Amino Acid Propensities for the Collagen Triple-Helix†

Anton V. Persikov,‡ John A. M. Ramshaw,§ Alan Kirkpatrick,§ and Barbara Brodsky*,‡

Department of Biochemistry, UMDNJ, Robert Wood Johnson Medical School, 675 Hoes Lane, Piscataway, New Jersey 08854, and CSIRO, 343 Royal Parade, Parkville, Australia

Received July 6, 2000; Revised Manuscript Received September 25, 2000

ABSTRACT: Determination of the tendencies of amino acids to form α-helical and β-sheet structures has been important in clarifying stabilizing interactions, protein design, and the protein folding problem. In this study, we have determined for the first time a complete scale of amino acid propensities for another important protein motif: the collagen triple-helix conformation with its Gly-X-Y repeating sequence. Guest triplets of the form Gly-X-Hyp and Gly-Pro-Y are used to quantitate the conformational propensities of all 20 amino acids for the X and Y positions in the context of a $(Gly\text{-}Pro\text{-}Hyp)_8$ host peptide. The rankings for both the X and Y positions show the highly stabilizing nature of imino acids and the destabilizing effects of Gly and aromatic residues. Many residues show differing propensities in the X versus Y position, related to the nonequivalence of these positions in terms of interchain interactions and solvent exposure. The propensity of amino acids to adopt a polyproline II-like conformation plays a role in their triple-helix rankings, as shown by a moderate correlation of triple-helix propensity with frequency of occurrence in polyproline II-like regions. The high propensity of ionizable residues in the X position suggests the importance of interchain hydrogen bonding directly or through water to backbone carbonyls or hydroxyprolines. The low propensity of side chains with branching at the C^{δ} in the Y position supports models suggesting these groups block solvent access to backbone C═O groups. These data provide a first step in defining sequence-dependent variations in local triple-helix stability and binding, and are important for a general understanding of side chain interactions in all proteins.

An understanding of how amino acid sequence determines the specific three-dimensional structure of a protein is important for structure prediction and protein design. One approach has been to experimentally quantitate the conformational preferences of individual amino acids, by measuring the tendency to form well-defined secondary structures. Host-guest designs of peptides or proteins, in which a single amino acid is changed in a constant framework, have been used for these studies, following the concept defined in the work of Scheraga (*1*). The tendency of each amino acid to form an α-helix was measured for protein helices (*2*, *3*), peptide models (*4*, *5*), and models of dimer α-helical coiled coil structures (*6*). A thermodynamic scale for the β-sheet forming tendencies was determined by studies on staphylococcal IgG binding protein G (*7*) and a zinc-finger host peptide (*8*). Here, the propensities of amino acids to adopt the collagen-like triple-helical conformation are quantitated using a host-guest model peptide system which has been established in our laboratory (*9*, *10*).

The collagen triple-helix consists of three extended left-handed polyproline II-like helices (PPII),[1] supercoiled about each other in a right-handed manner. A classic PPII-like helix consists of 3 residues/turn with 3.1 Å rise/residue and $\phi = -75°$, $\psi = 145°$ (*11*, *12*). The supercoiling of the PPII helices in collagen modifies these parameters to 2.9 Å rise/residue and 3.33 residue/turn, with similar ϕ, ψ angles (*13*−*15*). The close packing of three chains in the triple-helix requires every third residue to be Gly, resulting in a repeating $(Gly\text{-}X\text{-}Y)_n$ amino acid sequence. Replacement of the Gly by any other residue has a highly destabilizing effect (*16*). The X and Y positions can accommodate any amino acid, but in order to form a stable triple-helix, a significant fraction of these positions (typically about 20%) must be occupied by imino acids, to stabilize the extended nature of the individual chains. Proline residues are incorporated into both the X and Y positions during biosynthesis, and this is followed by enzymatic posttranslational hydroxylation of prolines in the Y positions to form hydroxyproline (Hyp). Gly-Pro-Hyp is the most stabilizing tripeptide unit present in collagen, and also represents the most common sequence. The three chains within the triple-helix are linked together by NH···CO hydrogen bonds and a water network that may play a role in bridging unsatisfied backbone carbonyls and hydroxyl groups of Hyp (*17*, *18*).

A greater appreciation of the broader biological importance of the triple-helix motif has emerged as its presence has been documented in an increasing number of proteins with a

* To whom correspondence should be addressed. Phone: (732) 235-4397. Fax: (732) 235-4783. E-mail: brodsky@umdnj.edu.

† This work was supported by Grants from NIH (GM60048 to B.B.), the Children's Brittle Bone Foundation (to B.B.), the National Science Foundation U.S.-Australia International Cooperative Research (to B.B.), and the Australian Technology Diffusion Program (to J.A.M.R.).

‡ Robert Wood Johnson Medical School.

§ CSIRO.

[1] Abbreviations: CD, circular dichroism; T_m, melting temperature; PPII, polyproline II; standard single letter and three letter codes have been used to denote common amino acids; hydroxyproline is denoted by O (single letter code) and Hyp (three letter code).

10.1021/bi001560d CCC: $19.00
Published on Web 11/03/2000

diverse range of binding activities. The PPII-like conformation is a secondary structure motif observed at low levels in many globular proteins (*19*, *20*), while the supercoiled collagen triple-helix conformation has a more specialized, but still widespread distribution. The triple-helix is the defining characteristic of all types of collagens (*21*−*23*). This includes the family of five fibril forming collagens and at least 14 types of nonfibrillar collagens (*21*, *24*). In addition, the collagen triple-helix is found as a domain required for self-association and ligand binding in proteins involved in host-defense (*23*, *25*).

The repeating unit in the collagen helix is the tripeptide Gly-X-Y sequence. Each of the three positions in this repeating unit has a distinctive environment. The Gly residues are buried in the middle, while the side chains of the X and Y residues are substantially exposed to solvent (*26*). The amide group of Gly is highly protected, as is the amide group of the X position (when not occupied by Pro), in contrast to the accessibility to bulk solvent of the amide group of a nonimino acid in the Y position (*27*). The Gly position is invariant, so determination of functional properties of the triple-helix, such as self-association and ligand binding, must relate to the identity of residues in the X and Y positions. Previous attempts to measure the tendency of different X and Y residues to stabilize the triple-helix have used various repeating polytripeptides and covalently linked trimers (*28*, *29*). These data, together with the frequency of occurrence of triplets, led to a proposed grouping of Gly-X-Y triplets that was used to predict local triple-helix stability (*30*, *31*).

