Prediction of Collagen Stability from Amino Acid Sequence*

Received for publication, February 14, 2005
Published, JBC Papers in Press, March 7, 2005, DOI 10.1074/jbc.M501657200

Anton V. Persikov‡‡, John A. M. Ramshaw‡, and Barbara Brodsky‡‡

From the ‡Department of Biochemistry, University of Medicine and Dentistry of New Jersey-Robert Wood Johnson Medical School, Piscataway, New Jersey 08854 and §Commonwealth Scientific and Industrial Research Organisation, 348 Royal Parade, Parkville, Victoria 3052, Australia

An algorithm was derived to relate the amino acid sequence of a collagen triple helix to its thermal stability. This calculation is based on the triple helical stabilization propensities of individual residues and their intermolecular and intramolecular interactions, as quantitated by melting temperature values of host-guest peptides. Experimental melting temperature values of a number of triple helical peptides of varying length and sequence were successfully predicted by this algorithm. However, predicted $T_m$ values are significantly higher than experimental values when there are strings of oppositely charged residues or concentrations of like charges near the terminus. Application of the algorithm to collagen sequences highlights regions of unusually high or low stability, and these regions often correlate with biologically significant features. The prediction of stability from sequence indicates an understanding of the major forces maintaining this protein motif. The use of highly favorable KGE and KGD sequences is seen to complement the stabilizing effects of imino acids in modulating stability and may become dominant in the collagenous domains of bacterial proteins that lack hydroxyproline. The effect of single amino acid mutations in the X and Y positions can be evaluated with this algorithm. An interactive collagen stability calculator based on this algorithm is available online.

The ability to predict structure and stability from amino acid sequence is an important step in the understanding of basic protein principles and the structural consequences of pathologica l mutations. The vast number of amino acid sequences available from DNA data contrasts with the smaller number of high resolution protein structures and the limited experimental data on protein stability. The ability to make predictions that are in good agreement with experimental data provides insight into the stabilizing interactions within proteins. In addition, there is much interest in computing the effect of single amino acid replacements on protein stability because destabilizing effects are associated with deleterious mutations that result in clinically detectable phenotypes (1–3). In contrast to globular proteins, the relation among sequence, structure, and stability is simpler and better defined for the linear collagen triple helix.

The collagen triple helix motif is found widely in structural proteins of the extracellular matrix and in an increasing set of non-collagenous proteins, many of which are involved in host-defense functions (4, 5). The close packing of three supercoiled polyproline II-like polypeptide chains in the collagen triple helix generates a requirement for Gly as every third residue (6–8). The observation of such a repeating (Gly-X-Y)$_n$ sequence pattern over a stretch of residues signifies a triple helix conformation. However, the collagen triple helix is not uniform in structure or stability. Crystal structures of collagen peptides show that variation in amino acid content leads to small but significant variations in the superhelix twist (9–11). Calorimetric results suggest the presence of multiple independent folding domains along a collagen molecule (12), and the presence of regions of different stability was confirmed by recent studies on recombinant collagen constructs (13). There are multiple binding domains in collagens (14), and regions of decreased triple helix stability have been implicated in binding in some cases (15–17). Self-association of type I collagen into fibrils is preceded by microfolding of specific triple helix regions (18, 19). Thus, specific residues along the (Gly-X-Y)$_n$ sequence determine functionally important modulation of structure and stability.

Experimental thermal stability data obtained from host-guest peptides is integrated here to produce an algorithm for predicting global melting temperatures of collagen triple helical peptides and short fragments and for detecting modulations in relative stability along a collagen chain. Good agreement is observed between predicted and observed stabilities of a number of collagen peptides. In cases in which the predicted $T_m$ is significantly different from that observed, interactions involving longer range electrostatic interactions or unraveling of the ends are suggested. The variations in stability along the collagen chain appear related to known functional sites, and high stability is achieved through a combination of stabilizing imino acid and KGE/D sequences.

