Structural Models of Osteogenesis Imperfecta-associated Variants in the *COL1A1* Gene*

Sean D. Mooney and Teri E. Klein‡

Osteogenesis imperfecta (OI) is a genetic disease in which the most common mutations result in substitutions for glycine residues in the triple helical domain of the chains of type I collagen. Currently there is no way to use sequence information to predict the clinical OI phenotype. However, structural models coupled with biophysical and machine learning methods may be able to predict sequences that, when mutated, would be associated with more severe forms of OI. To build appropriate structural models, we have applied a high throughput molecular dynamic approach. Homotrimeric peptides covering 57 positions in which mutations are associated with OI were simulated both with and without mutations. Our models revealed structural differences that occur with different substituting amino acids. When mutations were introduced, we observed a decrease in helix stability, as caused by fewer main chain backbone hydrogen bonds, and an increase in main chain root mean square deviation and specifically bound water molecules. ***Molecular & Cellular Proteomics 1:868–875, 2002.***

Computational tools for analyzing the molecular consequences of genetic variation are of current interest because they have the potential to characterize the molecular pathogenesis for disease (1–3). In this study, we applied a molecular dynamics method to the analysis of the structural consequences of 57 disease-associated mutations in the α1(I) chain of type I collagen. Our study revealed differences between mutant and wild type peptides and provides a theory for how mutations compensate for lost stability. This method is a promising approach that can be applied to other genes involving non-synonymous disease-associated mutations.

Although there are currently over 13,000 phenotypically annotated mutations in the Human Genome Mutation Database (20), we have little understanding of how most of these mutations result in the clinical picture associated with them. In combination with other methods, molecular dynamic methods can provide some insight into how disease-associated mutations confer a phenotype.

Type I collagen, the most abundant protein in animals, is a structural protein. Among other functions, it protects soft tissues, supports them, and, in vertebrates, connects them with the skeleton. Type I collagen provides animals with the ability to withstand forces such as pressure, torsion, and tension as well as to transmit forces from muscle to the skeleton. Its most distinctive structural feature is a triple helix characterized by *X*-*Y*-Gly repeating amino acid motifs in each of the three chains. This sequence naturally adopts a triple helix structure once nucleation has occurred.

Collagen molecules form a diverse family of 24 types encoded by over 30 genes. Fibrillar collagens such as type I, which is the major protein in bone, contain a long uninterrupted triple helix of over 1000 residues. Non-fibrillar collagens such as type IV, which is a basement membrane collagen, contain numerous disruptions across the length of the triple helix. Type I collagen is a heterotrimer formed by the products of two genes: *COL1A1* and *COL1A2*. The *COL1A1* gene located on the long arm of chromosome 17 (17q21.31–17q22.05), and the *COL1A2* gene is on chromosome 7 (7q21.3–7q22.1) (4). The coordinates of these genes in the draft human genome sequence are: chromosome 17:54653617–54671961 (*COL1A1*) and chr7:92756774–92793062 (*COL1A2*) (genome.ucsc.edu/).

Clinically, mutations in type I collagen genes are associated with osteogenesis imperfecta (OI)[1] and some forms of Ehlers-Danlos syndrome. Clinical, genetic, and radiological data have been used to classify OI into four types (for a review, see Ref. 5). OI type II is the most severe form and is usually lethal in the perinatal period. OI type I is characterized by multiple bone fractures, usually resulting from minimal trauma, autosomal dominant inheritance, and blue sclera. OI type III is a relatively severe form that is identified by very short stature, brittle bones, and blue sclera in infancy. OI type IV is intermediate in severity between OI type I and OI type III.

The majority of identified mutations are single nucleotide substitutions that result in alteration of glycine codons within the triple helical domain of either of the chains of type I procollagen. These mutations produce phenotypes that range from mild to lethal and appear to depend, in part, on the chain

From the Department of Genetics, School of Medicine, Stanford University, Stanford, California 94305

Received, September 29, 2002, and in revised form, October 11, 2002

Published, MCP Papers in Press, October 11, 2002, DOI 10.1074/mcp.M200064-MCP200

[1] The abbreviations used are: OI, osteogenesis imperfecta; r.m.s.d., root mean square deviation.

in which the substitution occurs, the position of the mutation in the chain, and the substituting amino acid. However, the rules that relate these characteristics, and the influence of additional factors, remain unspecified.

It has been difficult to understand and study the effects of these mutations in tissues or cells from individuals who are affected with them. In an effort to reduce the complexity of the molecule population, one group studied short peptides that incorporated sequences surrounding two mutations, the lethal mutation G913S and the non-lethal mutation G901S (6, 7) by NMR spectroscopy (numbers 901 and 913 refer to the position within the triple helix, and G and S are the single letter codes for glycine and serine, respectively). The thermal stability of the sequences flanking the lethal mutation was decreased in comparison to the stability of the sequences flanking the non-lethal mutation. Another study showed that the type of amino acid substituted for native glycine affected thermal stability (8). For example, substituting alanine or serine caused the melting point of a triple helix to decrease by 35 °C, while substituting arginine caused it to decrease more than 45 °C. The study showed a correlation between the level of destabilization and the severity of OI (8). Thus, it was proposed that stabilization apparently plays a major role in the severity of OI-associated mutations. This fact also suggests that short homotrimers can be used to model the naturally occurring heterotrimers in some cases. Structural NMR studies on similar peptides suggested that the lethal mutation altered the structure of the triple helix asymmetrically (7). This asymmetric loss of triple helical structure was attributed to disruption of the folding of the triple helix.

