IUBMB *Life*, 57(3): 161 – 172, March 2005

Taylor & Francis
Taylor & Francis Group

## *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.**

**IUBMB *Life*, 57: 161 – 172, 2005**

**Keywords** Collagen structure; triple helix; hydroxyproline; amino acid propensity.

## 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 $\alpha$ 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 $\alpha$ chains, described as $\alpha1$, $\alpha2$ and $\alpha3$, 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 $C^{\gamma}$ 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

Received 5 January 2005; accepted 23 February 2005

Address correspondence to: Manju Bansal, Molecular Biophysics Unit, Indian Institute of Science, Bangalore-560012, India.
Tel: 91 80 2293 2534. Fax: 91 80 2360 0535. E-mail: mb@mbu.iisc.ernet.in

ISSN 1521-6543 print/ISSN 1521-6551 online © 2005 IUBMB
DOI: 10.1080/15216540500090710

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 *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 $\alpha$-helical structure for polypeptides (not to be confused with the nomenclature of $\alpha$-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 $\alpha$-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 $\beta$-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$^{\alpha}$ bond was ideal for the formation of the five-member pyrollidine rings in case of Pro and Hyp.

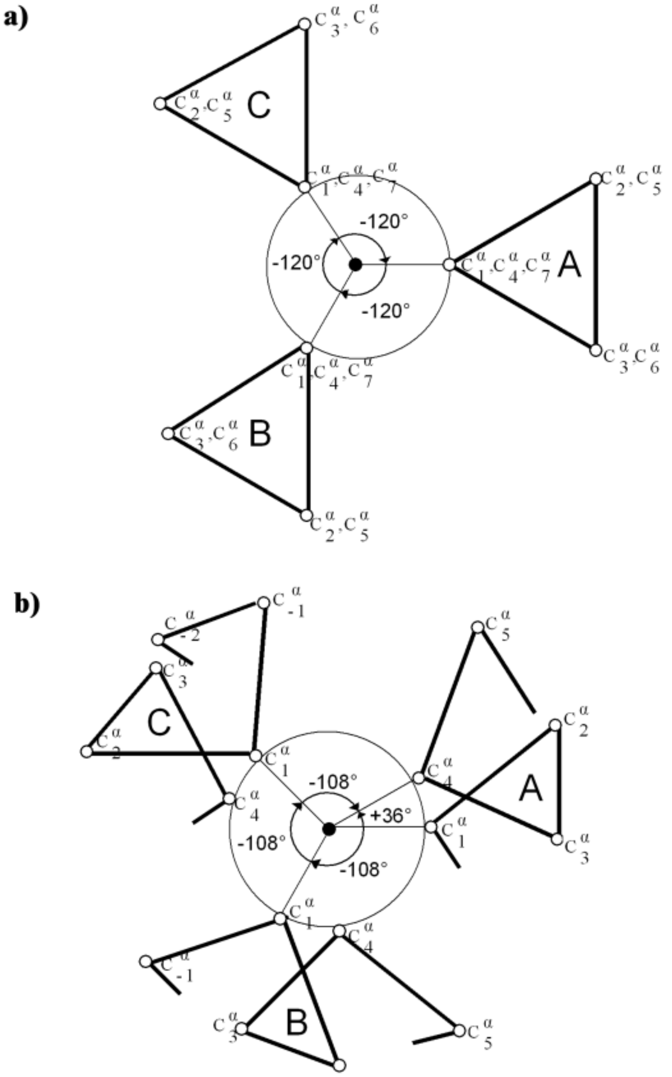
a)
$C^{\alpha}_{3}, C^{\alpha}_{6}$
$C^{\alpha}_{2}, C^{\alpha}_{5}$
C
$C^{\alpha}_{1}, C^{\alpha}_{4}, C^{\alpha}_{7}$
-120°
-120°
-120°
$C^{\alpha}_{2}, C^{\alpha}_{5}$
$C^{\alpha}_{1}, C^{\alpha}_{4}, C^{\alpha}_{7}$
A
$C^{\alpha}_{3}, C^{\alpha}_{6}$
$C^{\alpha}_{1}, C^{\alpha}_{4}, C^{\alpha}_{7}$
$C^{\alpha}_{3}, C^{\alpha}_{6}$
B
$C^{\alpha}_{2}, C^{\alpha}_{5}$
b)
$C^{\alpha}_{-1}$
$C^{\alpha}_{-2}$
$C^{\alpha}_{3}$
C
$C^{\alpha}_{2}$
$C^{\alpha}_{1}$
$C^{\alpha}_{4}$
-108°
-108°
-108°
+36°
$C^{\alpha}_{5}$
$C^{\alpha}_{2}$
$C^{\alpha}_{4}$
A
$C^{\alpha}_{1}$
$C^{\alpha}_{3}$
$C^{\alpha}_{1}$
$C^{\alpha}_{4}$
$C^{\alpha}_{-1}$
$C^{\alpha}_{3}$
B
$C^{\alpha}_{5}$

**Figure 1.** (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 ~3Å (*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 ~8.7Å about the common axis relating the three chains. Neighboring chains are related by a rotation of -108° and a translation of ~2.9Å (*4,5*).

