# Predicting the Clinical Lethality of Osteogenesis Imperfecta from Collagen Glycine Mutations†

Dale L. Bodian,‡ Balaraman Madhan,§ Barbara Brodsky,§ and Teri E. Klein*,‡

*Genetics Department, School of Medicine, Stanford University, Stanford, California 94305, and Department of Biochemistry, University of Medicine and Dentistry of New Jersey-Robert Wood Johnson Medical School, Piscataway, New Jersey 08854*

*Received January 6, 2008; Revised Manuscript Received February 22, 2008*

ABSTRACT: Osteogenesis imperfecta (OI), or brittle bone disease, often results from missense mutation of one of the conserved glycine residues present in the repeating Gly-X-Y sequence characterizing the triple-helical region of type I collagen. A composite model was developed for predicting the clinical lethality resulting from glycine mutations in the α1 chain of type I collagen. The lethality of mutations in which bulky amino acids are substituted for glycine is predicted by their position relative to the N-terminal end of the triple helix. The effect of a Gly → Ser mutation is modeled by the relative thermostability of the Gly-X-Y triplet on the carboxy side of the triplet containing the substitution. This model also predicts the lethality of Gly → Ser and Gly → Cys mutations in the α2 chain of type I collagen. The model was validated with an independent test set of six novel Gly → Ser mutations. The hypothesis derived from the model of an asymmetric interaction between a Gly → Ser mutation and its neighboring residues was tested experimentally using collagen-like peptides. Consistent with the prediction, a significant decrease in stability, calorimetric enthalpy, and folding time was observed for a peptide with a low-stability triplet C-terminal to the mutation compared to a similar peptide with the low-stability triplet on the N-terminal side. The computational and experimental results together relate the position-specific effects of Gly → Ser mutations to the local structural stability of collagen and lend insight into the etiology of OI.

Collagen is a fibrous protein that provides functional and structural integrity in the human body. Type I collagen, the major protein in bone, skin, and other tissues, is a heterotrimer comprised of two α1(I) and one α2(I) polypeptide chains, encoded by the COL1A1 and COL1A2 genes, respectively. Each chain includes 338 uninterrupted repeats of the Gly-X-Y triplet, where X and Y can be any amino acid but are most often proline and hydroxyproline (Hyp or O).[1] In the heterotrimeric protein, the Gly-X-Y repeats form the triple-helical structure that is characteristic of the collagen family. The conserved glycines in every third position are essential for the integrity of the protein since larger amino acids cannot be accommodated in the tightly packed core of the triple helix without disruption of the structure.

Missense mutation of a conserved Gly in the triple-helical region of type I collagen generally leads to osteogenesis imperfecta (OI). OI, often called brittle bone disease, is characterized by bone fragility and increased susceptibility to fracture, which may be accompanied by bone deformity, decreased life span, dentinogenesis imperfecta, hearing loss, and altered scleral hue (*1, 2*). The disease presents in a wide array of phenotypes ranging from mild to lethal. The severity depends on the chain in which the Gly substitution occurs, the location within the chain, and the substituting amino acid (*3*), but it is not currently possible to predict reliably the lethality from the sequence of the mutation. The ability to make such a prediction to guide genetic counseling decisions is becoming critical as diagnostic strategies move to direct sequencing of the genes (*4*).

Several models have been developed for relating Gly mutations to OI phenotype but each has met with limited success. The diversity of these models reflects the fact that the mechanism of phenotype development is unknown. According to the phenotype gradient model, disease severity increases with the position of the mutation along the triple helix from N to C (*5*). The identification of mutations not conforming to this pattern led to the proposal of the regional model, which states that OI severity reflects crucial and noncrucial regions interspersed along the collagen molecule (*6–8*). The critical regions are not well-defined but may represent ligand-binding sites (*9*). Computational methods have been applied to predict lethality based on physicochemical proper-

† This work was supported by NIH Grants AR051582 (to T.E.K.) and GM60048 (to B.B.), the Osteogenesis Imperfecta Foundation and the Children's Brittle Bone Society (to D.L.B.), and a BOYSCAST (DST, India) fellowship to B.M., on sabbatical from Central Leather Research Institute, Chennai 600020, India.

* To whom correspondence should be addressed: Genetics Department, 300 Pasteur Dr. L301, Stanford, CA 94305-5120. Phone: (650) 736-0156. Fax: (650) 725-3863. E-mail: teri.klein@stanford.edu.

‡ Stanford University.

§ University of Medicine and Dentistry of New Jersey-Robert Wood Johnson Medical School.

[1] Abbreviations: OI, osteogenesis imperfecta; C, carboxy; N, amino; Hyp or O, 4′-hydroxyproline; $T_m$, melting temperature; $T_m[+1]$, relative thermal stability (derived from experimental $T_m$ values on host-guest triple helical peptides) of the triplet adjacent and C-terminal to the triplet containing the Gly mutation; $T_m[-1]$, relative thermal stability of the triplet adjacent and N-terminal to the triplet containing the Gly mutation; ROC, receiver operating characteristic; CD, circular dichroism; DSC, differential scanning calorimetry; $MRE_{225}$, mean residue ellipticity at 225 nm.

10.1021/bi800026k CCC: $40.75 
Published on Web 04/16/2008

ties of individual residues surrounding the mutation site (*10–13*). These methods identified amino acid patterns partially correlating with lethality, but the results are not easily interpretable in biological terms. It has also been suggested that local helix stability correlates with phenotype, with mutation of higher-stability triplets resulting in more severe disease (*14, 15*). However, a statistically significant correlation between the stability of the mutated triplet and lethality was not found when this model was applied to an updated set of mutations (*9*). Others have examined the role of the global stability of the protein on phenotype, but a clear association has not yet been identified (*16–19*).

Evaluation of the role of stability in determining clinical phenotype has been hindered by difficulties in measuring this parameter. One approach has been to estimate sequence-specific effects on local stability using the melting temperature ($T_m$) of collagen-like peptides. $T_m$ values were determined systematically for two sets of host–guest peptides, one modeling amino acid substitutions for Gly (*20*) and a second measuring the relative stability of 82 different Gly-X-Y triplets (*21*). Using data from the former set, a correlation was found between OI phenotype of COL1A1 mutations and the destabilization induced by the substitutions (*20*). However, using data from the Gly-X-Y set, the local stability around a mutation site, approximated as the average $T_m$ of the neighboring triplets, did not show a clear association with clinical severity (*21*).

