Statistical Thermodynamics of the Collagen Triple-Helix/Coil Transition. Free Energies for Amino Acid Substitutions within the Triple-Helix

Andrew J. Doig*

Manchester Interdisciplinary Biocentre, The University of Manchester, 131 Princess Street, Manchester M1 7DN, U.K.

Received: July 7, 2008; Revised Manuscript Received: September 17, 2008

Collagen sequences frequently deviate from the most thermally stable $(Gly\text{-}Pro\text{-}Hyp)_n$ pattern, with many mutations causing osteogenesis imperfecta (or "brittle bone disease"). The effects of collagen mutations have been studied in short peptides. The analysis of this work is problematic, however, as triple-helices fray from their ends, making the coil/triple-helix equilibrium non-two-state. Here, I develop a statistical thermodynamic model to handle this equilibrium that is applicable to peptides that follow the $(G\text{-}X\text{-}Y)_n$ pattern, where Gly is present at every third position and where all three chains are identical. Parameters for substitutions at each position are included, as well as a penalty for initiating triple-helix formation. The model is applied to equilibrium experimental data at 37 °C to show that the extension of a triple-helix by a three residue unit stabilizes the triple-helix by 0.76 kcal/mol for Gly-Pro-Hyp and 0.33 kcal/mol for Gly-Pro-Pro. The replacement of Hyp by Arg, Asp, or Trp destabilizes the triple-helix by 1.5, 2.4, and 2.9 kcal/mol, respectively, where the substitution is present once in each chain. The model can thus be used to quantitatively interpret data on collagen peptides, giving free energies that can help rationalize mutations that affect collagen stability, and to design new collagen sequences.

Introduction

Collagen is the most abundant protein in vertebrates, forming a triple-helical structure and consisting of three supercoiled left-handed polyproline II-like helices. The close packing of three chains requires every third residue to be Gly, as any larger residue here is highly destabilizing.[1] The X and Y positions show strong preferences for Pro and hydroxyproline (Hyp), respectively, although deviations from this ideal sequence are common. In addition to natural variation in collagen sequences, many mutations are known that can lead to osteogenesis imperfecta (OI) or "brittle bone disease".[2] Understanding how mutations affect collagen stability can help rationalize their effects.

An attractive way to analyze sequence variations in collagen is to study short peptides of collagen-type sequences (i.e., $(Gly\text{-}Pro\text{-}Hyp)_n$, where n is typically 6−12) that form triple-helices in isolation. These are simple models for homotrimeric collagen. There are several difficulties with using peptides to study collagen structure and stability, however. First, equilibration is very slow, due to the large number of imino acids present. These form around 10% *cis* peptide bonds in the unfolded state, which interconvert with the *trans* conformation on a time scale of hours at room temperature.[3] Thermal denaturation experiments on these peptides, therefore, need to be performed very slowly, since the observed melting point (T_m) varies depending on the rate of change of temperature.[4] Second, peptide denaturation is not two-state between fully folded and fully unfolded conformations, since triple-helices often fray from their termini, giving a central folded triple-helix, with disordered ends. Identical substitutions will therefore have different energetic effects on triple-helix stability, depending on whether the substitution is in the middle of the chain or near a terminus. This problem can be handled by using a statistical thermodynamical model that calculates the stabilities and populations of all the conformations. Here, I develop such a model that can handle heterogeneous sequences of homotrimers, and I then apply it to previously published data to derive energetic terms for triple-helix elongation and substitutions. Although heterotrimers of collagen are common, they have been little studied in peptide models, so they are not considered further here.

TABLE 1: Alignment of G, X, and Y in Collagen

G	X	Y	G	X′	Y′	G	X″	Y″		
	G	X	Y	G	X′	Y′	G	X″	Y″	
		G	X	Y	G	X′	Y′	G	X″	Y″

The treatment of the triple-helix/coil equilibrium is similar to the Lifson−Roig model for the α-helix/coil equilibrium.[5,6] Every possible peptide conformation (i.e., helices of varying lengths and start positions) has a statistical weight that shows how stable it is. Individual residues are given statistical weights that depend on whether they are in a helix or not, where the greater the value of a statistical weight, the more stable the conformation. The weight of the conformation of an entire polypeptide is the product of the weights of the individual residues. Parameters are included for sites at helix termini and interiors, and for interactions between side chains.[7] Knowledge of these parameters allows the statistical weights and, hence, the populations of every conformation to be calculated.