The goal of the present study is to use a systematic host-guest peptide approach to establish a propensity scale for amino acids in the X and Y positions, introducing a single variable guest triplet embedded in the middle of a stabilizing Gly-Pro-Hyp repeating sequence. Previous studies have reported the effects of some nonpolar and charged residues on triple-helix stability (*9*, *10*). Here, characterization of additional peptides of the form $(Gly\text{-}Pro\text{-}Hyp)_3\text{-}Gly\text{-}X\text{-}Hyp\text{-}(Gly\text{-}Pro\text{-}Hyp)_4$ and $(Gly\text{-}Pro\text{-}Hyp)_3\text{-}Gly\text{-}Pro\text{-}Y\text{-}(Gly\text{-}Pro\text{-}Hyp)_4$ are used to determine the triple-helix propensity for all 20 possible amino acids in the X and Y positions. These data provide information on side chain interactions within the triple-helix and allow a comparison with propensities for other protein motifs.

MATERIALS AND METHODS

Peptide Synthesis and Purification. Peptides were synthesized by solid-phase chemistry on either an Applied Biosystems 430A synthesizer with the standard FastMoc method on Fmoc-RINK resin, or a PerSeptive Pioneer Peptide Synthesis System using Fmoc chemistry with an Fmoc-PAL-PEG-PS resin and 2% DBU/2% Piperidine as the Fmoc removal solution (*32*). On both instruments, side-chain protection was *tert*-butyl for the Hyp, Ser, Thr, and Tyr; *tert*-butyl ester for Asp and Glu; benzyloxycarbonyl for Lys, Trp; trityl for Cys, Asn and Gln; and pentamethylchromansulfonyl for Arg. Acetylation was performed by acetic anhydride and triethylamine in dimethylformamide. Then peptides were purified to >90% purity using a SHIMADZU reversed-phase HPLC system on a C-18 column, eluted in 0.1% trifluoroacetic acid with a binary gradient of 0 to 40% (v/v) water/acetonitrile. Laser desorption mass spectrometry (MALDI) confirmed the peptide identity. Peptides were dried in vacuo for 48 h prior to weighing. The preparation of some of the peptides has been previously reported (*9*, *10*, *33*).

Circular Dichroism Spectroscopy. CD measurements were made on an Aviv model 62DS spectrometer. Peptide solutions of concentrations 1 mg/mL in PBS buffer (0.15 M NaCl, 10 mM sodium phosphate, pH 7.0) were used, with peptides equilibrated at 5 °C for 48 h prior to analysis. For equilibrium-melting experiments, the ellipticity at 225 nm was monitored while the sample temperature was increased from 0 to 80 °C with an average heating rate of 0.1 °C/min. The melting curves were analyzed using a two-state reversible model:

$$T_3 \underset{k_2}{\overset{k_1}{\rightleftharpoons}} 3M \tag{1}$$

where T_3 is the triple-helical trimer state and M is the monomer state. Fraction folded was calculated from CD melting curves as a ratio

$$F(T) = \frac{\theta(T) - \theta_M(T)}{\theta_T(T) - \theta_M(T)} \tag{2}$$

where θ is the observed ellipticity and θ_T and θ_M are the ellipticities for the native and monomer forms respectively at temperature T. The equilibrium constant was calculated from the fraction folded temperature dependence as

$$K(T) = \frac{k_1}{k_2} = \frac{{c_M}^3}{c_T} = \frac{3{c_o}^2[1 - F(T)]^3}{F(T)} \tag{3}$$

where c_o is a molar peptide concentration (per mole of peptide chain), and was fitted by the function given by Engel et al. (*34*):

$$K(T) = \exp\left[\frac{\Delta H^\circ}{RT}\left(\frac{T}{T_m} - 1\right) - \ln\left(\frac{3{c_o}^2}{4}\right)\right] \tag{4}$$

Melting temperatures (T_m) and van't Hoff enthalpies (ΔH°) at this temperature were obtained from the best fitting of the experimental data (*3*) by eq 4. From repeated experiments on independently prepared samples, the error of determination of the melting temperature is estimated to be less than ±0.4 °C.

Calculation of Occurrences of Common Amino Acid Residues in Collagen Sequences. The amino acid sequences of fibril-forming collagens were obtained from SWISS-PROT annotated protein sequence database (http://www.expasy.ch/sprot/, May 17, 2000). The sequences of chains of collagens type I (α1 and α2), type II (α1), type III (α1), type V (α1, α2, α3), and type XI (α1 and α2) were used. Only the sequences in which every third amino acid residue is Gly were used for the calculations, giving a total of 3046 triplets. The frequency of occurrence of a given residue in the X (or Y) position is calculated as the number of times it occurs in the X (or Y) position divided by the total number of triplets.

Molecular Modeling. Molecular modeling was performed using the SYBYL 6.1 (Tripos Inc., St. Louis, MO) molecular modeling package. Coordinates for the collagen-like peptide $(Pro\text{-}Hyp\text{-}Gly)_4\text{-}Pro\text{-}Hyp\text{-}Ala\text{-}(Pro\text{-}Hyp\text{-}Gly)_5$ (*15*) were used