**MATERIALS AND METHODS**

The $T_m$ values of all host-guest peptides were measured under a set of standard conditions, with $c = 1$ mg/ml in phosphate-buffered saline, pH 7.0, and with a heating rate average of 0.1 °C/min, as previously reported (20). Small variations are seen at acid versus neutral pH, but all calculations are based on host-guest peptide data collected at pH 7.

The (Pro-Hyp-Gly)$_n$, peptides for $n = 6, 7, 8,$ and 12 were synthesized by Tufts Core Facility (Boston, MA) and purified using high pressure liquid chromatography; their identity was confirmed by matrix-assisted laser desorption ionization.

To extrapolate the dependence of the $T_m$ of the host peptides on peptide length, the experimental values for (Pro-Hyp-Gly)$_n$ and (Pro-Hyp-Gly)$_n$
Pro-Gly)_n versus n, where n is the number of tripeptide units, were fit to the exponential decay function

$$T_m^o = T_m^{max} - A \exp \left( -\frac{n}{n_0} \right)$$  (Eq. 1)

where $T_m^o (n)$ is defined as the base thermal stability of the repeating polytripeptide standard, $T_m^{max}$ is the maximum melting temperature, and the constant $n_0$ represents the length of the repeating peptide with $T_m = 0$.

RESULTS AND DISCUSSION

Experimental Stabilities of Host Guest Peptides

Experimental data on host-guest triple helical peptides, using a (Gly-Pro-Hyp)_n host, have provided information on the propensities of individual residues for the X and Y positions of Gly-X-Y triplets, the interactions within the triple helix for a given Gly-X-Y sequence, and the interactions resulting from neighboring tripeptide sequences (21–23). These data establish the basis for determining the loss of stability that will result from replacing Gly-Pro-Hyp tripeptide sequences by other Gly-X-Y sequences, essentially defining a set of rules for relating amino acid sequence and stability.

Individual Residue Propensities for X and Y Positions—The propensity measurements for all 20 residues in the X position in a Gly-X-Y context and all 20 residues in the Y position in a Gly-Pro-Y context were determined by measuring thermal stability of host-guest peptides (21). The most stable tripeptide unit is Gly-Pro-Hyp ($T_m = 47.3^\circ C$). Replacing Pro in the X position leads to a decrease in stability ranging from 4 °C for Gly-Glu-Hyp ($T_m = 42.9^\circ C$) to 15 °C for Gly-Trp-Hyp ($T_m = 31.9^\circ C$). Replacing Hyp in the Y position leads to a decrease in stability ranging from almost 0 °C for Gly-Pro-Arg ($T_m = 47.2^\circ C$) to 21 °C for Gly-Pro-Hyp ($T_m = 26.1^\circ C$).

Gly-X-Y Tripeptide Sequences—Direct intrachain interactions are not sterically possible between adjacent A and Y residues in the Gly-X-Y unit of a chain, but interchain interactions can take place between the Y residue in one chain and the X residue in an adjacent chain staggered by 1 residue (Fig. 1). Peptides with Gly-X-Y guest triplets were designed to model these interchain interactions. Only a restricted set of possible Gly-X-Y tripeptides are significantly populated in collagens (24), reflecting in part strong preferences for basic residues to be in the Y position and for Glu and hydrophobic residues to be in the X position and very low occurrence of Cys, Trp, and Tyr. A limited set of 41 guest Gly-X-Y sequences was selected to include the most common tripeptide sequences and to model a range of typical electrostatic and hydrophobic interactions. Because of the strong bias in collagen compositions, the selected 41 Gly-X-Y, 19 Gly-X-Hyp, 19 Gly-Pro-Y, and Gly-Pro-Hyp tripeptides cover about 80% of human fibrillar collagen sequences (22). Although Pro residues in the Y position are post-translationally modified to Hyp in multicellular animals, collagenous domains have recently been found in bacteria and viruses where there is no hydroxylation of Pro (25, 26). To model these sequences, Gly-Pro-Pro and Gly-Ala-Pro guest triplets were also included.