The limited success of previous predictive models suggests that a single method may be insufficient to explain all observed mutations. Collagen undergoes multiple post-translational modifications, secretion, fibrillogenesis, and mineralization (*3*) and interacts with a large number of proteins and proteoglycans (*22*). The phenotype may result from the effects of a mutation on multiple aspects of collagen structure, biosynthesis, and function, and different mutations could impact these to varying degrees. Our strategy for modeling such a complex system is to construct multiple predictive models, where each individual model represents the effects of a group of mutations likely to share a common mechanism. We applied this strategy to predicting the lethality of COL1A1 glycine mutations. The resulting model for branched and charged amino acid substitutions captures elements of previously proposed models in a quantitative framework. The model for Gly → Ser mutations is the first to identify a correlation between properties of the collagen molecule and lethality of these mutations. Experimental peptide studies confirm the hypothesis derived from the model that a sequence of low stability C-terminal to a triplet containing a Gly → Ser mutation has a significant and asymmetric effect on triple helix properties. These models represent the first components of a composite model for predicting OI severity and provide insight into the effect of a mutation on collagen structure and function and its role in disease etiology.

## MATERIALS AND METHODS

*Sequences and Data.* Sequences of the native COL1A1 and COL1A2 proteins are from RefSeq (*23*) entries NP_000079.2 and NP_000080.2, respectively, with Y-position prolines converted to hydroxyproline. Amino acids are numbered by their position in the triple-helical region.

Table 1: Frequency of OI Lethality by Substitution[a]

| substitution | no. of distinct mutations: total | lethal | nonlethal | % lethal |
|---|---|---|---|---|
| | (A) COL1A1 | | | |
| Ala | 21 | 4 | 17 | 19 |
| Cys | 49 | 19 | 30 | 39 |
| Asp | 22 | 15 | 7 | 68 |
| Glu | 6 | 3 | 3 | 50 |
| Arg | 31 | 17 | 14 | 55 |
| Ser | 61 | 18 | 43 | 30 |
| Val | 18 | 15 | 3 | 83 |
| total | 208 | 91 | 117 | 44 |
| | (B) COL1A2 | | | |
| Cys | 21 | 6 | 15 | 29 |
| Ser | 46 | 6 | 40 | 13 |

[a] The number of nonredundant lethal and nonlethal glycine mutations used in constructing the predictive models for COL1A1 and COL1A2.

Osteogenesis imperfecta is described in OMIM (*24*) entries 166200 (type I), 166210 (type IIA), 259420 (type III), and 166220 (type IV). A recent, inclusive collection of collagen Gly mutations published by the OI consortium, together with their associated lethality (*9*), served as the training data for the modeling. Mutations were used as published except for c.2516G>A, which was corrected to Asp following confirmation by the contributing author (P. Byers, personal communication). Mutations not associated with OI phenotypes were excluded. Since identical mutations were found in multiple patients, a nonredundant set was constructed based on the position of the mutation and the amino acid replacing Gly. Each unique mutation was assigned the lethality observed in the majority of patients sharing that mutation. The three mutations associated with equal numbers of lethal and nonlethal patients, Gly655Asp, Gly1009Ser, and Gly304Arg, were excluded. Table 1 summarizes the resulting data used for construction of the models. Newly sequenced, previously unreported, glycine missense mutations provided by T. F. Chan et al. (manuscript in preparation) and P. Byers (unpublished results) served as an independent test set. This set includes five novel Gly → Ser mutations in COL1A1 and one in COL1A2. Newly sequenced COL1A2 Gly → Cys mutations were not available.

Experimental relative thermostabilities for individual triplets in host–guest peptides were reported by Persikov et al. (*21*). The $T_m$ average values for five-triplet windows were computed with the collagen stability calculator (*21*). Relative thermostabilities for heterotrimeric sequences were computed as the weighted average of the two chains (*21*).

*Modeling and Statistical Analysis.* Machine learning calculations were performed with Weka version 3.4.8a (*25*). Decision trees were constructed with the J4.8 algorithm. Both single-attribute evaluators and attribute subset evaluation methods were used for feature selection, including CfsSubset Eval with the BestFirst search method, and SVMAttributeEval with the Ranker search method. Ten-fold cross validation was used for both model selection and feature selection.

Statistical analyses, including logistic regression, *t* test, and Wilcoxon rank-sum tests, were performed with R version 2.5.0 (*26*). The significance of differences between means was computed under the null hypothesis that samples were drawn from the same population. All tests were two-tailed, and $p$ values of $<0.05$ were considered significant. Permuta-

tion tests were conducted with the exactRankTests package, version 0.8–15 (*27*). Conversion of real valued data to integers, as required by the package, was performed two ways: using the default method and by computing ranks, with ties assigned the minimum value. Exact *p* values were computed when both samples had fewer than 50 values. Mutations lacking an experimentally determined $T_m[+1]$ were excluded from the statistical analyses. For COL1A2 Gly → Ser mutations, logistic regression was performed with weights of 2 for lethal and 1 for nonlethal.

Receiver operating characteristic (ROC) curves (*28*) were computed from the training sets using 10-fold cross validation. Lethality of novel mutations was predicted using the cutoff $T_m[+1]$ values determined from these curves.

*Peptide Synthesis and Purification.* The peptides were synthesized by the Tufts University Core Facility (Boston, MA) and were purified on a reverse-phase high-pressure liquid chromatography system (Shimadzu) with a C-18 column. Purity of all peptides was ensured by mass spectrometry using matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF). Peptide concentrations of starting solutions were measured by monitoring the absorbance at 275 nm corresponding to tyrosine. All experimental studies used peptides at a concentration of 1 mg/mL in 20 mM PBS (10 mM $NaH_2PO_4$, 10 mM $Na_2HPO_4$, and 150 mM NaCl) at pH 7.

*Circular Dichroism (CD) Spectroscopy.* All CD studies were carried out using an Aviv model 62DS spectrophotometer with a Peltier thermoelectric temperature controller attached. Wavelength scans of the peptides were carried out by collecting the signal from 250 to 215 at 0.5 nm intervals at 0 °C with an averaging time of 10 s.

*Determination of $T_m$ by CD.* The peptides were refolded for at least 40 h at 0 °C prior to the melting experiments. The denaturation of the peptides was monitored for temperatures between 0 and 60 °C and the melting temperatures determined using the procedure reported previously (*29*). From experiments repeated on independently prepared samples, the error of determination of the $T_m$ is ±0.5 °C.

*Folding Experiments.* CD folding experiments were carried out by heating the peptides at 45 °C for 15 min and then rapidly quenching the reactions in an ice–water bath and placing the mixtures in a pre-equilibrated CD cell at 0 °C. The dead time was on the order of 25 s. The ellipticity at 225 nm was monitored with a time constant of 2 s and a time interval of 10 s. The half-time of refolding, $t_{1/2}$, was determined as the time for the fraction folded to reach 0.5. The experimental error of determination of $t_{1/2}$ did not exceed 10%.

*Differential Scanning Calorimetry (DSC).* Samples were equilibrated at 0 °C for at least 48 h before the melting curves were recorded. DSC transition curves were recorded on a NANO-DSC II model 6100 (Calorimetry Sciences Corp.) calorimeter. The calorimetric enthalpy was calculated using the procedure reported previously (*29*). The experimental error of determination of the calorimetric enthalpy is less than 5%.