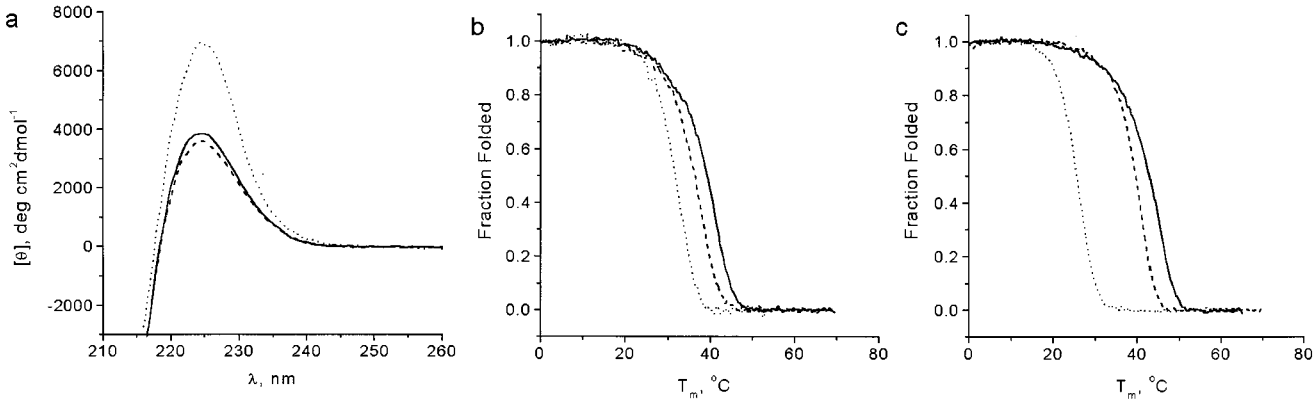

FIGURE 1: (a) CD spectra at 5 °C in PBS (pH 7) of host-guest peptides with guest triplets GPM (—), GPT (- - -) and GPW (•••). (b) Thermal transition curves for host-guest peptides, with GMO (—), GTO (- - -), and GWO (•••) guest triplets, showing the fraction folded versus temperature. (c) Thermal transitions for host-guest peptides, with GPM (—), GPT (- - -), and GPW (•••) guest triplets, showing the fraction folded versus temperature.

to obtain a model for the Ac-(Gly-Pro-Hyp)$_8$-Gly-Gly-CONH$_2$ host structure, which showed a regular triple-helix after energy minimization. Single replacements by guest X or Y residues were introduced into the central triplet of this structure, yielding Ac-(Gly-Pro-Hyp)$_3$-Gly-X-Hyp-(Gly-Pro-Hyp)$_4$-Gly-Gly-CONH$_2$ or Ac-(Gly-Pro-Hyp)$_3$-Gly-Pro-Y-(Gly-Pro-Hyp)$_4$-Gly-Gly-CONH$_2$ sequences. The structures were refined by energy minimization.

RESULTS

Host-Guest Peptide Design. The host peptide used in these studies, Ac-(Gly-Pro-Hyp)$_8$-Gly-Gly-CONH$_2$, consists of eight repeats of stabilizing Gly-Pro-Hyp triplets, with acetylation of the N-terminus and amidation of the C-terminus to eliminate destabilizing charge repulsion (*35*) and a C-terminal Gly-Gly sequence to eliminate the likelihood of diketopiperazine formation during synthesis. As previously reported, the length was selected to ensure formation of a stable triple-helix and yet be short enough so that the effect of a single guest triplet would not be masked by the constant part of structure (*9*, *10*, *33*). A guest residue is introduced in the central triplet of the host peptide, Ac-(Gly-Pro-Hyp)$_3$-Gly-X-Y-(Gly-Pro-Hyp)$_4$-Gly-Gly-CONH$_2$, at either the X or Y position indicated.

Effect of Residues in X Position. To obtain propensities, a set of peptides substituting 19 different amino acids for Pro in the X position of Ac-(Gly-Pro-Hyp)$_3$-Gly-X-Hyp-(Gly-Pro-Hyp)$_4$-Gly-Gly-CONH$_2$ was assembled. Studies on some of these peptides were previously reported by our laboratory (*9*, *10*, *16*, *33*, *36*). All of the host-guest peptides of this form have a circular dichroism spectrum at low temperature with a maximum near 225 nm, which is characteristic of a collagen triple-helix (Figure 1a). The magnitude of the mean residue ellipticity at the maximum varied, depending on the guest triplet, with highest values seen for the aromatic residues. Comparison of the thermal stabilities, as measured by CD spectroscopy, was used to obtain information about the triple-helix propensities of different amino acids. The 225 nm maximum changed with temperature, showing a sharp transition between triple-helical and monomer states (Figure 1b). The enthalpy values were calculated from the thermal transition curves (Table 1).

The host peptide forms a triple-helix with a melting temperature (T_m) of 47.3 °C, while replacing Pro in the guest X position by any other residue drops the stability, with the lowest T_m = 31.9 °C for X = Trp (Table 1). Pro is clearly the most stabilizing residue in the X position. It is striking that X = Pro has one of the lowest enthalpic contributions, supporting the concept that its high stability is of entropic origin. All charged residues (Glu, Lys, Arg, and Asp) are among the most stable peptides, with Ala and Gln also falling into this category. The peptides in this group all have higher enthalpic contributions than Pro, which could reflect side-chain interactions with available backbone carbonyl groups or solvent. Next lower in stability is a group including the nonpolar residues Leu, Val, Met, and Ile, along with Asn and Ser. Following down the stability scale, His, Thr, and Cys confer less stability, and the most destabilizing residues include the aromatic residues and Gly. There is no consistent pattern in the enthalpic values, with some unstable peptides, e.g., X = Tyr having a high enthalpy value, while more stable

Table 1: Melting Temperatures and Enthalpies of Host-Guest Peptides with All Common Amino Acids in the X Position, Together with Their Frequency of Occurrence in Fibril Forming Collagens[a]

Gly-X-Hyp	T_m (°C)	$\Delta H°$ (kJ/mol)	occurrence (%)
Pro	47.3	435	32.9
Glu[b]	42.9	590	13.0
Ala	41.7	480	11.1
Lys[b]	41.5	540	3.6
Arg[b]	40.6	520	2.8
Gln	40.4	565	2.9
Asp[b]	40.1	520	4.9
Leu[c]	39.0	437	7.8
Val	38.9	518	2.6
Met	38.6	452	0.9
Ile	38.4	624	2.0
Asn	38.3	502	2.1
Ser[d]	38.0	506	4.9
His	36.5	580	1.6
Thr	36.2	506	1.8
Cys	36.1	423	0.0
Tyr	34.3	629	0.5
Phe[c]	33.5	514	3.0
Gly[e]	33.2	575	1.6
Trp	31.9	593	0.0

[a] The amino acids are listed in the order of their stability. Human types I, II, III, V, and XI collagens were used for this calculation. The fibril forming collagens were selected because of their common form and function. [b] Ref *10*. [c] Ref *9*. [d] Ref *16*. [e] Ref *36*.

