Critical Review

Collagen Structure: The Madras Triple Helix and the Current Scenario

Arnab Bhattacharjee and Manju Bansal
Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India

Summary

This year marks the 50th anniversary of the coiled–coil triple helical structure of collagen, first proposed by Ramachandran’s group from Madras. The structure is unique among the protein secondary structures in that it requires a very specific tripeptide sequence repeat, with glycine being mandatory at every third position and readily accommodates the imino acids proline/hydroxyproline, at the other two positions. The original structure was postulated to be stabilized by two interchain hydrogen bonds, per tripeptide. Subsequent modeling studies suggested that the triple helix is stabilized by one direct inter chain hydrogen bond as well as water mediated hydrogen bonds. The hydroxyproline residues were also implicated to play an important role in stabilizing the collagen fibres. Several high resolution crystal structures of oligopeptides related to collagen have been determined in the last ten years. Stability of synthetic mimics of collagen has also been extensively studied. These have confirmed the essential correctness of the coiled-coil triple helical structure of collagen, as well as the role of water and hydroxyproline residues, but also indicated additional sequence-dependent features. This review discusses some of these recent results and their implications for collagen fiber formation.

INTRODUCTION

The fibrous protein collagen constitutes almost one quarter of the total protein content in most animals, being the major component of several connective tissues such as skin, tendons, ligaments, cartilage, bone, teeth, basement membranes, blood vessels etc. The term 'collagen' is in fact derived from the Greek word for glue and was initially used to describe that constituent of connective tissue which yields gelatin on boiling. However it was soon established that in some tissues, collagen is either heavily cross-linked or covalently bonded to some other stable structure so that it cannot be extracted by just heating. Thus, while the collagen in tendon forms long rope like structures, giving them great tensile strength, the hard rigid structure of bone and teeth arises due to calcification of the interstitial space between the molecules. Cross-links between fibers form flexible two dimensional sheets in skin, while a more complex arrangement in three dimensions is found in cartilage. The ubiquitous nature of this molecule has been confirmed by recent studies wherein it has been found that vertebrates have at least 27 collagen types (described as being type I to XXVII) with 42 distinct polypeptide chains. In addition, more than 20 other proteins have been reported to have collagen-like domains (1, 2). All collagens also possess non-collagenous domains in addition to the fibrous collagen domains. The collagen molecule consists of three polypeptide chains, called z chains, with the characteristic triplet repeat sequence Gly-X-Y and each chain is generally more than 1000 residue long. In some collagens all three chains are identical, while in others, the molecules contain two or even three different z chains, described as z1, z2 and z3, with the difference lying in the amino acids present in X and Y positions of the triplets.

The most striking feature of the collagen molecule is its unique tertiary structure, arising from its characteristic triplet repeat sequence and the presence of a large amount of the imino acid proline, which in about half the case is hydroxylated at its Cγ position. The first essentially correct structure for collagen was proposed 50 years ago by Ramachandran and Kartha and consists of three left handed polypeptide helices, held together by interchain hydrogen bonds (3, 4, 5). The history of how this structure was first established and its current status is worth reviewing, in view of its importance in relation to connective tissue disorders.