A complete table of the stability for all Gly-X-Y triplets was constructed using the experimental values for all frequent sequences and the predicted values for all others (Table I; experimental values are in bold). Predicted values were calculated on the basis of additivity of residues in the X and Y position (22).

$$T_m^{GXY} = T_m^{GXR} + T_m^{GPY} - T_m^{GPY}$$  (Eq. 2)

The predicted values gave good agreement (within ±3 °C) for the GAA and for 28 other guest triplets of the 41 Gly-X-Y triplets studied. The largest deviations were observed for GKD and GRD, which were more stable than predicted by 7 °C, suggesting some interchain electrostatic stabilization. Observed $T_m$ values of GXR sequences (GER, GAR, GKR, GQR, and GDR) were 2.5 °C smaller on average than predicted, indicating the need for an correction factor in a GXR context versus a GPR context.

Interactions between Adjacent Gly-X-Y Units—Interactions between adjacent Gly-X-Y tripeptides were included in the calculations. A recent study reported stabilities of a selection of host-guest peptides including residues in two adjacent tripeptide units, Gly-X-Y-Gly-X-Y’, covering possible direct interchain or intrachain interactions between residues that are separated by ≥3 residues in sequence (23) (Fig. 1). Significant

FIG. 1. Schematic illustration showing the location of X and Y residues within the triple helix and the possibility of intrachain and interchain interactions between residues in adjacent tripeptide units. The bottom panel shows that amino acids separated by ≥3 residues in sequence are unable to interact directly.
The rows of amino acids are listed in order of their X position propensity for triple helix formation, whereas the amino acids in columns are listed in order of their Y position propensity. Both Pro and Hyp (O) are included in the Y position.

### Table II

A list of $\Delta T_m$ corrections for the most significant stabilizing electrostatic and hydrophobic pairwise interactions between residues in adjacent tripeptide units (23)

<table>
<thead>
<tr>
<th>Sequence motif</th>
<th>$\Delta T_m$ (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>G_LGL_</td>
<td>+3.8</td>
</tr>
<tr>
<td>GL_GL_</td>
<td>+7.4</td>
</tr>
<tr>
<td>G_KGD_</td>
<td>+17.5</td>
</tr>
<tr>
<td>G_KGE_</td>
<td>+15.4</td>
</tr>
<tr>
<td>GE_G_K</td>
<td>+5.6</td>
</tr>
<tr>
<td>G_KGE_</td>
<td>+7.3</td>
</tr>
</tbody>
</table>

### Effect of Peptide Length and Blocking Groups on Stability

The methodology for determining the global stability of a peptide (see below) employs subtraction of the relative stability of the tripeptide sequences that make up a peptide from that expected for the repeating Gly-Pro-Hyp sequence of the same length. The $T_m$ value is seen to depend on $n$ (the number of tripeptide units) for repeating (Gly-Pro-Hyp) and (Gly-Pro-Hyp) peptides (27). The sharp dependence of stability on length lengthening with increasing $n$ can be fit to a single exponential decay (Fig. 2).

The effect of blocking groups on peptide stability was also taken into consideration. Studies at different pH values and on peptides with and without blocked termini are consistent with a reduction of stability by about 2 °C when unblocked charged N termini are present and by about 3 °C when unblocked charged C termini are present, for a peptide length of $n = 10$ (29). This destabilization is presumed to be due to repulsion when three charged termini are in close proximity, consistent with the unraveling of the termini observed in high resolution structures of collagen peptides (8, 30). End effects are more pronounced for short peptides than for longer ones, as seen for (Pro-Hyp-Gly), and for (Pro-Hyp-Gly)$_8$ (Fig. 2).