Substitutions for glycine within the triple helix are severely destabilizing and are usually considered to interrupt the triple helix (10). We have reproduced the destabilization caused by introducing alanine into the central glycine position of short idealized collagen-like peptide models (11). These models predict that alanine is destabilizing because of unfavorable steric and electrostatic interactions. As the interstitial region of the triple helix fills with mutating residue side chains, main chain hydrogen bonds break near the interstitial region of the triple helix (12). The serine residue side chain contains the potential to hydrogen bond with the carbonyl oxygen and amide nitrogen of the peptide backbone. This interaction can either be with an adjacent chain, with itself, or with a neighboring residue on the same chain (13).

This study continues our modeling of collagen-like peptides. We have built simplified homotrimeric models of 57 OI-associated mutations in the *COL1A1* gene and their corresponding wild type peptides. We analyzed energetic and structural relationships in abnormal collagen as well and identified structural features that may be important to collagen biology. When mutations were introduced, we observed a decrease in helical structure, specifically in main chain backbone hydrogen bonds, and an increase in main chain-bound water molecules. Because of the hydrogen bonding potential of the serine side chain, serine residue mutations are of particular interest. Mutant serine residues were usually observed hydrogen bonding with a backbone atom of an adjacent chain.

EXPERIMENTAL PROCEDURES

We chose 57 serine, cysteine, alanine, and valine mutations that are listed in the collagen mutation data base (14, 15) (www.le.ac.uk/genetics/collagen/). Because of computational limitations we worked with this subset of all of the mutations in the data base. To prepare the sequences for simulation, we collected the 8 amino acids centered on the mutant triplet position. We added the sequence "GPOG" to the N-terminal end and the sequence "POG" to the C-terminal end to help prevent unwinding (O is the single letter code for the modified proline residue, 4-hydroxyproline). All proline *Y* positions in the *X*-*Y*-Gly triplet were hydroxylated. Each peptide contained 31 residues, 15 on either side of the mutation. To summarize, each of the sequences had the following structure: (GPOG)(*XY*G)$_3$(*XYM*)(*XY*G)$_4$(POG) where *XY* are the native sequence amino acids and *M* is the mutating residue.

Homotrimeric structures were modeled using Gencollagen (16). Gencollagen models an arbitrary collagen-like sequence into a triple helical conformation and outputs a Protein Data Bank structure. These peptide models contained regular triple helical structures with extended side chains. Each of the 57 files that resulted from the modeling was copied and edited to give it the appropriate mutation position in the center of the mutation. This process gave 57 mutant structures and 57 native structures. Next we used the Leap program of the AMBER simulation package (17) to add hydrogen atoms to each of the structures and to add side chains for the mutant positions. All simulations were performed with AMBER version 6.0 (17). Counter-ions (Na^+ or Cl^-) were added to neutralize the simulation system.

Using Leap, each structure was placed in a box of solvent that extended at least 10 Å away from the triple helical peptide. The structures were equilibrated by performing 200 steps of conjugate gradient minimization on the entire system followed by a slow warm up to 280 K over 20 ps of solvent equilibration. Solvent equilibration was followed by a 3000-step minimization on the entire system and another slow warm up of the whole system to 280 K over 7.5 ps. Production dynamics were then performed for 400 ps at 280 K using a 2-fs time step. We chose 280 K because previous simulations showed that higher temperatures were often not suitable for simulating the more destabilized mutations.

After simulating each mutant and wild type peptide model, we calculated the structural differences between an idealized triple helix and the simulated structures. r.m.s.d. is used as a metric to determine the deviation from an idealized triple helical structure (defined from (POG)$_x$)and is defined as the square root of the average square distance between points in two models being compared. Triple helix ideality is defined using backbone parameters from a polyproline II helix (ϕ, ψ, ω of Gly-*X*-*Y* are −74°, 170°, 180°, −75°, 168°, 180°, −75°, 153°, and 180°). We calculate the r.m.s.d. by aligning the central region of a simulated structure with that of an idealized triple helix and then calculating the carbon α (Cα) r.m.s.d. using the 9 Cα atoms closest to the mutation site. We used central Cα atoms because movement at the end of a peptide is very noisy and therefore does not relate how much perturbation is caused by the mutation itself. This protocol was performed on each simulation by sampling structures every picosecond. Average r.m.s.d. values were calculated from 120–400 ps and compared between the wild type and mutated peptides, the residue identities, and the clinical OI type. r.m.s.d. values of 0.0 correspond to a perfectly idealized triple helix, and larger values represent larger deviations from ideality.

We also monitored the interruption of the regular hydrogen bonding network. This process identified all solute-solute hydrogen bonds that were present more than 80% of the time, beginning after 100 ps of