ORIGIN AND REFINEMENT OF THE COLLAGEN TRIPLE HELIX

Astbury (6) was the first to suggest a structure for collagen in 1938, which consisted of a mixture of trans and cis peptide units and the same feature was incorporated by Pauling and
Corey in the model proposed by them in 1951 (7), which had three co-axial helices. However neither of these structures were in agreement with the observed X-ray diffraction pattern of collagen fibers. It was Ramachandran’s group from Madras, in India, who first postulated a triple helical structure for collagen, containing only \textit{trans} peptide bonds, as in other natural proteins, in combination with the requirement that the structure should necessarily have one third the total number of residues as glycine (3). An additional requirement was that the structure should be able to accommodate a large proportion of imino acid residues (viz. proline and 4-hydroxyproline) which have their side chains folded back to form rigid five membered rings. This unique triple helical structure consisted of an assembly of three parallel helices, in which the special type of molecular packing is contributed by the occurrence of glycine as every third residue, i.e., a repeating Gly-Xaa-Yaa sequence in each polypeptide chain of the collagen protein. This prototype structure is stabilized by inter-chain hydrogen bonds, unlike the well established \textit{a}-helical structure for polypeptides (not to be confused with the nomenclature of \textit{a}-chains used for the single chains in the collagen triple helix), which is stabilized by intra-chain hydrogen bonds and all the main chain N-H and C = O groups are involved in these type of interactions (8). Each of the three chains in the collagen triple helix forms a left handed helix, with approximately three residues per turn and the three chains are related by a three-fold screw symmetry about a common axis (Fig. 1). Hence, unlike the \textit{a}-helix wherein all residues are in equivalent positions, in the collagen triple helix there can be distinctly different requirements for the three positions in the repeat unit. In particular, every third position, which lies towards the center of the triple helix cannot have any side chain attached to it, since presence of even a \textit{b}-carbon atom (as in alanine) leads to unacceptable inter-chain atomic contacts. Hence this position must necessarily have only glycine residues, thus providing a rational explanation for the unique amino acid composition and triplet repeat sequence (-Gly-X-Y-) of collagen. All other amino acids including the imino acids proline or hydroxyproline could be readily accommodated at the X and Y positions. The torsion angle about the N-C\text{a} bond was ideal for the formation of the five-member pyrroolidine rings in case of Pro and Hyp.

The original triple helical model, proposed by Ramachandran and Kartha in 1954 (3), was soon modified by the authors themselves (4, 5), to fit the more accurate helical parameters observed from x-ray diffraction of stretched collagen fibers (9), which indicated that the number of residues per turn is closer to 3.33, rather than 3. Consequently, they proposed a coiled-coil triple-helical structure for collagen (shown in Fig. 1b), with 10 residues in 3 turns of the left-handed minor helix of each chain. The major helix formed by this single helix is right-handed, accommodates 30 residues per turn and has a pitch of \(85.8\) Å. The neighboring helices in the triple helical assembly are thus related by a twist of -108° and rise of \(2.86\) Å. In this structure the requirement for glycine at every third position was even more stringent and it was also postulated that the imino acids will be preferentially accommodated at the Y position. The triple helix is stabilized by the formation of two inter-chain hydrogen bonds involving the amino groups of

\begin{figure}[h]
\centering
\includegraphics{fig1.png}
\caption{(a) Schematic diagram showing the projection down the helix axis of three left handed, parallel, polypeptide helices in the prototype collagen structure with 3 residues per turn and rise per residue of \(3\) Å (3). The positions 1, 4, 7… in all three chains can only accommodate Gly residues, while all other amino acids including proline can occur at other two positions (b). The modified coiled-coil triple helical structure for collagen with 3.33 residues per turn (5). Each helix undergoes a twist of +36° and a translation of \(8.7\) Å about the common axis relating the three chains. Neighboring chains are related by a rotation of -108° and a translation of \(2.9\) Å (4, 5).
\end{figure}
glycine residues as well as the amino acids at the X position. However, for the two hydrogen bonds to form, the neighboring chains have to approach quite close to each other and some of the atoms in the proposed structure come very close to each other leading to steric clashes, if standard van der Waal’s radii are assumed for the various atoms (10). These clashes can be avoided by a small rotation and translation of the three chains in the triple helix, leading to a structure that retains all the essential features of the Ramachandran coiled-coil triple-helix, but involves only one amino group per tripeptide, that from the glycyl residue, in the formation of inter-chain hydrogen bonds and allows imino acids to be present in both positions X and Y (shown in Fig. 2). Such a model was proposed by Rich and Crick (11), soon after the publication of the Ramachandran and Kartha coiled-coil structure (4, 5) in 1955 and has an inter chain hydrogen bond between the N-H

**Figure 2.** Ball and stick diagrams showing two projections of the currently accepted triple helical structure for collagen with one inter-chain hydrogen bond per tripeptide. The sequence shown is (Gly-Pro-Hyp)$_3$ and each chain in the triple helix has a different color ribbon drawn through the backbone.
group of glycine and C = O group of the amino/imino acid in
the X position. This structure differs only marginally from the
original two bonded coiled-coil triple helical structure, as is
obvious from a comparison of the various parameters
discussed later in this review. Minor variants of this structure
have been subsequently proposed ([12–15], but the essential
features remain invariant. One of the most interesting ideas to
resolve the one vs. two hydrogen bond debate came from
Ramachandran’s group ([16], wherein it was suggested that,
while the triple helix may be stabilized by only one direct
hydrogen bond involving the glycine amino group, additional
inter-chain hydrogen bonds may be formed via water
molecules (shown in Fig. 3a). This hypothesis was further
extended by Ramachandran’s group to suggest that one of
these water molecules could form additional hydrogen bonds
with the hydroxyl group of hydroxyproline residues (as well as
other hydroxyl group containing amino acids, such as serine
and threonine) present at the Y position (also shown in Fig.
3a) of the repeating sequence ([17–19], thus providing an
explanation for the observed correlation between the stability
and hydroxyproline content of various collagens ([20–22].
These fiber-diffraction based models have been shown to be
essentially correct by single crystal X-ray analysis of collagen
related oligopeptides and other biophysical studies.

