in all the lethal OI cases and b times in nonlethal cases, while non-Pro residues occur c and d times in lethal and nonlethal cases, respectively. The odds of Pro at position P_i leading to lethal phenotype is a/b ; and the odds for non-Pro is c/d . The odds ratio for Pro at position P_i leading to lethal phenotype is ad/bc . Following the above rule, the odds ratio of Pro leading to lethal OI at the P_2' position for the Ser mutation is calculated to be $(25 \times 156) / (41 \times 31) = 3.1$.

The odds ratio of the presence of Pro versus the absence of Pro at each position (P11-P11') for each type of substitution is calculated. The steps to calculate odds ratio are repeated for four other pairs: small, flexible residues vs bulkier, non-flexible residues, charged vs neutral residues, hydrophobic vs non-hydrophobic residues and polar vs non-polar residues. A significantly large odds ratio indicates that the residue type at that specified position is correlated to lethal OI, while a significantly small value indicates a correlation to nonlethal OI. Fisher's exact test is utilized to test the statistical significance considering the small sample size. All of the tests are two-tailed, and a p-value of <0.05 is considered significant. All of the analyses are performed by the software SAS (Statistical Analysis System, SAS Institute Inc., NC).

Predictive Modeling Decision tree models are built to predict OI lethality based on the information from the odds ratio analysis. The critical positions identified to be significantly correlated with OI lethality by odds ratio analysis are included as input variables. Due to the limit of the sample size, decision tree modeling is focused on the critical positions in the P1-P5 and P1'-P5' regions. Each input variable has five nominal values: Pro, Small, Hydrophobic, Polar or Charged. Lethal OI phenotype is defined as the output variable. Ten-fold cross validation is used for model selection and feature selection. All of the decision tree analysis is performed using SAS Enterprise Miner package (SAS Institute Inc., NC).

RESULTS

Odds analysis of different regions in the $\alpha 1(I)$ chain

A total of 615 cases of Gly missense mutations with known OI lethality were included for bioinformatics analysis in this study. According to the gradient model, the $\alpha 1(I)$ chain of collagen tends to have more lethal cases when moving from the N-terminal end to the C-terminal end. We therefore divided the collagen chain into 5 regions, each containing ~200 amino acids (Figure 2A). The N-terminal region (Region 1, position<200) contains predominantly non-lethal mutations, but a few exceptions are observed for Asp, Val, Glu, Arg and Ser mutations (Figure 2B). The other regions (Region 2-5) contain mostly lethal mutations when the substituting residue is Arg, Glu, Asp or Val, while they display much more inter-dispersed non-lethal mutations if the

Figure 2. Distribution of different types of Gly mutations (Ala, Ser, Cys, Glu, Arg, Asp and Val) along the $\alpha 1(I)$ chain. The triple helix collagen chain is divided into five regions from the N-terminal to the C-terminal: region 1 (position <200), region 2 (position 200-400), region 3 (position 400-600), region 4 (position 600-800) and region 5 (position >800) (A). Blue and red circles are used to represent non-lethal and lethal mutations, respectively (B). Region 1 contains predominantly non-lethal mutations. Regions 2-5 contain mostly lethal mutations for the Arg, Glu, Asp and Val mutations.

To clarify if the gradient model well describes the relationship of sequence to OI phenotype, the odds of the five regions leading to lethal vs nonlethal cases were calculated (Figure S1). The N-terminal region (Region 1) has tiny odds, suggesting essentially nonlethal OI cases, which is consistent with earlier observations (Figure S1). However, from the N-terminal to the C-terminal, the odds do not increase linearly; instead, regions 2 and 4 display larger odds than regions 3 and 5 (Figure S1). The odds analysis suggests that the OI mutations show a region-dependent pattern rather than a gradient pattern.

Amino acid distribution analysis of different regions in the $\alpha 1(I)$ chain and relationship to OI phenotype

To investigate the largely variable odds leading to lethal OI, the distribution of amino acid types in the five regions was calculated (Figure 3). Amino acids were classified into five categories: Pro, Small, Charged, Hydrophobic and Polar. First, the amino acid distribution of residues at the X-positions in the (Gly-X-Y)_n triplets in each region was calculated. No statistical difference was found in the distribution of X-position residues between different regions (p=0.26) (Figure 3A). For example, the X-position has ~35% of amino acid types as Pro and ~24% small amino acids in each region, respectively. The studies on the Y-position show that all the regions have similar distributions of Y-position residues (data not shown). This indicates that the amino acid distribution in each region is similar along the collagen chain.

Figure 3. The distribution of amino acid types in the five regions in the $\alpha 1(I)$ chain. The frequency of amino acid types in the X position of the (Gly-X-Y)_n triplets in each region is compared (p=0.26) (A). The frequency of amino acid types in the P2 position of the mutations in each region is compared (p<0.001) (B). Collagen peptide chain is divided into five regions from the N-terminal to the C-terminal: <200(light gray), 200-400(gray), 400-600(white), 600-800(dark gray),>800(sparse white). Amino acids are classified into five categories: Pro, Small, Charged, Hydrophobic and Polar. The frequency of each type of amino acids at the X position and the P2 position is counted for each region, respectively. The X position has uniformly ~35% of amino acid types as Pro and 24% small amino acids in each region (A). Region 5 (Position>800) has the most Pro at the P2

position, while region 1 (Position<200) has the most hydrophobic residues (B).

The local amino acid distribution around the mutation sites was calculated to determine if there is a correlation between local sequence and OI phenotype. The distribution of amino acid types at the P2 position of the mutants (the site 2-residues N-terminal to the mutation) is quite different between the 5 regions ($p<0.001$) (Figure 3B). Region 5 (Position>800) has the most Pro at the P2 position, while region 1 (Position<200) has the most hydrophobic residues at that position. In contrast, small and charged residues are more occupied at the P2 position in region 4 (position 600-800) and region 2 (position 200-400), respectively. Analysis of amino acid types at other P1-P11 and P1'-P11' positions show similarly that mutations in different regions have different local amino acid distributions (data not shown). This indicates that although the average amino acid distribution is similar in each region of collagen, the local amino acid sequence environment around the mutations is significantly different.

