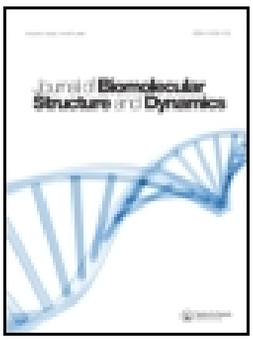

Taylor & Francis
Taylor & Francis Group

Journal of Biomolecular Structure and Dynamics

ISSN: 0739-1102 (Print) 1538-0254 (Online) Journal homepage: http://www.tandfonline.com/loi/tbsd20

Relative orientation of collagen molecules within a fibril: a homology model for homo sapiens type I collagen

Thomas A. Collier, Anthony Nash, Helen L. Birch & Nora H. de Leeuw

To cite this article: Thomas A. Collier, Anthony Nash, Helen L. Birch & Nora H. de Leeuw (2018): Relative orientation of collagen molecules within a fibril: a homology model for homo sapiens type I collagen, Journal of Biomolecular Structure and Dynamics, DOI: 10.1080/07391102.2018.1433553

To link to this article: https://doi.org/10.1080/07391102.2018.1433553

Journal of Biomolecular Structure and Dynamics, 2018
https://doi.org/10.1080/07391102.2018.1433553

Relative orientation of collagen molecules within a fibril: a homology model for *homo sapiens* type I collagen

Thomas A. Collier[a], Anthony Nash[b], Helen L. Birch[c] and Nora H. de Leeuw[d]*

[a]*Institute of Natural and Mathematical Sciences, Massey University, Auckland 0632, New Zealand;* [b]*Department of Physiology, Anatomy and Genetics, University of Oxford, South Parks Road, Oxford OX1 3QX, UK;* [c]*Institute of Orthopaedics and Musculoskeletal Science, UCL, RNOH Stanmore Campus, London, UK;* [d]*School of Chemistry, Cardiff University, Main Building, Park Place, Cardiff CF10 3AT, UK*

Communicated by Ramaswamy H. Sarma

(*Received 9 October 2017; accepted 23 January 2018*)

Type I collagen is an essential extracellular protein that plays an important structural role in tissues that require high tensile strength. However, owing to the molecule's size, to date no experimental structural data are available for the *Homo sapiens* species. Therefore, there is a real need to develop a reliable homology model and a method to study the packing of the collagen molecules within the fibril. Through the use of the homology model and implementation of a novel simulation technique, we have ascertained the orientations of the collagen molecules within a fibril, which is currently below the resolution limit of experimental techniques. The longitudinal orientation of collagen molecules within a fibril has a significant effect on the mechanical and biological properties of the fibril, owing to the different amino acid side chains available at the interface between the molecules.

Keywords: collagen; molecular dynamics; extracellular matrix protein; computational biology; homology modelling; protein structure; fibril; orientation

Introduction

Collagen is the most abundant protein in the human body, constituting over a quarter of the dry mass of the human body (Kadler, Baldock, Bella, & Boot-Handford, 2007). Found primarily in the extracellular matrix (ECM), it provides strength, elasticity and functionality to connective tissues (Ottani, Raspanti, & Ruggeri, 2001). Currently, there are 28 different members of the collagen family (Kadler et al., 2007), although the fibril-forming type I collagen is the most abundant in tissues and organs that require tensile strength, such as tendon, ligament and bone. The mechanical functions of the supramolecular structure in collagenous tissues are optimised for the direction and magnitude of load. For example, in the skin the fibres form an anisotropic network to respond effectively to multidirectional forces (Ottani et al., 2001), whereas in tendons the fibres align in one direction to maximise their effectiveness to respond to a uniaxial load (Silver, Freeman, & Seehra, 2003).

Type I collagen molecules are 300 nm in length, 1.5 nm in diameter and comprise two α1 and one α2 polypeptide chains twisted into a continuous triple helix, flanked on both ends by non-helical telopeptides. Under physiological conditions, solvated collagen molecules spontaneously form long thin fibrils in a process called fibrillogenesis, which sees the molecules aligned parallel yet staggered according to the Hodge-Petruska model (Petruska & Hodge, 1964). Through aligning in this way an observable periodicity is created known as the D-band, which is composed of a gap region (0.54D) and an overlap region of higher protein density (0.46D). Further association occurs laterally to form fibrils which have diameters varying between 20 and 500 nm and a length in the millimetre scale (Kannus, 2000; Pingel et al., 2014), depending on the organism and the location of the tissue.

Much attention has been paid previously to investigating the way in which the collagen fibrils align and orientate within collagen fibres, and how collagen fibres align in fasicles. A variety of techniques have been employed to do this; scanning electron microscopy (Pannarale, Braidotti, D'Alba, & Gaudio, 1994), small angle X-ray scattering (Liao, Yang, Grashow, & Sacks, 2005; Moger et al., 2007), polarised light microscopy (Ugryumova, Jacobs, Bonesi, & Matcher, 2009), infrared and polarised Raman spectroscopy (Bi, Li, Doty, & Camacho, 2005; Galvis, Dunlop, Duda, Fratzl, & Masic, 2013; Masic et al., 2011; Schrof, Varga, Galvis, Raum, & Masic, 2014). However, as yet no method is capable of probing below the fibrillar level to determine the

*Corresponding author. Email: DeLeeuwN@cardiff.ac.uk

orientation of the collagen molecules. It is worth noting that, although the spectroscopic techniques (Raman and FT-IRIS) currently offer the closest to atomic-scale detail, the response of the amide I band (~1620 to 1700 cm^{-1}) is made up of contributions from all of the amide I scattering centres present in the structure, and thus it will consist of multiple responses by the collagen molecule to the incident light. For this reason, there is still a need to develop more advanced techniques and methodologies to be able to sample the orientation of the individual collagen molecules within the fibril.