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 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

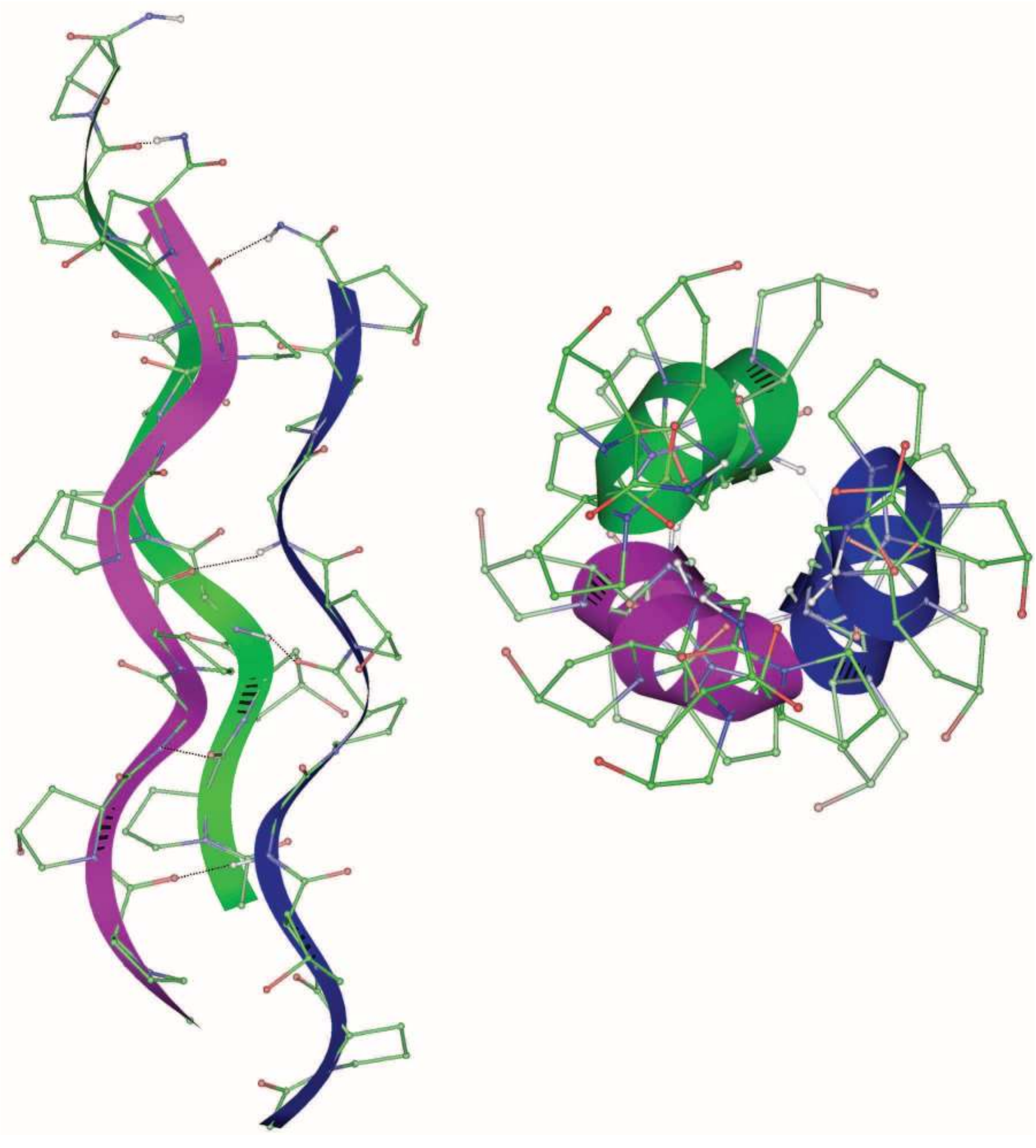

**Figure 2.** Ball and stick diagrams showing two projections of the currently accepted triple helical structure for collagen with one inter-chain hydrogen bond per tripeptide. The sequence shown is (Gly-Pro-Hyp)$_3$ and each chain in the triple helix has a different color ribbon drawn through the backbone.

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 OLIGOPEPTIDES

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 C$\alpha$ 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 are 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,