Table 2: Melting Temperatures and Enthalpies of Host-Guest Peptides with All Common Amino Acids in the Y Position, Together with Their Frequency of Occurrence in Fibril Forming Collagens[a]

Gly-Pro-Y	T_m (°C)	$\Delta H°$ (kJ/mol)	occurrence (%)
Hyp	47.3	435	34.0
Arg	47.2	610	11.4
Met	42.6	436	9.0
Ile	41.5	559	2.1
Gln	41.3	559	6.9
Ala	40.9	502	10.6
Val	40.0	481	4.3
Glu[b]	39.7	630	2.0
Thr	39.7	647	4.2
Cys	37.7	471	0.0
Lys[b]	36.8	400	9.0
His	35.7	497	0.5
Ser[c]	35.0	435	4.0
Asp	34.0	776	4.8
Gly	32.7	665	0.7
Leu[d]	31.7	514	1.7
Asn	30.3	640	2.1
Tyr	30.2	657	0.0
Phe[d]	28.3	557	0.2
Trp	26.1	670	0.0

[a] The amino acids are listed in the order of their stabilities for the Y position, which differs from the order seen in Table 1 for the X position. [b] Ref (*10*). [c] Ref (*16*). [d] Ref (*9*).

peptides such as X = Leu, having relatively low enthalpies.

Effect of Residues in Y Position. Another set of 20 peptides containing the host residue, Hyp, and all 19 other amino acids in the Y position was also characterized in terms of CD spectra and thermal stability. Studies on some of these peptides were previously reported by our laboratory (*9*, *10*, *16*, *33*). All peptides show the characteristic triple-helix maximum and have varying stabilities depending on the identity of the guest Y residue. The host peptide with Y = Hyp has T_m = 47.3 °C, and a similar stability (T_m = 47.2 °C) is found when Y = Arg (Table 2). The greater enthalpy of the peptide with Y = Arg suggests a difference in the mechanism of stabilization. The Y = Met peptide drops the stability by 4.7 °C compared to Y = Hyp, while the other amino acids cause greater decreases in thermal stability (Table 2), showing a continuous scale ranging down to the lowest value T_m = 26.1 °C for Y = Trp. The least stable group includes Gly, Leu, Asn, Tyr, Phe, and Trp. While aromatic residues appear to be particularly destabilizing, the charged, nonpolar, and other residues are mixed throughout the stability range.

Comparison of the Stability Scales for the X and Y Positions. The Gibbs energy change upon unfolding (ΔG) is usually used as a measure of stability of a macromolecule (*37*). Using the two-state model, this energy could be estimated as $\Delta G(T) = -RT \ln K(T)$. The Gibbs energy change is proportional to the melting temperature (T_m) for our peptides (unpublished results). Although the Gibbs energy calculated from the thermal transition curves could be used to construct a propensity scale, as in previous publications (*7*), T_m values were chosen in this case because they are directly measurable from the unfolding experiment with a low error.

The relative stability scale for the Y position shows some correlation with that for the X position, with a correlation coefficient of R = 0.74 (Table 3). Common features of both the X and Y scales are the maximal stabilization by X = Pro and Y = Hyp, and the extreme destabilizing effect of aromatic residues and Gly. However, significant differences are observed between the two scales. These are likely to be related to the nonequivalence of the X and Y sites, in terms of interchain interactions and solvent exposure. The range of thermal stabilities in the Y position is broader (21 °C) than for the X position (15 °C), and there are clusters of residues with similar propensities for the X position, while the thermal stabilities for residues in the Y position shows a more continuous distribution. For most residues, the tendency to stabilize the triple-helix in the X position in a Gly-X-Hyp triplet is not the same as in the Y position of a Gly-Pro-Y triplet (Figure 2). Eight of the residues (Glu, Lys Asp, Asn, Leu, Phe, Tyr, and Trp) show a T_m that is higher by about 4−6 °C when in the X position than the same residue in the Y position. Some residues show similar stability in both positions (Gln, Val, Ala, His, Cys, Gly, and Ser). Only four residues (Arg, Met, Ile, and Thr) are more favorable in the Y position than in the X position.

Table 3: Comparison of the Rank Ordering of Amino Acids Measured by Thermal Stability in the X (TH-X) and Y Positions (TH-Y) of the Triple Helix with Their Rankings for PPII-Like Conformation, β-Sheet, α-Helix, and Coiled Coil α-Helix[a]

	TH-X	TH−Y	PP II[b]	β-sheet[c]	α-helix[d]	coiled coil[e]
correlation with TH-X	1.00	0.74	0.68	−0.59	−0.43	−0.45
correlation with TH−Y	0.74	1.00	0.52	−0.38	−0.29	−0.33
	Pro	Hyp	Pro	Tyr	Ala	Ala
	Glu	Arg	Gln	Thr	Leu	Arg
	Ala	Met	Arg	Ile	Met	Lys
	Lys	Ile	Lys	Phe	Ile	Leu
	Arg	Gln	Thr	Trp	Gln	Met
	Gln	Ala	Leu	Val	Arg	Trp
	Asp	Val	Asp	Ser	Lys	Phe
	Leu	Glu	Met	Met	Tyr	Ser
	Val	Thr	Ala	Cys	Val	Gln
	Met	Cys	Cys	Leu	Phe	Glu
	Ile	Lys	Val	Arg	Trp	Cys
	Asn	His	Glu	Asn	His	Ile
	Ser	Ser	Asn	His	Thr	Tyr
	His	Asp	Phe	Gln	Glu	Asp
	Thr	Gly	Ser	Lys	Ser	Val
	Cys	Leu	Ile	Glu	Asp	Thr
	Tyr	Asn	Trp	Ala	Cys	Asn
	Phe	Tyr	Tyr	Asp	Asn	His
	Gly	Phe	His	Gly	Gly	Gly
	Trp	Trp	Gly	Pro	Pro	Pro

[a] At the top is shown the correlation of each ranking with that of the X position of the triple-helix and that of the Y position of the triple-helix [b] Frequency of occurrence of amino acids in regions with PPII-like parameters in known globular protein structures. (*20*) [c] Amino acid propensities for β-sheet (*7*). [d] Propensity scale for α-helix determined by Blaber et al. (*2*). [e] Propensities for α-helix as determined on α coiled coil dimer system (*6*).