Methods

All calculations were performed using the program Mathematica 6.0 (Wolfram Research Inc.) using a PC. Figures 1 and 2 were made using Mathematica and Figures 3 and 4 using Origin 7.5.

Theory

Representation of Triple-helix Formation. Collagen forms a triple-helix from peptides of typical sequence $(Gly\text{-}Pro\text{-}Hyp)_n$.

* To whom correspondence should be addressed. Phone: +44-161-3064224; fax: +44-161-236-0409; e-mail: andrew.doig@manchester.ac.uk.

10.1021/jp805979k CCC: $40.75
Published on Web 11/01/2008

TABLE 2: Coil Conformation Alignment

C	C	Y	G	X′	Y′	G	X″	Y″		
	C	X	Y	G	X′	Y′	G	X″	C	
		G	X	Y	G	X′	Y′	G	C	C

TABLE 3: Statistical Weight of Columns in Triple-Helix/Coil Model

column pair	weight of 2nd column
AA	1
AC	0
BA	1
BC	w
CA	1
CC	0
DA	1
DC	0
AB	v
AD	0
BB	0
BD	0
CB	0
CD	w
DB	w
DD	0

TABLE 4: Energetic Preferences for the Amino Acids at the Y Positions within the Collagen Triple-Helix at 37°C

residue Y	w	ΔG for addition of Gly-Pro-Y to triple-helix ($-RT$ ln w) (kcal/mol)	ΔG for substitution of 3 Hyp with 3 Y in triple-helix (kcal/mol)
Pro	1.7	−0.3	0.4
Arg	0.30	0.70	1.5
Asp	0.076	1.6	2.4
Trp	0.035	2.1	2.9

Gly occupies a unique site within the triple-helix that other amino acids are too large to occupy. The three polypeptides are offset so that a Gly, a Pro, and a Hyp site in different chains are adjacent. The sites are called G, X, and Y and are aligned within collagen as shown in Table 1.

More accurately, the collagen structure is not stable unless all three sites are occupied. Hence, the isolated G, XG, Y″X″ or Y″ structures shown in Table 3 will not be stable, and residues in these locations, at the end of a triple-helix, will instead be in a disordered coil conformation (C). The conformation in Table 1 is therefore better described as shown in Table 2.

If the polypeptide forms less than the maximum possible length of triple-helix, additional columns with three residues in C conformations can be added to the beginning or end of the triple-helix. We assume that staggered helices, with one chain offset by 3, 6, 9, etc. units are not formed. This is a reasonable assumption as they will be much less stable, as the triple-helices are shorter. Collagen structures can thus be described as series of columns, each containing sites for three residues:

$$\mathrm{A} = \begin{pmatrix} \mathrm{C} \\ \mathrm{C} \\ \mathrm{C} \end{pmatrix} \quad \mathrm{B} = \begin{pmatrix} \mathrm{G} \\ \mathrm{Y} \\ \mathrm{X} \end{pmatrix} \quad \mathrm{C} = \begin{pmatrix} \mathrm{X} \\ \mathrm{G} \\ \mathrm{Y} \end{pmatrix} \quad \mathrm{D} = \begin{pmatrix} \mathrm{Y} \\ \mathrm{X} \\ \mathrm{G} \end{pmatrix}$$

The B, C, and D columns are defined by three residues in different chains, each adopting the polyproline II conformation and being in contact, as in collagen; all other conformations are within column A. A triple-helix can thus be described as a repeating pattern of the columns BCDBCD... Since the polypeptides are identical, it is arbitrary which of the columns—B, C, or D—is adopted first at the N-terminus of a helix; here, I use B for this position. A triple-helix cannot start with either B, C, or D, since this would mean each triple-helix is counted three times instead of once. A triple-helix can terminate after either B, C, or D at the C-terminus. The conformation of a polypeptide of N residues forming one triple-helical segment can therefore be written as:

$$\mathrm{A}_a(\mathrm{BCD})_b\mathrm{A}_c \text{ or } \mathrm{A}_a(\mathrm{BCD})_b\mathrm{BCA}_c \text{ or } \mathrm{A}_a(\mathrm{BCD})_b\mathrm{BA}_c$$

where $a + 3b + c = N - 2$; $N - 2 - 3b - c \geq a \geq 0$; $N - 2 - 3b - a \geq c \geq 0$; $b = (N - 2 - a - c)/3$

The rules for adjacent columns are:

A can be followed by A or B;

B can be followed by A or C;

C can be followed by A or D;

D can be followed by A or B.