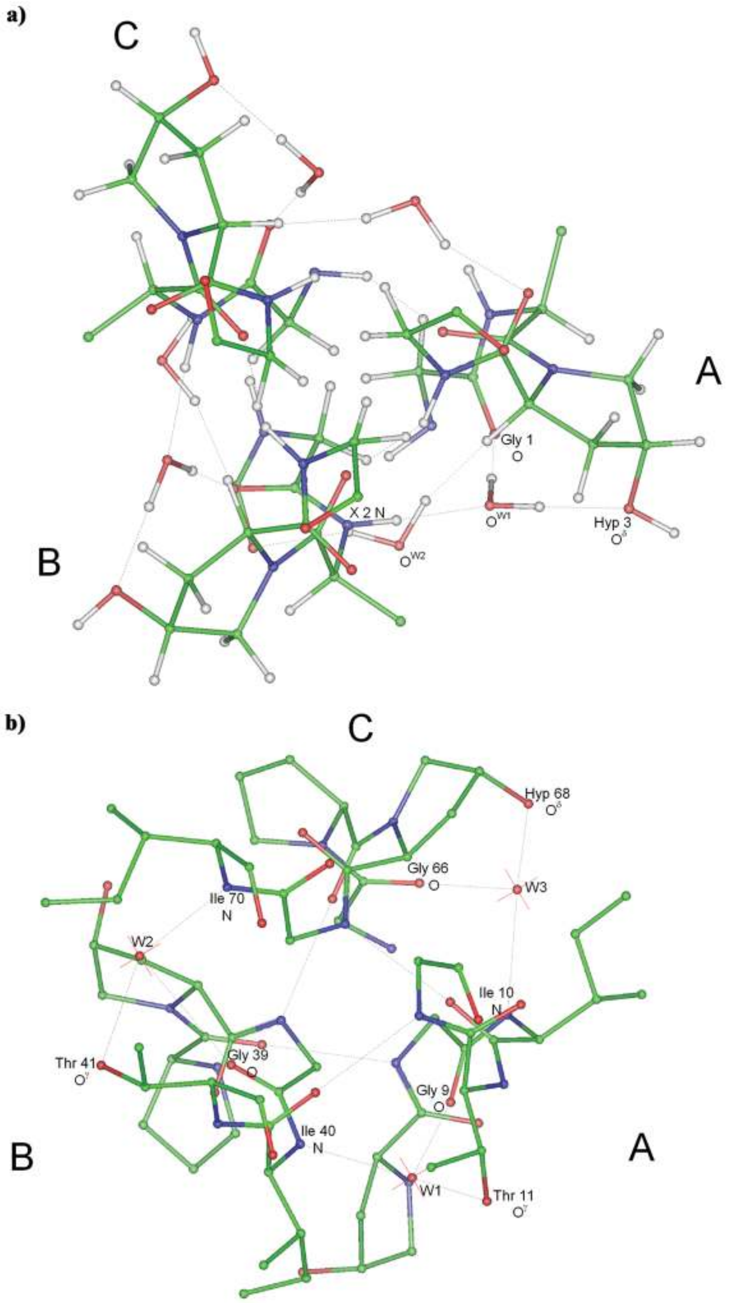

**Figure 3**. (a) Projection down the helix axis of the one hydrogen-bonded fiber model (*13*), showing details of the direct interchain hydrogen bond, between the glycyl amide group and the carbonyl oxygen of residue at X position in the neighboring chain of a triple helix with Gly-X-Hyp sequence. The modeled water molecules ($O^{W1}$ and $O^{W2}$) linking the backbone of two neighboring chains, as well as the $O^{\delta}$ 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.

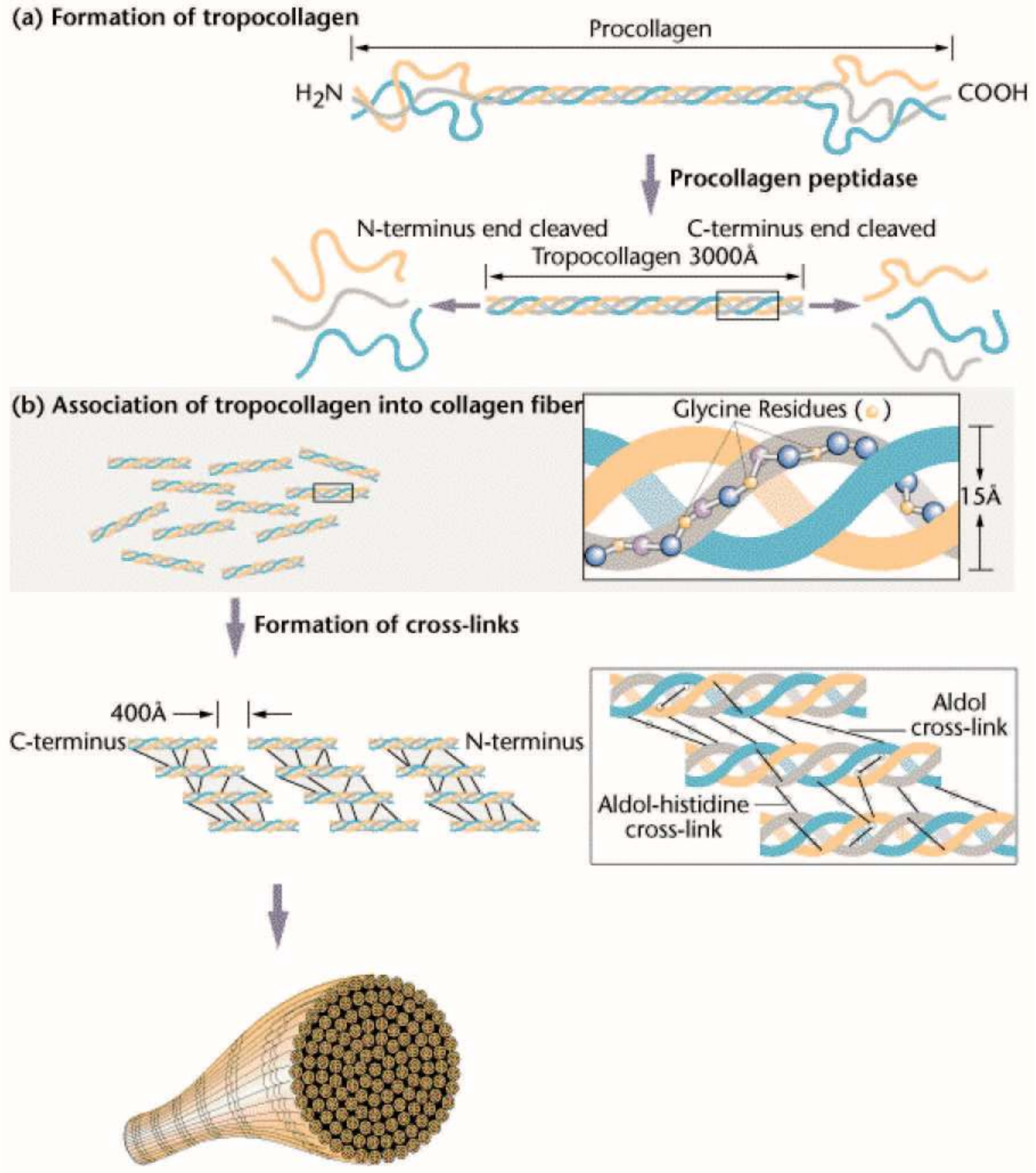

**Figure 4.** The individual triple helices or tropocollagen molecules, are arranged to form fibrils which are of high tensile strength and flexibility and can be further assembled and cross-linked. The figure has been reproduced from Klug, W. S. and Cummings, M. R. (*23*).