## RESULTS

*Predicting Lethality from Position and Amino Acid Substitution.* The phenotype resulting from a glycine mutation in COL1A1 has been proposed to depend on the position of the mutation along the triple helix and the amino acid replacing Gly (*3*), but the effectiveness of these two attributes in predicting lethality has not been quantitated. A decision tree was constructed based on position and substituting residue to model the relationship between these two features and clinical lethality. The model was trained on the 208 distinct mutations comprising the nonredundant set of COL1A1 Gly mutations identified in 389 patients by the OI consortium (*9*). The data (Table 1A) confirm that both features are needed in the model. The residue replacing Gly is required since the fraction of lethal positions varies by amino acid, ranging from 19% for Ala to 83% for Val, and since different substitutions at the same position can yield different clinical outcomes (*9*). The specific glycine replaced is also important since each substituting residue is associated with both lethal and nonlethal phenotypes.

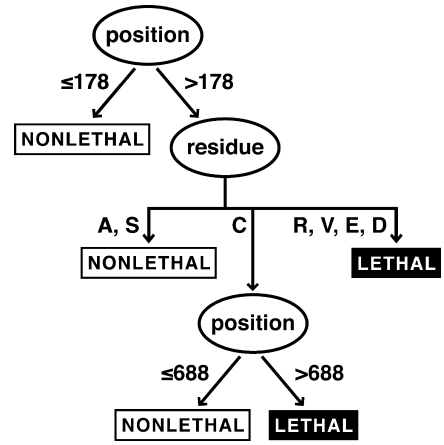


FIGURE 1: COL1A1 decision tree. Schematic representation of the rules for classifying a COL1A1 Gly mutation as lethal or nonlethal based on the position of the mutation in the triple helix and the residue replacing glycine. Amino acids are represented by their one-letter code.

The structure of the resulting decision tree is illustrated in Figure 1. The model predicts all Gly mutations within the 178 N-terminal residues of the triple helix to be nonlethal. This mimics the phenotype gradient model (*5*), and the cutoff at position 178 agrees well with the observation that substitutions in the first 200 residues are generally nonlethal (*9*). The decision tree classifies mutations C-terminal to Gly 178 by substituting residue. Ala and Ser substitutions are all predicted to be nonlethal, whereas Arg, Val, Glu, and Asp are predicted to be lethal. Cys substitutions are classified as lethal when they occur C-terminal to position 688.

For mutations C-terminal to position 178, the lethality of the mutations predicted by the decision tree correlates with the destabilization introduced by the substituting residue. Thermostability measurements of peptides modeling Gly mutations (*20*) revealed that the substitutions predicted to be lethal by the decision tree (Arg, Val, Glu, and Asp) are more disruptive to triple-helical structure than those predicted to be nonlethal (Ala and Ser). Cys substitutions, predicted to have both lethal and nonlethal mutations in this region, have intermediate thermostability. Interestingly, use of the measured thermostability values of these peptides as a feature in the modeling instead of the substituting residues results in an essentially identical decision tree.

The performance of the decision tree is summarized in Table 2. This simple model captures >80% of the informa-

Table 2: Decision Tree Performance[a]

| | training set | cross validation |
|---|---|---|
| % correct | 82% | 73% |
| TP[b] rate nonlethal | 0.96 | 0.88 |
| TP[b] rate lethal | 0.63 | 0.54 |
| FP[c] rate nonlethal | 0.37 | 0.46 |
| FP[c] rate lethal | 0.04 | 0.12 |

[a] The performance of the decision tree illustrated in Figure 1, evaluated on the training data and by 10-fold cross validation. [b] True positive. [c] False positive.

Table 3: Decision Tree Performance by Residue Replacing Gly and Lethality[a]

| | lethal | | | nonlethal | | | |
|---|---|---|---|---|---|---|---|
| substitution | no. correct | no. wrong | % correct | no. correct | no. wrong | % correct | overall % correct |
| Ala | 0 | 4 | 0.0 | 17 | 0 | 100.0 | 81.0 |
| Ser | 0 | 18 | 0.0 | 43 | 0 | 100.0 | 70.5 |
| Cys | 11 | 8 | 57.9 | 27 | 3 | 90.0 | 77.6 |
| Arg | 16 | 1 | 94.1 | 13 | 1 | 92.9 | 93.5 |
| Val | 15 | 0 | 100.0 | 3 | 0 | 100.0 | 100.0 |
| Glu | 3 | 0 | 100.0 | 3 | 0 | 100.0 | 100.0 |
| Asp | 13 | 2 | 86.7 | 6 | 1 | 85.7 | 86.4 |

[a] Performance for the decision tree illustrated in Figure 1, computed from the training set.

tion in the training set. Prediction accuracy for new data is estimated to be 73% by cross validation. The model tends to overpredict nonlethal since nearly 90% of the nonlethal mutations are predicted correctly, compared to 54% of the lethals. The false positive rate of 0.12 for lethal suggests that, overall, a lethal prediction is likely to be correct.

Table 3 shows the performance broken down by substitution and lethality. For nonlethal mutations, the accuracy is approximately 90% or better for all amino acids in the training set. Lethal mutations proved to be more difficult to model, with performance varying by amino acid. Accuracy for lethal substitutions of large and branched substitutions, Arg, Val, Glu, and Asp, was 87–100%, similar to the performance on these substitutions in nonlethal mutations. In contrast, lethal Ala, Ser, and Cys substitutions are not well captured, particularly Ser, for which none of the 18 lethal mutations are predicted correctly. Although none of the lethal Gly → Ala mutations are predicted correctly, there are only four in the current data set and the overall accuracy of the decision tree for this substituting residue is >80% (Table 3).

The results suggest that Arg, Val, Glu, and Asp substitutions are likely to share a mechanism of lethality since these mutations are well-represented by the same decision tree rules. The relatively weak overall performance of the decision tree for Gly → Ser mutations suggests that these substitutions have a different mechanism of lethality, which may be better predicted by an alternate model.