The individual triple helices or tropocollagen molecules, as
they are sometimes called, are arranged to form fibrils which
are of high tensile strength and flexibility and can be further
assembled and cross-linked (Fig. 4) so as to support stress
efficiently ([23]. Abnormalities in the collagen molecular
structure or its organization into mature fibers lead to different
diseases associated with connective tissues, such as Ethlers-
Danlos syndrome, osteogenesis imperfecta and some types of
osteoporosis and arthritis ([1, 2]. Most common mutations in
the collagen gene are single base substitutions that convert the
codon of the critical glycine residue to that of a bulkier
residue, which causes considerable distortion of the triple helix
or even prevent its formation beyond this point. Amino acid
changes in the other two positions of the triplet have milder
effects. Also since there are regions of high and low stability
within the collagen triple helix and stability within the collagen
triple helix depends on the amino acids present in the other
two positions of the repeating triplet, so mutations in different
regions can have different effects. In addition, only in fibril
forming collagens is a single continuous helix mandatory,
while other collagens that normally contain several interruptions
in the triplet repeat, can readily accommodate additional
disruptions with no major ill effects. Interestingly mutations
that produce some structural alterations in the polypeptide
chain but still allow the chains to assemble into a triple helix,
generally manifest as more severe phenotypes than those that
prevent triplex formation altogether ([1]). This is because the
triple helices containing the mutated chain will have an
abnormal structure, which will affect the formation of higher
order structure or alter their assembly and function. However
the exact relationship between the amino acid sequence change
and its effect on the structure and lethality of a mutation in the
collagen molecule is still not clear. The sequence dependent
variation observed in the structural parameters of collagen
triple helical structure and other biophysical studies on
stability of various collagen mimics can provide useful insights
into how differences in amino acid distribution may affect the
stability and higher order assembly of the triple helical
molecules in collagenous tissues.