The alignment of the collagen molecules has been well studied, with the D-banding periodicity being the subject of many research articles (Cameron, Cairns, & Wess, 2007; Fraser, MacRae, Miller, & Suzuki, 1983; Kukreti & Belko, 2000; Orgel, Irving, Miller, & Wess, 2006; Ottani et al., 2001). A variety of techniques have been utilised to probe this D-banding periodicity, although the most common method is through the use of small angle X-ray diffraction, with a series of sharp X-ray peaks present parallel to the fibre axis. However, to the best of our knowledge, no study to date has looked at the rotational orientation of the individual collagen molecules within the fibril.

A low-resolution crystal structure was first determined for type I collagen taken from the tail of *Rattus Norvegicus* by Orgel et al. in 2006, after early attempts to use X-ray diffraction data were thwarted by the electron density map being un-interpretable in the gap regions (Orgel et al., 2006). Particular attention was paid to the lateral packing of collagen molecules into the quasi-hexagonal packing structure. However, no mention was made as to the orientation of the collagen molecules around the longitudinal axis, possibly due to the resolution of the crystal structure being too low. The presence of a single molecule of collagen within the resulting published structure (PDB code: 3HR2) suggests that no deviation in orientation was observed between the collagen molecules, essentially suggesting all collagen molecules within the collagen fibril, extracted from a rat tail tendon, exhibit an orientation around the longitudinal axis of 0°.

The orientation of the collagen molecules about their principal axis will greatly influence the structural properties of collagen fibrils, as the orientation of collagen molecules will determine mechanical properties, owing to the different possible intermolecular forces that occur at the interface between the collagen molecules. The most significant influence of orientation will be on the biological properties, with orientation determining the accessibility of the biomolecule binding sites and presentation of key amino acid residues. The availability of certain amino acids at the interface between the collagen molecules will alter the tissue properties, owing to the different possible inter-molecular interactions. Examples include the side chains available to form non-enzymatic advanced glycation end product cross-links between the molecules (Monnier et al., 2014), and the side chains available to form inter-molecular hydrogen-bonds, either directly or through a water-mediated process (Streeter & de Leeuw, 2011).

Initially, the collagen molecules aggregate as a result of the intermolecular forces, before later forming the covalent interactions via the mature enzymatic cross-link. This scenario could therefore mean that the driver for the determination of the orientation of the collagen molecules will be to maximise the number of favourable inter-molecular interactions to form a low energy fibril. To investigate the lowest energy orientations of the collagen molecules we will use a novel two stage modelling approach, which takes its inspiration from the work by Adams, Arkin, Engelman, & Brünger, 1995 on computational method development for the determination of conformation and rotation angles of the pentameric transmembrane domain of phospholamban (Adams et al., 1995). Our approach begins by conducting a comprehensive single-point energy search of all of the possible orientations at small rotation intervals of 6°, using the *Homo sapiens* sequence for type I collagen. The results of the single point energy search are then used to conduct short molecular dynamics searches of the 150 lowest energy orientations, for further sampling of the potential energy landscape, to find the lowest energy orientation. Validation is then conducted by testing some of the lowest energy orientations within a fibrillar environment, to develop a new homology model for *Homo sapiens* type I collagen that takes into account the orientation of the collagen molecules about the longitudinal axis.

Results and discussion

First, a Blast scoring search was conducted to determine a suitable template structure to use for the production of the supramolecular structure of the new homology model (Altschul, Gish, Miller, Myers, & Lipman, 1990). A template sequence search was conducted with the BlastP software suite, using the human target sequence of the α1 and α2 chains given in the Uniprot entries; CO1A1_human (P02452) and CO2A1 (P08123), respectively. From the results of the BlastP database search, it was clear that the highest scoring sequence is that of collagen α1 and α2 for the *Rattus norvegicus,* which had BLAST max scores 1.8 and 2 times larger than the next sequence with an experimentally determined structure 3HQV. As such, we decided to use the structure of the *Rattus norvegicus* sequence, as used in our previous work (Collier, Nash, Birch, & de Leeuw, 2015), as the reference structure to generate the model structure for the *Homo sapiens* sequence. We considered that this approach would generate a reliable method, owing to the

strong sequence identity similarity of 91% between the two sequences.

A straight-chained structure of the *Homo sapiens* collagen molecule, with the correct helical propensity, was generated using the Triple Helical Building Script (THeBuScr) (Rainey & Goh, 2004). To accurately apply the rotation of the collagen molecule, the assumption is made that the collagen molecule must be considered as a straight rod/cylinder, and thus the straight molecule from the THeBuScr programme was used directly. The strands are generated within the fibrillar environment by positioning a box of length 360 nm at the beginning of the collagen molecule, so that the strand contains a full collagen molecule, a gap region and a short 110 residue triple helical and telopeptide section. The second strand is generated by placing the end of the box at the end of the collagen molecule, such that this strand also includes a short triple helical region, telopeptide, a gap region and a full collagen molecule, as seen in Figure 1.