More than one triple-helical segment can be easily represented as strings of BCD... columns, separated by one or more A columns, showing a disordered region between the triple-helices. Any isolated triple-helix can thus be described as a one-dimensional string of letters. For example, a peptide that has 12 residues, with a triple-helix running from residue 4 to 8 in one chain, 5 to 9 in a second chain, and 6 to 10 in a third chain, is written as AAABCDBCAA.

Statistical Weights. Statistical weights show the stability of a peptide conformation. Peptide conformations differ depending on where their triple-helices start and how long the helices are. Every possible conformation of the polypeptide has a statistical weight, and the population of each conformation is equal to its statistical weight divided by the sum of the statistical weights for every conformation (the partition function, Z). The statistical weight of a peptide conformation is equal to the product of the statistical weights of each column in that conformation. This is a valid assumption for structures that lack long-range interactions, such as a linear triple-helix. If any substitution perturbs the unfolded state, such as by introducing hydrogen bonds between backbone groups or by affecting coil flexibility (by removing an imino acid, for example), these effects will be subsumed within the parameter w. The column weights are defined as follows:

All coil columns (A) have a weight of 1, as a reference. The weights of columns can be regarded as equilibrium constants compared to coil. A weight of more than 1 means that the column is more stable than coil and vice versa.

Triple-helical columns (B, C, and D) have weights of w. This is the weight for adding an additional triple-helical column to the end of an existing triple-helix. The number of columns with a weight of w in a triple-helical segment thus shows the length of the triple-helix and how many residues are in the triple-helix ($3w$). The single parameter w is the product of weights for residues to be in the G, X, and Y positions.

Forming the first column incurs a penalty (v) since it is entropically disadvantageous for three residues to find and bind to each other. Since a triple-helix always initiates with a B column following an A, an AB pair has a weight of vw, with w showing that there is a triple-helical column present and v the initiation penalty. The value of v is concentration dependent, being proportional to the concentration squared, since it is easier to form a triple-helix from 3 monomers at higher concentration. No other weights are concentration dependent. The conformation AAABCDBCAA will therefore have a statistical weight of vw.[5]

The weights for each column pair are summarized in Table 1. These weights can be rewritten as a matrix, **M**:

$$\mathbf{M} = \begin{array}{c} \\ \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \\ \mathrm{D} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{A} & \mathrm{B} & \mathrm{C} & \mathrm{D} \end{array} \\ \begin{pmatrix} 1 & \upsilon w & 0 & 0 \\ 1 & 0 & w & 0 \\ 1 & 0 & 0 & w \\ 1 & w & 0 & 0 \end{pmatrix} \end{array}$$

The state of the first column in a pair is shown as a row, and the state of the second column in a pair is shown above the matrix. Each statistical weight applies to the second column in a pair. For example, the statistical weight for the second residue in a DB pair can be read in **M** as w. Now, the partition function for triple-helix formation (Z) for a homopolymer can be calculated as:

$$Z = (1 \quad 0 \quad 0 \quad 0)\mathbf{M}^{N-2}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \tag{1}$$

The initial vector ensures that the N-terminus of the polypeptide is in a coil and the first column has to be an A or B column. The final vector allows the last column to have any conformation. N is in the number of residues, and the $N - 2$ term arises because a minimum of 3 residues is required to form a triple-helix. This matrix method is discussed in detail by Poland and Scheraga.[8]