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

**Table 1.**

List of triple helical collagen structures proposed from X-ray fiber diffraction as well as single crystal studies. The sequence of the fragments used for the structural analysis is indicated in each case. 'X' indicates any amino acid in the second position of the triplets and 'O' indicates Hydroxyproline. The root mean square deviation (rmsd) value of the backbone atoms in each structure when compared to the fiber model (RDBF) are also listed.

| Serial no. | Structure description used | PDB id | Sequence of the fragment* | rmsd value w.r.t RDBF structure | Reference nos. |
|---|---|---|---|---|---|
| 1 | GNR1 | Fiber model | (GXPGXPGXP) | 0.84 | *4,12* |
| 2 | ARFC | Fiber model | (GPOGPOGPO) | 0.47 | *10,11* |
| 3 | GNR2 | Fiber model | (GPOGPOGPO) | 0.66 | *13* |
| 4 | RDBF | Fiber model | (GPOGPOGPO) | – | *14* |
| 5 | HMB1 | 1CAG | (G$_{12}$POAPOGPO20)<br>(G42POAPOGPO50)<br>(G72POAPOGPO80) | 1.12 | *24,25* |
| 6 | HMB2 | 1QSU | (G$_{18}$POGPOGPO$_{26}$)<br>(G$_{48}$POGPOGPO$_{56}$)<br>(G$_{78}$POGPOGPO$_{86}$) | 0.93 | *27* |
| 7 | HMB3 | 1QSU | (G$_{9}$POGEKGPO$_{17}$)<br>(G$_{39}$POGEKGPO$_{47}$)<br>(G$_{69}$POGEKGPO$_{77}$) | 0.65 | *27* |
| 8 | HMB4 | 1BKV | (G$_{9}$ITGARGLA$_{17}$)<br>(G$_{39}$ITGARGLA$_{47}$)<br>(G$_{69}$ITGARGLA$_{77}$) | 0.34 | *28,29* |
| 9 | JEMS | 1DZI | (G$_{7}$FOGERGPO$_{15}$)<br>(G$_{7}$FOGERGPO$_{15}$)<br>(G$_{7}$FOGERGPO$_{15}$) | 1.06 | *30* |
| 10 | ZAGA | 1K6F | (G$_{3}$PPGPPGPP$_{11}$)<br>(G$_{3}$PPGPPGPP$_{11}$)<br>(G$_{3}$PPGPPGPP$_{11}$) | 0.74 | *31* |
| 11 | OKUY | 1IIT | (G$_{1}$PPGPP$_{6}$)<br>(G$_{2}$PPGPP$_{7}$)<br>(G$_{3}$PPGPP$_{8}$) | 0.34 | *32* |

* In case of crystal structures 5–9, these fragments are middle parts of longer molecules with flanking GPO or GPP sequences.

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 $C^{\gamma}$ 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 $O^{\gamma}$ 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 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.

| Parameters | | | | | | |
|---|---|---|---|---|---|---|
| Structures | *n* | h | t | r ($C^{\alpha}_{G}$ ) | r ($C^{\alpha}_{X}$ ) | r ($C^{\alpha}_{Y}$) |
| GNR1 | 3.27 | 2.91 | −110 | 1.15 | 3.5 | 3.5 |
| ARFC | 3.33 | 2.86 | −108 | 1.6 | 4.07 | 3.43 |
| GNR2 | 3.27 | 2.91 | −110 | 1.4 | 3.95 | 3.6 |
| RDBF | 3.33 | 2.98 | −108 | 1.93 | 4.06 | 2.9 |
| HMB1 | 3.35 (0.14) | 2.77 (0.14) | −107.5 (24.1) | 2.32 (0.4) | 4.53 (0.4) | 3.68 (0.5) |
| HMB2 | 3.53 (0.02) | 2.82 (0.02) | −101.9 (2.9) | 1.80 (0.1) | 4.10 (0.1) | 3.18 (0.1) |
| HMB3 | 3.45 (0.03) | 2.85 (0.04) | −104.4 (4.4) | 1.70 (0.1) | 4.08 (0.1) | 3.24 (0.04) |
| HMB4 | 3.34 (0.03) | 2.91 (0.03) | −107.9 (4.6) | 1.74 (0.1) | 4.06 (0.1) | 3.19 (0.1) |
| JEMS | 3.37 (0.04) | 2.86 (0.1) | −106.8 (8.1) | 1.79 (0.2) | 4.07 (0.2) | 3.22 (0.1) |
| ZAGA | 3.45 (0.01) | 2.84 (0.03) | −104.3 (2.4) | 1.84 (0.1) | 4.07 (0.1) | 3.07 (0.03) |
| OKUY | 3.50 (0.00) | 2.9 (0.01) | −102.9 (0.9) | 1.85 (0.01) | 4.06 (0.04) | 3.07 (0.02) |

*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 ($C^{\alpha}_{G}$) = Average radius (in Angstroms) of $C^{\alpha}$ atom of the Gly residues.
r ($C^{\alpha}_{X}$) = Average radius (in Angstroms)of $C^{\alpha}$?atoms?of the residues in X position of the triple helix.
r ($C^{\alpha}_{Y}$) = Average radius (in Angstroms)of $C^{\alpha}$ atoms?of the residues in Y position of the triple helix.