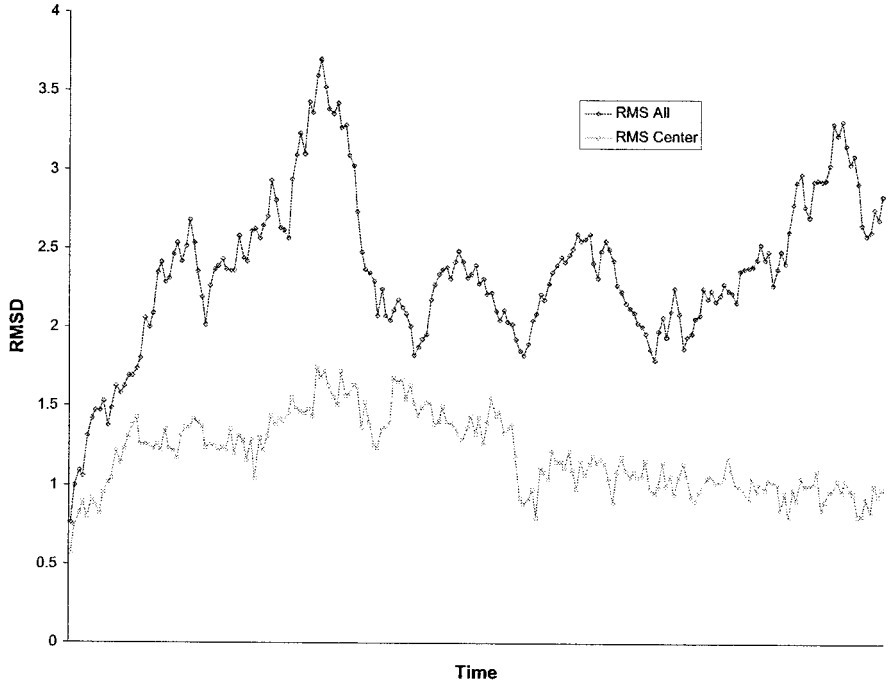

FIG. 1. **Comparison of Cα r.m.s.d. values using the central region of the helix *versus* the entire peptide.** Data series "*RMS All*" shows the r.m.s.d. difference between the entire peptide structure and an idealized triple helix, while "*RMS Center*" shows the same structure compared only using the central Cα atoms in a protocol analogous to that described in the text. *RMS*, root mean square; *RMSD*, root mean square deviation.

dynamics. Hydrogen bonds were labeled "main chain" if both donor and acceptor were part of the main chain of the triple helix and thus were internal to the triple helix. All others were labeled side chain and thus were external to the triple helix itself. The first two triplets on each end were not included in the analysis. To determine whether solvent hydrogen bonding was compensating for the lost interchain hydrogen bonds, we counted the number of solvent backbone-solute hydrogen bonds in the wild type and mutant peptides. Because serines in the place of glycines in the triple helix have increased hydrogen bonding potential, we performed a special analysis of side chains when these residues were involved.

Due to the large volume of data, we were not able to rigorously characterize every mutation. We chose three simulations for visualization and characterization. The models that we chose involved many specifically bound solvent molecules, lost interchain hydrogen bonds, and peptides with interesting serine binding patterns.

RESULTS

After scripts were developed for automatically running the required software packages described under "Experimental Procedures," each peptide was simulated on an SGI Origin 3800 class supercomputer. The total time for all simulations was just under 3 days with each running on an individual processor. Each of the 114 simulations showed density equilibration well within 50 ps of the 400-ps simulation. Fig. 1 shows the difference between using the internal Cα values and using all Cα values for calculating r.m.s.d. values. The internal coordinates give more information on the structural integrity of the mutation environment because the peptide ends are not considered, while all the Cα atoms showed the overall dynamics of the peptide ends. Table I shows all the mutations used with the sequences, OI phenotypes, total charge, and average r.m.s.d. of all mutant and wild type peptide models.

The average idealized r.m.s.d. of all mutated peptides is 1.34 ± 0.07, and the average for each wild type peptide is 0.79 ± 0.03. The average r.m.s.d. for each simulated peptide is shown in Table I. The 4 alanine mutations showed the lowest r.m.s.d. difference, which is 1.17 ± 0.21, with the wild type forms showing a r.m.s.d. of 0.77 ± 0.16. The 23 serine residues showed an average r.m.s.d. of 1.25 ± 0.21 with wild type forms showing 0.75 ± 0.14. The 24 cysteine residues showed a slightly greater r.m.s.d. of 1.39 ± 0.33 with the wild type forms showing a r.m.s.d. of 0.81 ± 0.25. The seven analyzed valine residues showed average r.m.s.d. values of 1.56 ± 0.32 with the wild type forms showing average r.m.s.d. values of 0.93 ± 0.32. r.m.s.d. values of all peptides correlated with molecular weight of the side chain of the substituting residue. These results are summarized in Table II.

To separate by lethality of OI, a total of five mutations (4 serine and 1 valine) were not analyzed because they had been associated with both a lethal (Type II) and a non-lethal form of OI (Type I, III, or IV). The results are summarized in Tables I and V. Surprisingly, in all but one case, mutations of lethal forms showed r.m.s.d. values less than that of the non-lethal forms for the same substituting residue. While the reason for this difference is not clear, one possibility is that the more severe mutations may be "stiffer" and less able to compensate for the perturbing nature of the mutation.

Table I summarizes the average number of hydrogen bonds for the central region of the helix. In 79% of the peptides, the number of backbone hydrogen bonds present at least 80% of the time decreased when the mutation was introduced. The wild type peptides showed an average of 15.4 ± 0.8 hydrogen bonds involving only main chain atoms that were present more than 80%, while the mutant peptides showed an aver-

TABLE I

Comparison of the mutations modeled showing structural differences between each of the disease-associated mutations

Columns are as follows. POS, identity of mutation and position along the chain. OI, osteogenesis imperfecta clinical phenotype. r.m.s.d., average r.m.s.d. of wild type peptides (standard deviation). r.m.s.d. MT, average r.m.s.d. of mutant peptides (standard deviation). HB, all hydrogen bonds between main chain atoms in wild type peptides. HB MT, all hydrogen bonds between main chain atoms in mutant peptides. SHB, all solvent-solute hydrogen bonds between main chain atoms in wild type peptides. SHB MT, all solvent hydrogen bonds between main chain atoms in mutant peptides.