Comparison of Stability Scale with Frequency of Occurrence in Fibrillar Collagen Sequences. The propensities of all amino acid residues in the X and Y positions to form triple-helices, as measured by stabilities of host-guest peptides, were compared with the number of occurrences of individual residues in the X and Y positions in the family of human fibril forming collagens (types I, II, III, V, and XI). Fibrillar collagen sequences have a high content of imino

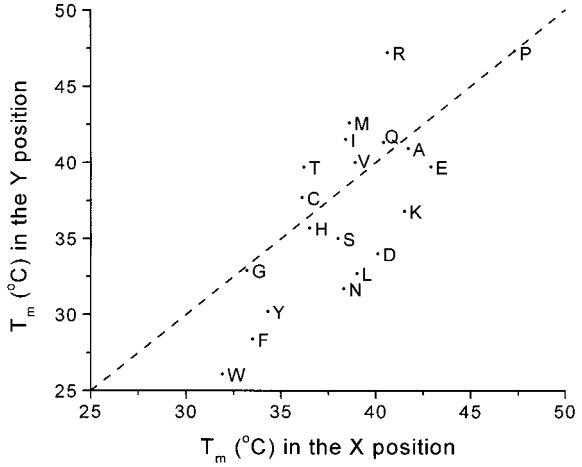

FIGURE 2: T_m values of all 20 amino acids as guests in the Y position of Gly-Pro-Y triplets are plotted against the T_m values of the same amino acid in the X position of Gly-X-Hyp guest triplets. The dashed line has a slope of 1, and represents the location expected if the same values of T_m were found for X and Y positions.

acids in the X position (Pro) and the Y position (Hyp), while Ala occurs frequently in both sites (~11%). Most other residues are found at very low frequency (<3%) in collagens, and no Cys and Trp residues are observed. Some residues show strong preferences for either the X or Y position. For example, 121 out of 140 Glu are in the X position, while 94 out of 127 Arg and 125 of 160 Lys are in the Y position (*38*).

The occurrences of amino acids in the X positions in collagen sequences are compared with the relative thermal stabilities of the peptide set with 20 different residues in the X position of the Gly-X-Hyp guest triplet (Table 1). Triplets of the form Gly-X-Hyp are a reasonable choice as guests, since these represent the most common sequences; typically, Hyp is found in the Y position for at least one-third of all triplets in collagen. A logarithmic plot of the occurrence in the X position vs thermal stability of host-guest peptides (Figure 3a) shows a moderately good correlation ($R^2 = 0.92$). A linear fit shows a lower correlation. The most stabilizing residue of the host-guest peptide set, Pro, is also the most frequent residue, constituting ~33% of the X positions in collagen. Glu and Ala, the next most stabilizing X guest residues, have the second and third highest occurrences in the X position (13 and 11%, respectively). There is also some similarity between the most destabilizing residues and residues which are rare or absent in collagen. For example, Trp is the most destabilizing residue in the host-guest series and never occurs in fibrillar collagens (Table 1).

In a similar manner, the occurrence of amino acids in the Y positions in fibrillar collagen sequences are compared with the relative thermal stabilities of the peptide set with 20 different residues in the Y position of the Gly-Pro-Y guest triplet (Table 2). Again, a best fit was obtained with a logarithmic plot of the occurrence in the Y position vs thermal stability of host-guest peptides ($R^2 = 0.85$) (Figure 3b). The most stabilizing and the most frequent residue in the Y position is Hyp, analogous to Pro in the X position. Arg, which shows a T_m similar to Hyp in the Y position (*33*), is the second most frequent Y position residue (11%), but is still much less common than Hyp (34%). The next most stable guest residue, Met, also shows a high frequency

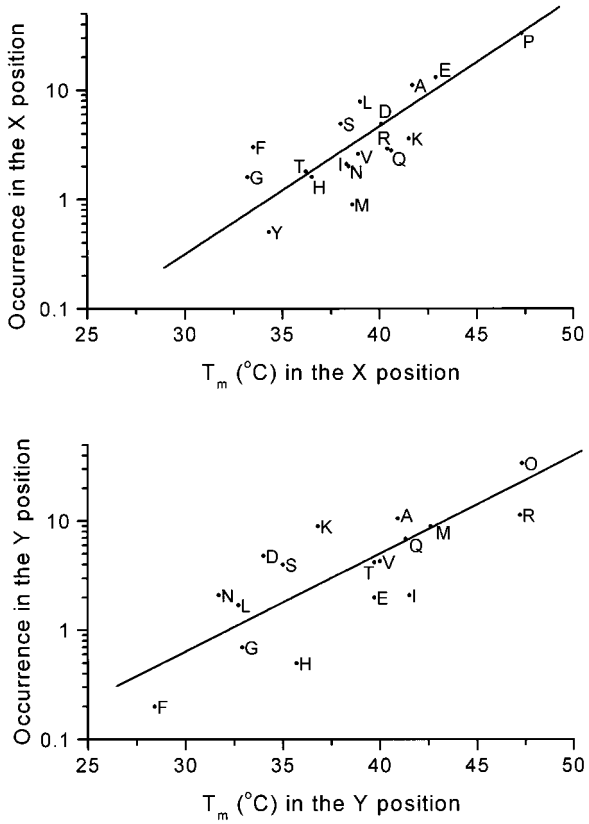

FIGURE 3: (a) Frequency of occurrence (%) of each residue in the X position of all fibril forming collagens is plotted against the T_m values observed for peptides with GXO guest triplets. The correlation is $R^2 = 0.92$, and the best fit line is shown. (b) The frequency (%) of each residue in the Y position of all fibril forming collagens is plotted against the T_m values observed for peptides with GPY guest triplets. The correlation is $R^2 = 0.85$, and the best fit line is shown. Because of the logarithmic scale, residues with zero occurrence are not displayed.

(9%). The three aromatic residues have practically zero occurrence in the Y position, which correlates with their being at the bottom of the stability scale.

DISCUSSION

The host-guest triple-helical peptide system described here has many features that make it appropriate for establishing a propensity scale. These peptides show a sharp transition from a triple-helical trimer to an unfolded monomer state, consistent with a two-state model. The melting temperature derived from this transition is an accurate measure of stability and propensity. All peptides show characteristic triple-helical CD spectra, and it is likely that all of the host-guest peptides adopt the same uniform triple-helical conformation regardless of the substitutions in the X or Y position. Crystal structures of collagen-like peptides indicates that $(Pro\text{-}Hyp\text{-}Gly)_4$-Glu-Lys-Gly-$(Pro\text{-}Hyp\text{-}Gly)_5$, as well as $(Pro\text{-}Hyp\text{-}Gly)_{10}$, adopt a triple-helix with a precise $^7/_2$ superhelical symmetry (*39*, *40*), and this well-defined conformation is expected for all host-guest peptides. This work evaluated the effect of each residue in the most common environment, with an imino acid in the adjacent position. It is likely that the derived propensity scales will be generally valid, since preliminary studies suggest that there is relatively little influence of the identity of the neighboring residue within one Gly-X-Y triplet (unpublished data).