Odds analysis of different types of Gly->X mutations in the $\alpha 1(I)$ chain

Ser, Arg and Cys are the most frequent types of mutations, accounting for 3/4 of all the mutations (Table 1). Consistent with earlier observations, different Gly-substituting residues show very different odds leading to lethal OI (Figure S2). Ala has the smallest odds of 0.20, suggesting that the chance of Ala leading to lethal cases is approximately 1/5 times the chance of Ala leading to nonlethal cases (Table 1). Compared with Ala, the odds of Ser leading to lethal vs nonlethal cases is 1.4 times greater (odds ratio=0.28/0.20). The charged or branched residues (Glu, Arg, Asp and Val) have relatively larger odds of leading to lethal OI than other amino acid substitutions. Val and Asp clearly have the largest odds as they are the only two kinds of mutants that have larger chances of leading to lethal cases versus nonlethal cases. The huge difference between charged residues Asp and Glu/Arg is also notable. The diverse odds of different Gly substitutions indicate that they may have different sequence preferences for OI.

Mutation Types	Lethal Cases	Nonlethal Cases	Total	Odds
Ala	6	30	36	0.20
Ser	56	197	253	0.28
Cys	31	65	96	0.48
Glu	6	9	15	0.67
Arg	51	60	111	0.85
Asp	36	17	53	2.12
Val	35	16	51	2.19
Total	221	394	615	0.56

Table 1. Summary of Gly mutations in the $\alpha 1(I)$ chain of collagen. The Gly mutations with known OI lethality are included for bioinformatics analysis. For each type of mutations (Ala, Ser, Cys, Arg, Glu, Asp and Val), the number of lethal and nonlethal cases is counted. Their odds leading to lethal vs nonlethal OI phenotypes are calculated. Different Gly-substituting residues show very different odds leading to lethal OI.

Mutation specific local consensus sequences that are correlated to lethal and non-lethal OI using odds ratio analysis

Odds ratio analysis is utilized across an 8-triplet region around a Gly substitution in the $\alpha 1(I)$ chain to determine if a specific residue type and a specific position near the mutation is significantly correlated to lethal OI. It is immediately clear that different Gly-> X mutations have different local consensus sequences that correlate with lethal and non-lethal OI (Figure 4). The OI lethality of different Gly->X mutations shows a different dependence on the local sequence environment (Figure 4). Generally speaking, for Gly-> Ser mutations, the presence of Pro at P1, P2 and P2' sites and the absence of Pro at the P5' site significantly increases the risk of lethal OI; while for Gly to Arg mutations, the presence of Pro at the P1' site is significantly correlated to lethal OI. Only a few positions, rather than the whole region, are significantly correlated to OI phenotype.

Figure 4. Consensus sequence patterns for Gly to Ser, Cys and Arg mutations in the $\alpha 1(I)$ chain that are correlated with lethal and nonlethal OI. Dependency of OI lethality on the type and position of neighboring residues within 4-triplets of Gly substitutions by Ser, Cys or Arg are calculated based on odds ratio analysis in the $\alpha 1(I)$ chain. The relative positions of the residues are indicated by Pi (the ith residue N-terminal to the mutation site) or Pi' (the ith residue C-terminal to the mutation site). When the presence of certain types of amino acids at some specific positions around the mutation site is significantly correlated to lethal or nonlethal OI phenotypes, the position and the amino acid type is highlighted with colored boxes (Pro colored in yellow, small flexible residues (Sma) in red, charged residues (+/-) in green, Polar residues (Pol) in olive, and Hydrophobic residues (Hyd) in gray).

Ala has the smallest dataset and has the fewest positions significantly correlated to OI lethality (data not shown). This is likely due to the limited sample size. Ser, Arg and Cys have the three largest datasets (Table 1); therefore, detailed odds ratio analyses are focused on these three types of mutations (Figure 4, S3). First, the importance of Pro residues near the Ser/Cys/Arg mutations was tested to evaluate the correlation between its position relative to the mutation site and OI lethality (Figure S3A). The odds ratio analysis indicates that the Pro in the nearest triplet (P2-P2') is particularly important for all the substitutions. The presence of Pro at these positions tends to result in lethal phenotype (indicated by +). In addition, the presence of Pro further away from the sites (P11-P7, P5'-P11') seems more likely to result in non-lethal phenotype (indicated by -).

Small residues nearby the mutation sites are found to be crucial for OI lethality at a few positions (Figure S3B). Essentially, the absence of small residues (P5-P2') is correlated with lethal OI, while the presence of small residues further away from the site (P11-P7, P4'-P11') tends to result in lethal phenotype. Small residues near

the mutation sites may give the mutant more room to adapt, therefore resulting in nonlethal OI; while the presence of small residues further away may disfavor nucleation or renucleation, therefore resulting in lethal OI (Xiao et al., 2011b). It is worth noting that the pattern of lethal and nonlethal OI arising from the positions of small residues show a very complementary pattern with the positions of Pro and their correlation to lethal and nonlethal OI (Figure 4, S3). Particularly in the P2-P2' region, the presence of Pro tends to result in lethal OI at certain positions (+), while the absence of small residues (-) tends to result in lethal OI.

Hydrophobic residues are generally favored at sites (P2-P7') for nonlethal OI, while Arg shows the strictest requirement (Figure 4, S3C). Polar residues show a more random pattern, while the presence of polar residues around Ser substitutions (P4, P7') is correlated with lethal OI (Figure 4, S3D). For the charged residues, Arg shows the strictest requirement (Figure 4, S3E): nearby charged residues at three sites N-terminal to the mutation, particularly at the P2 site, are correlated with lethal OI; nearby charged residues at three sites C-terminal to the mutation are correlated with nonlethal OI.

