doi:10.1016/j.jmb.2007.05.076 J. Mol. Biol. (2007) 372, 1097–1107

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

The Variability in Type I Collagen Helical Pitch Is Reflected in the D Periodic Fibrillar Structure

G. J. Cameron[1], D. E. Cairns[1]* and T. J. Wess[2]

[1]*Department of Computing Science and Mathematics University of Stirling Stirling, FK9 4LA, UK*

[2]*Structural Biophysics Group School of Optometry and Vision Science, Cardiff University Redwood Building King Edward VII Avenue Cardiff, CF11 3NB, UK*

**Corresponding author*

The variability in amino acid axial rise per residue of the collagen helix is a potentially important parameter that is missing in many structural models of fibrillar collagen to date. The significance of this variability has been supported by evidence from collagen axial structures determined by electron microscopy and X-ray diffraction, as well as studies of the local sequence-dependent conformation of the collagen helix. Here, sequence-dependent variation of the axial rise per residue was used to improve the fit between simulated diffraction patterns derived from model structures of the axially projected microfibrillar structure and the observed X-ray diffraction pattern from hydrated rat tail tendon. Structural models were adjusted using a genetic algorithm that allowed a wide range of structures to be tested efficiently. The results show that variation of the axial rise per residue could reduce the difference metric between model and observed data by up to 50%, indicating that such a variable is a necessary part of fibril model structure building. The variation in amino acid translation was also found to be influenced by the number of proline and hydroxyproline residues in the triple helix structure.

Keywords: type I collagen; genetic algorithm; fibril; structure; X-ray

Introduction

The triple helical polyproline (PP) II helix is regarded as a rigid protein secondary structure element, where the axial rise per amino acid residue is highly restricted. The classical description of the collagen helix is as a PP-II helix where the amino acid sequence Gly-Pro-Hyp predominates. Collagen gene products are more diverse in their sequence and Gly-Pro-Hyp only accounts for approximately 10% of collagen sequences. Our hypothesis is that collagen sequence variability translates to local deviations of collagen helical structure from the ideal PP-II helix, which may be critical to satisfy the topological, cross-linking and dynamic properties of supramolecular collagen structures.

The majority of collagen in animals is in a fibrillar form, in which collagen triple helices interact with neighbouring molecules leading to a highly cross-linked structure. This structural framework provides a connection between the molecular strength of the triple helix and the mesoscopic properties of a fibril as well as the collagen tissue itself. The axis of each collagen fibril exhibits long-range order, with each of the approximately 300 nm long molecules being staggered by 67 nm (*D*) with respect to its neighbour. This corresponds to approximately 234.2(±0.5) amino acids per *D* period for type I collagen.[1] Due to the non-integer relationship between molecular length and the *D* period, the projection of the axial structure repeat for a type I collagen fibril consists of a gap region and an overlap region in the ratio of 0.54:0.46 *D*, respectively.[2] The ends of the triple helical region are defined by telopeptides, which do not conform to the repeating Gly–X_{aa}–Y_{aa} structure of the helical region. The telopeptide region is known to be responsible for the formation of inter-molecular cross-links.[3]

Determination of fibrillar collagen structure has previously relied on X-ray diffraction and electron microscopic techniques. The broad structural features that have been observed experimentally can be qualitatively reproduced in appropriate models by translating the amino acid sequence into an electron density profile. These models are often constructed by placing the residues at positions dictated by the formal collagen helical co-ordinates of either a 3_{10} or 7_2 helix, both of which have a very similar axial translation per residue. Studies that seek to explain experimentally observed features at higher resolution typically fail, as the correlation between the predicted

Abbreviation used: GA, genetic algorithm.