### Algorithm Relating Amino Acid Sequence to Triple Helix Stability

The relative stability of each Gly-X-Y tripeptide compared with Gly-Pro-Hyp and the interaction between adjacent Gly-X-Y tripeptides were used to derive an algorithm for predicting triple helix stability. The $T_m$ values, rather than Gibbs free energy values, were used for calculating peptide stability. The extremely long times needed to reach equilibrium and the lack of agreement of the equilibrium curve with a two-state model presented practical and theoretical limitations to thermodynamic characterization (20). Fortunately, the use of $T_m$ values obtained under standardized conditions has proved to be useful as an empirical measure of triple helix stability (20). Additivity of $T_m$ values was observed for peptides with residues that cannot interact. Thus, $T_m$ values
are seen to be a good measure of relative stability, as long as standard conditions of buffer, pH, and rate of heating are maintained (20). The algorithm predicts a global $T_m$ value for collagen model peptides between 6 and 20 tripeptides in length and predicts a relative stability for collagen sequences.

The global thermal stability of homotrimeric triple helical peptides with length $6 \leq n \leq 20$ is predicted by an algorithm consisting of the following steps.

1) For the total number of triplets $n$ in a given peptide, the base $T_m$ ($n$) for (Pro-Hyp-Gly), or (Pro-Pro-Gly), is calculated from the length dependence (Eq. 1), including any effect of blocking groups.

2) The melting temperature value is decreased for every triplet in the sequence that is not Gly-Pro-Hyp, subtracting a correction value of $\Delta T_m^{*}$ (Table I). The N-terminal and C-terminal tripeptide units are excluded from the calculation due to the staggering of the chains and the reported disorder for the peptide ends (8, 10, 30, 31).

3) The final value for the peptide melting temperature is adjusted using the $\Delta T_m^{corr}$ values for interactions between neighboring tripeptides (Table II).

The algorithm can be formulated as follows.

$$T_m = T_m^0 - \sum_{i=1}^{n-1} \Delta T_m^{*} + \Sigma \Delta T_m^{corr} \quad (Eq. 3)$$

The collagen stability algorithm is available to all users for calculation of global stability of peptides and local stability variations in collagens and collagen-like domains (rwjm, umdju.biochemistry/collagen).

**Prediction of $T_m$ Values for Collagen-like Peptides**

The stability algorithm was applied to 40 synthetic collagen-like peptides whose $T_m$ values have been experimentally determined under the same defined standard conditions (Table III). Most of the peptides are $n = 10$ tripeptide units in length, and some have unblocked ends, whereas others have terminal blocking groups. Excellent agreement was found between the calculated and observed $T_m$ values for peptides with GPO tripeptide units on both ends. For instance, for the unblocked peptide T1–785, the predicted $T_m$ value is $17.1 \pm 0.8$ °C, in close agreement with the observed $T_m$ of $15.9 \pm 0.8$ °C. When KGE or KGD sequences are present, the good agreement is dependent on the inclusion of $\Delta T_m^{corr}$ correction values for interactions between adjacent triplets. For instance, peptide T1–655, which has GPO caps on both ends, has an observed $T_m$ value of $42.8 \pm 0.8$ °C. If each independent triplet is considered, one would subtract $15.5 \pm 0.8$ °C for GPA, $15.7 \pm 0.8$ °C for GPO, and $15.8 \pm 0.8$ °C for GKD, yielding $30.8 \pm 0.8$ °C as the predicted $T_m$ value. However, there is a KGD sequence, which gives $+17.5 \pm 0.8$ °C, and an increase of $5 \pm 0.8$ °C because the ends are blocked, giving a net predicted value of $38.3 \pm 0.8$ °C.
of 42.7 °C, which is very close to the observed value of 42.8 °C. The set of peptides related to T1–892 with GPA sequences on the N-terminal ends also show excellent agreement with predictions. It is notable that the “reverse” peptide, T1–892r, which has the same tripeptide composition but in a different order, has the same \( T_m \) as T1–892, supporting the dependence of thermal stability on tripeptide unit composition when there are no interactions present (Table III) (32).