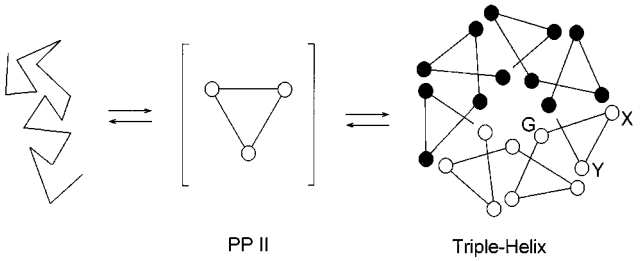

FIGURE 4: A schematic drawing of the folding of monomer chains to the triple-helix with $^7/_2$ symmetry. The PPII-like structure shown in brackets is used to indicate that the triple-helix propensity relates to the frequency in PPII-like structures, even though there is no stable single chain PPII-like intermediate.

In this study, the thermal stabilities of peptides containing different amino acids as guest residues were used to rank the tendency of all 20 amino acids in the X and Y positions to adopt the triple-helical conformation. The scale of stabilities for these two nonequivalent positions can be used to clarify the interactions stabilizing collagen triple-helices. There are several levels at which amino acid side chains could play a role in triple-helix stability (Figure 4): (1) conformational features of the unfolded chain with respect to the native form; (2) stabilization of the extended PPII-nature of individual chains; and (3) interactions between the three chains in one molecule. All of these appear to play a role in determining the observed amino acid propensities for the triple-helix.

The unfavorable entropy of side chains in going from an unfolded to helical form has been useful in explaining the α-helix propensity of different amino acids (*41*), and is likely to play a role in the triple-helix as well. Gly is near the bottom of the triple-helix rankings for both X and Y positions, as well as the α-helix and β-sheet propensity scales (*4*, *7*). This is consistent with the unfavorable entropic consequences of ordering a Gly which has an unordered form with a large amount of accessible ϕ, ψ space (*5*, *6*). The unfavorable nature of Gly in the X and Y positions contrasts with the requirement for Gly as every third residue in the collagen triple-helix, where the unfavorable entropy must be overcome by hydrogen bonding and van der Waals interactions. Other residues may have favorable interactions in the unfolded chain, resulting in a lower stability. For example, the greater stability of Glu and Gln residues in both the X and Y positions relative to Asp and Asn could reflect an entropic factor, similar to that seen in α-helices, which favors the participation of shorter side chains in intrachain interactions in the monomer unfolded form (*4*, *5*).

Because of the PPII-like nature of the individual chains, it is interesting to compare triple-helix propensity with the propensity for PPII formation, but no direct experimental data is available. However, recent studies show about 50% of all known globular protein structures contain at least four to five residues with PPII-like ϕ, ψ angles, and the frequencies of residues found in these PPII-like regions have been collated (*20*, *42*). Comparison of the triple-helix propensity scales for X and Y positions with the frequency of occurrence of residues in PPII-like conformations shows a moderate correlation, in contrast with the lack of correlation seen for α-helix and β-sheet propensity scales (Table 3). This supports the concept that the propensity of a given residue for the triple-helical structure is driven in part by its tendency

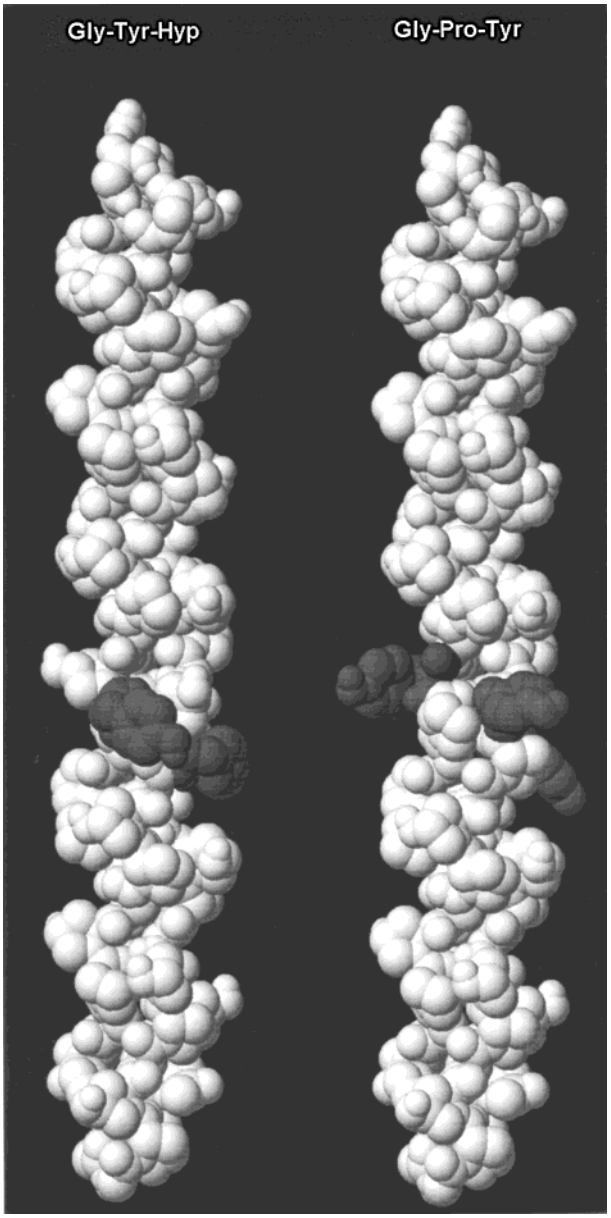

FIGURE 5: Computer models of host-guest peptides with guest triplet Gly-Tyr-Hyp (left) and Gly-Pro-Tyr (right). Both molecules are shown in the same orientation and the Tyr residues are shown in darker shading.

to adopt a PPII-like conformation. In a PPII helix, all residues are fully exposed and all carbonyl groups are available for solvent interactions. However, in the triple-helix, Gly is fully buried, while the side chains of the X and Y residues are largely, but not totally exposed to solvent, and may interact with a neighboring chain (*26*) (Figure 4). The correlation of triple-helix propensity with PPII frequency is better for the X position than the Y (Table 3), perhaps because of its greater exposure to solvent (*26*).