Heteropolymers will have different values of the triple-helix/coil parameters. The simplest situation, previously studied experimentally, is to have a single (guest) substitution within a $(\text{Gly-Pro-Hyp})_n$ host, where one of the Pro or Hyp sites is replaced by another amino acid. For example, consider the Gly-Pro-Hyp-Gly-Lys-Hyp-Gly-Pro-Hyp sequence, where the guest is a Lys replacing a Pro at an X position. A single substitution will affect the statistical weights of three successive columns. In this example, the Lys at position X′ will affect the w-values of successive C, D, and B columns. The matrix **M** will differ at each of the positions where Lys is at X′ from when Pro is at X or X″. The partition function, Z, for this sequence will therefore be:

$$\mathrm{Z} = (1 \quad 0 \quad 0 \quad 0)\cdot\mathbf{M}\cdot\mathbf{M}\cdot\mathbf{M}_\mathrm{g}\cdot\mathbf{M}_\mathrm{g}\cdot\mathbf{M}_\mathrm{g}\cdot\mathbf{M}\cdot\mathbf{M}\cdot\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \tag{2}$$

where

$$\mathbf{M}_\mathrm{g} = \begin{pmatrix} 1 & \upsilon w_\mathrm{g} & 0 & 0 \\ 1 & 0 & w_\mathrm{g} & 0 \\ 1 & 0 & 0 & w_\mathrm{g} \\ 1 & w_\mathrm{g} & 0 & 0 \end{pmatrix}$$

w is the product of the statistical weights for Gly at the G position, Pro at the X position, and Hyp at the Y position, and w_g is the product of the statistical weights for Gly at the G position, Lys at the X position, and Hyp at the Y position. Any substitution can be handled in the same way, with a different matrix, a guest w-value of w_g and three successive columns containing the guest matrix when calculating Z.

Triple-helix Content. The overall fraction of residues that form triple-helix (f) is the population of residues with a w weighting, divided by the maximum possible population of residues with a w weighting. For homopolymers this is given by eq 3.

$$f = \frac{\partial(\ln Z)/\partial(\ln w)}{N - 2} \tag{3}$$

When a guest residue is present, with w_g replacing w:

$$f = \frac{\partial(\ln Z)/\partial(\ln w) + \partial(\ln Z)/\partial(\ln w_\mathrm{g})}{N-2} \tag{4}$$

The probability that position i is in a triple-helix is given by eq 5. In this equation, the matrix for position i has the first column all set to zero, making the probability of position i being in a coil conformation zero. The equation is therefore the sum of the statistical weights for all the conformations where position i is helical divided by the sum of the statistical weights for all conformations. In other words, this equation is the sum of the weights of all the conformations where residue i is helical divided by the sum of the weights for all conformations, thus giving the fraction of the time that residue i is helical.

$$p\ (\text{position } i \text{ is in a triple-helix}) = \frac{1}{\mathrm{Z}}(1 \quad 0 \quad 0 \quad 0)\prod_{j=1}^{j=i-1}\begin{pmatrix} 1 & \upsilon w & 0 & 0 \\ 1 & 0 & w & 0 \\ 1 & 0 & 0 & w \\ 1 & w & 0 & 0 \end{pmatrix}\begin{pmatrix} 0 & \upsilon w & 0 & 0 \\ 0 & 0 & w & 0 \\ 0 & 0 & 0 & w \\ 0 & w & 0 & 0 \end{pmatrix}\prod_{j=i+1}^{j=N-2}\begin{pmatrix} 1 & \upsilon w & 0 & 0 \\ 1 & 0 & w & 0 \\ 1 & 0 & 0 & w \\ 1 & w & 0 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \tag{5}$$

Results

Determination of Amino Acid Preferences in Collagen. The triple-helix/coil parameters can be determined by calculating values of parameters that give a fraction helix equal to that measured experimentally. It is essential that the experimental data is at equilibrium. Although the stabilities of many collagen peptides have been reported as melting points, these are rarely at equilibrium, due the very slow equilibration between the numerous *cis*/*trans* isomers in Pro and Hyp. Reported melting points thus usually depend on the rate of heating. Equilibrium data has been reported, however, by Persikov et al.[9] This data can be analyzed with the model to obtain triple-helix/coil parameters and hence free energies for triple-helix formation.