COMPARISON OF THE STRUCTURAL PARAMETERS OF
COLLAGEN RELATED OLIGOPETIDES

Several crystal structures of collagen mimics ([24–32] are
now available in the Protein Data Bank and we have
compared them with the various fiber models as well as with
each other, in order to get some insight into sequence
dependent variation in the structure. The details of the various
fiber models and tripeptide fragments from the crystal
structures that have been analyzed here, are given in Table
1. In order to avoid the end effects, only the middle regions of
these oligopeptides have been considered for this analysis. The
Gly→Ala mutant structure (1CAG) and the non imino acid
containing fragments from 1BKV and 1DZI are of particular
interest. In case of the fiber models, co-ordinates of the atoms
in a triplet were obtained from the published literature ([12, 14,
15], and the intra, as well as inter helical twist and rise values
were used to generate the 9 amino acid long triple helices. The
various structural parameters in the crystal structures, helical
twist (t), helical rise (h), number of residues per helical turn
(n), as well as radii of Cz atom cylinders, were obtained using
our in house program HELANAL ([33]) and are listed in Table
2. It is interesting to note that the parameters for the non-
imino acid containing fragments are similar to those for the
fiber models. The phi (\(\phi\)) and psi (\(\psi\)) values of the triple helical
structures have been plotted on a Ramachandran map (Fig. 5)
to get an idea of the variation in backbone torsion angles,
corresponding to the Gly, X and Y positions in the structures,
particularly in case of non Gly-Pro-Hyp type sequences. The
rmsd (Root Mean Square Deviation) values between the
various structures, in the case of a single chain as well as for
the triple chain assembly, when all the heavy atoms in the
backbone were considered, were also calculated. It was found
that the single chain conformation is fairly invariant in all the
structures, except in the case of HMB1, in which one of the
glycine residues in each chain has been replaced by an alanine.
Hence even though an alanine can be accommodated in lieu of
glycine in the single chain of the collagen triple helix, the steric
constraints imposed on an assembly of the three chains, forces
local distortion in individual helices. This is reflected in the
greater rmsd values for this structure as compared to all other
structures as well as larger dispersion of \(\phi\) and \(\psi\) values in this
structure (see Fig. 5). Similarly the small bend in the Integrin
bound collagen (JEMS) structure, following the G-E-R triplet,
Figure 3. (a) Projection down the helix axis of the one hydrogen-bonded fiber model (13), showing details of the direct interchain hydrogen bond, between the glycyl amide group and the carbonyl oxygen of residue at X position in the neighboring chain of a triple helix with Gly-X-Hyp sequence. The modeled water molecules (O^W1 and O^W2) linking the backbone of two neighboring chains, as well as the O^i of Hyp residues at Y position (16, 17) are also shown. (b) Interchain hydrogen bonds observed in the 1BKV crystal structure (29). Only the fragment Hyp 8 to Gly 12 in chain A and corresponding regions Pro 37 to Thr 41 in chain B and Gly 66 to Ile 70 in chain C are shown here. The water mediated hydrogen bonds involving Thr 11 in chain A, Thr 41 in chain B and Hyp 68 in chain C are also shown. The water molecules W1, W2 and W3 are indicated as +. The various atoms are color coded as: carbon-green, nitrogen-blue, oxygen-red and hydrogen-white.
gives rise to slightly higher rmsd values for this triple helix. Interestingly the rmsd between the original two hydrogen bonded model of Ramachandran (GNR1) and the other structures is only marginally higher than that found between the various crystal structures and other fiber models.

It has recently been suggested that the small systematic differences in the $(\phi, \psi)$ values at X and Y positions lead to differences in the preference for the proline ring pucker when the imino acids occurs at these positions (34). This in turn is related to the occurrence of hydroxyproline at Y position, stabilizing the triple helix, since the pyrollidine ring in Hyp prefers only one type of pucker ($C^\gamma$-exo or up conformation). However since hydroxylation of proline occurs before helix formation and the $\phi, \psi$ variation is within the range

Figure 4. The individual triple helices or tropocollagen molecules, are arranged to form fibrils which are of high tensile strength and flexibility and can be further assembled and cross-linked. The figure has been reproduced from Klug, W. S. and Cummings, M. R. (23).
observed for standard deviation values, additional data is required to accept this rationale for the occurrence of hydroxyproline at Y position. Some other experimental results have suggested that inductive effects, due to attachment of an electronegative atom at C\text{g}" are the basis for the contribution of Hyp residues to the stability of collagen (35 – 37). It should however be noted that, the earlier hypothesis on the role of hydroxyproline in stabilizing the triple helical structure through additional water mediated hydrogen bonds (17, 18), has also been recently confirmed (shown in Fig 3b) by the presence of such features in both the structures containing (Gly-X-Hyp) type sequence in their local regions (25, 29) as well as other biophysical studies (38, 39). Similar hydrogen bonds are found to occur (29) in the presence of threonine residues (also shown in Fig 3b). The Thr residues have also been reported to stabilize fibrillar aggregates through glycosylation at their O\text{g}" position (40).

The oligopeptide fragments with mixed sequences show values of helical parameters (particularly twist) that are closer to those suggested from diffraction studies of native collagen fibres (-108°) than the oligomers with (Gly-Pro-Pro) type of repeats (which have twist values of about – 102°). Since even a small variation in twist has implications for the intermolecular interactions between triple helices and their assembly (41), we examined the length distribution of fragments consisting of the four types of triplets, (Gly-X-Y), (Gly-Pro-Y), (Gly-X-Hyp) and (Gly-Pro-Hyp) in the amino acid sequences of five types of collagen chains and the data is shown in Fig. 6. Complete sequences have been reported for these collagen chains, with Type I, II, and III belonging to fibrillar collagen (42 – 45) and Type IV to basement membrane collagen (46). It is clearly seen that imino acid containing triplets are found scattered along the collagen molecule, and rarely occur in stretches of more