Two explicit collagen molecules were defined as shown in the bottom image of Figure 1, and these molecules were rotated independently by 6° increments from 0–354°, with the energy computed for each of the 3600 possible combinations. Next, the energies of the resulting systems were plotted, showing the energy as a function of its orientation in the AC and BD strands, which can be seen in Figure 2. Less than 2% of the simulations yielded close contacts or steric clashes between bulky side chains on the collagen molecules, which resulted in a large increase in energy, i.e. orders of magnitudes greater than the average energy. Due to the relatively low proportion of these high-energy structures, depicted by the white regions in Figure 2, it was decided to omit these from the results of the simulations. The potential energy reported is made up of contributions resulting from a wide variety of inter-molecular interactions within and between the collagen molecules, i.e. primarily salt bridges (Keshwani, Banerjee, Brodsky, & Makhatadze, 2013; Persikov, Ramshaw, Kirkpatrick, & Brodsky, 2005; Yang, Chan, Kirkpatrick, Ramshaw, & Brodsky, 1997), direct hydrogen bonding (Brodsky, 1999; Brodsky & Ramshaw, 1997; Persikov, Ramshaw, Kirkpatrick, & Brodsky, 2002) and water-mediated hydrogen bonding (De Simone, Vitagliano, & Berisio, 2008; Kuznetsova, Rau, Parsegian, & Leikin, 1997; Streeter & de Leeuw, 2011).

From Figure 2, we can see that there is a wide distribution of energies throughout, with a number of regions in white exhibiting potential energies higher than -6.85×10^6 kcal/mol, with no significant clustering of low energy regions. However, what can be seen in Figure 2 is the large number of smaller regions of low energy configurations, illustrated by the dark blue regions. Through comparison of the energies of these regions we were able to identify the lowest 150 orientations from the single energy point scan.

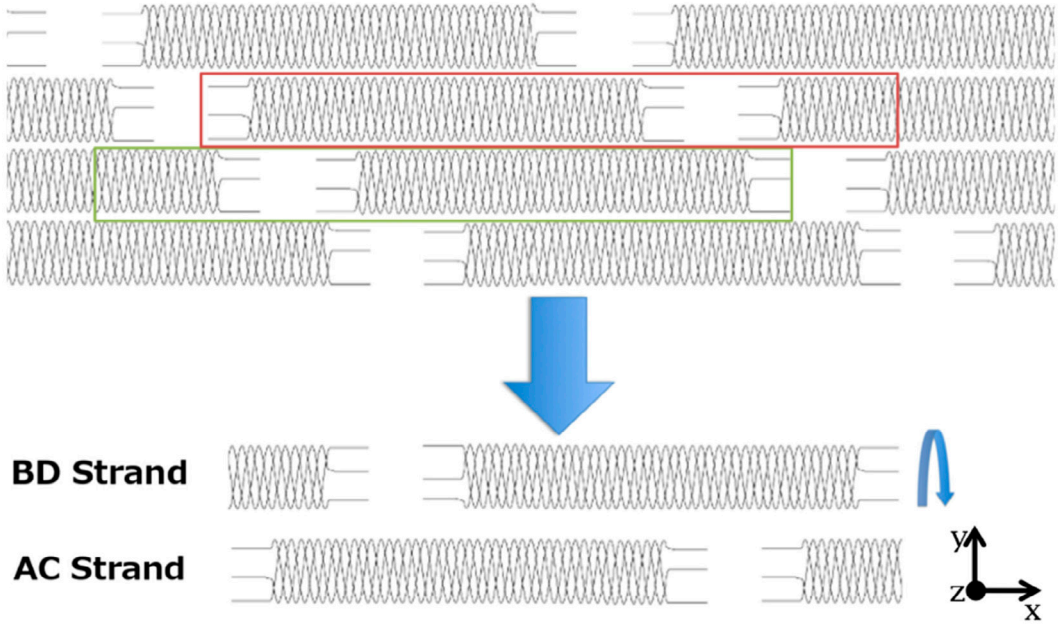

Figure 1. Schematic of the fibril (Top), with the red box (AC) and green box (BD), illustrating the regions of the collagen fibril used in the orientation study.
Notes: After generation of the two strands, alignment to the x-axis, rotation about the x-axis, followed by translation, we obtain the quarter-staggered two collagen molecule model illustrated at the bottom of this figure, with the AC strand on the bottom and the BD strand above.