Locardi et al. have previously also developed a model for the triple-helix/coil equilibrium,[10] based on an earlier model of Schwarz and Poland.[11] They included an initiation parameter, σ, that is similar to the initiation parameter υ, used here, although their treatment of helix propagation was rather different. For the peptide $\text{Ac-(Gly-Pro-Hyp)}_5\text{-CONH}_2$ they found that σ is 0.0202 $(\text{mol/L})^{-2}$ at a concentration of 1 mM. Persikov et al. used a peptide concentration (c_0) of 0.369 mM for $\text{Ac-(Pro-Hyp-Gly)}_{10}\text{-CONH}_2$.[9] If we assume that nucleation is independent of chain length and temperature, then υ for this peptide is $0.0202 \times (0.369 \times 10^{-3})^2 = 2.75 \times 10^{-9}$. Multiplying by the concentration squared converts σ to the dimensionless υ. It is assumed that the same value of υ is applicable to all sequences, due to the current lack of experimental data to determine varying υ-values.

The peptide Ac-(Pro-Hyp-Gly)$_{10}$-CONH$_2$ has a T_m of 330.1 K and $\Delta H°$ for the coil/triple-helix transition of 96 kcal/mol (the mean of the two van't Hoff enthalpies reported) at a concentration of 1 mg/ml. I choose a temperature of 310.1 K (37 °C) to analyze from now on, as blood temperature, which is most relevant to collagen in vivo. The fraction helix for this peptide, can be found by equating two expressions for the equilibrium constant for triple-helix formation:[12,13]

$$\frac{3c_0^2(1-f)^3}{f} = \exp\left[\frac{\Delta H^0}{RT}\left(\frac{T}{T_m}-1\right)-\ln\left(\frac{3c_0^2}{4}\right)\right] \qquad (6)$$

Solving, we find that $f = 0.971$. The fraction helix, f, is calculated from helix-coil theory using eq 1, with $N = 30$ (10 triplets) and $v = 2.75 \times 10^{-9}$. The only unknown in eq 3 is w, which is 3.46, after the expression for f is set to 0.971. 3.46 is the product of the statistical weights for adding a Gly, a Pro, and a Hyp to the G, X, and Y positions of an existing triple-helix at 310.1 K. The free energy for extending a triple-helix by one Pro-Hyp-Gly unit is, therefore, $-RT \ln 3.46 = -0.76$ kcal/mol at 37 °C. These results are dependent on the value of the nucleation parameter obtained from the Locardi et al. data. Although it is somewhat questionable to use this data, since it is not certain that their data was at equilibrium and the ease of nucleation may vary with peptide length and temperature, I found that the helix content, and hence the w-value, was highly insensitive to varying the value of v (data not shown). Errors in w-values arising from uncertainty in the value of v are thus small.

The peptide (Pro-Pro-Gly)$_{10}$ has a T_m of 297.6 K and $\Delta H°$ for the coil/triple-helix transition of 63 kcal/mol at a concentration of 1 mg/mL.[9] From eq 6, its fraction helix (f) at 310.1 K is 0.041. Solving eq 3 as above gives $w = 1.70$. The free energy for extending a triple-helix by one Pro-Pro-Gly unit is, therefore, $-RT \ln 1.70 = -0.33$ kcal/mol at 37 °C. The replacement of Hyp by Pro at the X position of a triplet therefore destabilizes the collagen triple-helix by $0.76 - 0.33 = 0.43$ kcal/mol at 37 °C. Because there are three X positions within each triple-helix, each individual Hyp-to-Pro replacement destabilizes the triple-helix by 0.15 kcal/mol.

The w-value for Asp at position Y is determined as follows: the peptide Ac-(Gly-Pro-Hyp)$_3$-Gly-Pro-Asp-(Gly-Pro-Hyp)$_4$-CONH$_2$ has a T_m of 301.2 K and $\Delta H°$ for the coil/triple-helix transition of 97 kcal/mol at a concentration of 0.460 mM.[14] The fraction helix (f) for this peptide at 310.1 K, found using eq 6, is 0.039. We have a guest residue at the 12th position of the peptide. The partition function will, therefore, be:

$$Z = (1 \quad 0 \quad 0 \quad 0)\cdot\mathbf{M}^9\cdot\mathbf{M}_g\cdot\mathbf{M}_g\cdot\mathbf{M}_g\cdot\mathbf{M}^{10}\cdot\begin{pmatrix}1\\1\\1\\1\end{pmatrix} \qquad (7)$$