*Prediction of Lethal Gly → Ser Mutations.* The decision tree implies that, for mutations C-terminal to residue 178 (other than Gly → Cys), the lethality depends on the substituting residue and not the position. Accurate predictions for Arg, Val, Asp, and Glu, the most disruptive substitutions, suggest that for these residues the destabilization resulting from the mutation itself leads to lethality. In contrast, the relatively poor performance for the less destabilizing Ser substitutions suggests that the lethality of these mutations is

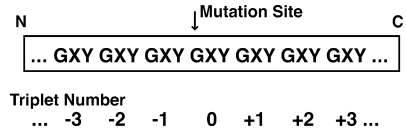


FIGURE 2: Numbering system for triplets surrounding a mutation site. The Gly-X-Y triplet containing the mutation is numbered 0; triplets to the N-terminal side of the mutation are given negative numbers and those to the C-terminal side positive numbers.

more sensitive to their environment. To develop an alternative model that performs better for these mutations, we hypothesized that a lethal Gly → Ser mutation results in a stability lower than that expected from the substitution itself due to modulation by the stability of the surrounding sequence. We modeled the relative stability of individual triplets neighboring a Gly → Ser mutation by their $T_m$ in host–guest peptides (*21*). Feature selection revealed that, of the seven triplets surrounding the mutation site (Figure 2), the $T_m$ of the triplet on the C-terminal side of the mutation ($T_m[+1]$) is most predictive of lethality, being consistently ranked highest by all methods. The thermostability of the triplet containing the mutation itself was not predictive of lethality, consistent with previous findings (*9*). The calculations were repeated using averaged $T_m$ values incorporating thermostability contributions from multiple neighboring triplets, both in the same chain and in COL1A2, since the stability of a single triplet may not be sufficient to represent the stability of the region containing the mutation. However, none of these averaged $T_m$ values was as predictive of lethality as $T_m[+1]$.

Several tests were performed to assess whether the correlation between $T_m[+1]$ and lethality is statistically significant. By *t* test, the difference in the mean $T_m[+1]$ of 4.3 °C between lethals and nonlethals is significant, with a *p* value of 0.029 (Table 4). Since the *t* test assumes that the $T_m[+1]$ values are normally distributed but the true distribution is unknown, the significance was also assessed with nonparametric tests. The *p* value from a Wilcoxon rank sum test is 0.020, by logistic regression 0.017, and in exact permutation tests 0.011. All these tests agree that the difference in the mean $T_m[+1]$ between lethal and nonlethal Gly → Ser mutations in COL1A1 is statistically significant. Interestingly, the mean $T_m[+1]$ of 35.9 °C for lethals is lower than that of 40.2 °C for nonlethals (Table 4), consistent with a correlation between low stability and lethality. The complete list of mutations and their associated $T_m[+1]$ values is provided in Table S1 of the Supporting Information.

To use $T_m[+1]$ to classify new mutations as lethal or nonlethal, a cutoff value must be chosen. Mutations with a $T_m[+1]$ at or below the cutoff would be predicted to be lethal, whereas those with a $T_m[+1]$ above the cutoff would be classified as nonlethal. A cutoff is selected by maximizing the true positive rate while maintaining the false positive rate at an acceptable level. The ROC curve in Figure 3A plots the percentage of true positives (lethals correctly predicted) and false positives (incorrectly predicted nonlethals) at various $T_m[+1]$ cutoff values. The indicated point represents a cutoff of 37.7 °C, corresponding to an estimated accuracy of 71% for lethals and 78% for nonlethals (Table 4).

*$T_m[+1]$ Model Validation.* The results suggest that $T_m[+1]$ correlates with lethality of Gly → Ser mutations in COL1A1.

Table 4: $T_m[+1]$ Models Summary[a]

| | | mean $T_m[+1]$[b] (°C) | | $N$ | | | | % correct[e] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| chain | substitution | lethal | nonlethal | lethal | nonlethal | $p$[c] | $T_m[+1]$ cutoff (°C)[d] | lethal | nonlethal | total | % coverage[f] |
| COL1A1 | Ser | 35.9 | 40.2 | 14 | 36 | 0.03 | 37.7 | 71 | 78 | 76 | 82 |
| COL1A2 | Ser | 30.9 | 40.0 | 4 | 30 | 0.02 | 32.9 | 75 | 86 | 85 | 74 |
| COL1A2 | Cys | 35.8 | 41.9 | 5 | 14 | 0.008 | 37.7 | 80 | 87 | 85 | 90 |

[a] Summary of the models relating the lethality of Gly mutations to the relative thermostability of triplets to the C-terminal side of the mutations, as approximated by the melting temperatures ($T_m$) of model peptides. [b] $T_m$ of a peptide containing the triplet C-terminal to the mutated triplet. [c] $p$ value computed by a two-sided $t$ test. [d] The minimum $T_m$ with the specified true and false positive rates chosen from the ROC curves in Figure 3. [e] Estimated by 10-fold cross validation. [f] Percent of mutations with a measured $T_m[+1]$.

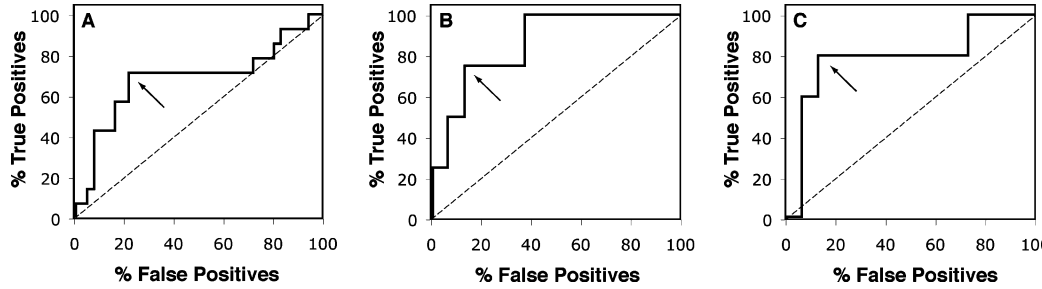


FIGURE 3: ROC curves for $T_m[+1]$ models. (A) COL1A1 Gly → Ser model. (B) COL1A2 Gly → Ser model. (C) COL1A2 Gly → Cys model. Solid lines show true and false positive rates determined from the training data by cross validation. Dotted lines indicate the line of no discrimination, representing a random classifier. Good classifiers yield points above the dotted line, with points closer to the top left corner representing better performance. A perfect classifier would give a point in the top left-hand corner, representing 0% false positives and 100% true positives. Arrows indicate true and false positive rates for the cutoff $T_m[+1]$.

Table 5: Lethality Prediction for Novel Gly → Ser Mutations

| chain | position | lethality | triplet [+1] | $T_m[+1]$ | prediction | correct? |
|---|---|---|---|---|---|---|
| COL1A1 | 16 | nonlethal | GPQ | 41.3 | nonlethal | yes |
| COL1A1 | 706 | lethal | GPS | 35 | lethal | yes |
| COL1A1 | 763 | nonlethal | GAO | 41.7 | nonlethal | yes |
| COL1A1 | 916 | lethal | GET | 35.9 | lethal | yes |
| COL1A1 | 967 | nonlethal | GAO | 41.7 | nonlethal | yes |
| COL1A2 | 718 | nonlethal | GPA | 40.9 | nonlethal | yes |

To validate this model, it was tested on previously unobserved mutations from newly sequenced patients. Of the five novel COL1A1 Gly → Ser mutations, 100% are predicted correctly by the $T_m[+1]$ model (Table 5). In addition to validating the model, these data show that the model is capable of extrapolation, since two of the predictions are based on $T_m$ values for triplets that are not present in the training set (GPS and GET; Table S1 of the Supporting Information). This demonstrates that $T_m[+1]$ not only correlates with lethality but also is predictive of lethality.