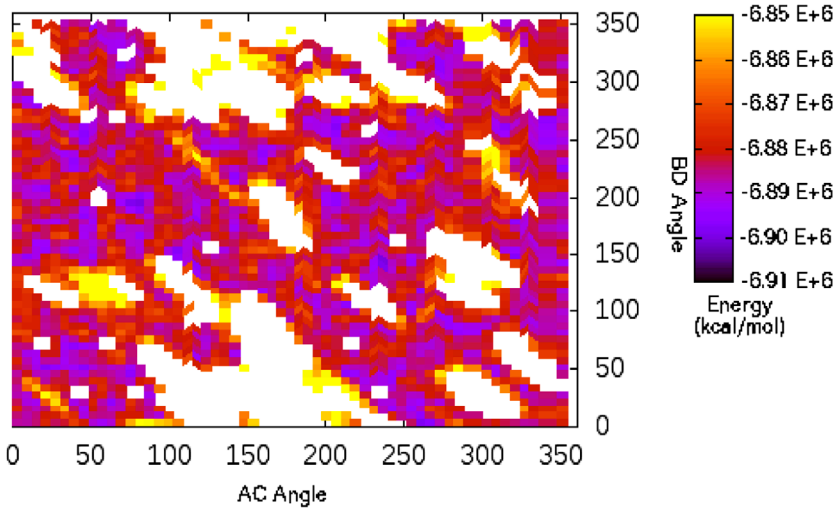

Figure 2. A plot of the potential energy as a function of the orientation angle of the AC strand and the orientation of its corresponding BD strand, with the scale plotted reduced for increased resolution.
Notes: Potential energy is defined by the colour on a sliding scale from yellow – high energy – to blue – low energy – with white representing values significantly off scale due to very high energies.

Beginning with the 150 lowest energy structures identified from the single point energy search, short molecular dynamics simulations (MD) were run to identify both the most abundant orientations, as well as the lowest energy orientations. During thermostatted MD simulations the free energy of a system tends to a minimum and hence the lower energy states are more probable, although random thermal fluctuations will introduce occasional higher energy states. This has two consequences for our investigation; the first is that a sub-optimal orientation will tend towards a thermodynamic equilibrium, therefore rotating into a state of optimal interaction with the second molecule. The second is that we can use the frequency of orientations as a measure of the stability of that particular orientation. Therefore, as the simulations proceed, higher energy structures will move towards a lower energy state. This was indeed observed, as illustrated by an example of the 240–84° initial orientation in Figure 3, where we see a change in the orientation observed until a relatively steady state is obtained, in this case for the AC strand. The BD strand in Figure 3, began at an initial good approximation of the lower energy orientation, and hence it remains at steady state. After an initial relaxation period, which we excluded from our results, we were able to monitor the average orientation to determine the most frequent orientations present.

Through monitoring the frequency of particular orientations within the 100 ps MD simulations, we were able to collate orientation frequency data, as shown in Figure 4. The data were collated into bins such that only integer values were used for the remainder of the study;

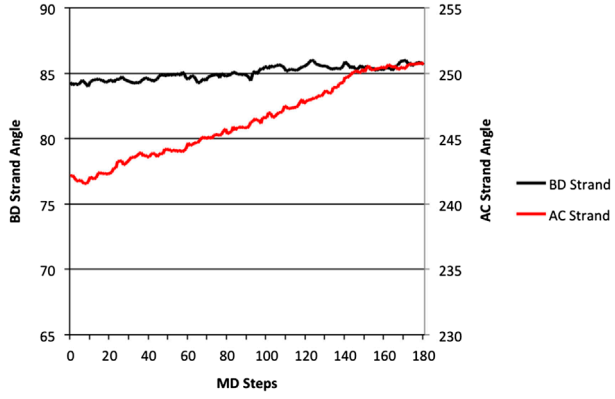

Figure 3. The equilibration of both explicit collagen molecules from the initial 240–84° orientation, AC strand in red moving to equilibration and the BD strand in black beginning at approximate equilibrium.

this reduced the number of possible configurations to 129,600 possible discrete orientations, making the data more manageable and reliable within the limits of the calculated ±0.30° standard error of the mean. What can immediately be seen from Figure 4 is that a significant proportion of the possible orientations remain unpopulated. Instead, there is a clustering of frequently populated orientations and an almost complete exclusion of states elsewhere. Of particular note are the four key exclusion regions from the 340–0° to 20–359°; 160–0° to 200–359°; 0–340° to 359–20°; and 0–160° to 359–200°, which leads to a cross-shaped region through the plot. This observation supports the idea presented in