The fraction helix is given by eq 4. The w-value used in the matrix $\mathbf{M}$ is 3.46, since the replacement is within a Gly-Pro-Hyp background. The only unknown in this equation is w_g. Solving, we find that $w_g = 0.076$, equivalent to a free energy of $-RT \ln w_g = 1.6$ kcal/mol. Replacing a Hyp at position Y with an Asp in a triple-helix will therefore destabilize the triple-helix by $1.6 + 0.8 = 2.4$ kcal/mol. Each Asp will be present three times in the triple-helix, so an individual Asp is destabilizing to the triple-helix by 0.8 kcal/mol. Data on the peptides

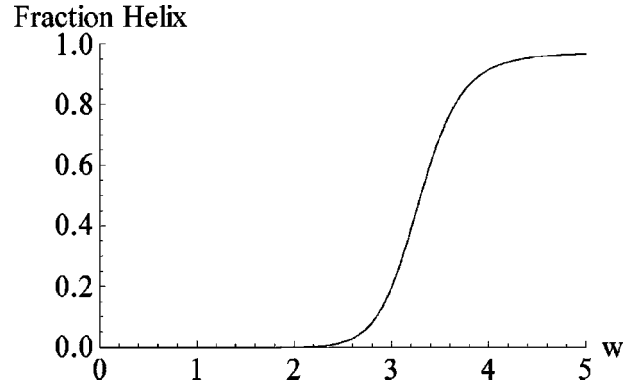

Figure 1. The triple-helix fraction for (Gly-Pro-Hyp)$_6$ as a function of w, with $v = 2.75 \times 10^{-9}$.

Ac-(Gly-Pro-Hyp)$_3$-Gly-Pro-Trp-(Gly-Pro-Hyp)$_4$-CONH$_2$ and Ac-(Gly-Pro-Hyp)$_3$-Gly-Pro-Arg-(Gly-Pro-Hyp)$_4$-CONH$_2$ can be used in the same way to give w_g for Arg at the Y position as 0.30 and w_g for Trp at the Y position as 0.035, equivalent to 0.7 and 2.1 kcal/mol, respectively. Complete results are shown in Table 4. Equilibrium data for the T1−892 peptide is also known,[14] but it contains too many substitutions from the Gly-Pro-Hyp pattern, and hence too many unknown parameters, to quantitatively analyze.

The partition function and fraction helix were calculated for an 18 residue host peptide (Gly-Pro-Hyp)$_6$ using eqs 1 and 3. Figure 1 shows the fraction triple-helix for this sequence as a function of w, with $v = 2.75 \times 10^{-9}$. A w-value above 2.8 is required for significant triple-helix formation; when w is above 4 triple-helix formation is essentially complete. Figure 2 shows the fraction triple-helix for this sequence as a function of v, with $w = 3.46$. A smooth increase in triple-helix content with increasing v is seen. Figure 3 shows the fraction triple-helix for a peptide (Gly-Pro-Hyp)$_n$ as a function of the number of residues in the peptide. At least 14 residues are needed for significant triple-helix formation with these parameters. As the number of residues increases, the fraction triple-helix will level off at a value less than 100%, due to terminal fraying and long helices breaking in the middle to form more than one helical segment. Figure 4 shows the fraction helix as a function of position for (Gly-Pro-Hyp)$_6$, with $v = 2.75 \times 10^{-9}$ and $w = 3.46$, calculated from eq 5. Fraying is apparent at the termini, with the first and last few residues being less likely to be helical than those in the center. Fraying is equally likely from either ends. The mean value of fraction triple-helix for each residue is equal to the mean triple-helix content of the individual residues (0.661). Lower values of w, which will be the case for deviations from the Gly-Pro-Hyp pattern, will result in greater fraying, as well as a lower mean triple-helix content.