Applying the algorithm to the test peptide set in Table III, the predicted \( T_m^{\text{pred}} \) values show excellent agreement (within ±2 °C) for 14 peptides and an overall correlation coefficient of \( r = 0.83 \) (Table III, Fig. 3). However, in a number of cases, the predicted values differed from experimental \( T_m \) values by >4 °C. Predicted values are consistently higher than observed ones for peptides containing consecutive strings of positively and negatively charged residues: T7–2058 (GER-GEK-GER-GEQ), T1–904 (GPR-GDK-GET), MBL (GKD-GRD-GTK-GEK), and MSR-1 (GPK-GQK-GEK). This suggests that there are long-range effects in strings of residues of opposite charge such that simple additivity of individual triplets plus KGE/D effects no longer applies. Examination of peptides including highly basic sequences from the heparin binding region of the collagenous tail of the asymmetric form of acetylcholinesterase points to a potential destabilizing effect of charge repulsion at the uncapped N terminus. When these highly basic sequences are included in a host-guest context, with GPO caps at both ends, there is very good agreement between predicted and observed \( T_m \) values. However, when there is an uncapped N terminus, the experimental \( T_m \) values are often lower than those predicted. It is likely that charge repulsion is leading to unraveling at the N terminus and a lower than expected stability. Remaining discrepancies are likely pointing to other effects that have not been taken into consideration in the stability algorithm.

**Calculation of Relative Stability for Collagen-like Domains and Full-length Collagens**

The thermal stability algorithm must be modified for collagens because of their length and the presence of multiple cooperative units during unfolding. Following the averaging approach first set forth previously (33, 34), the algorithm can be applied to discover thermally stable and labile domains along the triple helix. A stability coefficient is assigned for every GXY triplet (Table I) and corrected for the interaction between triplets (Table II). The stability is averaged over a window of 5 tripeptide units, with the average relative stability value for the triplet \( i \) equal to the average of the stability coefficients in the interval \([i − 2, i + 2]\), inclusive. The averaged relative stability values are plotted against the tripeptide number in collagen sequence. For heterotrimer sequences, the values of the three individual collagen chain sequences are averaged.

To illustrate the application of the algorithm to collagens and collagen-like domains, stability profiles were calculated for the type I collagen heterotrimer and the homotrimers of type II collagen, the collagenous domain (ColQ) of the asymmetric form of acetylcholinesterase and the collagen domain of the bacterial protein Scil of *Streptococcus pyogenes* (Fig. 4). The profiles show that the average stability along most of the molecules stays in an intermediate range on a relative scale (Fig. 4), with a small number of local regions of high and low stability. Examination of type I collagen shows the C-terminal region has the highest stability, whereas two regions of low stability are identified as the cross-linking sites KGHR, at residues 87 and 930. For type II collagen, the highest stability peaks are seen at both ends, together with a very strong peak near residue 271, the site of immunodominant T-cell epitope in type II collagen implicated in rheumatoid arthritis (residues 261–273) (35). The bacterial Scil protein (26) shows two peaks at regions rich in KGE/D sequences, and the two heparin binding sites can be located along the stability profile of the ColQ tail of acetylcholinesterase (17). Previously, it has been assumed that regions of high imino acid content will be the most stable, whereas regions deficient in imino acids will be less stable. Examination of the stability profiles shows that regions lacking imino acids often contain...
KGE/D sequences and thus are quite stable. A wide range of KGE/D contents is observed in different collagens, ranging from 3% in type I and II collagens to 10% in type IV collagen in basement membranes (Table IV). The high KGE/D content may provide stability to compensate for the numerous destabilizing interruptions present in type IV collagen. The very high KGE/D content of some bacterial proteins, such as 20% for ScI (26), suggests the importance of electrostatic stabilization when the imino acid content is low and Hyp is absent (Table IV).

The ability to calculate stability profiles from amino acid sequence opens the door to the analysis of collagen mutations. It is possible to determine whether a mutation is occurring in a region of low or high stability because it was suggested that this will affect the clinical consequences (34). In the case of mutations in the X and Y positions, it is also possible to recalculate the profile with the amino acid change to see whether they lead to significant destabilization, as proposed for deleterious mutations in globular proteins. The sites of two mutations in type II collagen, one leading to achondrogenesis-hypochondrogenesis (36) and the other to spondyloepiphysseal dysplasia congenita (37), are both shown to lead to local destabilization (Fig. 5).