*Extension of the $T_m[+1]$ Model to COL1A2.* We also tested whether the correlation between $T_m[+1]$ and lethality extends to Gly → Ser mutations in COL1A2. Although there are only four lethal Gly → Ser mutations in COL1A2 with available $T_m[+1]$ data, the relationship between lethality and $T_m[+1]$ is maintained (Table 4 and Table S2 of the Supporting Information). As in COL1A1, lethal Gly → Ser mutations in COL1A2 have lower average $T_m[+1]$ values than nonlethals (30.9 °C vs 40 °C). The 9.1 °C difference is statistically significant, with $p$ value <0.05 by both parametric and nonparametric tests. Specifically, the $p$ values are 0.023, 0.0079, 0.040, and <0.004 for the $t$ test, Wilcoxon rank sum test, logistic regression, and permutation tests, respectively. With a cutoff $T_m[+1]$ set to 32.9 °C (Figure 3B), the estimated prediction accuracy is 75% for lethals, 86% for nonlethals, and 85% overall (Table 4). The one novel COL1A2 Gly → Ser mutation available was predicted correctly with this model (Table 5). This supports the proposed association between $T_m[+1]$ and lethality for Gly → Ser mutations and shows that it applies to both chains.

The relationship between $T_m[+1]$ and lethality is also observed for Gly → Cys mutations in COL1A2 (Table 4 and Table S3 of the Supporting Information). The mean $T_m[+1]$ for lethals is 35.8 °C and for nonlethals 41.9 °C. The 6.1 °C difference between means is statistically significant by a $t$ test, with a $p$ value of 0.008. The difference is also significant by nonparametric tests, with a $p$ value of 0.011 for the Wilcoxon rank sum test, 0.034 by logistic regression, and 0.007 for permutation tests. This is consistent with the hypothesis that $T_m[+1]$ is predictive of lethality for certain mutations and suggests that the model can be applied to Gly → Cys mutations in COL1A2. Using a cutoff $T_m[+1]$ of 37.7 °C (Figure 3C), the accuracy for prediction of new mutations is estimated to be 85%, 80% for lethals and 87% for nonlethals (Table 4).

*Experimental Peptide Studies.* The modeling results imply that there is an asymmetric effect of neighboring triplets on the stability of collagen containing a Gly → Ser mutation, since Ser substitutions with low $T_m[+1]$ values are associated with a lethal phenotype but a similar correlation was not seen with low $T_m[-1]$ mutations. Peptides were designed with the aim of investigating the structural consequences of having a low-stability triplet C-terminal to a Gly → Ser mutation, compared with a peptide having the same low-stability triplet N-terminal to the mutation. There is a lethal Gly → Ser mutation (Gly631Ser) in COL1A1 with a low-stability triplet (GFA) C-terminal to the mutation and a high-stability triplet (GPO) N-terminal to the mutation. To model this site, a 34-mer peptide was synthesized representing residues 628–636 of COL1A1 (GPOSPAGFA) flanked on both sides by highly stabilizing $(GPO)_4$ sequences, and a C-terminal Tyr for concentration determination (Table 6). The N-terminal end was acetylated, and the C-terminus was amidated to increase stability (*30*). The related peptide

Table 6: Experimental Characterization of the Designed Peptides

| peptide | sequence | $MRE_{225}$ (deg $cm^2$ $dmol^{-1}$)[a] | $T_m$ (°C)[b] | $\Delta H_{cal}$ (kJ/mol)[c] | $t_{1/2}$ (min)[d] |
|---|---|---|---|---|---|
| GPOSPAGFA | $(GPO)_4GPOSPAGFA(GPO)_4Y$ | 3570 | 16.2 | 275 | 58.7 |
| GFASPAGPO | $(GPO)_4GFASPAGPO(GPO)_4Y$ | 3700 | 20.3 | 340 | 40.7 |

[a] Triple-helix content. [b] Stability. [c] Calorimetric enthalpy. [d] Half-time of refolding.

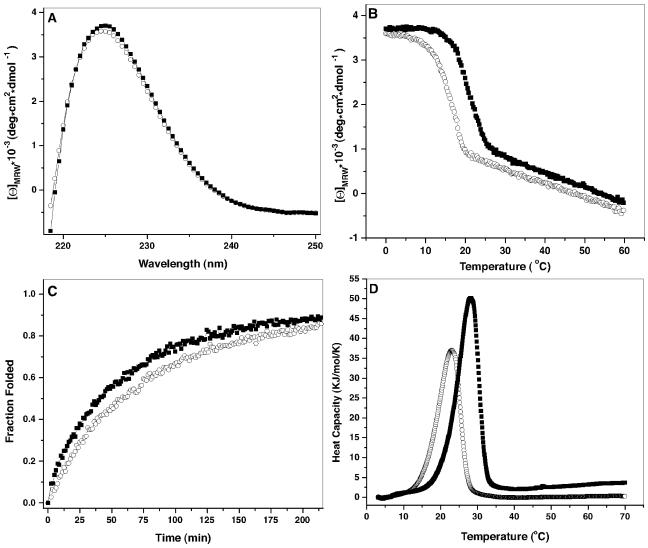


FIGURE 4: Stability, folding, and thermodynamics for peptides GPOSPAGFA and GFASPAGPO. (A) CD spectra recorded at 0 °C. (B) CD thermal transition profiles showing mean residue ellipticity at 225 nm as a function of temperature. (C) CD folding profiles represented as the fraction folded vs time. (D) DSC thermograms. (○) GPOSPAGFA and (■) GFASPAGPO.

GFASPAGPO was synthesized which reverses the location of the low-stability GFA triplet and the high-stability GPO triplet (Table 6). Previous studies suggested that in the absence of a mutation, the interchange of triplets located at different positions does not affect thermal stability unless favorable charge interactions are created (*31, 32*).

The CD spectra of both peptides exhibit a maximum at 225 nm, which is characteristic of the triple-helix conformation (Figure 4A), but peptide GFASPAGPO (3700 deg $cm^2$ $dmol^{-1}$) exhibited a slightly higher mean residue ellipticity ($MRE_{225}$) compared to peptide GPOSPAGFA (3570 deg $cm^2$ $dmol^{-1}$). The high $MRE_{225}$ of both peptides indicates the Gly → Ser mutation has been incorporated into a fully triple-helical molecule (*33*). The melting profiles of both peptides, as monitored by the decrease in the $MRE_{225}$ with increasing temperature, show a highly cooperative thermal transition (Figure 4B). Interestingly, the $T_m$ of peptide GFASPAGPO is 20.3 °C, while peptide GPOSPAGFA shows a lower $T_m$ of 16.2 °C. The observation of lower thermal stability for the peptide having the low-stability triplet C-terminal to the mutated triplet is consistent with the proposed model.