Discussion

Other groups have studied the statistical thermodynamics of the formation of a triple-helix, deriving parameters for triple-helix initiation and propagation as functions of temperature, concentration, and chain length.[10,11] Previous work has been limited to triplet repeat sequences, such as Ac-(Gly-Pro-Hyp)$_n$-NH$_2$ and H-(Gly-Pro-Pro)$_n$-OH, where every unit is assigned the same initiation and propagation parameters. The model developed here is applicable to heteropolymers of any sequence within the (GXY)$_n$ framework.

Persikov et al. have successfully developed algorithms to predict the stability of triple-helices by quantifying the effects of substitutions on T_m values.[15] Their method is not easy to

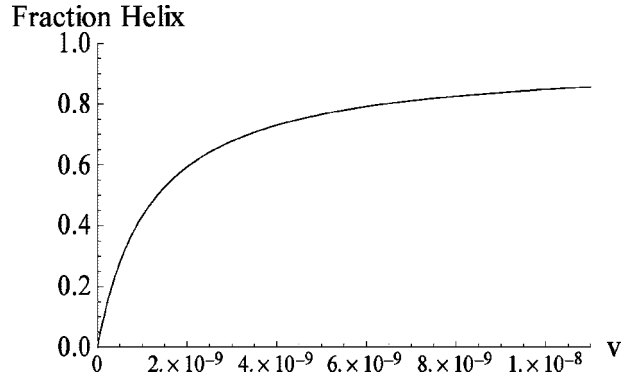

Figure 2. The triple-helix fraction for (Gly-Pro-Hyp)$_6$ as a function of v, with $w = 3.46$.

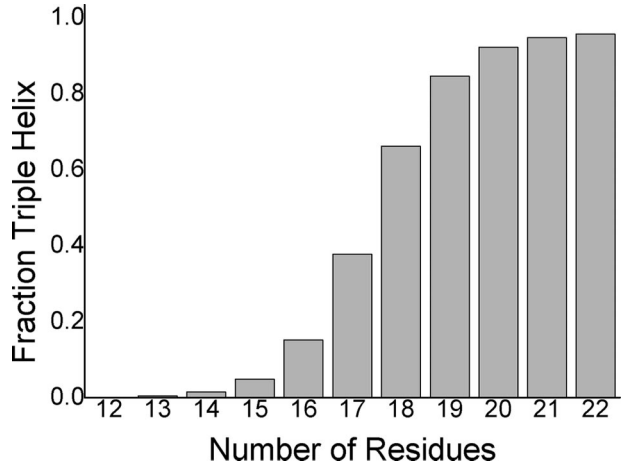

Figure 3. The triple-helix fraction for (Gly-Pro-Hyp)$_n$ as a function of the number of residues in the peptide, with $w = 3.46$ and $v = 2.75 \times 10^{-9}$.

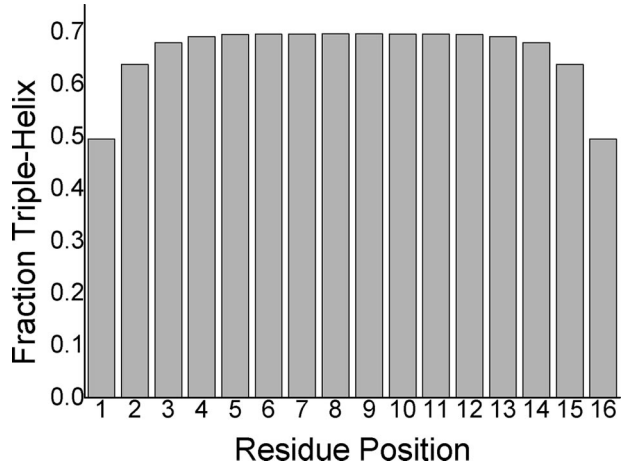

Figure 4. The probability of triple-helix formation for each unit in (Gly-Pro-Hyp)$_6$, with $v = 2.75 \times 10^{-9}$ and $w = 3.46$. There are 16 residue positions, each with a Gly, a Pro and a Hyp in the triple-helix, due to the offset arrangement of the triple-helix.

transfer to different collagen sequences, however, since identical substitutions will have different effects on T_m in different contexts, and the T_m values vary with rate of heating of the peptide. In contrast, the ΔG values determined here are generally applicable to any collagen sequence at equilibrium that does not have unknown side chain interactions energies and that does not form interactions with another triple-helix.