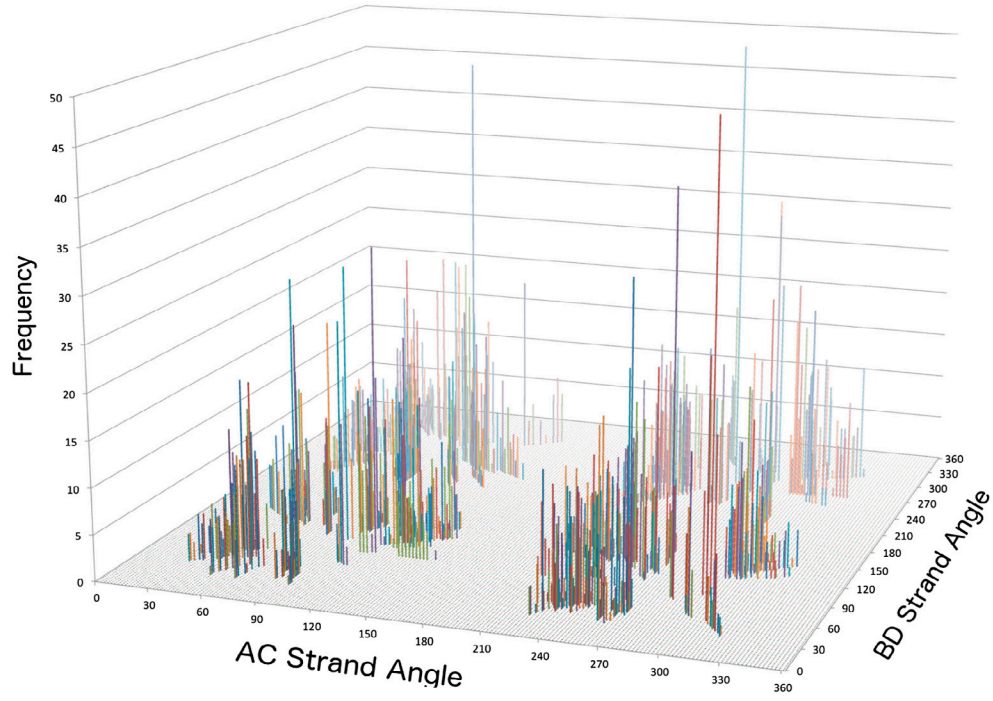

Figure 4. 3D frequency histogram plot of the relative orientations of the two collagen model strands, with angle of the AC strand on the x axis, angle of BD strand on the y axis, and the frequency of the orientation on the z axis.

Figure 3, that despite a large number of potential orientations being identified in the SPE scan, within that region equilibration of the system results into the molecules adjusting into a lower energy orientation, resulting in the frequency at which particular orientations occur reducing to zero during the MD simulations.

Within these clusters of high occurrence orientations observed in Figure 4, we see a small number of very high occurrence orientations, or two very closely related orientations. To gain a better understanding of the favourable orientations that collagen molecules like to adopt, we identified the 30 most frequent orientations, which are presented in Figure 5, along with their accompanying frequency as a percentage of the total number of calculated orientations. Although the frequencies reported look relatively low, with the largest frequency being 0.21%, when you consider that nearly 42,000 orientations were calculated from the MD simulations in which a possible 129,600 orientations are possible, the significance of these values becomes apparent.

Upon extracting the most frequently occurring 30 orientations, we first wanted to investigate the distribution of these orientations. The distribution of the 30 values can be seen in Figure 6, as red squares. Additionally we overlaid the position of the lowest 150 orientations from the single point energy searches. It is apparent that the most frequently occurring orientations are located in four distinct regions of the orientation plot, with the same four exclusion regions present for the SPE identified orientations, but slightly extended. This finding is of significance, as it indicates that interaction between either the top and bottom surface (320–40° and 130–240° regions) of the collagen molecule likely results in unfavourable interactions.

The final stage in identifying the preferential interactions between the collagen molecules for packing in a microfibril is to investigate their effect on the energetics of the system. To do this we calculated the potential energy for the 30 most frequently occurring orientations, for the duration of the 100 ps simulations relative to the 0–0° model energy, the results of which are presented in Figure 7. What is immediately apparent is that all of these orientations have lower energies than those calculated for the 0–0° interaction model. Within these most frequent orientations we have three orientations: 106–258°; 110–254°; 124–302°, with values 50% lower than the average values for the other 27 orientations, making them the optimum interaction orientations. It is seen that the orientations with higher frequencies tend to have lower energies, but with a couple of exceptions. In these cases, a neighbouring orientation is also abundant, for example the 298–268° orientation reports a very low energy but a relatively small frequency owing to the neighbouring 298–264° orientation also exhibiting a high frequency. Considering these data, we can use the clusters of the 30 highest frequency orientations to

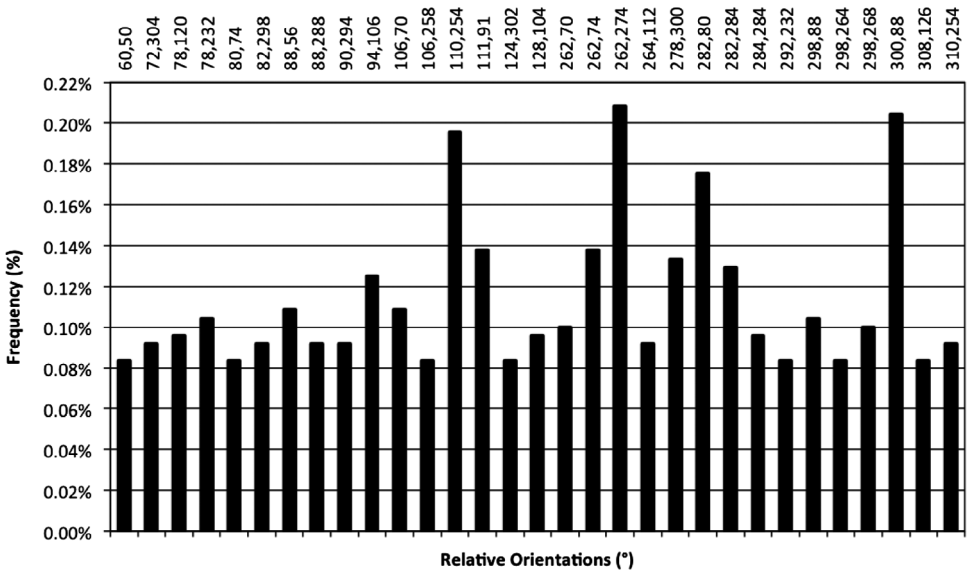

Figure 5. Thirty most frequent orientations identified from the molecular dynamics simulations accompanied with their frequency as a percentage of the total calculated orientations.