Predictive Modeling for different Gly->X mutations in the $\alpha 1(I)$ chain

Decision tree analysis was performed to build predictive models of OI phenotypes and genotypes. For Gly to Ser mutations, P4, P2, P1, P2' and P5' sites are identified as the critical OI sites from odds ratio analysis (Figure 4) and are included for further decision tree modeling. Position P4 is considered as the first important node of the decision tree (Figure 5). If the residue at the P4 site is Pro, small or charged, the OI phenotype depends on the residue type at the P1 position. If the P1 residue is Pro, small or charged at the same time, the OI case tends to be nonlethal; otherwise, the OI phenotype would depend on the P2 position. If the P2 residue is charged at the same time, it tends to result in lethal OI. Otherwise, the OI phenotype would depend on the residue type at the P2' position. The decision tree model indicates that the amino acid types at the P4, P2, P1 and P2' positions collectively determine the OI lethality for Ser mutations.

Figure 5. Decision tree model of Ser substitution leading to lethal OI in the $\alpha 1(I)$ chain. L and N in the circles represent Lethal and Nonlethal OI cases, respectively. Critical positions are identified as nodes in the tree graph. Position P4 is considered as the first important node of the decision tree. The amino acid types at the P4, P2, P1 and P2' positions collectively determine the OI lethality for Ser mutations.

For Arg and Cys mutations in the $\alpha 1(I)$ chain, positions (P2, P1', P2' and P4') and (P5', P1, P4 and P4') are considered as the important nodes of the decision tree models, respectively (Figure S4, S5). It shows that these three types of mutations have a different preference for positions and residue types leading to lethal OI. The performance of the decision tree models by Ser, Arg and Cys substitutions is summarized in Table 2. For Ser and Cys mutations, ~98% of the nonlethal mutations are predicted correctly. Compared with nonlethal mutations, lethal mutations are more challenging to predict by the model. Accuracy for lethal substitutions of Ser and Cys is about 39% and 65%, respectively. The modeling behaves best for Arg mutations, with accuracy of ~90% for both lethal and nonlethal mutations. The model does not capture the information of lethal Ser and Cys substitutions very well. However, the overall accuracy of the decision tree for all three substituting residues is 85%-89%, indicating that approximately 90% of all information about lethal and nonlethal mutations is captured in this model.

Mutant Type	Lethal Cases			Non-lethal Cases			Overall Corr. %
	Correct Pred.	Wrong Pred.	Correct Pred. %	Correct Pred.	Wrong Pred.	Correct Pred. %	
Ser	22	34	39%	194	3	98%	85%
Arg	45	6	88%	54	6	90%	89%
Cys	20	11	65%	63	2	97%	86%

Table 2. Decision tree performance by Gly substitutions leading to lethal OI in the $\alpha 1(I)$ chain. For Ser and Cys mutations, ~98% of the nonlethal mutations are predicted correctly, while the accuracy for lethal substitutions is about 39% and 65%, respectively. Arg mutations show accuracy of ~90% for both lethal and nonlethal mutations.

Overview of Gly mutations in the $\alpha 2(I)$ chain

A total of 488 cases of Gly missense mutations resulting in OI in the $\alpha 2(I)$ chain were included in this study (Table 3). Unlike the $\alpha 1(I)$ chain, the $\alpha 2(I)$ chain displays much more inter-dispersed non-lethal versus lethal mutations (Figure 6). Similar to the $\alpha 1(I)$ chain, different residues that replace Gly, as well as different regions in the $\alpha 2(I)$ chain, show very different odds leading to lethal OI (Figure S6). Furthermore, the amino acid distribution of X-position residues displays no statistical difference between different regions ($p=0.91$), while the distribution of residue types at the P2 position of a mutant is significantly different ($p<0.001$) between the regions (Figure 7). Similar to the $\alpha 1(I)$ chain, the regions have similar amino acid distributions, but the mutations in the $\alpha 2(I)$ chain display very different local amino acid sequence environments in each region.

Mutation Types	Lethal Cases	Nonlethal Cases	Total	Odds
Ala	0	9	9	0.00
Ser	17	195	212	0.09
Arg	10	41	51	0.24
Cys	9	33	42	0.27
Val	16	44	60	0.36
Asp	31	54	85	0.57
Glu	12	17	29	0.71
Total	95	393	488	0.24

1 Table 3. Summary of Gly mutations in the $\alpha 2(I)$ chain of collagen. The Gly mutations with known OI lethality are included for
2
3 bioinformatics analysis. For each type of mutation (Ala, Ser, Cys, Arg, Glu, Asp and Val), the number of lethal and nonlethal
4
5 cases are counted, and their odds leading to lethal vs nonlethal OI are calculated.
6

32 Figure 6. Distribution of different types of Gly mutations (Ala, Ser, Cys, Glu, Arg, Asp and Val) along the $\alpha 2(I)$ chain. The
33
34 triple helix collagen chain is divided into five regions from the N-terminal to the C-terminal: region 1 (position<200), region 2
35
36 (position 200-400), region 3 (position 400-600), region 4 (position 600-800) and region 5 (position>800) (A). Blue and red
37
38 circles are used to represent non-lethal and lethal mutations, respectively (B). The $\alpha 2(I)$ chain displays much more inter-
39
40 dispersed non-lethal versus lethal mutations.
41

Figure 7. The distribution of amino acid types in the five regions in the $\alpha 2(I)$ chain. Collagen peptide chain is divided into five regions from the N-terminal to the C-terminal: <200(light gray), 200-400(gray), 400-600(white), 600-800(dark gray), >800(sparse white). Amino acids are classified into five categories: Pro, Small, Charged, Hydrophobic and Polar. At each position, the frequency of each type of amino acids is counted. The frequency of amino acid types in the X-position of the (Gly-X-Y)_n triplets in each region is compared (p=0.91) (A). The frequency of amino acid types in the P2 position of the mutations in each region is compared (p<0.001) (B).