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 occuring 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*).

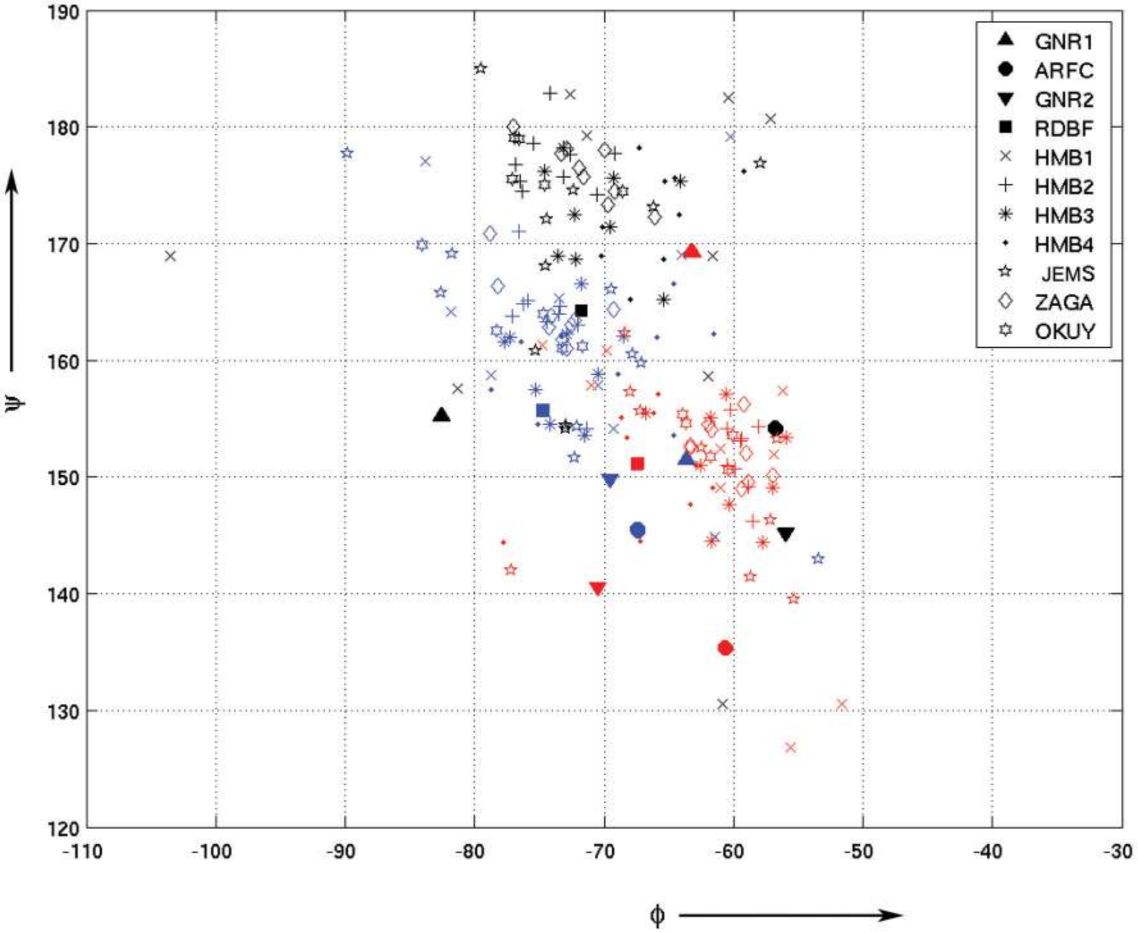

**Figure 5.** Value of the backbone torsion angle phi ($\phi$) is plotted against the psi ($\psi$) value for each residue is the collagen fiber models and oligopeptide crystal structures listed in Table 1. The values corresponding to the three positions in the triplet sequence Gly-X-Y are shown in black, blue and red colors respectively. Each structure is represented by a different symbol, as indicated in top right corner.