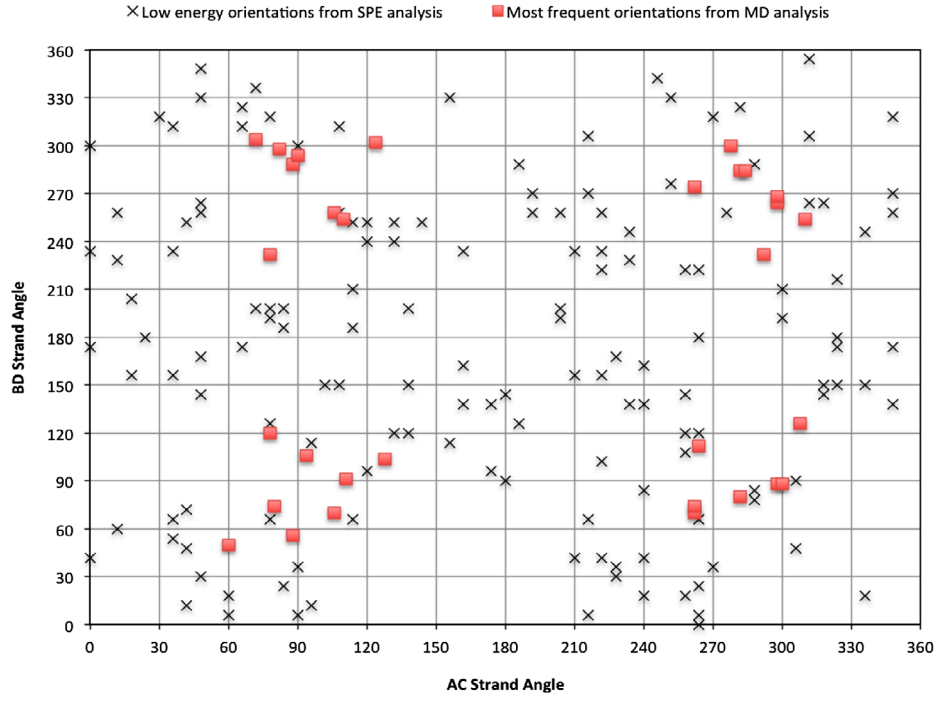

Figure 6. Plot illustrating the angles of the AC and BD collagen strands for the thirty most frequent orientations identified from the molecular dynamics simulations, as red squares, and the 150 lowest energy orientations determined from the single point energy rotation search.

determine the most favourable interactions sites. If we transpose the clusters consisting of the 30 most frequent orientations onto a representative collagen molecule, we see a 'bow' shape of favourable values in Figure 8(A). If we then do the same for the 15 lowest energy orientations we see a narrowing of the left hand 'bow' as

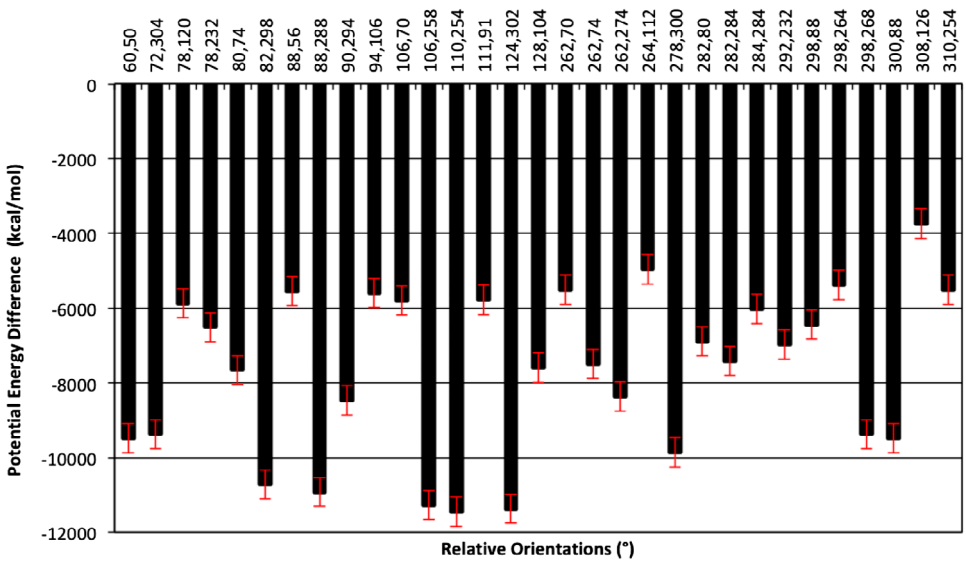

Figure 7. Figure showing the average potential energy difference of the 30 most frequent orientations identified from the molecular dynamics simulations relative to the average potential energy of the 0-0° orientation system.
Notes: Error bars illustrate standard error in reported values.