Development of mutation specific local consensus sequences in the $\alpha 2(I)$ chain

Odds ratio analysis is utilized across an 8-triplet region around a Gly substitution in the $\alpha 2(I)$ chain to determine if a specific residue type and a specific position near the mutation is significantly correlated to lethal OI (Figure 8). Generally speaking, the presence of Pro and small residues at these positions plays the most important role in OI lethality (Figure 8). It is worth noting that small flexible residues at the P2-P2' positions are generally correlated with nonlethal OI (Figure 8, S7). The presence of small residues further away from the site (P11-P7, P5'-P11') mostly tends to result in lethal phenotype, which is similar to the $\alpha 1(I)$ chain (Figure 8, S7). The presence of Pro shows a much more complex pattern in the $\alpha 2(I)$ chain. For Ser mutations, Pro at position P1' and P2' tends to result in lethal OI, while Pro at position P1 and P2 tends to result in nonlethal OI (Figure 8). In contrast, for Arg and Asp mutations, Pro at the P2' position tends to cause nonlethal OI (Figure S7). Small

1 amino acids are generally favored near Gly mutations for nonlethal OI, while Pro residues have a position-to-
2 position effect.
3
4

21 Figure 8. Consensus sequence patterns for Gly to Ser, Cys and Arg mutations in the α 2(I) chain that are correlated with lethal
22 and nonlethal OI. Dependency of OI lethality on the type and position of neighboring residues within 4-triplets of Gly
23 substitutions by Ser, Cys or Arg are calculated based on odds ratio analysis in the α 2(I) chain. The relative positions of the
24 residues are indicated by Pi (the ith residue N-terminal to the mutation site) or Pi' (the ith residue C-terminal to the mutation
25 site). When the presence of certain types of amino acids at some specific positions around the mutation site is significantly
26 correlated to lethal or nonlethal OI phenotypes, the position and the amino acid type is highlighted with colored boxes (P`ro
27 colored in yellow, small flexible residues (Sma) in red, charged residues (+/-) in green, Polar residues (Pol) in olive, and
28 Hydrophobic residues (Hyd) in gray).

41 Predictive Modeling for the α 2(I) chain

43 A decision tree model was built for the most abundant Ser mutations to predict OI lethality based on local
44 amino acid sequences in the α 2(I) chain. Positions P4', P5', P1' and P2 are identified as the important nodes of
45 the decision tree model (Figure S8). If the residue at the P4' site is polar, the OI phenotype tends to be nonlethal;
46 otherwise, the OI phenotype depends on other positions. The model of Ser mutations correctly predicts ~53% of
47 lethal substitutions, while the overall accuracy is approximately 95%. It indicates that 95% of all the
48 information about lethal and nonlethal mutations is captured in this model (Table S3).

60 Model validation using novel Gly substitutions in the α 1(I) and α 2(I) chains

To validate the decision tree models, they were tested on previously unreported mutations from the newly updated cases in the Database of Collagen Mutations (www.le.ac.uk/genetics/collagen) (Dalglish, 1997). Of the 13 new cases of Gly-Ser/Arg/Cys mutations in the $\alpha 1(I)$ chain, 85% can be correctly predicted using local amino acid sequences around the mutation site (Table S4). Of the eight new cases of Gly-Ser substitutions in the $\alpha 2(I)$ chain, 88% can be correctly predicted (Table S4). 15 novel cases occur at positions that have not been observed previously in the collagen chain, indicating that the models are capable of extrapolation. The results demonstrate that the decision tree models can predict novel mutations using the local amino acid sequences reasonably well.

OI phenotype prediction in the MLBR in the $\alpha 1(I)$ chain and the lethal clusters in the $\alpha 2(I)$ chain

When the overall distribution of ligand binding sites were investigated in type I collagen, three hot spots of interactions were observed, and they were identified as Major Ligand Binding Regions (MLBR) (Di Lullo et al., 2002). The second and third major ligand binding regions (MLBR2 and MLBR3), located at positions 680-830 and 920-1014 in the $\alpha 1(I)$ chain, soon attracted a lot of attention, as the OI phenotypes resulting from the mutations in these two regions were exclusively lethal. However, nonlethal mutations were recently observed in the MLBR. We compared the distribution of residue types in the X position and the P2 position of all of the mutations in the MLBR in the $\alpha 1(I)$ chain (Figure 9). The amino acid distribution of X-position residues displays no statistical difference between the MLBR and other regions ($p=0.81$) (Figure 9A), while the distribution of residue types at the P2 position of a mutant is significantly different ($p<0.001$) (Figure 9B). It indicates that though the MLBR have similar amino acid distributions overall, the mutations in the MLBR region have very different local amino acid sequence environments. The differences in the sequence environment for a mutant are primarily different distribution patterns of Pro and small residues (Figure 9B). Similarly, eight lethal clusters were identified in the $\alpha 2(I)$ chain (Marini et al., 2007). The distribution analysis of the amino acids in the X and the P2 positions demonstrated that the mutations in the lethal clusters in the $\alpha 2(I)$ chain also displayed very different local amino acid sequence environments (Figure S9).

Figure 9. The distribution of residue types in the X and P2 position in the MLBR in the $\alpha 1(I)$ chain. Collagen peptide chain is divided into four regions: MLBR1 (light gray), MLBR2 (gray), MLBR3 (white), and other (dark gray). At each position, the frequency of each type of residues (Pro, Small, Charged, Hydrophobic and Polar) is counted. The frequency of residue types in the X-position in the four regions is compared ($p=0.81$) (A). The frequency of residue types in the P2 position of an observed mutation in the four regions is compared ($p<0.001$) (B). Local sequence environment of a mutant in the MLBR is important for OI lethality.

Furthermore, we investigated the role of local sequence in predicting OI lethality in the MLBR in the $\alpha 1(I)$ chain and the lethal clusters in the $\alpha 2(I)$ chain. To predict lethal mutations in the MLBR using the local sequence model, the prediction accuracy is ~47% for Ser mutations and ~79% for Cys mutations (data not shown). For non-lethal mutations, the model performed much better, and the prediction accuracy is 7 out of 7 for the $\alpha 1(I)$ chain and 17 out of 18 for the $\alpha 2(I)$ chain (Table S5, S6).