After heat denaturation, the refolding profiles of the peptides were monitored using $MRE_{225}$. Both peptides folded to a completely triple-helical form after 30 h. However, peptide GFASPAGPO exhibited a faster folding rate ($t_{1/2}$ = 40.7 min) than peptide GPOSPAGFA ($t_{1/2}$ = 58.7 min) (Figure 4C). DSC thermograms (Figure 4D) of the peptides confirm the CD observations that peptide GFASPAGPO has greater thermal stability than GPOSPAGFA, although the observed $T_m$ values are higher because of the faster DSC heating rate under these nonequilibrium conditions (*29*). The calorimetric enthalpy of peptide GFASPAGPO (340 kJ/mol) was higher than that found for peptide GPOSPAGFA (275 kJ/mol). These data are consistent with a significant decrease in stability, hydrogen bonding, and folding rate when the destabilizing triplet is C-terminal to the triplet with the mutation and provide experimental support for the $T_m[+1]$ model.

## DISCUSSION

Missense mutations in type I collagen lead to OI, but the relationship between the mutation and the resulting disease phenotype is poorly understood. Here, we present computational models relating glycine mutations in COL1A1 to clinical lethality and experimental data supporting conclusions derived from the models.

The models can be used to predict the lethality of OI based on the mutation. A decision tree predicts substitutions of Gly with Arg, Val, Asp, and Glu to be nonlethal in the N-terminal region of the triple helix, and lethal otherwise. Ala substitutions are predicted to be nonlethal. The lethality of Gly → Ser mutations is modeled by $T_m[+1]$, the relative thermostability of the triplet C-terminal to the mutated triplet. Mutations with high $T_m[+1]$ values are predicted to be nonlethal and those with low $T_m[+1]$ values lethal. The $T_m[+1]$ model is supported by accurate predictions for six novel mutations and by the experimental peptide results.

In addition to their use as prediction methods, the models provide insight into the etiology of OI. Both models implicate stability as a critical determinant of lethality, with lower stability associated with more severe disease. Substitutions of Arg, Val, Asp, and Glu for Gly have been shown to be the most destabilizing to the triple helix (*20*). The decision tree implies that, for these most disruptive mutations in the C-terminal region of the triple helix, the destabilization caused by the mutation itself generally leads to lethal disease, regardless of the local sequence environment. In contrast, the $T_m[+1]$ model suggests that the lethality of a relatively mild Ser substitution depends on its environment, since the lethality is modeled more accurately when the relative thermostability of the triplet C-terminal to the mutated triplet is considered. Furthermore, the impact on stability by triplets neighboring a Gly → Ser mutation is predicted to be asymmetric, since a statistically significant correlation with lethality was found for $T_m[+1]$ but not $T_m[-1]$.

The implication that Gly → Ser mutations with low $T_m[+1]$ values result in lower-than-expected stability is supported by experimental peptide results reported here (Table 6) and elsewhere (*33*). Lower stability, folding rate, and calorimetric enthalpy were observed for peptide GPOSPAGFA, with low $T_m[+1]$ and high $T_m[-1]$ values, than for peptide GFASPAGPO, with the positions of the high- and low-stability triplets reversed. These observed differences support an asymmetric synergistic effect of a low-stability triplet C-terminal to a Gly → Ser mutation which may contribute to the lethal phenotype. The lower calori-

metric enthalpy of GPOSPAGFA compared to that of GFASPAGPO is likely to reflect a loss of direct backbone H bonds C-terminal to the mutation site or to less direct hydration effects. The crystal structure of a Gly → Ala peptide, in an all Gly-Pro-Hyp environment, shows a very localized perturbation with the direct NH···CO bond replaced by a water-mediated bond and a local untwisting of the superhelix (*34*); it is possible that the presence of a less stable triplet C-terminal to the mutation site of the Gly → Ser substitution may further perturb the triple helix in an asymmetric manner. The peptide studies were conducted on homotrimers with a glycine substitution in all three chains, but it is likely that similar consequences will occur when a low-stability triplet is C-terminal to a glycine mutation site in one or two chains of type I collagen.

The structural effect of replacement of a Gly with Arg, Val, Glu, or Asp is not yet known. Although charged residues may promote triple helix stability in the X and Y positions of Gly-X-Y triplets, they appear to be destabilizing when in the normal Gly position. Electrostatic interactions could play a role in destabilization, but Gly → Val substitutions also lead to lethal phenotypes, suggesting that other factors, e.g., size, must also be important. This is consistent with the observation that the stability of peptides modeling charged substitutions for Gly was not strongly affected by varying salt concentration and pH (*20*).

The correlation between $T_m[+1]$ and lethality seen for Gly → Ser mutations in COL1A1 was also observed for Gly → Ser and Gly → Cys mutations in COL1A2, suggesting that these three sets of mutations share a mechanism of lethality. The higher $T_m[+1]$ cutoff between lethal and nonlethal for Gly → Cys mutations compared to Gly → Ser mutations in COL1A2 (Table 4) implies that greater stability near the mutation site is required to maintain the integrity of the triple helix in the presence of this substitution. This is consistent with the observation that, when replacing Gly in model peptides, Cys causes a greater decrease in thermostability than Ser (*20*). The difference in cutoff $T_m[+1]$ between Gly → Ser mutations in the two chains (Table 4) may explain why there are fewer lethal Ser substitutions in COL1A2 than in COL1A1 (Table 1). Of the 338 possible Gly → Ser mutations in COL1A1, 33% of those with a measured $T_m[+1]$ have relative thermostability values at or below the cutoff of 37.7 °C, and thus are predicted to be lethal. In contrast, although a similar fraction (34%) of the triplets in COL1A2 also have $T_m[+1]$ values at or below 37.7 °C, the cutoff for this chain was set at the lower value of 32.9 °C, resulting in lethal predictions for only 14% of the postions. The lower cutoff for COL1A2 suggests that this chain has a weaker effect on the thermostability of the heterotrimer, possibly due to the protein composition of one COL1A2 chain versus two COL1A1 chains. We also tested whether the $T_m[+1]$ model applies to Gly → Ala and Gly → Cys mutations in COL1A1; however, the results were not definitive. For Ala, there are too few lethal mutations with $T_m[+1]$ values to build a robust model. The ambiguous results for Cys substitutions may reflect their capacity to form intramolecular disulfide bonds (*35*).

Exceptions to the models may provide clues to additional factors that could contribute to the development of a lethal phenotype, which can then be incorporated into future, refined models. One factor that may be important for clinical lethality is ligand-binding sites, and two regions with only lethal mutations overlap proposed major ligand-binding regions in COL1A1 (*9*). Lethal mutations that were incorrectly predicted by the decision tree or $T_m[+1]$ models were examined to assess whether their phenotype might result from interfering with the function of one of these sites. Although no lethal mutations in these regions were incorrectly predicted for Arg, Val, Glu, Asp, Cys, or Ser substitutions, two lethal Gly → Ala mutations do lie within these sites. This result suggests that interfering with ligand binding may be a mechanism of lethality only for the least destabilizing mutations and is consistent with the proposal derived from the modeling that the destabilization induced by Ala substitutions may be insufficient to lead to a lethal phenotype. Additional data are required to test these possibilities.