POS	OI[a]	r.m.s.d.	r.m.s.d. MT	HB[b]	HB MT	SHB	SHB MT
G910A	*	1.33 (0.16)	0.80 (0.14)	16	15	1	1
G928A	*	1.10 (0.13)	0.66 (0.08)	16	15	5	9
G256V	*	1.36 (0.16)	0.69 (0.07)	17	12	1	3
G586V	*	1.63 (0.15)	1.41 (0.39)	12	12	3	3
G637V	*	1.59 (0.20)	0.77 (0.13)	14	12	3	4
G802V	*	1.44 (0.23)	0.72 (0.11)	16	13	6	6
G844V	*	1.58 (0.18)	0.89 (0.20)	15	12	0	9
G973V	*	2.01 (0.44)	0.86 (0.12)	17	12	2	3
G451S	*	1.22 (0.17)	0.73 (0.13)	16	13	1	2
G478S	*	1.14 (0.15)	0.77 (0.15)	15	14	5	5
G565S	*	1.33 (0.15)	0.76 (0.12)	16	16	3	4
G598S	*	1.23 (0.15)	0.71 (0.09)	15	14	1	1
G631S	*	1.22 (0.16)	0.72 (0.10)	16	13	2	6
G913S*	*	1.25 (0.13)	0.68 (0.09)	16	15	3	3
G964S	*	1.39 (0.13)	0.75 (0.09)	16	12	0	3
G691C	*	1.69 (0.26)	0.71 (0.09)	16	13	6	1
G718C	*	1.33 (0.11)	0.69 (0.09)	16	14	2	0
G748C	*	1.17 (0.12)	0.80 (0.12)	16	14	2	3
G244C	*	1.27 (0.10)	0.59 (0.06)	16	15	1	5
G448C	*	1.93 (0.20)	0.73 (0.10)	16	15	0	1
G904C	*	1.06 (0.11)	0.73 (0.10)	16	14	0	0
G946C	*	1.22 (0.12)	0.71 (0.10)	16	15	3	2
G988C	*	1.27 (0.20)	0.77 (0.11)	15	13	2	9
G85V	I	1.31 (0.13)	1.18 (0.26)	15	11	3	10
G901S	I	1.25 (0.11)	0.63 (0.07)	15	15	0	1
G43C	I	1.23 (0.16)	0.72 (0.11)	18	16	1	2
G46C	I	1.31 (0.16)	0.74 (0.12)	17	17	1	1
G94C	I	1.64 (0.25)	0.95 (0.14)	14	13	4	4
G205C	I	1.60 (0.25)	0.75 (0.13)	16	14	0	0
G223C	I, IV	1.01 (0.08)	0.87 (0.11)	15	16	3	6
G154A	III	1.21 (0.23)	0.91 (0.17)	17	15	0	5
G247S	*, III	1.16 (0.15)	0.75 (0.13)	16	14	1	1
G541S	III	1.36 (0.13)	0.76 (0.13)	15	17	5	4
G589S	III, IV	1.35 (0.10)	0.70 (0.09)	16	14	0	2
G601S	III, IV	1.58 (0.26)	0.73 (0.12)	16	14	4	7
G643S	III	1.00 (0.10)	0.85 (0.23)	16	14	7	5
G661S	III	1.40 (0.16)	0.79 (0.13)	16	15	1	1
G415S	*, III, IV	1.40 (0.16)	0.90 (0.18)	16	14	5	7
G844S	III	1.01 (0.09)	0.72 (0.08)	16	14	4	6
G862S	III, *	1.19 (0.12)	0.73 (0.11)	15	11	6	5
G871S	III	1.08 (0.12)	0.75 (0.11)	16	15	0	0
G898S	III	1.36 (0.20)	0.74 (0.11)	16	14	2	3
G973S	III	1.19 (0.20)	0.81 (0.16)	17	14	1	0
G175C	III, IV	1.16 (0.19)	0.69 (0.08)	16	16	3	8
G211C	III, IV	2.10 (0.34)	1.27 (0.19)	14	13	0	2
G226C	III, IV	1.15 (0.20)	0.68 (0.10)	16	16	5	3
G415C	III, IV	1.42 (0.21)	0.90 (0.21)	14	16	4	4
G526C	III	1.30 (0.16)	0.80 (0.17)	18	18	3	5
G868C	III	1.31 (0.19)	0.70 (0.09)	16	13	2	1
G220A	IV	1.04 (0.18)	0.71 (0.08)	16	14	5	4
G382S	IV	1.28 (0.14)	0.76 (0.10)	15	11	3	0
G448S	IV	1.50 (0.21)	0.77 (0.17)	17	16	4	2
G832S	IV	1.04 (0.17)	0.75 (0.09)	14	13	5	5
G880S	IV	1.15 (0.09)	0.64 (0.07)	16	15	2	0
G178C	IV	1.36 (0.18)	0.70 (0.09)	16	16	2	2
G349C	IV	1.30 (0.15)	0.83 (0.22)	15	16	3	4
G382C	IV	1.49 (0.29)	1.45 (0.53)	15	10	3	1
G523C	IV	1.62 (0.28)	0.76 (0.10)	19	17	1	1

[a] Lethal OI (Type II) is represented by an asterisk.

[b] All hydrogen bond counts are those present greater than 80% of the time.

age of 13.6 ± 1.5. Hydrogen bonds involving side chains were much rarer and were more exchangeable, showing an average of 0.34 ± 0.79 side chain hydrogen bonds in wild type peptides and an average of 0.57 ± 0.84 side chain hydrogen bonds in the mutant peptides for hydrogen bonds that were present more than 80% of the time. To determine whether these lost hydrogen bonds were compensated for by solvent hydrogen bonding, we determined the number of solvent and backbone solute hydrogen bonds. Because many of these hydrogen bonds exchange quickly, we sampled all hydrogen bonds that were present more than 10, 20, 30, . . . and 90% of the time. The results are summarized in Tables I and II. Table III suggests that the mutant species have more specific solvent hydrogen bonds that exchange less frequently, further suggesting that the fixing of a few specific water molecules helps to compensate for lost solute hydrogen bond stability. Table IV shows the structural parameters grouped by osteogenesis imperfecta phenotypes.