DISCUSSION

Local amino acid distribution around OI mutations is correlated to OI phenotypes

Though a large number of OI mutations have been identified, it remains challenging to accomplish a reliable genotype-phenotype correlation for OI. Our odds analysis of the mutations in the five different regions of the $\alpha 1(I)$ chain supports a region-dependent pattern of OI (Figure S1). The five regions show a very similar general

distribution of amino acids. However, the local distribution of amino acids near a mutation site is significantly different among the regions (Figure 3). A similar phenomenon is also observed in the $\alpha 2(I)$ chain (Figure 7), again suggesting that the local amino acid sequence environment of a mutant, rather than the sequence along the whole chain, may modulate the OI lethality. In addition, the eight lethal clusters in the $\alpha 2(I)$ chain have no significant difference of amino acid distribution from other regions; however, the local amino acid sequence environment of a mutation is quite different between the clusters and other regions (Figure S9). Our analysis suggests strongly that the local sequence environment around a mutation plays a fundamental role in OI lethality. Earlier gradient and regional models have suggested different distribution patterns of OI mutation sites, while our bio-statistical analysis has revealed that the distribution of local amino acid sequences around the mutation sites may determine OI profiles.

Consensus complementary local sequence pattern of Gly mutations highlights the importance of neighboring Pro and small amino acids for OI phenotypes

A number of important trends about which factors are related to OI phenotypes have been revealed by odds ratio analysis. The complementary relationship of the placement of Pro and small residues and their correlation to OI phenotype is noted (Figure 4, S3). When Pro is present or small flexible residues are absent near a mutation, the OI case tends to be lethal; when Pro is present or small flexible residues are absent further away from the mutation, the OI case tends to be nonlethal. NMR and CD studies of collagen model peptides have shown that conformational flexibility near the mutation site is important for structural adaptation, while a strong stabilizing Pro-containing sequence is needed for re-nucleation further downstream from the mutation (Bryan et al., 2011; Xiao et al., 2011b). It was shown that peptides with Arg mutations achieved full folding only after increasing the triple-helix stability of the N-terminal sequence as well as by introducing flexible residues near the mutation site (Xiao et al., 2011b). The presence of Pro near the mutation site may give the mutation less room to adapt, therefore resulting in lethal OI, while the presence of Pro further away may play an important role in nucleation or renucleation, therefore resulting in nonlethal OI (Hyde et al., 2006; Xiao et al., 2011b). The complementary sequential pattern of Pro and small residues seen in the odds ratio analysis is consistent with

those biophysical studies.

Different Gly->X mutations have different local sequence patterns that are correlated with lethal and nonlethal OI phenotypes

Consensus sequence patterns have been developed for individual Gly mutations that are correlated with lethal and nonlethal OI (Figure 4, 8). Synergy between the type of residue that is mutated and the sequences of the neighboring residues in determining the outcome of OI phenotype is observed. For the $\alpha 1(I)$ chain, Pro in the P5-P5' region plays the largest role for Ser mutations, while small residues are the determinant factors for Arg mutations. Pro and small residues, instead, play a more balanced role for Cys mutations. Meanwhile, hydrophobic, charged and polar residues exert additional effects on OI lethality, while different types of mutations show different preferences. Different Gly substitutions show different sequence patterns with respect to phenotype outcome, which helps to promote an understanding of why different substitutions at the same Gly position may have very different phenotypes (Marini et al., 2007). Previous Circular Dichroism (CD) and Differential Scanning Calorimetry (DSC) profiling of 41 different Gly substitutions from 47 patients has shown variable ΔT_m (the change of the melting temperature) for the same type of substituting residues at different mutation sites along the collagen chain (Makareeva et al., 2008). The structural heterogeneity of these OI mutations seems likely to result from the synergy of the identity of the substituting residue and its neighboring sequence.

Different local sequence pattern for Ser mutations in the $\alpha 1(I)$ and $\alpha 2(I)$ chains

In contrast to the Ser mutations in the $\alpha 1(I)$ chain, the Ser mutations in the $\alpha 2(I)$ chain correlate the Pro residues at the P1 and P2 positions with nonlethal OI (Figure 4, 8). It may imply that Ser mutations are easier to incorporate into the triple helix in the $\alpha 2(I)$ chain than in the $\alpha 1(I)$ chain, which is consistent with the observation of more nonlethal substitutions in the $\alpha 2(I)$ chain. Hartgerink has revealed that the Ser mutation at position 247 in either of the $\alpha 1(I)$ and $\alpha 2(I)$ chains has a differential destabilizing effect using his versatile

heterotrimer model peptide system (Gaub and Hartgerink, 2008; O'Leary et al., 2011; Fallas et al., 2012). The different local sequence pattern of the OI mutations we have identified for the $\alpha 1(I)$ and $\alpha 2(I)$ chains provides novel insights into the effects of neighboring residues on the OI mutations and may help us design model peptides to investigate the interactive effects of OI mutations and local sequences in the $\alpha 1(I)$ and $\alpha 2(I)$ chains. The consensus sequence pattern of each type of substitutions furthers our understanding of the molecular mechanism of OI.

Predictive modeling of OI phenotypes from local amino acid sequence around a Gly mutation site

The odds ratio analysis is very powerful and straightforward to interpret and can test many factors at a time. However, it may not consider different factors simultaneously. We therefore extended the bio-statistical approaches to decision tree modeling to build comprehensive models that include all potentially important factors and predict the OI phenotypes. The established decision models based on local amino acid sequences perform pretty well for the prediction of OI lethality for the dominant Ser/Arg/Cys mutations in the $\alpha 1(I)$ chain and Ser mutations in the $\alpha 2(I)$ chain with a high accuracy of at least 85% (Table 2). A composite decision tree model was built based on the identity of the substituting residue and the location of the mutant in the collagen chain and achieved an accuracy of 70% and 78% for Gly-Ser and Gly-Cys substitutions, respectively (Bodian et al., 2008). The composite model highlighted the importance of the locations of the mutations in the collagen chain, generally supporting the gradient model. Recent advances of OI studies have provided novel insight into the crucial role of higher-order interaction in OI phenotypes, highlighting the importance of the regional model (Marini et al., 2007; Sweeney et al., 2008; Forlino et al., 2011; Parkin et al., 2011; Xiao et al., 2011a). Our local sequence model has discovered, for the first time, that those special 'lethal' regions may distinguish themselves from other regions by displaying special local amino acid sequences around the mutation sites. Collagen has a seemingly uniform triple helix structure, and these special local sequences may provide a useful link to higher-order interactions and a definitive feature for OI phenotypes.