E-mail address of the corresponding author: dec@cs.stir.ac.uk

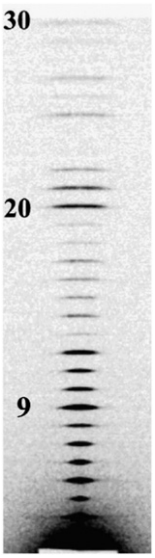

Figure 1. X-ray diffraction image of the meridional region from fibrillar type I collagen. The meridional Bragg peaks are formed from the interactions of the X-rays with the regular repeating structure of the axial packing scheme of the collagen molecules. This Figure shows the resultant Bragg peaks of diffracted rat tail tendon with sample-to-detector distances of 2000 mm (ID02, ESRF). For guidance, a number of the more intense peaks have been labelled.

position of amino acids, based on the ideal helical structures and electron density profile, is not sufficient. This implies the existence of regions of possible rarefaction and compaction of electron density/axial translation per residue in the fibrillar structure, rather than the expected constant of 0.286 nm per residue. The electron density profile and molecular packing structure that was produced by Orgel *et al.*,[4] addressed the above issues and suggested a general approach to collagen structure model building. Interestingly, the study only obtained an optimal fit of the electron density by allowing a fold in the telopeptide region and utilising a non-uniform distribution of amino acid scattering objects, a feature that was also noted by the electron microscopy studies of Chapman *et al.*[5]

Here, we investigated the modelling of the non-uniform axial translation of amino acids in the collagen helix through the utilisation of a genetic algorithm (GA)-based approach.[6] The aim of this work was to evolve a sequence-based structure that obtains the best fit to the observed meridional X-ray diffraction data (Figure 1). Three structural elements of the model were investigated, the telopeptide inter-residue spacing, folding of the telopeptide and the helical inter-residue spacing.

Results

In the initial model structure, the distances between amino acid residues within the helical region of the molecule were fixed at 0.286 nm and the telopeptide regions at either end of the helix were allowed to evolve. In particular, the distances between residues were free to change and random folding positions within the telopeptides were encouraged, as a previous study by Orgel *et al.*[7] indicated the possibility of a fold in the C-terminal telopeptide. The resultant set of model intensities scored over 30 orders, gave an *R*-factor value of 0.541 when compared to the observed intensities. This was not a significant improvement upon previous studies such as that by Hulmes *et al.*[11]

As discussed earlier, the distances between amino acid residues within the helical part of the collagen molecule may not be uniform.[5] This can be quantified by measuring the breadth of the meridional helical diffraction peak corresponding to the axial rise per residue in the X-ray diffraction images. The breadth of this peak defines the limits feasible for the distribution of axial rises per residue that could be used within the genetic algorithm. For the models under investigation, the measured minimum and maximum spacings were 0.270 nm and 0.304 nm. A further criterion for the helical region of the collagen molecule was that the average inter-residue spacing should be equal to 0.286 nm and that a distribution plot of inter-residue distances should closely follow a trace of the helical diffraction peak that led to the defined limits.

The evolution of the structure of the extrahelical telopeptides was less restricted than the helical region of the molecule due to the relative lack of information about these regions of the collagen molecule. It was decided to allow any inter-residue spacing for the amino acids of the telopeptide up to a distance of 0.38 nm, the maximum possible size of a peptide bond.[8] Previous studies report that these regions are compressed with an average inter-residue spacing of approximately 0.2 nm.[9]

Table 1 indicates the recorded values of the first 30 orders of observed intensity and the values for the telopeptide model discussed above. Columns 4 and 5 in Table 1 record the intensity values generated for the best-fit models scored over 30 orders of intensity using two separate methods to evolve the inter-residue spacing,designated as the A3 and A6 methods. The A3 method incorporated the effects of three amino acids in a triplet to determine spacing. The A6 method used six amino acids, three from the triplet before the inter-residue step and three after it.

The A6 method gives the best fit to the observed data when scored over the first 30 orders of intensity with an *R*-factor value of 0.173. The A3 method of evolving the inter-residue spacing produced models with *R*-factor values for 30 orders of 0.247. Although this fit was weaker than for the A6 approach, the A3 model was deemed more biologically feasible as it maintained a better match with the structural criteria that had been stipulated within the software. Figures 2 and 3 show the fit of the A3 and A6 models scored over 30 orders of meridional intensity.

The biological structures of these models must be examined against the previously defined criteria in order to determine whether they are feasible. The length of each molecule is of interest, as collagen