<table>
<thead>
<tr>
<th>Serial no.</th>
<th>Structure description used</th>
<th>PDB id</th>
<th>Sequence of the fragment*</th>
<th>rmsd value w.r.t RDBF</th>
<th>Reference nos.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GNR1 Fiber model</td>
<td></td>
<td>(GXPGXPGXP)</td>
<td>0.84</td>
<td>4,12</td>
</tr>
<tr>
<td>2</td>
<td>ARFC Fiber model</td>
<td></td>
<td>(GPOGPOGPO)</td>
<td>0.47</td>
<td>10,11</td>
</tr>
<tr>
<td>3</td>
<td>GNR2 Fiber model</td>
<td></td>
<td>(GPOGPOGPO)</td>
<td>0.66</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>RDBF Fiber model</td>
<td></td>
<td>(GPOGPOGPO)</td>
<td>–</td>
<td>14</td>
</tr>
<tr>
<td>5</td>
<td>HMB1 1CAG</td>
<td></td>
<td>(G12POAPOGPO20)</td>
<td>1.12</td>
<td>24,25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(G42POAPOGPO50)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(G72POAPOGPO80)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>HMB2 1QSU</td>
<td></td>
<td>(G18POPGPOGPO26)</td>
<td>0.93</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(G48POPGPOGPO56)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(G78POPGPOGPO80)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>HMB3 1QSU</td>
<td></td>
<td>(G3POGEKPOGPO17)</td>
<td>0.65</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(G39POGEKPOGPO47)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(G69POGEKPOGPO77)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>HMB4 1BKV</td>
<td></td>
<td>(G3ITGARGLA17)</td>
<td>0.34</td>
<td>28,29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(G39ITGARGLA47)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(G69ITGARGLA77)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>JEMS 1DZI</td>
<td></td>
<td>(G3FOGERGPO13)</td>
<td>1.06</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(G3FOGERGPO15)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(G3FOGERGPO15)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>ZAGA 1K6F</td>
<td></td>
<td>(G3PPGPPGPP11)</td>
<td>0.74</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(G3PPGPPGPP11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(G3PPGPPGPP11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>OKUY 1IIT</td>
<td></td>
<td>(G3PPGPP8)</td>
<td>0.34</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(G3PPGPP8)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* In case of crystal structures 5 – 9, these fragments are middle parts of longer molecules with flanking GPO or GPP sequences.
than two contiguous triplets. Hence the average helical structure is expected to be closer to that seen in the (Gly-X-Y) regions of oligopeptides rather than that observed in the oligopeptides containing (Gly-Pro-Pro/Hyp) triplet repeats, thus validating the earlier fiber models proposed for the collagen molecule.

Interestingly the Type IV collagen that is known to have more inter-triple helix interactions, contains a much higher content of Hyp residues at Y position and a very significant number of them occur as (Gly-X-Hyp) triplets, with no increase in frequencies of Gly-Pro-Hyp triplets. This feature also manifests itself in the higher frequency of occurrence of 3 and 4 mers of (Gly-X-Hyp) repeats, suggesting that hydroxyproline, while stabilizing the triple helical structure, can also play a role in intermolecular interactions, as indicated by model building and crystal structure studies (18, 25, 47). Similar interactions can also occur via other side chains and there are also indications that several amino acids selectively prefer X or Y position in natural collagen chains (37, 48).

### AMINO ACID PREFERENCES IN COLLAGEN CHAINS OF TYPE I-IV

A recent analysis of collagen sequence data has reported the distribution of various triplets using sequence data from an assortment of collagen chains of varying lengths (37, 48, 49). We have carried out a detailed analysis to assess the propensities of the amino acids to occur in X and Y positions considering each of the five collagen chains, for which full length sequences are available (42–46) and the results are shown in Table 3.

It is well known that alanine, which constitutes the most frequently occurring amino acid after glycine and proline/hydroxyproline, shows equal preference for both X and Y positions in the Gly-X-Y triplets, in all types of collagens. Since synthetic polypeptides rich in alanine do not take up collagen like helical structures, it indicates that alanine is well tolerated in collagen structures, but does not provide any additional stability. This is confirmed by the melting temperature studies on host guest peptides wherein Ala occurs in the middle region of the stability scale (37, 49). There is a clear bias for Glu, Leu and Phe in the X position whereas Arg and Lys are seen to be preferred in the Y position of the Gly-X-Y triplets in all the five collagen chains. Interestingly Gln and Thr also show a preference for Y position in Type I and Type II collagens while Met shows a similar preference in Type I, II and IV collagens. The residues showing preference for Y position can help stabilize triple helices as well as assemblies of triple helices through additional interactions. Threonine can mimic the Hyp residue (19, 29) by forming water mediated H-bonds. Arginine in Y position has been shown to form inter chain hydrogen bonds with a carbonyl oxygen of a neighboring chain (29, 30) in the crystal structures HMB4 and JEMS. It has also been reported that replacement of Hyp with Arg in this position results in a triple helix of nearly equal stability, whereas all other amino acid substitutions, including Lys, result in triple helices with lower melting temperatures (37, 49).