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

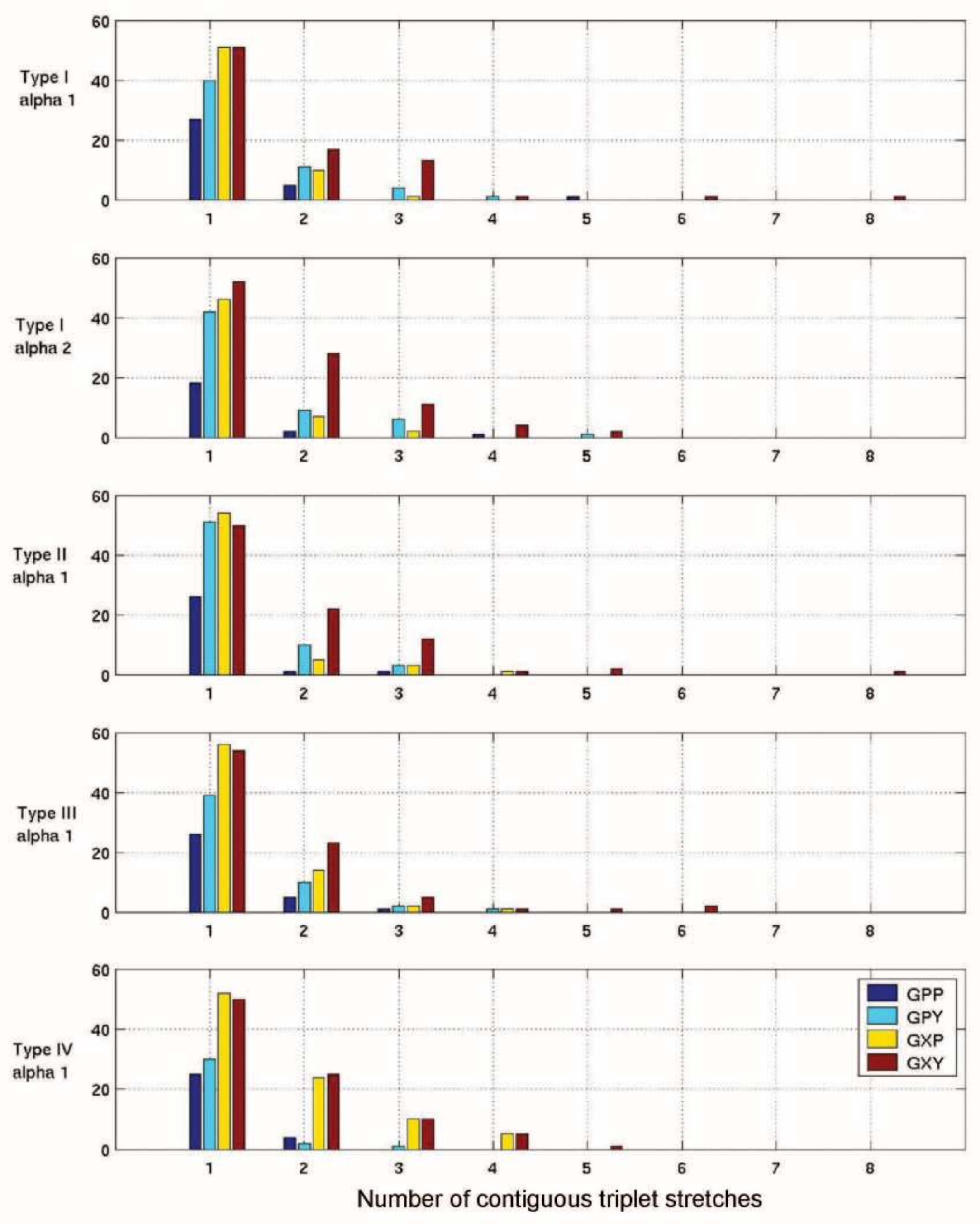

**Figure 6.** Bar diagrams showing the length distribution of the four type of triplet stretches among representative examples of Collagen types I to IV. The amino acid sequences correspond to type I α1 chain from skin tendon and bone of human (N_P000079, 338, *42*), type I α2 chain from placenta and 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.

**Table 3.**
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.

| Amino Acids | Type I alpha 1 | Type I alpha 2 | Type II alpha 1 | Type III alpha 1 | Type IV alpha 1 |
|---|---|---|---|---|---|
| Ala | eq | eq | eq | eq | eq |
| Arg | *Y* | *Y* | *Y* | *Y* | Y |
| Asn | eq | eq | eq | Y | X |
| Asp | eq | eq | eq | Y | X |
| Gln | *Y* | *Y* | *Y* | eq | eq |
| Glu | *X* | *X* | *X* | *X* | *X* |
| Ile | eq | eq | X | X | X |
| Leu | *X* | *X* | *X* | *X* | X |
| Lys | Y | Y | Y | Y | *Y* |
| Met | Y | Y | Y | X | Y |
| Phe | *X* | *X* | *X* | *X* | X |
| Pro/Hyp | eq | eq | eq | eq | Y |
| Ser | eq | eq | Y | X | X |
| Thr | *Y* | Y | *Y* | eq | X |
| Val | eq | Y | eq | eq | eq |