OI phenotype prediction in the MLBR is complex: non-lethal mutations can be predicted based on local

amino acid sequence

The observation of almost exclusively lethal mutations in the two Major Ligand Binding Regions (MLBR) lead to the proposal that the interference of the mutations with higher-order interactions of collagen may be the basis of lethal OI in these regions (Marini et al., 2007). The recent discovery of nonlethal mutations in the MLBR, however, complicated the phenotype-genotype relationship in the MLBR. Our sequence analysis demonstrated that the mutations in the MLBR showed very different local amino acid environment when compared with other regions, suggesting that the local sequence of those mutations may play a role in the underlying molecular basis of OI phenotypes. Our local decision tree models performed extremely well for nonlethal mutations with a high accuracy of ~96% (Table S5, S6). For lethal mutations in the MLBR, the prediction accuracy becomes much poorer. These analyses may indicate that in the MLBR, non-lethal mutations are highly predicted by local sequence considerations alone, whereas lethal mutations are not as easily predicted and may be a result of sequence effects and more complex higher order interactions.

Considering the complexity of the diagnosis of OI phenotypes, some simplifications are made in our analysis that could affect the performance of our models. First, the lethal or nonlethal OI outcome is likely confounded with many factors such as the site of the patient birth, the qualification of medical personnel, and other general factors known to affect survival of infants with severe genetic defects (e.g., ethnicity) (Ries-Levavi et al., 2004). The limited availability of OI cases in the current database is insufficient for these corrections. Second, the lethality of some mutations may also be affected by other genes such as BMP1, CRTAP, WNT1 and LEPRE1 (Forlino et al., 2011). Third, the amino acids in the $\alpha 1(I)$ and $\alpha 2(I)$ sequences may not be absolutely conserved. Common variants may affect the interpretation of local sequence effects. In addition, patients with exactly the same mutation may exhibit full range of Sillence phenotypes from mild/moderate OI I/IV to lethal OI II. For example, the Gly88Glu and Gly688Ser mutations in the $\alpha 1(I)$ chain resulted in the full range of OI I, II, III, and IV phenotypes (Marini et al., 2007). The single outcome from our predictive model is probably accounting for the dominant outcome. These complexities underline the challenges to discover the underlying molecular bases of OI.

Elucidating the sequential bases of OI phenotypes is a key to understanding the etiology of OI. The current multiple models highlight the complexity of OI. Our local amino acid sequence model complements the current OI models and provides a consensus sequence bases for OI lethality. It bridges the gap between theoretical models and biophysical experimental results by highlighting the importance of the local amino acid sequence environment. In addition, it provides a guideline for the design of model peptides that would contribute to our biophysical understanding of OI. This novel odds ratio and decision tree modeling method may be applied to other collagen-related diseases and provide novel insights to the pathogenesis and diagnosis.

ACKNOWLEDGEMENTS

This work was supported by grants from the National Institutes of Health GM45302 and the National Science Foundation DBI-0403062 and DBI-0320746 to J.B, and the National Natural Science Foundation of China (Grant No. 21305056) to J.X. We would like to thank Maria Janowska for many helpful discussions and Cody Hoop for critical reading of the manuscript.

APPENDIX. SUPPORTING MATERIAL

Figure S1-9 and Table S1-6.

REFERENCES

- Beck, K., Chan, V. C., Shenoy, N., Kirkpatrick, A., Ramshaw, J. A., Brodsky, B., 2000. Destabilization of osteogenesis imperfecta collagen-like model peptides correlates with the identity of the residue replacing glycine. *Proc Natl Acad Sci U S A* 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.
- Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Sayers, E. W., 2013. GenBank. *Nucleic Acids Res* 41, D36-42.
- Bertini, I., Fragai, M., Luchinat, C., Melikian, M., Toccafondi, M., Lauer, J. L., Fields, G. B., 2012. Structural basis for matrix metalloproteinase 1-catalyzed collagenolysis. *J Am Chem Soc* 134, 2100-2110.
- Bodian, D. L., Klein, T. E., 2009. COLdb, a database linking genetic data to molecular function in fibrillar collagens. *Hum Mutat* 30, 946-951.
- Bodian, D. L., Madhan, B., Brodsky, B., Klein, T. E., 2008. Predicting the clinical lethality of osteogenesis imperfecta from collagen glycine mutations. *Biochemistry* 47, 5424-5432.