### Table 2.

Average values of the structural parameters, for the triple helical structures listed in Table 1. The standard deviation values for the various parameters of the crystal structures are given within parenthesis.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Structures</th>
<th>n</th>
<th>h</th>
<th>t</th>
<th>$r(C^a_G)$</th>
<th>$r(C^a_X)$</th>
<th>$r(C^a_Y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNR1</td>
<td>3.27</td>
<td>2.91</td>
<td>110</td>
<td>1.15</td>
<td>3.5</td>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td>ARFC</td>
<td>3.33</td>
<td>2.86</td>
<td>108</td>
<td>1.6</td>
<td>4.07</td>
<td>3.43</td>
<td></td>
</tr>
<tr>
<td>GNR2</td>
<td>3.27</td>
<td>2.91</td>
<td>110</td>
<td>1.4</td>
<td>3.95</td>
<td>3.6</td>
<td></td>
</tr>
<tr>
<td>RDBF</td>
<td>3.33</td>
<td>2.98</td>
<td>108</td>
<td>1.93</td>
<td>4.06</td>
<td>2.9</td>
<td></td>
</tr>
<tr>
<td>HMB1</td>
<td>3.35 (0.14)</td>
<td>2.77 (0.14)</td>
<td>107.5 (24.1)</td>
<td>2.32 (0.4)</td>
<td>4.53 (0.4)</td>
<td>3.68 (0.5)</td>
<td></td>
</tr>
<tr>
<td>HMB2</td>
<td>3.53 (0.02)</td>
<td>2.82 (0.02)</td>
<td>101.9 (2.9)</td>
<td>1.80 (0.1)</td>
<td>4.10 (0.1)</td>
<td>3.18 (0.1)</td>
<td></td>
</tr>
<tr>
<td>HMB3</td>
<td>3.45 (0.03)</td>
<td>2.85 (0.04)</td>
<td>104.4 (4.4)</td>
<td>1.70 (0.1)</td>
<td>4.08 (0.1)</td>
<td>3.24 (0.04)</td>
<td></td>
</tr>
<tr>
<td>HMB4</td>
<td>3.34 (0.03)</td>
<td>2.91 (0.03)</td>
<td>107.9 (4.6)</td>
<td>1.74 (0.1)</td>
<td>4.06 (0.1)</td>
<td>3.19 (0.1)</td>
<td></td>
</tr>
<tr>
<td>JEMS</td>
<td>3.37 (0.04)</td>
<td>2.86 (0.1)</td>
<td>106.8 (8.1)</td>
<td>1.79 (0.2)</td>
<td>4.07 (0.2)</td>
<td>3.22 (0.1)</td>
<td></td>
</tr>
<tr>
<td>ZAGA</td>
<td>3.45 (0.01)</td>
<td>2.84 (0.03)</td>
<td>104.3 (2.4)</td>
<td>1.84 (0.1)</td>
<td>4.07 (0.1)</td>
<td>3.07 (0.03)</td>
<td></td>
</tr>
<tr>
<td>OKUY</td>
<td>3.50 (0.00)</td>
<td>2.9 (0.01)</td>
<td>102.9 (0.9)</td>
<td>1.85 (0.01)</td>
<td>4.06 (0.04)</td>
<td>3.07 (0.02)</td>
<td></td>
</tr>
</tbody>
</table>

$n$ = Number of residues per turn of the triple helix.

h = Helical rise (in Angstroms) for the triple helix.

$ t$ = Helical twist (in degrees) for the triple helix.

$r(C^a_G)$ = Average radius (in Angstroms) of $C^a$ atom of the Gly residues.

$r(C^a_X)$ = Average radius (in Angstroms) of $C^a$ atoms of the residues in X position of the triple helix.