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

1. Myllyharju, J. and Kivirikko, K. I. (2004) Collagens, modifying enzymes and their mutations in humans, flies and worms. *Trends Genet*. **20**, 33 – 43.
2. Kielty, C. M., and Grant, M. E. (2002) The collagen family: structure, assembly, and organisation in the extracellular matrix. Connective tissue and its heritable disorders. *Molecular Genetics in Medical Aspects*, 2nd ed., pp. 159 – 221, Wiley Liss, New York.
3. Ramachandran, G. N., and Kartha, G. (1954) Structure of collagen. *Nature* **174**, 269 – 270.
4. Ramachandran, G. N., and Kartha, G. (1955) Structure of collagen. *Nature* **176**, 593 – 595.
5. Ramachnadran, G. N., and Kartha, G. (1955) Studies on collagen. *Proc. Indian Acad. Sci*. **A42**, 215 – 234.
6. Astbury, W. T. (1938) X-ray adventures among the proteins. *Trans. Faraday Soc*. **34**, 378 – 388.
7. Pauling, L., and Corey, R. B. (1951) The structure of fibrous proteins of the collagen gelatin group. *Proc. Natl. Acad. Sci*. **37**, 272 – 281.
8. Pauling, L., and Corey, R. B. (1951) Atomic coordinates and structure factors for two helical configurations of polypeptide chains. *Proc. Natl. Acad. Sci*. **37**, 235 – 240.
9. Ramachandran, G. N., and Ambady, G. K. (1954) Elements of the helical structure of collagen. *Curr. Sci*. **23**, 349 – 350.
10. Sasisekharan, V. (1962) Stereochemical criteria for polypeptide and protein structures. *Collagen* 39 – 78 Interscience Publishers, New York.
11. Rich, A., and Crick, F. H. C. (1955) The structure of collagen. *Nature* **176**, 915 – 916.
12. Rich, A., and Crick, F. H. C. (1961) The molecular structure of collagen. *J. Mol. Biol*. **3**, 483 – 506.
13. Ramachandran, G. N., and Sasisekharan, V. (1965) Refinement of the structure of collagen *Biochim. Biophys. Acta*. **109**, 314 – 316.
14. Ramachandran, G. N. (1967) Structure of collagen at the molecular level. *Treatise on Collagen* **1**, 103 – 183 Academic Press, New York.
15. Fraser, R. D. B., Macrae, T. P., and Suzuki, E. (1979) Chain conformation of the collagen molecule. *J. Mol. Biol*. **129**, 463 – 481.
16. Ramachandran, G. N., and Chandrasekharan, R. (1968) Interchain hydrogen bonds via bound water molecules in the collagen triple helix. *Biopolymers* **6**, 1649 – 1658.
17. Ramachandran, G. N., Bansal, M., and Bhatnagar, R. S. (1973) A hypothesis on the role of hydroxyproline in stabilizing the collagen structure. *Biochim. Biophys. Acta* **322**, 166 – 171.
18. Bansal, M., Ramkrishnan, C., and Ramachandran, G. N. (1975) Stabilization of the collagen structure by hydroxyproline residues. *Proc. Indian Acad. Sci*. **A82**, 152 – 164.
19. Bansal, M. (1977) Stereochemical restrictions on the occurrence of amino acid residues in the collagen structure. *Int. J. Pept. Prot. Res*. **9**, 224 – 234.
20. Berg, R. A., and Prockop, D. J. (1973) The thermal transition of a non-hydroxylated form of collagen. Evidence for a role of hydroxyproline in stabilizing the triple helix of collagen. *Biochem. Biophys. Res. Commun*. **52**, 115 – 120.
21. Burjanadze, T. V. (2000) New analysis of the phylogenetic change of collagen thermostability. *Biopolymers* **53**, 523 – 528.
22. Mizuno, K., Hayashi, T., and Bachinger, H. P. (2003) Hydroxylation-induced stabilization of the collagen triple helix. *J. Biol. Chem*. **278**, 32373 – 32379.
23. Klug, W. S., and Cummings, M. R. (1997) Genes and proteins. In *Concepts of Genetics*, 5th ed. Prentice-Hall, Upper Saddle River, New Jersey.
24. Bella, J., Eaton, M., Brodsky, B., and Berman, H. M. (1994) Crystal and molecular structure of a collagen-like peptide at 1.9 Å resolution. *Science* **266**, 75 – 81.