Serine residues are of particular interest because they have a hydrogen bonding side chain. Table V shows hydrogen bonding patterns for each of the 24 mutant serine residues. Fig. 2 shows only hydrogen bonds physically involving the central, mutant serine residue. As we observed in the other peptides, most of the hydrogen bonding patterns were due to highly exchanging hydrogen bonds that were present less than 10% of the time. Most of the low exchanging hydrogen bonds were with the backbone of an adjacent chain. In a few cases, such as G415S, G601S, G631S, and G898S, there were specific water molecules that bound directly to the serine side chain. Fig. 2 illustrates the range of interactions in which serine residues participated.

The two peptides studied previously, G901S and the lethal G913S (6), showed different hydrogen bonding patterns. The overall hydrogen bond counts between the solute and solvent were similar. The differences between hydrogen bonding patterns of the serine side chains were, however, very different. In the non-lethal peptide, G901S, we observed 157 total hydrogen bonds with the mutant serine side chains, while in the lethal peptide, G913S, we observed 76 total hydrogen bonds.

TABLE II

Comparison of the four amino acids substituted in different osteogenesis imperfecta-associated peptides

The structural parameters for each of the peptides averaged by mutant amino acid identity identify valine as the most perturbing and alanine as the least perturbing. See text for discussion. WT, wild type.

	No. of sequences	r.m.s.d.	r.m.s.d. WT	H-bonds	H-bonds WT	H-bond difference	Solvent H-bonds	Solvent H-bonds WT
Ala	4	1.17	0.77	16.3	14.3	−1.5	4.8	2.8
		(0.21)	(0.16)	(0.5)	(0.5)	(0.7)	(3.3)	(2.6)
Val	7	1.56	0.93	15.1	12.0	−3.1	5.4	2.6
		(0.32)	(0.32)	(1.8)	(0.6)	(1.9)	(3.0)	(1.9)
Ser	24	1.25	0.75	15.9	14.8	−1.1	3.0	2.7
		(0.21)	(0.14)	(1.2)	(1.8)	(2.2)	(2.3)	(2.1)
Cys	24	1.39	0.81	15.8	14.0	−1.7	2.8	2.2
		(0.33)	(0.25)	(0.7)	(1.4)	(1.6)	(2.5)	(1.6)

TABLE III

Percentage of mutant peptides with more solvent-solute backbone hydrogen bonds than the corresponding wild type peptides

Standard deviations are shown in parentheses. The percentage was defined as the percentage of mutant peptides with more solvent-solute hydrogen bonds than the wild type peptide. Mut, mutant.

	Solvent percentage								
	10	20	30	40	50	60	70	80	90
H-bonds Mut	158.6	61.0	30.3	17.1	10.7	7.1	4.8	3.4	2.1
	(19.7)	(12.4)	(8.2)	(6.0)	(4.8)	(4.0)	(3.1)	(2.6)	(2.1)
H-bonds	159.1	59.6	29.7	16.6	10.4	6.5	4.0	2.5	1.7
	(18.4)	(12.6)	(9.3)	(6.5)	(4.8)	(3.9)	(2.8)	(1.9)	(1.4)
Percentage	50.0	50.0	48.3	51.7	60.3	56.9	69.0	75.9	69.0

TABLE IV

Comparison of mutants associated with the four forms of osteogenesis imperfecta

Structural parameters for each of the four types of osteogenesis imperfecta-associated peptides. Type II is lethal, while Types I, III, and IV are non-lethal. See text for discussion. WT, wild type.

	No. of sequences	r.m.s.d.	r.m.s.d. WT	H-bonds	H-bonds WT	H-bonds difference	Solvent H-bonds	Solvent H-bonds WT
I	6	1.39	0.83	14.3	15.8	−1.5	3.0	1.5
		(0.25)	(0.24)	(2.2)	(1.5)	(2.6)	(3.7)	(1.6)
II	21	1.38	0.77	13.6	15.7	−2.0	3.9	2.3
		(0.31)	(0.20)	(1.3)	(1.0)	(1.6)	(3.0)	(1.8)
III	12	1.22	0.78	14.9	16.3	−1.4	2.9	2.6
		(0.22)	(0.16)	(1.5)	(0.8)	(1.7)	(2.2)	(2.4)
IV	10	1.30	0.83	14.0	15.8	−1.8	2.1	3.0
		(0.28)	(0.33)	(2.5)	(1.5)	(2.9)	(2.0)	(1.4)

A vast majority of those hydrogen bonds were highly exchangeable nonspecific hydrogen bonds with solvent.

The mutant peptides that had the mutations G244C, G601S, and G844V were chosen for further analysis. G244C (18) is associated with OI type II, and the peptide had one less low exchange backbone hydrogen bond. Its mutant form had five specific solvent hydrogen bonds. An analysis of solvent molecules surrounding the mutation site showed that a single water molecule bridged two of the disrupted chains, while a second water molecule, which was fully solvated, was hydrogen bonding strongly with one of the mutant cysteines (Fig. 3*a*). This result suggests that solvent molecules may bind specifically both in a water bridging pattern or in a configuration with only a single hydrogen bond to the peptide.

G601S is associated with OI types III and IV (19). Analysis of this molecule showed a solvent molecule bridging the two peptide chains and hydrogen bonding to a serine residue with a side chain that was hydrogen bonding to the opposite chain (Fig. 3*b*). G844V is associated with OI type II and is interesting because it showed a greater difference in specifically bound solvent molecules than any other peptide. The solvent hydrogen bonds were conferred through five specifically bound water molecules. One of the water molecules was bridging near the valine residues, while four were hydrogen bonding far from the mutation site in approximately the region where structural disruption begins (Fig. 3*c*).