$r(C^a_Y)$ = Average radius (in Angstroms) of $C^a$ atoms of the residues in Y position of the triple helix.

<table>
<thead>
<tr>
<th>Structures</th>
<th>n</th>
<th>h</th>
<th>t</th>
<th>$r(C^a_G)$</th>
<th>$r(C^a_X)$</th>
<th>$r(C^a_Y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNR1</td>
<td>3.27</td>
<td>2.91</td>
<td>110</td>
<td>1.15</td>
<td>3.5</td>
<td>3.5</td>
</tr>
<tr>
<td>ARFC</td>
<td>3.33</td>
<td>2.86</td>
<td>108</td>
<td>1.6</td>
<td>4.07</td>
<td>3.43</td>
</tr>
<tr>
<td>GNR2</td>
<td>3.27</td>
<td>2.91</td>
<td>110</td>
<td>1.4</td>
<td>3.95</td>
<td>3.6</td>
</tr>
<tr>
<td>RDBF</td>
<td>3.33</td>
<td>2.98</td>
<td>108</td>
<td>1.93</td>
<td>4.06</td>
<td>2.9</td>
</tr>
<tr>
<td>HMB1</td>
<td>3.35 (0.14)</td>
<td>2.77 (0.14)</td>
<td>107.5 (24.1)</td>
<td>2.32 (0.4)</td>
<td>4.53 (0.4)</td>
<td>3.68 (0.5)</td>
</tr>
<tr>
<td>HMB2</td>
<td>3.53 (0.02)</td>
<td>2.82 (0.02)</td>
<td>101.9 (2.9)</td>
<td>1.80 (0.1)</td>
<td>4.10 (0.1)</td>
<td>3.18 (0.1)</td>
</tr>
<tr>
<td>HMB3</td>
<td>3.45 (0.03)</td>
<td>2.85 (0.04)</td>
<td>104.4 (4.4)</td>
<td>1.70 (0.1)</td>
<td>4.08 (0.1)</td>
<td>3.24 (0.04)</td>
</tr>
<tr>
<td>HMB4</td>
<td>3.34 (0.03)</td>
<td>2.91 (0.03)</td>
<td>107.9 (4.6)</td>
<td>1.74 (0.1)</td>
<td>4.06 (0.1)</td>
<td>3.19 (0.1)</td>
</tr>
<tr>
<td>JEMS</td>
<td>3.37 (0.04)</td>
<td>2.86 (0.1)</td>
<td>106.8 (8.1)</td>
<td>1.79 (0.2)</td>
<td>4.07 (0.2)</td>
<td>3.22 (0.1)</td>
</tr>
<tr>
<td>ZAGA</td>
<td>3.45 (0.01)</td>
<td>2.84 (0.03)</td>
<td>104.3 (2.4)</td>
<td>1.84 (0.1)</td>
<td>4.07 (0.1)</td>
<td>3.07 (0.03)</td>
</tr>
<tr>
<td>OKUY</td>
<td>3.50 (0.00)</td>
<td>2.9 (0.01)</td>
<td>102.9 (0.9)</td>
<td>1.85 (0.01)</td>
<td>4.06 (0.04)</td>
<td>3.07 (0.02)</td>
</tr>
</tbody>
</table>
The amino acid preferences seen for X position cannot be readily explained on the basis of any explicit interactions, except for Glu in X position forming a stabilizing ion-pair with Lys/Arg in Y position in a neighboring chain (29, 30, 50). It has been suggested that in the case of Leu and Phe it is a negative preference for Y position, that makes X the position of preference for these residues. However it may be pertinent to point out that both these residues also show a preference to sequester in (Gly-X-Hyp) rather than the general (Gly-X-Y) triplet sequence, perhaps indicating a more specific hydrophobic interaction between the two side chains.

Interestingly, many of the sequence features observed in the collagen sequences can be easily rationalized on the basis of simple stereochemical restrictions on the occurrence of amino acid residues in the triple helical molecules. One such study by our group (19, 51) had suggested that Leu and Phe can be accommodated at both locations X and Y, but when a proline residue occurs at X position, it restricts the freedom of orientation of these side chains at Y position in the neighboring chain. It is also not possible to accommodate the inter-chain hydrogen bond forming water molecules if these large non-polar residues are present at position Y. In contrast at X position, they have much greater freedom of orientation and their presence will also help to shield the hydrogen bonded water molecules as well as the direct interchain hydrogen bonds from disturbance by the solvent medium (19). Hence from the stability point of view, their presence in conjunction with the Hyp residues at Y position could make these triplet sequences almost as favorable as Gly-Pro-Hyp triplets.