- 1 Bosanquet, D. C.,Glasbey, J. C.,Williams, I. M.,Twine, C. P., 2014. Systematic Review and Meta-analysis of
2 Direct Versus Indirect Angiosomal Revascularisation of Infrapopliteal Arteries. *Eur J Vasc Endovasc Surg*.
- 3 Brodsky, B.,Thiagarajan, G.,Madhan, B.,Kar, K., 2008. Triple-helical peptides: an approach to collagen
4 conformation, stability, and self-association. *Biopolymers* 89, 345-353.
- 5 Bryan, M. A.,Brauner, J. W.,Anderle, G.,Flach, C. R.,Brodsky, B.,Mendelsohn, R., 2007. FTIR studies of
6 collagen model peptides: complementary experimental and simulation approaches to conformation and
7 unfolding. *J Am Chem Soc* 129, 7877-7884.
- 8
9 Bryan, M. A.,Cheng, H.,Brodsky, B., 2011. Sequence environment of mutation affects stability and folding in
10 collagen model peptides of osteogenesis imperfecta. *Biopolymers* 96, 4-13.
- 11 Byers, P. H., 1989. Inherited disorders of collagen gene structure and expression. *Am J Med Genet* 34, 72-80.
- 12 Byers, P. H.,Cole, W. G. 2002. Osteogenesis Imperfecta. In: Royce P. M., Steinmann B., editors. *Connective*
13 *tissue and its hereditary disorders*. New York: Wiley-Liss. p 385-430.
- 14 Cohn, D. H.,Byers, P. H., 1990. Clinical screening for collagen defects in connective tissue diseases. *Clin*
15 *Perinatol* 17, 793-809.
- 16 Dalgleish, R., 1997. The human type I collagen mutation database. *Nucleic Acids Res* 25, 181-187.
- 17 Dalgleish, R., 1998. The Human Collagen Mutation Database 1998. *Nucleic Acids Res* 26, 253-255.
- 18 den Dunnen, J. T.,Antonarakis, S. E., 2001. Nomenclature for the description of human sequence variations.
19 *Hum Genet* 109, 121-124.
- 20 Di Lullo, G. A.,Sweeney, S. M.,Korkko, J.,Ala-Kokko, L.,San Antonio, J. D., 2002. Mapping the ligand-
21 binding sites and disease-associated mutations on the most abundant protein in the human, type I collagen. *J*
22 *Biol Chem* 277, 4223-4231.
- 23 Emsley, J.,Knight, C. G.,Farndale, R. W.,Barnes, M. J., 2004. Structure of the integrin alpha2beta1-binding
24 collagen peptide. *J Mol Biol* 335, 1019-1028.
- 25 Emsley, J.,Knight, C. G.,Farndale, R. W.,Barnes, M. J.,Liddington, R. C., 2000. Structural basis of collagen
26 recognition by integrin alpha2beta1. *Cell* 101, 47-56.
- 27 Fallas, J. A.,Gaub, V.,Hartgerink, J. D., 2009. Solution structure of an ABC collagen heterotrimer reveals a
28 single-register helix stabilized by electrostatic interactions. *J Biol Chem*.
- 29 Fallas, J. A.,Lee, M. A.,Jalan, A. A.,Hartgerink, J. D., 2012. Rational design of single-composition ABC
30 collagen heterotrimers. *J Am Chem Soc* 134, 1430-1433.
- 31 Forlino, A.,Cabral, W. A.,Barnes, A. M.,Marini, J. C., 2011. New perspectives on osteogenesis imperfecta. *Nat*
32 *Rev Endocrinol* 7, 540-557.
- 33 Gaub, V.,Hartgerink, J. D., 2007. Self-assembled heterotrimeric collagen triple helices directed through
34 electrostatic interactions. *J Am Chem Soc* 129, 2683-2690.
- 35 Gaub, V.,Hartgerink, J. D., 2008. Synthetic collagen heterotrimers: structural mimics of wild-type and mutant
36 collagen type I. *J Am Chem Soc* 130, 7509-7515.
- 37 Hohenester, E.,Sasaki, T.,Giudici, C.,Farndale, R. W.,Bachinger, H. P., 2008. Structural basis of sequence-
38 specific collagen recognition by SPARC. *Proc Natl Acad Sci U S A* 105, 18273-18277.
- 39 Hyde, T. J.,Bryan, M. A.,Brodsky, B.,Baum, J., 2006. Sequence dependence of renucleation after a Gly
40 mutation in model collagen peptides. *J Biol Chem* 281, 36937-36943.
- 41 Kramer, R. Z.,Bella, J.,Mayville, P.,Brodsky, B.,Berman, H. M., 1999. Sequence dependent conformational
42 variations of collagen triple-helical structure. *Nat Struct Biol* 6, 454-457.
- 43 Lauer-Fields, J. L.,Fields, G. B., 2002. Triple-helical peptide analysis of collagenolytic protease activity. *Biol*
44 *Chem* 383, 1095-1105.
- 45 Li, Y.,Brodsky, B.,Baum, J., 2009. NMR conformational and dynamic consequences of a gly to ser substitution
46 in an osteogenesis imperfecta collagen model Peptide. *J Biol Chem* 284, 20660-20667.
- 47 Li, Y.,Kim, S.,Brodsky, B.,Baum, J., 2005. Identification of partially disordered peptide intermediates through
48 residue-specific NMR diffusion measurements. *J Am Chem Soc* 127, 10490-10491.
- 49 Lu, K. G.,Stultz, C. M., 2013. Insight into the degradation of type-I collagen fibrils by MMP-8. *J Mol Biol* 425,
50 1815-1825.
- 51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