**CONCLUSIONS**

Analyses have been carried out to assess propensities of residues to form α-helices, β sheets, and coiled coil α-helices (38–41), but, to the best of our knowledge, this report represents the first case in which it is feasible to use amino acid sequence to predict \( T_m \) values of peptides and to predict stability variations along proteins in a quantitative manner. Calculation of the relationship between amino acid sequence and stability is possible for the collagen triple helix because of (1) its linear nature, which limits interactions to be local, involving residues close in sequence; (2) the small size of the repeating unit Gly-X-Y; and (3) the strong preferential occurrence of a limited number of possible sequence combinations. The algorithm derived from peptide studies gives good predictions for the \( T_m \) values of many collagen-like peptides, suggesting that the important propensities and interactions are well described. The cases in which the agreement is not good point to the need for a better understanding of interactions. These include multiple like and unlike charges and charge repulsion that may unravel ends, as well as factors that were not explored in this study, such as the stabilizing effects of glycosylation of threonine in the Y positions (42, 43).

Application of this algorithm to collagens suggests that regions of unusually high or low stability are likely to be of biological importance. The establishment of the rules of collagen stability highlights the different strategies used for stabilization of the triple helix in bacteria. About 80% of mutations in disease have been found to lead to protein destabilization in globular proteins (1), and destabilization appears to correlate well with collagen disease and severity as well (3, 44). The prediction of the effect of single amino acid replacements in the X or Y positions on collagen stability is now possible, as well as evaluation of the stability of the region in which the mutation occurs. This approach may also be useful in the design of novel triple-helical constructs for production in recombinant systems and applications in biomaterial and tissue engineering structures (45).

**REFERENCES**


**TABLE IV**

<table>
<thead>
<tr>
<th>Protein</th>
<th>Total no.</th>
<th>% of X+Y</th>
<th>% of Y+K</th>
<th>Total no.</th>
<th>% of KGE + KGD</th>
<th>% of triplets</th>
</tr>
</thead>
<tbody>
<tr>
<td>COL1A1</td>
<td>236</td>
<td>35%</td>
<td></td>
<td>6 + 6</td>
<td>3.6%</td>
<td></td>
</tr>
<tr>
<td>COL2A1</td>
<td>224</td>
<td>33%</td>
<td></td>
<td>9 + 3</td>
<td>3.6%</td>
<td></td>
</tr>
<tr>
<td>COL4A1</td>
<td>307</td>
<td>33%</td>
<td></td>
<td>25 + 23</td>
<td>10.3%</td>
<td></td>
</tr>
<tr>
<td>AchE</td>
<td>38</td>
<td>33%</td>
<td></td>
<td>5 + 3</td>
<td>13.8%</td>
<td></td>
</tr>
<tr>
<td>ScII</td>
<td>15</td>
<td>15%</td>
<td></td>
<td>6 + 4</td>
<td>20.4%</td>
<td></td>
</tr>
</tbody>
</table>

**FIG. 5.** Effect of two mutations resulting in cartilage disorders on the calculated relative stability of type II collagen (residues 625–850 are shown). The altered stability as a result of an Arg to Cys replacement at position 704 (X position, leading to a case of achondrogenesis-hypochondrogenesis type II) and at position 789 (Y position, leading to a case of spondyloepiphysseal dysplasia congenita) are indicated by dotted lines (36, 37).
Prediction of Collagen Stability from Amino Acid Sequence
Anton V. Persikov, John A. M. Ramshaw and Barbara Brodsky

doi: 10.1074/jbc.M501657200 originally published online March 7, 2005

Access the most updated version of this article at doi: 10.1074/jbc.M501657200

Alerts:
- When this article is cited
- When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts

This article cites 51 references, 15 of which can be accessed free at
http://www.jbc.org/content/280/19/19343.full.html#ref-list-1