25. Kramer, R. Z., and Berman, H. M. (1998) Patterns of hydration in crystalline collagen peptides. *J. Biomol. Struct. Dyn.* **16**, 367 – 380.
26. Kramer, R., Vitagliano, L., Bella, J., Berisio, R., Mazarella, L. Brodsky, B., Zagari, A., and Berman, H. M. (1998) X-ray crystallographic determination of a collagen-like peptide with the repeating sequence (Pro-Pro-Gly). *J. Mol. Biol.* **280**, 623 – 638.
27. Kramer, R. Z., Venugopal, M. G., Bella, J., Mayville, P., Brodsky, B., and Berman, H. M. (2000) Staggered molecular packing in crystals of a collagen-like peptide with a single charged pair. *J. Mol. Biol.* **301**, 1191 – 1205.
28. Kramer, R. Z., Bella, J., Mayville, P., Brodsky B., and Berman, H. M. (1999) Sequence dependent conformational variations of collagen triple-helical structure. *Nat. Struct. Biol.* **6**, 454 – 457.
29. Kramer, R., Bella, J., Brodsky, B., and Berman, H.M. (2001) The crystal and molecular structure of a collagen like peptide with a biologically relevant sequence. *J. Mol. Biol.* **311**, 131 – 147.
30. Emsley, J., Knight, C. G., Farndale, R.W., Barnes, M. J., and Liddington, R. C. (2000) Structural basis of collagen recognition by integrin $\alpha 2\beta 1$. *Cell* **100**, 47 – 56.
31. Berisio, R., Vitagliano, L., Mazarella, L., and Zagari, A. (2002) Crystal structure of the collagen triple helix model [(Pro-Pro-Gly)$^{10}$]$^{3}$. *Prot. Sci.* **11**, 262 – 270.
32. Hongo, C., Nagarajan, V., Noguchi, K., Kamitori, S., Okuyama, K., Tanaka, Y., and Nishino, N. (2001) Average crystal structure of (Pro-Pro-Gly)9 at 1.0 angstroms resolution. *Plym. J.* **33**, 812.
33. Bansal, M., Kumar, S., and Velavan, R. (2000) HELANAL: A program to characterize helix geometry in proteins. *J. Biomol. Struct. Dyn.* **17**, 811 – 819.
34. Vitagliano, L., Berisio, R., Mazarella, L., and Zagari, A. (2001) Structural basis of collagen stabilization induced by proline hydroxylation. *Biopolymers* **58**, 459 – 464.
35. Holmgreen, S. K., Bretscher, L. E., Taylor, K. M., and Raines, R. T. (1999) A hyperstable collagen mimic. *Chem. Biol.* **6**, 63 – 70.
36. Mooney, S. D., Kollman, P. A., and Klein, T. E. (2002) Conformational preferences of substituted prolines in the collagen triple helix. *Biopolymers* **64**, 63 – 71.
37. Jenkins, C. L., and Raines R. T. (2002) Insights on the conformational stability of collagen *Nat. Prod. Rep.* **19**, 49 – 59.
38. Miles, C.A., and Bailey, A. J. (2001) Thermally labile domains in the collagen molecule. *Micron* **32**, 325 – 332.
39. Slatter, D. A., Miles, C. A., and Bailey, A. J. (2003) Asymmetry in the triple helix of collagen-like heterotrimers confirms that external bonds stabilize collagen structure. *J. Mol. Biol.* **329**, 175 – 183.
40. Mann, K., Mechling, D. E., Bachinger, H. P., Eckerskorn, C., Gaill, F., and Timpl, R. (1996) Glycosylated threonine but not 4-hydroxyproline dominates the triple helix stabilizing positions in the sequence of a hydrothermal vent worm cuticle collagen. *J. Mol. Biol.* **261**, 255 – 266.
41. Kadler, K. E., Holmes, D. F., Trotter, J. A., and Chapman, J. A. (1996) Collagen fibril formation. *Biochem. J.* **316**, 1 – 11.
42. Ruza, E., Sotillo, E., Sierrasesumaga, L., Azcona, C., and Patino-Garcia, A. (2003) Analysis of polymorphisms of the vitamin D receptor, estrogen receptor, and collagen I alpha1 genes and their relationship with height in children with bone cancer. *J. Pediatr. Hematol. Oncol.* **25**, 780 – 786.
43. De Wet, W., Bernard, M., Benson-Chanda, V., Chu, M. L., Dickson, L., Weil, D., and Ramirez, F. (1987) Organization of the human pro-alpha 2(I) collagen gene. *J. Biol. Chem.* **262**, 16032 – 16036.
44. Metsaranta, M., Toman, D., De Crombrugghe, B., and Vuorio, E. (1991) Mouse type II collagen gene. Complete nucleotide sequence, exon structure, and alternative splicing. *J. Biol. Chem.* **266**, 16862 – 16869.
45. Toman, P. D., and De Crombrugghe, B. (1994) The mouse type-III procollagen-encoding gene: genomic cloning and complete DNA sequence. *Gene* **147**, 161 – 168.
46. Lauer-Fields, J. L., Malkar, N. B., Richet, G., Drauz K., and Fields, G. B. (2003) Melanoma cell CD44 interaction with the alpha 1(IV) 1263-1277 region from basement membrane collagen is modulated by ligand glycosylation. *J. Biol. Chem.* **278**, 14321 – 14330.
47. Berisio, R., Vitagliano, L., Mazarella, L., and Zagari, A. (2001) Crystal structure of a collagen-like polypeptide with repeating sequence Pro-Hyp-Gly at 1.4 Å resolution: Implications for collagen hydration. *Biopolymers* **56**, 8 – 13.
48. Ramshaw, J. A. M., Shah, N. K., and Brodsky, B. (1998) Gly-X-Y tripeptide frequencies in collagen: a context for host-guest triple helical peptides. *J. Struct. Biol.* **122**, 86 – 91.
49. Persikov, A. V., Ramshaw, J. A. M., Kirkpatrick, A., and Brodsky, B. (2000) Amino acid propensities for the collagen triple-helix. *Biochemistry* **39**, 14960 – 14967.
50. Vitagliano, L., Nemethy, G., Zagari, A., and Scherega, H. A (1993) Stabilization of the triple- helical structure of natural collagen by side-chain interactions. *Biochemistry* **32**, 7354 – 7359.
51. Bansal, M., and Ramachandran, G. N. (1978) A theoretical study on the structure of (Gly-Pro-Leu)n and (Gly-Leu-Pro)n. *Int. J. Pept. Prot. Res.* **11**, 73 – 81.
52. Miles, C. A., Sims, T. J., Camacho, N. P., and Bailey, A. J. (2002) The role of the alpha2 chain in the stabilization of the collagen type I heterotrimer: a study of the type I homotrimer in oim mouse tissues. *J. Mol. Biol.* **321**, 797 – 805.
53. Sacca, B., Fiori, S., and Moroder, L. (2003) Studies on the local conformational properties of the cell-adhesion domain of collagen type IV in synthetic heterotrimeric peptides. *Biochemistry* **42**, 3429 – 3436.
54. Sacca, B., Renner, C., and Moroder, L. (2002) The chain register in heterotrimeric collagen peptides affects triple helix stability and folding kinetics. *J. Mol. Biol.* **324**, 309 – 318.
55. Cabral, W. A., Mertts, M. V., Makareeva, E., Colige, A., Tekin, M., Pandya, A., Leikin, S., and Marini, J. C. (2003) Type I collagen triplet duplication mutation in lethal osteogenesis imperfecta shifts register of alpha chains throughout the helix and disrupts incorporation of mutant helices into fibrils and extracellular matrix. *J. Biol. Chem.* **278**, 10006 – 10012.