DISCUSSION

The use of structural models to analyze non-synonymous disease-associated mutations shows great promise for dis-

TABLE V
Serine residue hydrogen bonding patterns

The number of observed hydrogen bonds between γ oxygen of a mutated serine and solvent, main chain, and side chain atoms. POS, identity of mutation and position along the chain.

POS	OI[a]	Total HBs[b]	Solvent 20%[c]	Solvent 80%	Total solute[b]	Main chain 20%[c]	Side chain 20%	Main chain 80%	Side chain 80%
G247S	*, III	99	0	0	10	3	0	1	0
G382S	IV	184	1	0	11	4	0	1	0
G415S	*, III, IV	132	2	1	13	2	0	1	0
G448S	IV	244	2	0	8	1	0	1	0
G451S	*	128	4	0	8	4	0	0	0
G478S	*	168	1	0	9	1	1	0	0
G541S	III	172	0	0	9	1	0	0	0
G565S	*	72	1	0	12	1	0	1	0
G589S	III, IV	110	4	0	11	2	1	1	0
G598S	*	185	1	0	10	3	0	1	0
G601S	III, IV	103	1	1	9	2	1	1	0
G631S	*	108	4	1	9	1	0	1	0
G643S	III	95	3	0	9	2	3	2	0
G661S	III	165	1	0	10	2	0	1	0
G832S	IV	125	0	0	10	3	0	1	0
G844S	III	196	0	0	7	1	0	0	0
G862S	*, III	177	0	0	6	2	0	1	0
G871S	III	204	1	0	10	2	0	0	0
G880S	IV	153	2	0	9	2	0	1	0
G898S	III	131	2	1	11	1	0	1	0
G901S	I	157	0	0	5	1	0	1	0
G913S*	*	76	0	0	13	3	2	1	1
G964S	*	85	2	0	9	1	0	1	0
G973S	III	158	1	0	8	4	0	1	0

[a] Lethal OI (Type II) is represented by an asterisk.

[b] Total number of hydrogen bonds represents the total observed hydrogen bonds observed regardless of rate of exchange. The overwhelming majority of observed hydrogen bonds are fast-exchanging hydrogen bonds with solvent.

[c] Solvent 20% and 80% represents the number of observed hydrogen bonds involving the γ oxygen of a mutated serine and any solvent atom that are present in at least 20% and 80% of the snapshots analyzed, respectively. Main chain 20% and 80% and side chain 20% and 80% represent the number of hydrogen bonds involving serine side chains and the peptide main chain and other side chains, respectively.

covering the molecular pathogenesis of many diseases, that is, the manner in which the altered sequence in the protein leads to altered function. In this study, we have shown that collagen models can be built, equilibrated, and simulated and further that they show structural differences between the mutant and wild type peptides.

From our studies, structural parameters, by themselves, are not enough to predict the phenotype of a disease-associated mutation in type I collagen. Although the observed structural differences correlate with the identity of the perturbing residue, the relationships to severity of disease are still elusive. The lack of correlation, while disturbing, may not be surprising because disease severity in collagen disorders results from several factors. These factors likely include, in addition to folding interruptions, alterations in the secretion of molecules that contain abnormal chains, abnormal molecular assembly into fibril in the matrix, and, finally, defective mineralization of fibrils in bone. While each of these could reflect an effect of altered structure, other factors in the cell could modulate them so that structural studies alone probably cannot predict all these variations.

Valine residues showed the largest r.m.s.d. difference from their respective wild type peptides. Valine has the largest volume of the substituting residues we analyzed, and it is likely that the r.m.s.d. difference is attributable to the larger volume required in the interstitial region of the triple helix. There was great variability in r.m.s.d. differences within a given class of amino acid mutations, giving more evidence that the amino acid environment around a mutation is as important as the identity of the mutation itself.

Serine can hydrogen bond with serine residues on an opposite chain, the main chain, solvent, or another nearby side chain. We found that serine residues usually formed highly exchangeable hydrogen bonds with the solvent and occasionally formed low exchanging hydrogen bonds with the adjacent main chain or, more rarely, with solvent. The hydrogen bond patterns serine residues adopt confer the structure on the surface of the triple helix. Since we see variability within a given substituting amino acid type, it is likely that neighboring residues confer this difference. These differences suggest that when stability of the triple helix is not significantly altered and folding of the triple helical domain is not disrupted, surface structural differences on the triple helix can confer differences in function.

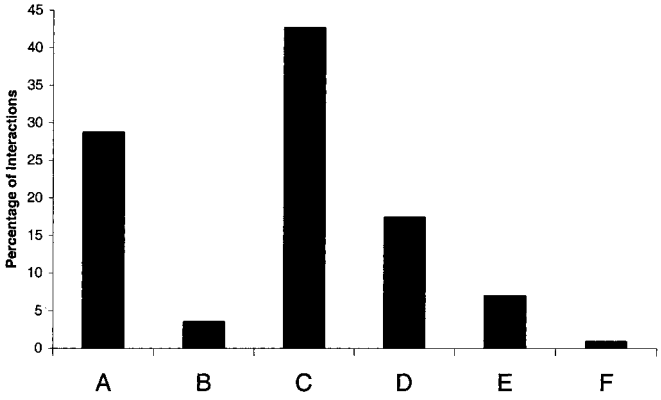

FIG. 2. **Interactions specific to serine residues.** Interactions preferred by serine residues and the percentage of simulation time they are observed participating in them. The data labels refer to the following interactions between the mutant serine side chains and the following: *A*, solvent hydrogen bonds that are present at least 20% of the simulation time; *B*, solvent hydrogen bonds that are present at least 80% of the simulation time; *C*, main chain hydrogen bonds that are present at least 20% of the simulation time; *D*, main chain hydrogen bonds that are present at least 80% of the simulation time; *E*, adjacent serine side chain hydrogen bonds that are present at least 20% of the simulation time; and *F*, adjacent serine side chain hydrogen bonds that are present at least 80% of the simulation time. The production simulation time for this analysis was 24 300-ps simulations. 20% is equivalent to 60 ps, and 80% is equivalent to 240 ps.