Interactions between the three chains in a supercoiled triple-helix play a dominant role in stability. Information on side chain interactions comes from computational analyses (*43*−*45*) and from recent crystal structures of triple-helical peptides (*40*, *46*). Each guest X or Y residue in the peptide sequence results in the presence of three guest residues within the triple-helix, but there is no interaction between three X residues (or three Y residues) because of their large angular separation (103°) around the triple-helix (Figure 5). In

contrast, a side chain of an X residue can interact with the Y residue of the neighboring chain, or with the available backbone C═O groups, either directly or through water (*40*, *45*, *46*). Such interactions can help explain our observed rankings of host-guest peptide stability in the X and Y positions.

The present study indicates the most stable residues in the Y position are Hyp and Arg, followed by Met. Hyp is known to be the most stabilizing residue in the Y position, while the high stability of Arg is likely to be attributed to its hydrogen bonding to a C═O in a neighboring strand as well as van der Waals interactions with the backbone, as seen in X-ray and computational studies (*40*, *45*, *46*). Computational modeling shows similar favorable nonbonded interactions between the Met side chain and the backbone (*45*). The least stable residues as guests in the Y position are those with branching at the δ carbon, e.g. His, Asp, Leu, Asn, Tyr, Phe, and Trp. Others with branches at the C^{β} are not as unstable, e.g., Val and Thr. Computational analyses show that the occurrence of residues with side chains branching at the C^{δ} prevent the binding of water molecules to the collagen backbone when in the Y position, but not when located in the X position (*43*). This suggests that a mechanism similar to that proposed recently by Luo and Baldwin, where bulky side chains block access to solvent, is operating in the triple-helix as well (*47*).

In the X position, the most stabilizing Pro residue is followed by the ionizable residues, Ala and Gln. The participation of polar side chains in hydrogen bonding directly or through water to available backbone carbonyls or Hyp hydroxyl groups may contribute to their favorable nature (*40*, *46*, *48*). Interactions of ionizable residues with a dipole moment of the peptide backbone, as seen for the α-helix, are not a consideration because of the lack of dipole moment in the triple-helix structure (*49*). Modeling shows that residues with side chains branching at the C^{δ} can be more easily accommodated in the X position than in the Y position, and all C^{δ}-branched residues show a greater propensity when guests are in the X position (Figure 5).

A reasonable correlation was seen between stability of host-guest triple helical peptides and the observed frequency of residues in collagen. The strict amino acid sequence constraint of having Gly as every third residue and the high Pro and Hyp content in the X and Y positions are the dominant features promoting triple-helix formation. Despite these strict sequence and compositional constraints, there appears to be a modest correlation ($R^2 \sim 0.9$) between the frequency of a given amino acid in the collagen sequence and its propensity to adopt this conformation. The observed exponential relationship between frequency in collagen sequences and stability in host-guest peptides suggests a small incremental increase in local stability could be the basis for increased occurrence. The best fit curve for occurrence vs stability was almost indistinguishable for the X and Y residues. This suggested that the frequency of occurrence of a residue is related to triple-helix stability which can be modulated by varying either X or Y residues. Introduction of nonimino acids into the X and Y positions lowers the stability, while providing functional groups for self-association and binding. The sequence of the triple-helix must contain the information to dictate intermolecular association for fibril formation and binding of various integrins and other extracellular matrix components, as well as ensure the formation of a triple-helix of the appropriate thermal stability, close to the upper limit of the body temperature (*17*, *23*). Our results suggest the selection of residues involved in binding and self-association could be influenced by their propensity to stabilize the triple-helix.

Some residues do not fit the general correlation between occurrence and stability. For example, Phe and Gly show a higher frequency of occurrence in the X position than predicted for an exponential fit with the host-guest stability. Phe residues are observed preferentially in the X position and are suggested to interact with other molecules, especially in the overlap region of collagen fibrils (*50*). Thus, their functional interactions may require a larger number of these residues than expected. Gly in the X position, generating triplets of the form Gly-Gly-Y, are rare in all collagens, except type III collagen. These triplets have been shown to result in local destabilization of the triple-helix and have been suggested to play a role in recognition and flexibility of tissues with high type III collagen contents (*36*). In addition, Cys in its reduced form has a moderate stability and yet is never found in fibril forming collagens, except in pathological conditions (*51*). Its absence in collagens is likely to relate to its capacity for disulfide bond formation.

The establishment of X and Y triple-helix propensity scales will help to clarify the forces stabilizing the triple-helix, to predict regions that will form stable triple-helices, and aid in the design of triple-helical peptides and proteins with functional properties. We have observed that individual amino acids have definite conformational preferences for triple-helix formation that can determine global collagen stability. In addition, the experimental establishment of the relative stabilizing effect of different Gly-X-Hyp and Gly-Pro-Y tripeptide sequences confirms that there will be local variations in stability along the collagen sequence. The data provide a firm experimental basis for construction of an algorithm for predicting the relative stability along a collagen molecule, using a methodology such as that proposed by Bächinger and Davis (*31*). Such local changes could relate to energetic changes or modulations in structural parameters such as the pitch of the triple-helix, which could provide a basis for recognition and binding.

ACKNOWLEDGMENT

We thank Mr. Nick Bartone for amino acid analysis, Teresita Silva for assistance with HPLC purification of the peptides, and Mr. Nigel Shenoy for studies on the Gln and Asn containing peptides. We thank Dr. Konrad Beck for the data on peptides with His.

REFERENCES

1. Scheraga, H. (1978) *Pure Appl. Chem. 50*, 315−324.
2. Blaber, M., Zhang, X.-J., and Matthews, B. W. (1993) *Science 260*, 1637−1640.
3. Blaber, M., Zhang, X.-J., Lindstrom, J. D., Pepiot, S. D., Baase, W. A., and Matthews, B. W. (1994) *J. Mol. Biol. 235*, 600−624.
4. Myers, J. K., Pace, C. N., and Scholtz, J. M. (1997) *Biochemistry 36*, 10923−10929.
5. Pace, C. N., and Scholtz, J. M. (1998) *Biophys. J. 75*, 422−427.
6. O'Neil, K. T., and DeGrado, W. F. (1990) *Science 250*, 646−651.