Our hypothesis that the lethality of Gly → Ser mutations is related to disruption of triple-helical structure on the C-terminal side is supported by NMR studies (*36*) and molecular dynamics simulations (*13*). However, it has been suggested that residues on the N-terminal side may contribute to lethality by influencing renucleation of the triple helix following disruption by a mutation (*36, 37*). The results presented here suggest that the thermostabilities of the three triplets N-terminal to a Ser substitution do not systematically correlate with lethality but do not exclude the possibility that regions either farther upstream or variable in distance from the mutation play important roles.

Four simplifications were made in constructing the models that could affect their performance. (i) The continuum of disease phenotypes (*1*) was discretized into lethal and nonlethal classes. Subjectivity inherent in disease diagnosis could lead to classification errors. One possible example is Gly874Ser, which is apparently misclassified by the model (Table S1 of the Supporting Information); however, the true lethality is unknown due to insufficient clinical data for a definitive diagnosis (P. Byers, personal communication). (ii) The models use hard cutoffs for prediction but represent biological processes likely to be gradual transitions. An alternative is to output a probability of lethality; however, the usefulness of such a score will depend on the application and the data used in its construction. (iii) The models represent the complex process of disease development by a very small number of features. Additional factors not yet included in the model, such as ligand-binding sites (*9*) and individual variation, are likely to modulate the phenotype. For example, Gly1003Ser in COL1A1 lies in a region required for nucleation of the triple helix (*38*), and its lethal phenotype may result from disruption of this important function. (iv) Local stability was approximated by the relative thermostability of homotrimeric peptides. Multiple factors, including the mutation itself, the environment in the native protein, and the conditions of the experiment (*39*), could alter the relative $T_m$ values. Furthermore, the set of triplets with measured $T_m$ values is incomplete, and there may be unintentional bias in the selected sequences.

Another factor limiting the models is that the set of mutations used in their construction is incomplete. With the available data, it is not possible to distinguish whether the lethality of N-terminal Gly → Ser mutations should be predicted by the $T_m[+1]$ model or by the decision tree. Since all known Ser substitutions at or N-terminal to position 178 are nonlethal, they are equally well predicted by both models.

The set of available mutations also limits the precision to which the cutoffs in the models can be specified. In the decision tree, the cutoff between N- and C-terminal regions is estimated at residue 178, but with the given data, any of the residues between positions 155 and 201 could be selected without introducing more than one additional misclassified mutation. The data also support a model with cutoffs specific for each amino acid, a possibility that had been raised previously (*40*), and trees with this structure are apparent in the cross-validation runs. The availability of appropriate data will also facilitate the extension of the models to discriminating types of nonlethal OI.

The analysis presented here of the effects of collagen mutations was aided by distinguishing sets of substitutions likely to share a mechanism of lethality. The groups may differ in the process affected, or the degree of effect. The identification of groups of Gly mutations, including N-terminal mutations, C-terminal bulky substitutions, Gly → Ser mutations with low and high $T_m$[+1] values, and Ala substitutions, will be useful in guiding experiments to determine the biological mechanism by which the mutations and their effect on collagen stability result in OI phenotypes. This divide-and-conquer approach is not specific for OI. Genotype−phenotype association studies are revealing a myriad of disease-associated genes. Determining the impact of sequence variations on the encoded proteins will be a necessary next step in understanding the mechanisms by which they lead to disease.

## ACKNOWLEDGMENT

We thank Michael Walker and Randall Radmer (Stanford University, Stanford, CA) for helpful discussions, Russ Altman (Stanford University) for critical reading of the manuscript, and Peter Byers (University of Washington, Seattle, WA), Linda DiMeglio (Indiana University, Bloomington, IN), and TingFung Chan and Pui-Yan Kwok (University of California, San Francisco, CA) for providing unpublished mutations.

## SUPPORTING INFORMATION AVAILABLE

$T_m$[+1] values for COL1A1 Gly → Ser mutations (Table S1), $T_m$[+1] values for COL1A2 Gly → Ser mutations (Table S2), and $T_m$[+1] values for COL1A2 Gly → Cys mutations (Table S3). This material is available free of charge via the Internet at http://pubs.acs.org.