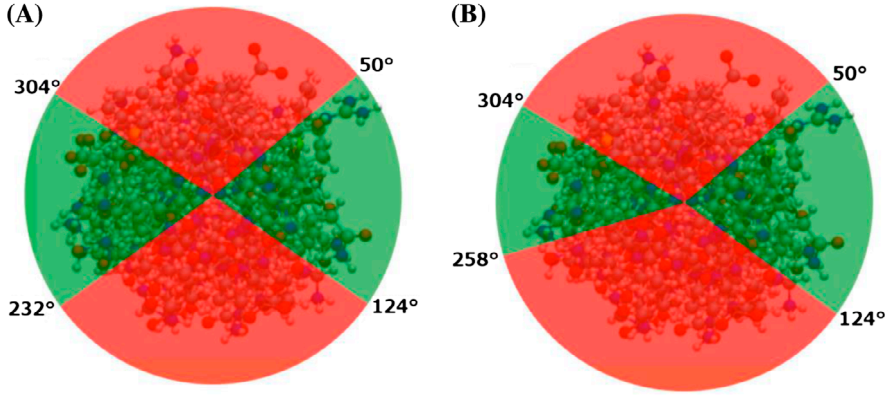

Figure 8. Figures showing the calculated favourable interaction regions shown in green and unfavourable shown in red, based on (A) frequency data and (B) energetics data.

seen Figure 8(B). The red regions are unfavourable orientations for interaction with any orientation of the neighbouring collagen molecules.

To obtain the lowest energy packing of the collagen molecules, the system needs to consist of interactions between the collagen molecules based on the optimum orientations. Collagen molecules within a microfibril can pack in a quasi-hexagonal manner (Hulmes & Miller, 1979), as shown in Figure 9. The current implementation in collagen models, taken from the crystal structure for the *Rattus norvegicus* sequence (Orgel et al., 2006), for the packing, results in 12 interactions within the hexagonal arrangement; one 90–70°, three 270–90°, four 150–330° and four 30–210°. Four of these lie within the high frequency cluster range, whilst the remaining eight lie outside of this area. To obtain the optimum packing for our *Homo sapiens* sequence, rotation of the explicit collagen molecule needs to be conducted so that the number of favourable interactions are optimised. Owing to the periodic nature of the arrangement of collagen molecules within the fibril, the rotation can only be applied to one molecule so that the periodicity is conserved. Rotation of the explicit collagen by 26° in either direction results in a 50% reduction in the number of unfavourable orientation interactions and a subsequent 100% increase in the favourable orientation interactions. The angles of interaction are now one 64–244°, three 244–64°, four 124–304°, and four 4–184° interactions within the hexagonal

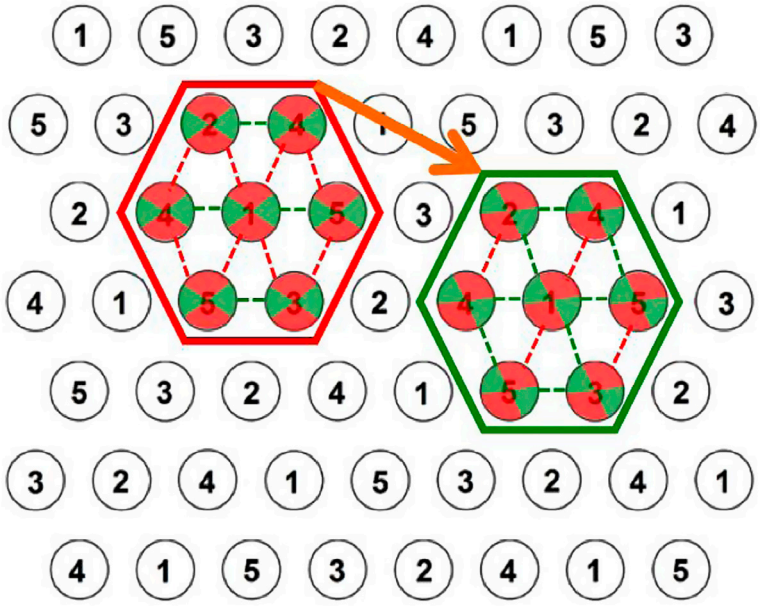

Figure 9. Image depicting the impact of a 26° clockwise rotation of the collagen molecules within a hexagonal closed packed unit. Notes: The red hexagonal unit shows the orientation present within the *Rattus norvegicus* unit cell and the green hexagonal unit shows the configuration after the clockwise rotation. Green and red areas on the collagen molecules illustrate the favourable and unfavourable interactions, respectively, as previously described in Figure 8, with the dashed line similarly coloured showing the interaction orientations of the collagen molecules.

unit after the 26° rotation of the collagen molecules. The impact of such a rotation, seen in Figure 9, results in a doubling in the number of favourable interactions within the hexagonal close-packed unit. In addition to increasing the number of favourable interactions by this 26° rotation, the rotation also had the added effect that the second lowest energy orientation could be adopted, i.e. 124–304°, thus having a significant stabilising effect on the fibril.