- 1 Makareeva, E., Mertz, E. L., Kuznetsova, N. V., Sutter, M. B., DeRidder, A. M., Cabral, W. A., Barnes, A.
2 M., McBride, D. J., Marini, J. C., Leikin, S., 2008. Structural heterogeneity of type I collagen triple helix and
3 its role in osteogenesis imperfecta. *J Biol Chem* 283, 4787-4798.
- 4 Marini, J. C., Forlino, A., Cabral, W. A., Barnes, A. M., San Antonio, J. D., Milgrom, S., Hyland, J. C., Korkko,
5 J., Prockop, D. J., De Paepe, A., Coucke, P., Symoens, S., Glorieux, F. H., Roughley, P. J., Lund, A. M., Kuurila-
6 Svahn, K., Hartikka, H., Cohn, D. H., Krakow, D., Mottes, M., Schwarze, U., Chen, D., Yang, K., Kuslich,
7 C., Troendle, J., Dagleish, R., Byers, P. H., 2007. Consortium for osteogenesis imperfecta mutations in the
8 helical domain of type I collagen: regions rich in lethal mutations align with collagen binding sites for
9 integrins and proteoglycans. *Hum Mutat* 28, 209-221.
- 10 O'Leary, L. E., Fallas, J. A., Hartgerink, J. D., 2011. Positive and negative design leads to compositional control
11 in AAB collagen heterotrimers. *J Am Chem Soc* 133, 5432-5443.
- 12 Parkin, J. D., San Antonio, J. D., Pedchenko, V., Hudson, B., Jensen, S. T., Savage, J., 2011. Mapping structural
13 landmarks, ligand binding sites, and missense mutations to the collagen IV heterotrimers predicts major
14 functional domains, novel interactions, and variation in phenotypes in inherited diseases affecting basement
15 membranes. *Hum Mutat* 32, 127-143.
- 16 Persikov, A. V., Ramshaw, J. A., Brodsky, B., 2000. Collagen model peptides: Sequence dependence of triple-
17 helix stability. *Biopolymers* 55, 436-450.
- 18 Prockop, D. J., Constantinou, C. D., Dombrowski, K. E., Hojima, Y., Kadler, K. E., Kuivaniemi, H., Tromp,
19 G., Vogel, B. E., 1989. Type I procollagen: the gene-protein system that harbors most of the mutations
20 causing osteogenesis imperfecta and probably more common heritable disorders of connective tissue. *Am J*
21 *Med Genet* 34, 60-67.
- 22 Ramachandran, G. N., Kartha, G., 1955. Structure of collagen. *Nature* 176, 593-595.
- 23 Raynal, N., Hamaia, S. W., Siljander, P. R., Maddox, B., Peachey, A. R., Fernandez, R., Foley, L. J., Slatter, D.
24 A., Jarvis, G. E., Farndale, R. W., 2006. Use of synthetic peptides to locate novel integrin alpha2beta1-binding
25 motifs in human collagen III. *J Biol Chem* 281, 3821-3831.
- 26 Reuter, M. S., Schwabe, G. C., Ehlers, C., Marschall, C., Reis, A., Thiel, C., Graul-Neumann, L., 2013. Two novel
27 distinct COL1A2 mutations highlight the complexity of genotype-phenotype correlations in osteogenesis
28 imperfecta and related connective tissue disorders. *Eur J Med Genet* 56, 669-673.
- 29 Rich, A., Crick, F. H., 1955. The structure of collagen. *Nature* 176, 915-916.
- 30 Ries-Levavi, L., Ish-Shalom, T., Frydman, M., Lev, D., Cohen, S., Barkai, G., Goldman, B., Byers, P., Friedman, E.,
31 2004. Genetic and biochemical analyses of Israeli osteogenesis imperfecta patients. *Hum Mutat* 23, 399-400.
- 32 Shoulders, M. D., Raines, R. T., 2009. Collagen structure and stability. *Annu Rev Biochem* 78, 929-958.
- 33 Sillence, D. O., Senn, A., Danks, D. M., 1979. Genetic heterogeneity in osteogenesis imperfecta. *J Med Genet* 16,
34 101-116.
- 35 Stultz, C. M., 2002. Localized unfolding of collagen explains collagenase cleavage near imino-poor sites. *J Mol*
36 *Biol* 319, 997-1003.
- 37 Sweeney, S. M., Orgel, J. P., Fertala, A., McAuliffe, J. D., Turner, K. R., Di Lullo, G. A., Chen, S., Antipova,
38 O., Perumal, S., Ala-Kokko, L., Forlino, A., Cabral, W. A., Barnes, A. M., Marini, J. C., San Antonio, J. D., 2008.
39 Candidate cell and matrix interaction domains on the collagen fibril, the predominant protein of vertebrates. *J*
40 *Biol Chem* 283, 21187-21197.
- 41 Sykes, B., 1985. The molecular genetics of collagen. *Bioessays* 3, 112-117.
- 42 Sztrolovics, R., Glorieux, F. H., van der Rest, M., Roughley, P. J., 1993. Identification of type I collagen gene
43 (COL1A2) mutations in nonlethal osteogenesis imperfecta. *Hum Mol Genet* 2, 1319-1321.
- 44 Tsai, M. I., Xu, Y., Dannenberg, J. J., 2005. Completely geometrically optimized DFT/ONIOM triple-helical
45 collagen-like structures containing the ProProGly, ProProAla, ProProDAla, and ProProDSer triads. *J Am*
46 *Chem Soc* 127, 14130-14131.
- 47 Van Dijk, F. S., Sillence, D. O., 2014. Osteogenesis imperfecta: clinical diagnosis, nomenclature and severity
48 assessment. *Am J Med Genet A* 164A, 1470-1481.
- 49 Wang, Q., Orrison, B. M., Marini, J. C., 1993. Two additional cases of osteogenesis imperfecta with substitutions
50 for glycine in the alpha 2(I) collagen chain. A regional model relating mutation location with phenotype. *J*
51 *Biol Chem* 268, 25162-25167.

- 1 Xiao, J.,Addabbo, R. M.,Lauer, J. L.,Fields, G. B.,Baum, J., 2010. Local conformation and dynamics of
2 isoleucine in the collagenase cleavage site provide a recognition signal for matrix metalloproteinases. *J Biol*
3 *Chem* 285, 34181-34190.
- 4 Xiao, J.,Baum, J., 2009. Structural insights from (15)N relaxation data for an anisotropic collagen peptide. *J Am*
5 *Chem Soc* 131, 18194-18195.
- 6 Xiao, J.,Cheng, H.,Silva, T.,Baum, J.,Brodsky, B., 2011a. Osteogenesis imperfecta missense mutations in
7 collagen: structural consequences of a glycine to alanine replacement at a highly charged site. *Biochemistry*
8 *50*, 10771-10780.
- 9 Xiao, J.,Madhan, B.,Li, Y.,Brodsky, B.,Baum, J., 2011b. Osteogenesis imperfecta model peptides:
10 incorporation of residues replacing Gly within a triple helix achieved by renucleation and local flexibility.
11 *Biophys J* 101, 449-458.
- 12 Xu, K.,Nowak, I.,Kirchner, M.,Xu, Y., 2008a. Recombinant collagen studies link the severe conformational
13 changes induced by osteogenesis imperfecta mutations to the disruption of a set of interchain salt bridges. *J*
14 *Biol Chem* 283, 34337-34344.
- 15 Xu, P.,Huang, J.,Cebe, P.,Kaplan, D. L., 2008b. Osteogenesis imperfecta collagen-like peptides: self-assembly
16 and mineralization on surfaces. *Biomacromolecules* 9, 1551-1557.
- 17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65