Application of the model to additional experimental equilibrium data would allow additional w-values to be determined. This would allows us to calculate the stability of any triple-helix of any length at 37 °C, provided that triple-helix/coil parameters for all the residues present are known, there are no interactions between triple-helices, no side-chain interactions within the triple-helix and the three polypeptides align correctly in a triple-helix. The model can deal with more than one triple-helical segment present in the same polypeptide. One long triple-helix is the most stable single conformation. In very long chains, however, there are so many ways of making multiple helical segments of a reasonable length, that they occasionally get populated. One difficulty with multiple helical segments is that initiation will be easier if a triple-helical segment is already present elsewhere in the chain. The initiation of the first triple-helix is difficult as three monomers need to find each other. If three chains are already bonded by a triple-helix, it is easier for the chains to associate elsewhere in the sequence, so v will be higher for initiating the second triple-helix and depend on the length of the disordered region between the helical segments. Such effects are beyond the scope of this work.

As the w-values for the most frequently observed Gly-Pro-Hyp and Gly-Pro-Pro triplets are close to 1 at 37 °C, ΔG for adding a unit to the end of an existing triple-helix is close to zero, leading to fraying at the ends of the triple-helix (Figure 4). As a large number of conformations are populated, a statistical thermodynamical model that takes account of this equilibrium is more accurate than a two-state, all-or-nothing, analysis, particularly if a region where fraying is frequent is perturbed.

Conclusions

A statistical thermodynamic model for triple-helix formation from identical polypeptides has been derived that allows the quantitative interpretation of mutations within triple-helices. When applied to experimental data, free energies for the propagation of a triple-helix and the replacement of Hyp by Pro, Arg, Asp, and Trp in a (Gly-Pro-Hyp)$_n$ background are derived. I suggest that future experimental data on triple-helix formation at equilibrium is analyzed in this way, leading to free energy scales that can be used to design new triple-helices. This data may improve our understanding of mutations in collagen that cause triple-helix destabilization and disease, though this should preferably be in peptide models that are closer to natural sequences.

Acknowledgment. I thank Richard Kammerer and Karl Kadler for helpful suggestions on the manuscript.

References and Notes

(1) Beck, K.; Chan, V. C.; Shenoy, N.; Kirkpatrick, A.; Ramshaw, J. A. M.; Brodsky, B. *Proc. Nat. Acad. Sci. U.S.A.* **2000**, *97*, 4273.
(2) Kuivaniemi, H.; Tromp, G.; Prockop, D. J. *FASEB J.* **1991**, *5*, 2052.
(3) Engel, J.; Prockop, D. J. *Annu. Rev. Biophys. Biophys. Chem.* **1991**, *20*, 137.
(4) Miles, C. A.; Bailey, A. J. *J. Mol. Biol.* **2004**, *337*, 917.
(5) Lifson, S.; Roig, A. *J. Chem. Phys.* **1961**, *34*, 1963.
(6) Qian, H.; Schellman, J. A. *J. Phys. Chem.* **1992**, *96*, 3987.
(7) Doig, A. J. *Biophys. Chem* **2002**, *101*−*102*, 281.
(8) Poland, D.; Scheraga, H. A. Academic Press: New York and London, 1970.
(9) Persikov, A. V.; Xu, Y.; Brodsky, B. *Protein Sci.* **2004**, *13*, 893.
(10) Locardi, E.; Kwak, J.; Scheraga, H. A.; Goodman, M. *J. Phys. Chem. A* **1999**, *103*, 10561.
(11) Schwarz, M.; D., P. *Biopolymers* **1974**, *13*, 687.
(12) Persikov, A. V.; Ramshaw, J. A. M.; Kirkpatrick, A.; Brodsky, B. *Biochemistry* **2000**, *39*, 14960.
(13) Engel, J.; Chen, H. T.; Prockop, D. J.; Klump, H. *Biopolymers* **1977**, *16*, 601.
(14) Persikov, A. V.; Xu, Y.; Brodsky, B. *Protein Sci.* **2007**, *13*, 893.
(15) Persikov, A. V.; Ramshaw, J. A. M.; Brodsky, B. *J. Biol. Chem.* **2005**, *280*, 19343.

JP805979K