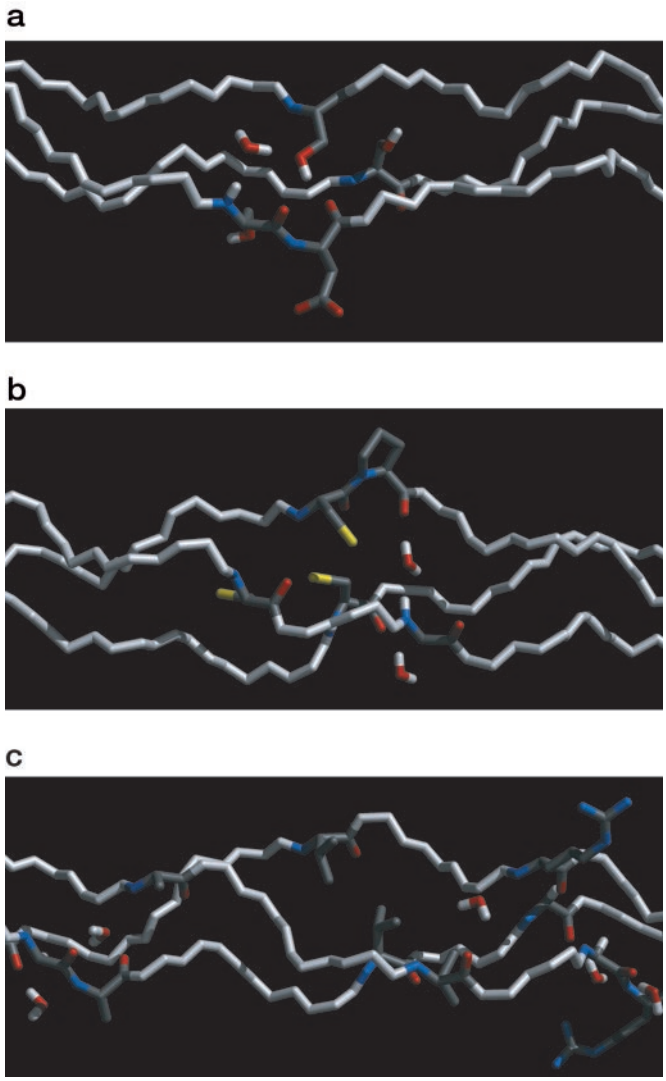

FIG. 3. **Solvent binding examples (see text for detailed descriptions).** *a*, serine mutant G601S with specifically bound water; *b*, cysteine mutant G244C with two specifically bound water molecules; *c*, valine mutant G844V with five specifically bound water molecules.

Most of the mutated molecules compensated for loss of some of the intrinsic hydrogen bonds with solvent hydrogen bonds that had a low probability of exchange. Many of these hydrogen bonds were with specifically bound water molecules. Their presence supports the hypothesis proposed from earlier studies of idealized peptides that solvent molecules compensate for lost main chain hydrogen bonds (9, 12). We never observed more than 10 specific solvent hydrogen bonds and observed zero hydrogen bonds in seven mutant structures (12%). That some simulations show decreases in solvent binding when the mutation is present further illustrates how an analysis of many mutations simultaneously is necessary to give insight into general structural changes that occur when mutations are introduced.

The thermodynamic effects of mutations are hard to analyze structurally without an accurate model of the unfolded state. We do, however, observe some differences between lethal and non-lethal peptides. Lethal mutations had slightly fewer backbone hydrogen bonds than non-lethal mutations and, perhaps surprisingly, were slightly less perturbing than non-lethal mutations. We think that more severe mutations may induce a loss of flexibility so that disruptions along the chain cannot be regularly compensated. To show stronger correlations than this, we are using machine learning methods to predict the phenotype of mutations since phenotype is an accumulation of a large number of properties.

The next step for analyzing these mutations is to build an accurate energetic model of collagen-like peptides. Free energy techniques using thermodynamic integration have been successful in studying mutations in the glycine positions of collagen (11). These methods are challenging because they require a large amount of computational time to calculate in a high throughput manner. The Molecular Mechanics Poisson Boltzmann Surface Area (MM/PBSA) method has been applied to proteins and is a powerful tool for analyzing the energetic properties of protein structures. Although it is much faster than other methods, its main drawback is its association with a larger random error than thermodynamic integration calculations. To analyze a large number of peptides, we are calculating the MM/PBSA energy of both the unfolded and folded state of both the mutant and wild type peptides. These energies can then be compared with the structural results and be incorporated into machine learning methods to predict phenotype.

There are many ways that alterations of structure and folding can cause a phenotypic change on the COL1 gene products. We believe the most significant alterations in structure and stability will have a substantial impact on a phenotype. It is difficult to speculate, however, on where these causes are most likely to occur in the functional pathway of collagen gene products without further experimental evidence. A structural change may have an effect on the kinetics of folding or may inhibit critical inter- and intrahelical interactions. A mutated triple helix is likely to cause uneven packing of the triple helices into a fibril, including solvent, and subsequently a fiber. As we begin to model collagen fibrils at the atomic level, we may gain better insights into the downstream structural effects of a single point mutation.

Selected structures from these simulations are being stored to build a data base of collagen mutation models for the use of researchers. This data base could aid experimental researchers interested in characterizing disease-associated mutations as well as researchers investigating other structural features of collagen, such as enzymatic binding. Eventually we would like to incorporate these models into an algorithm that can predict the phenotype of disease-associated mutation.