## REFERENCES

1. Byers, P. H., and Cole, W. G. (2002) Osteogenesis Imperfecta, in *Connective Tissue and its Heritable Disorders* (Royce, P. M., and Steinmann, B., Eds.) pp 385–430, Wiley-Liss, New York.
2. Marini, J. C., and Letocha, A. D. (2005) Osteogenesis Imperfecta, in *Disease of Bone and Mineral Metabolism* (Arnold, A., Ed.) endotext.com.
3. Byers, P. H. (2001) Disorders of Collagen Biosynthesis and Structure, in *Online Metabolic and Molecular Bases of Inherited Disease*, pp 5241–5285.
4. Cabral, W. A., Milgrom, S., Letocha, A. D., Moriarty, E., and Marini, J. C. (2006) Biochemical screening of type I collagen in osteogenesis imperfecta: Detection of glycine substitutions in the amino end of the α chains requires supplementation by molecular analysis. *J. Med. Genet. 43*, 685–690.
5. Byers, P. H. (1989) Inherited disorders of collagen gene structure and expression. *Am. J. Med. Genet. 34*, 72–80.
6. Marini, J. C., Lewis, M. B., Wang, Q., Chen, K. J., and Orrison, B. M. (1993) Serine for glycine substitutions in type I collagen in two cases of type IV osteogenesis imperfecta (OI). Additional evidence for a regional model of OI pathophysiology. *J. Biol. Chem. 268*, 2667–2673.
7. Wang, Q., Orrison, B. M., and Marini, J. C. (1993) Two additional cases of osteogenesis imperfecta with substitutions for glycine in the α2(I) collagen chain. A regional model relating mutation location with phenotype. *J. Biol. Chem. 268*, 25162–25167.
8. Sztrolovics, R., Glorieux, F. H., van der Rest, M., and Roughley, P. J. (1993) Identification of type I collagen gene (COL1A2) mutations in nonlethal osteogenesis imperfecta. *Hum. Mol. Genet. 2*, 1319–1321.
9. Marini, J. C., Forlino, A., Cabral, W. A., Barnes, A. M., San Antonio, J. D., Milgrom, S., Hyland, J. C., Korkko, J., Prockop, D. J., De Paepe, A., Coucke, P., Symoens, S., Glorieux, F. H., Roughley, P. J., Lund, A. M., Kuurila-Svahn, K., Hartikka, H., Cohn, D. H., Krakow, D., Mottes, M., Schwarze, U., Chen, D., Yang, K., Kuslich, C., Troendle, J., Dalgleish, R., and Byers, P. H. (2007) Consortium for osteogenesis imperfecta mutations in the helical domain of type I collagen: Regions rich in lethal mutations align with collagen binding sites for integrins and proteoglycans. *Hum. Mutat. 28*, 209–221.
10. Hewett, R., Leuchner, J., Mooney, S., and Klein, T. E. (2003) Analysis of mutations in the COL1A1 gene with second-order rule induction. *Int. J. Pattern Recognit. 17*, 721–740.
11. Hunter, L., and Klein, T. E. (1993) Finding Relevant Biomolecular Features in Osteogenesis imperfecta. *The Proceedings of the 1993 Conference on Intelligent Systems and Molecular Biology 1*, 190–197.
12. Klein, T. E., and Wong, E. (1992) *Proceedings of the International Conference on System Science*, Vol. 1, IEEE Press, Piscataway, NJ.
13. Radmer, R. J., and Klein, T. E. (2004) Severity of osteogenesis imperfecta and structure of a collagen-like peptide modeling a lethal mutation site. *Biochemistry 43*, 5314–5323.
14. Bachinger, H. P., and Davis, J. M. (1991) Sequence specific thermal stability of the collagen triple helix. *Int. J. Biol. Macromol. 13*, 152–156.
15. Bachinger, H. P., Morris, N. P., and Davis, J. M. (1993) Thermal stability and folding of the collagen triple helix and the effects of mutations in osteogenesis imperfecta on the triple helix of type I collagen. *Am. J. Med. Genet. 45*, 152–162.
16. Bateman, J. F., Moeller, I., Hannagan, M., Chan, D., and Cole, W. G. (1992) Characterization of three osteogenesis imperfecta collagen α1(I) glycine to serine mutations demonstrating a position-dependent gradient of phenotypic severity. *Biochem. J. 288* (Part 1), 131–135.
17. Westerhausen, A., Kishi, J., and Prockop, D. J. (1990) Mutations that substitute serine for glycine alpha 1–598 and glycine alpha 1–631 in type I procollagen. The effects on thermal unfolding of the triple helix are position-specific and demonstrate that the protein unfolds through a series of cooperative blocks. *J. Biol. Chem. 265*, 13995–14000.
18. Kuivaniemi, H., Tromp, G., and Prockop, D. J. (1991) Mutations in collagen genes: Causes of rare and some common diseases in humans. *FASEB J. 5*, 2052–2060.
19. Makareeva, E., Mertz, E. L., Kuznetsova, N. V., Sutter, M. B., Deridder, A. M., Cabral, W. A., Barnes, A. M., McBride, D. J., Marini, J. C., and Leikin, S. (2008) Structural heterogeneity of type I collagen triple helix and its role in osteogenesis imperfecta. *J. Biol. Chem. 283*, 4787–4798.
20. 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.
21. Persikov, A. V., Ramshaw, J. A., and Brodsky, B. (2005) Prediction of collagen stability from amino acid sequence. *J. Biol. Chem. 280*, 19343–19349.
22. Di Lullo, G. A., Sweeney, S. M., Korkko, J., Ala-Kokko, L., and San Antonio, J. D. (2002) Mapping the ligand-binding sites and disease-associated mutations on the most abundant protein in the human, type I collagen. *J. Biol. Chem. 277*, 4223–4231.
23. Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007) NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res. 35*, D61–D65.

24. Online Mendelian Inheritance in Man, OMIM (2007) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, and National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD (http://www.ncbi.nlm.nih.gov/omim/).
25. Witten, I. H., and Frank, E. (2005) *Data Mining: Practical machine learning tools and techniques*, 2nd ed., Morgan Kaufman, San Francisco.
26. Team, R. D. C. (2007) *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna.
27. Hothorn, T., and Hornik, K. (2006) exactRankTests: Exact Distributions for Rank and Permutations Tests.
28. Fawcett, T. (2003) *ROC Graphs: Notes and Practical Considerations for Data Mining Researchers*, HP Laboratories, Palo Alto, CA.
29. Persikov, A. V., Xu, Y., and Brodsky, B. (2004) Equilibrium thermal transitions of collagen model peptides. *Protein Sci. 13*, 893–902.
30. Venugopal, M. G., Ramshaw, J. A., Braswell, E., Zhu, D., and Brodsky, B. (1994) Electrostatic interactions in collagen-like triple-helical peptides. *Biochemistry 33*, 7948–7956.
31. Buevich, A. V., Silva, T., Brodsky, B., and Baum, J. (2004) Transformation of the mechanism of triple-helix peptide folding in the absence of a C-terminal nucleation domain and its implications for mutations in collagen disorders. *J. Biol. Chem. 279*, 46890–46895.
32. Persikov, A. V., Ramshaw, J. A., Kirkpatrick, A., and Brodsky, B. (2005) Electrostatic interactions involving lysine make major contributions to collagen triple-helix stability. *Biochemistry 44*, 1414–1422.
33. Yang, W., Battineni, M. L., and Brodsky, B. (1997) Amino acid sequence environment modulates the disruption by osteogenesis imperfecta glycine substitutions in collagen-like peptides. *Biochemistry 36*, 6930–6935.
34. Bella, J., Brodsky, B., and Berman, H. M. (1996) Disrupted collagen architecture in the crystal structure of a triple-helical peptide with a Gly → Ala substitution. *Connect. Tissue Res. 35*, 401–406.
35. Torre-Blanco, A., Adachi, E., Romanic, A. M., and Prockop, D. J. (1992) Copolymerization of normal type I collagen with three mutated type I collagens containing substitutions of cysteine at different glycine positions in the α1 (I) chain. *J. Biol. Chem. 267*, 4968–4973.
36. 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 bone disease. *Biochemistry 37*, 15528–15533.
37. Wenstrup, R. J., Shrago-Howe, A. W., Lever, L. W., Phillips, C. L., Byers, P. H., and Cohn, D. H. (1991) The effects of different cysteine for glycine substitutions within α2(I) chains. Evidence of distinct structural domains within the type I collagen triple helix. *J. Biol. Chem. 266*, 2590–2594.
38. McLaughlin, S. H., and Bulleid, N. J. (1998) Molecular recognition in procollagen chain assembly. *Matrix Biol. 16*, 369–377.
39. Chan, V. C., Ramshaw, J. A., Kirkpatrick, A., Beck, K., and Brodsky, B. (1997) Positional preferences of ionizable residues in Gly-X-Y triplets of the collagen triple-helix. *J. Biol. Chem. 272*, 31441–31446.
40. Byers, P. H. (1990) Brittle bones−fragile molecules: Disorders of collagen gene structure and expression. *Trends Genet. 6*, 293–300.

BI800026K