Thus interchain interactions appear to play a crucial role in determining the variable local stability of different regions of a collagen molecule (50, 52, 53). They have also

**Figure 5.** Value of the backbone torsion angle phi ($\phi$) is plotted against the psi ($\psi$) value for each residue is the collagen fiber models and oligopeptide crystal structures listed in Table 1. The values corresponding to the three positions in the triplet sequence Gly-X-Y are shown in black, blue and red colors respectively. Each structure is represented by a different symbol, as indicated in top right corner.
Figure 6. Bar diagrams showing the length distribution of the four type of triplet stretches among representative examples of Collagen types I to IV. The amino acid sequences correspond to type I α1 chain from skin tendon and bone of human (N_P000079, 338, 42), type I α2 chain from placenta and skin of human (P08123, 342, 43), type II α1 chain from cartilaginous tissue of house mouse (P28481, 340, 44), type III α1 chain from brain tissue of house mouse (P08121, 343, 45) and type IV α1 chain from basement membrane of human (NP_001836, 375, 46) collagens. The Genbank accession no. of the sequences, the number of amino acid triplets and the corresponding reference number are given above within parenthesis, for each collagen chain.
Preference of various amino acids for the X or Y position in the amino acid sequence of type I – IV collagen chains.

Amino acids which constitute at least 0.5% of the composition in each type of collagen chain are taken into account when reporting the preference of amino acids for the X or Y position in the Gly-X-Y triplets.

Table 3.

<table>
<thead>
<tr>
<th>Amino Acids</th>
<th>Type I alpha 1</th>
<th>Type I alpha 2</th>
<th>Type II alpha 1</th>
<th>Type III alpha 1</th>
<th>Type IV alpha 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>eq</td>
<td>eq</td>
<td>eq</td>
<td>eq</td>
<td>eq</td>
</tr>
<tr>
<td>Arg</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Asn</td>
<td>eq</td>
<td>eq</td>
<td>eq</td>
<td>Y</td>
<td>X</td>
</tr>
<tr>
<td>Asp</td>
<td>eq</td>
<td>eq</td>
<td>eq</td>
<td>eq</td>
<td>Y</td>
</tr>
<tr>
<td>Gln</td>
<td>Y</td>
<td>Y</td>
<td>Y eq</td>
<td>eq</td>
<td>eq</td>
</tr>
<tr>
<td>Glu</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>X</td>
</tr>
<tr>
<td>Ile</td>
<td>eq</td>
<td>eq</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Leu</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>Y</td>
<td>X</td>
</tr>
<tr>
<td>Lys</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Met</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>Phe</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Pro/Hyp</td>
<td>eq</td>
<td>eq</td>
<td>eq</td>
<td>eq</td>
<td>eq</td>
</tr>
<tr>
<td>Ser</td>
<td>eq</td>
<td>eq</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Thr</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>eq</td>
<td>X</td>
</tr>
<tr>
<td>Val</td>
<td>eq</td>
<td>Y</td>
<td>eq</td>
<td>eq</td>
<td>eq</td>
</tr>
</tbody>
</table>

eq: Indicates that the amino acid has equal preference for both X and Y positions.
X: Indicates that the amino acid has preference for X position (> 60%), those showing strong preference (> 80%) are underlined.
Y: Indicates that the amino acid has preference for Y position (> 60%), those showing strong preference (> 80%) are underlined.

been similarly implicated in a recent study, wherein it was found that a shift in the register of collagen chains with respect to each other in the triple helix, without interrupting the triplet sequence, has profound influence on the triple helix stability and fibril formation (54, 55). This could explain the rare type of collagen mutation in which a duplication or deletion of one or two Gly-X-Y triplet occurs, leading to severe clinical consequences (55). Thus it appears that even after 50 years, there is still a lot more we need to understand about the finer features of the collagen structure and get a complete picture of the conformational micro heterogeneity responsible for collagen fiber formation as well as selective recognition of collagen molecules by other proteins.

ACKNOWLEDGEMENTS

The authors acknowledge the financial support received from Department of Biotechnology, Govt. of India.

REFERENCES