7. Smith, C. K., Withka, J. M., and Regan, L. (1994) *Biochemistry 33*, 5510−5517.
8. Kim, C. A., and Berg, J. M. (1993) *Nature 362*, 267−270.
9. Shah, N. K., Ramshaw, J. A. M., Kirkpatrick, A., Shah, C., and Brodsky, B. (1996) *Biochemistry 35*, 10262−10268.
10. Chan, V. C., Ramshaw, J. A. M., Kirkpatrick, A., Beck, K., and Brodsky, B. (1997) *J. Biol. Chem. 272*, 31441−31446.
11. Cowan, P. M., and McGavin, S. (1955) *Nature 176*, 501−503.
12. Sasisekharan, V. (1959) *Acta Crystallogr. 12*, 897−903.
13. Fraser, R. D. B., and MacRae, T. P. (1973) *Conformation of Fibrous Proteins*, Academic Press, New York.
14. Rich, A., and Crick, F. H. C. (1961) *J. Mol. Biol. 3*, 483−506.
15. Bella J., Eaton, M., Brodsky, B., and Berman, H. M. (1994) *Science 266*, 75−81.
16. Beck, K., Chan, V. C., Shenoy, N., Kirkpatrick, A., Ramshaw, J. A. M., and Brodsky, B. (2000) *Proc. Natl. Acad. Sci. U.S.A. 97*, 4273−4278.
17. Privalov, P. L. (1982) *Adv. Protein Chem. 35*, 1−104.
18. Bella, J., Brodsky, B., and Berman, H. M. (1995) *Structure 3*, 893−906.
19. Adzhubei, A. A., and Sternberg, M. J. E. (1993) *J. Mol. Biol. 229*, 472−493.
20. Stapley, B. J., and Creamer, T. P. (1999) *Protein Sci. 8*, 587−595.
21. Kielty, C. M., Hopkinson, I., and Grant, M. E. (1993) in *Connective Tissue and its Heritable Disorders* (Royce, P. M., and Steinmann, B., Eds.) pp 103−148, John Wiley and Sons, New York.
22. Brown, J. C., and Timpl, R. (1995) *Int. Arch. Allergy Immunol. 107*, 484−490.
23. Brodsky, B., and Ramshaw, J. A. M. (1997) *Matrix Biol. 15*, 545−554.
24. Kadler, K., (1994) *Protein Profile 1*, 519−638.
25. Hoppe, H. J., and Reid, K. B. (1994) *Structure 2*, 1129−1133.
26. Jones, E. Y., and Miller, A. (1991) *J. Mol. Biol. 218*, 209−219.
27. Fan, P., Li, M.-H., Brodsky, B., and Baum, J. (1993) *Biochemistry 32*, 13299−13309.
28. Doyle, B. B., Traub, W., Lorenzi, G. P., and Blout, E. R. (1971) *Biochemistry 10*, 3052−3060.
29. Heidemann, E., and Roth, W. (1982) *Adv. Polym. Sci. 43*, 143−203.
30. Dölz, R., and Heidemann, E. (1986) *Biopolymers 25*, 1069−1080.
31. Bächinger, H. P., and Davis, J. M. (1991) *Int. J. Biol. Macromol. 13*, 152−156.
32. Dettin, M., Pegoraro, S., Rovero, P., Bicciato, S., Bagno, A., and Di Bello, C. (1997) *J. Pept. Res. 49*, 103−111.
33. Yang, W., Chan, V. C., Kirkpatrick, A., Ramshaw, J. A. M., and Brodsky, B. (1997) *J. Biol. Chem. 272*, 28837−28840.
34. Engel, J., Chen, H. T., Prockop, D. J., and Klump, H. (1977) *Biopolymers 16*, 601−622.
35. Venugopal, M. G., Ramshaw, J. A. M., Braswell, E., Zhu, D., and Brodsky, B. (1994) *Biochemistry 33*, 7948−7956.
36. Shah, N. K., Sharma, M., Kirkpatrick, A., Ramshaw, J. A. M., and Brodsky, B. (1997) *Biochemistry 36*, 5878−5883.
37. Marky, L. A., and Breslauer, K. J. (1987) *Biopolymers 26*, 1601−1620.
38. Ramshaw, J. A. M., Shah, N. K., and Brodsky, B. (1998) *J. Struct. Biol. 122*, 86−91.
39. Nagarajan, V., Kamitori, S., and Okuyama, K. (1999) *J. Biochem. 125*, 310−318.
40. Kramer, R. Z., Venugopal, M. G., Bella, J., Mayville, P., Brodsky, B., and Berman, H. M. (2000) *J. Mol. Biol. 301*, 1191−1205.
41. Creamer, T. P., and Rose, G. D. (1994) *Proteins 19*, 85−97.
42. Sreerama, N., and Woody, R. W. (1994) *Biochemistry 33*, 10022−10025.
43. Bansal, M. (1977) *Int. J. Pept. Protein Res.* 9, 224−234.
44. Bansal, M., and Ramachandran, G. N. (1978) *Int. J. Pept. Protein Res. 11*, 73−81.
45. Vitagliano, L., Nemethy, G., Zagari, A., and Scheraga, H. A. (1993) *Biochemistry 32*, 7354−7359.
46. Kramer, R. Z., Bella, J., Mayville, P., Brodsky, B., and Berman, H. M. (1999) *Nat. Struct. Biol. 6*, 454−457.
47. Luo, P., and Baldwin, R. L. (1999) *Proc. Natl. Acad. Sci. U.S.A. 96*, 4930−4935.
48. Emsley, J., Knight, C. G., Farndale, R. W., Barnes, M. J., and Liddington, R. C. (2000) *Cell 101*, 47−56.
49. Wada, A. (1976) *Adv. Biophys. 9*, 1−63.
50. Jones, E. Y., Miller, A., Fraser, R. D. B., Macrae, T. P., and Suzuki, E. (1985) in *Extracellular Matrix: Structure and Function* (Reddi, A. R., Ed.) pp 359−365, A. R. Liss Inc., New York.
51. Ala-Kokko, L., Baldwin, C. T., Moscowitz, R. W., and Prockop, D. J. (1990) *Proc. Natl. Acad. Sci. U.S.A. 87*, 64565-6568.

BI001560D