We have applied a method of high throughput mutation analysis using molecular dynamics. This method shows the range of structural interactions that occur when a glycine in the triple helical domain is mutated in collagen-like peptides. These peptides seemed to compensate for mutation-induced lost stability with 1) a large number of solvent-backbone hydrogen bonds that have a high rate of exchange and 2) a small number of solvent-backbone hydrogen bonds that exchange very slowly. We believe that our method will have application in analyzing the molecular consequences of disease-associated mutations in other systems. One of the limitations in the use is the extent to which mutations like these alter the initial folding of molecules because one of the assumptions we make is that at least some of these molecules can complete, although imperfectly, the formation of triple helix. There are currently many phenotypically annotated mutations where the underlying molecular basis for the disease association is not known. In combination with machine learning methods and experimental results, molecular mechanic methods show promise in providing insight into the underlying causes of disease.

Acknowledgments—We thank Professor Peter Byers (University of Washington) for helpful discussions and comments on the manuscript, the Biocomputation Core Facility at Stanford University, and the University of California, San Francisco Computer Graphics Laboratory for the use of its resources.

* This work was funded by National Institutes of Health Grants AR47720-01 (to T. E. K., principal investigator) and LM05652 (to R. B. A., principal investigator). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ To whom correspondence should be addressed: Dept. of Genetics, Stanford University, 251 Campus Dr., MSOB x-215, Stanford, CA 94305-5479. Tel.: 650-736-0156; Fax: 650-725-7944; E-mail: teri.klein@stanford.edu.

REFERENCES

1. Chasman, D., and Adams, R. (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.* **307,** 683–706
2. Sunyaev, S., Ramensky, V., and Bork, P. (2000) SNP frequencies in human genes: an excess of rare alleles and differing modes of selection. *Trends Genet.* **16,** 335–337
3. Wacey, A. I., Cooper, D. N., Liney, D., Hovig, E., and Krawczak, M. (1999) Perturbational effects of amino acid substitutions in the DNA-binding domain of p53. *Hum. Genet.* **104,** 15–22
4. Retief, E., Parker, M., and Retief, A. (1985) Regional chromosomal mapping of human collagen genes α2(I) and α1(I) (COLIA2 and COLIA1). *Hum. Genet.* **69,** 304–308
5. Byers, P. (1993) in *Connective Tissue and Its Heritable Disorders: Molecular, Genetic and Medical Aspects* (Royce, P., and Steinmann, B., eds) pp. 317–350, Wiley-Liss, New York
6. Wei, Y., Madhavi, B., and Brodsky, B. (1997) Amino acid sequence environment modulates the disruption by osteogenesis imperfecta glycine substitutions in collagen-like peptides. *Biochemistry* **36,** 6930–6935
7. Liu, X., Kim, S., Dai, Q. H., Brodsky, B., and Baum, J. (1998) Nuclear magnetic resonance shows asymmetric loss of triple helix in peptides modeling a collagen mutation in brittle bones disease. *Biochemistry* **37,** 15528–15533
8. Beck, K., Chan, V. C., Shenoy, N., Kirkpatrick, A., Ramshaw, J. A., and 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
9. Bella, J., Eaton, M., Brodsky, B., and Berman, H. M. (1994) Crystal and molecular structure of a collagen-like peptide at 1.9 Å resolution. *Science* **266,** 5182
10. Long, C. G., Braswell, E., Zhu, D., Apigo, J., Baum, J., and Brodsky, B. (1993) Characterization of collagen-like peptides containing interruptions in the repeating Gly-X-Y sequence. *Biochemistry* **32,** 11688–11695
11. Mooney, S. D., Huang, C. C., Kollman, P. A., and Klein, T. E. (2001) Computed free energy differences between point mutations in a collagen like peptide. *Biopolymers* **58,** 347–353
12. Klein, T., and Huang, C. (1999) Computational investigations of structural changes resulting from point mutations in a collagen-like peptide. *Biopolymers* **49,** 167–183
13. Lewis, P., Momany, F., and Scheraga, H. (1973) *Isr. J. Chem.* **11,** 121–152
14. Dalgleish, R. (1997) The human type I collagen mutation database. *Nucleic Acids Res.* **25,** 181–187
15. Dalgleish, R. (1998) The human collagen mutation database 1998. *Nucleic Acids Res.* **26,** 253–255
16. Huang, C. C., Couch, G. S., Pettersen, E. F., Ferrin, T. E., Howard, A. E., and Klein, T. E. (1998) in *Pacific Symposium on Biocomputing '98* (Hunter, L., and Klein, T. E., eds) pp. 349–361, World Scientific Publishing, Singapore
17. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Jr., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995) A second generation force field for the simulation of proteins and nucleic acids. *J. Am. Chem. Soc.* **117,** 5179–5197
18. Fertala, A., Westerhausen, A., Morris, G., Rooney, J. E., and Prockop, D. J. (1993) Two cysteine substitutions in procollagen I: a glycine replacement near the N-terminus of α1(I) chain causes lethal osteogenesis imperfecta and a glycine replacement in the α2(I) chain markedly destabilizes the triple helix. *Biochem. J.* **289,** 195–199
19. Lund, A. M., Nicholls, A. C., Schwartz, M., and Skovby, F. (1997) Parental mosaicism and autosomal dominant mutations causing structural abnormalities of collagen I are frequent in families with osteogenesis imperfecta type III/IV. *Acta Paediatr.* **86,** 711–718
20. Krawczak, M., and Cooper, D. N. (1997) The Human Gene Mutation Database. *Trends Genet.* **13,** 121–122