To test this hypothesis further a number of explicit rotations within the fibrillar model were simulated. Five different rotations were employed: 90° and 26° anti-clockwise rotations, 0° rotation, and both 90° and 26° clockwise rotations. MD simulations were performed for 60 ns and the energetics analysed. The relative energies with respect to 0° rotation (the orientation present in *Rattus norvegicus*) are reported in Table 1. What is immediately apparent is that employing the same orientation present in the template structure would result in a higher energy fibril compared to all but one of the other rotations tested, i.e. the clockwise 90° rotation which exhibits a higher energy. Our hypothesised 26° rotation exhibits energies 79 and 48 kcal/mol lower than any other order of rotation, for the anti-clockwise and clockwise rotation, respectively. Therefore, it is most probable that the Homo sapiens collagen molecules exhibit a 26° anti-clockwise rotation about their principal axis relative to the template structure, thus experiencing different

Table 1. Energies of the rotated molecules within the fibrillar environment, values reported relative to the template structure of *Rattus Norvegicus*. Standard error in values is 4.5 kcal/mol.

Orientation	Energy (kcal/mol)
90° ACW	−3219.1
26° ACW	−3298.5
0°	0
26° CW	−3267.4
90° CW	+328.6

inter-molecular interactions. However, the homology model's macroscopic structure is the same as the crystal structure, with the same undulations seen within the gap regions.

The current models (Gautieri, Vesentini, Redaelli, & Buehler, 2011; Streeter & de Leeuw, 2010) used to model type I collagen molecules within a fibrillar environment are based upon the use of the structures determined by Orgel et al. in 2006 (Orgel et al., 2006). As we have shown above, this may not be the most accurate assumption, given that a 26° clockwise rotation of the explicit collagen molecule will result in a significant reduction in the energy. However, there are a number of other factors that also dictate the possible orientations within the fibril. For example, mature enzymatic cross-links will likely reduce the number of

possible orientations further. Taking these into account, the current study is probably most comparable to the possible orientations of collagen molecules within a collagen gel or artificial construct, in the absence of mature cross-links. Further studies would be required, using a three- or multi-collagen molecule model with explicit cross-links, to see how these would influence the possible orientations, which is, however, beyond the capabilities of current computational resources. In addition, collagenous tissues are rarely homogeneous; even tendons, which consist of 65–80% (dry weight) collagen contain other ECM molecules such as decorin on the surface of the fibril, which may alter the alignment of the collagen molecules at this interface (22, 311). However, considering the widespread use of collagen gels as tissue engineering scaffolds, we can confidently say that the orientations exhibited in such samples, which have much lower concentrations of other proteins and in the absence of enzymatic cross-links, are likely to be those identified in this study.

Additional studies confirming the cell dimensions of the model, as well as further validation of the new homology model are available in the accompanying supplementary information.

Conclusion

Given the absence of a crystal structure for fibrillar *Homo sapiens* type I collagen, we have developed a homology model using the crystal structure of the *Rattus norvegicus* sequence as a template. The orientation of collagen molecules packed within a collagen fibril potentially has significant implications on the fibril's mechanical and biological properties. However, their determination remains unresolved. In this work, we have used a single point energy scan of 6° rotation increments of two staggered collagen strands, each consisting of a full collagen molecule, a gap region and a short collagen peptide, to identify the low energy interaction regions. The lowest energy orientations identified from this single point energy scan were then used as starting configurations for short MD simulations. The frequency of orientations and energies were computed over these MD simulations to determine the most favourable orientations. Clustering of low energy and high-frequency orientations was observed, in such a way that the interactions were optimum within two small windows of orientation, between 50 and 124° and 232–304°, with respect to the orientation of the collagen molecule in the Orgel crystal structure (21). Given the hexagonal close packing of collagen molecules, we identified that a 26° anti-clockwise rotation of the explicit collagen molecule in current models would result in an increased number of favourable interactions. This proposition was verified by implementation of the 26° anti-clockwise rotation within the fibrillar model, with the energies reported being over 3000 kcal/mol lower in energy than for the orientation present in the template structure, thus making this rotation the most probable orientation of collagen molecules within a human fibril, given the findings from our model.

Methods

Identification of target structure

The web portal version of Standard Protein BLAST, part of the blastp suite from the US National Center for Biotechnology Information, is used for database searching for the template sequence. The accession numbers for the human target sequence used are CO1A1_human (P02452) and CO2A1 (P08123), which includes all the hydroxyproline and hydroxylysine residues, as designated in the post-translational modification section of the entries. A number of databases were used for the search, including the Protein Data Bank (Berman, Henrick, & Nakamura, 2003; Berman et al., 2000; Bernstein et al., 1977), UniProt (The UniProt Consortium, 2014), SwissProt (Bairoch & Apweiler, 2000) and NCBI own libraries (Pruitt, Tatusova, & Maglott, 2007). For gene sequence data only, a manual search was conducted to identify if an available crystal or experimentally derived PDB file is obtainable.

Building the rotation model

The model was constructed using the amino acid sequence for *Homo sapiens*. A straight-chained structure of a collagen molecule with the correct helical propensity was generated using the Triple Helical Building Script (THeBuScr) (Rainey & Goh, 2004). The primary sequences of the collagen peptide chains α1 and α2, translated from the genes COL1A1_human (P02452) and COL1A2_human (P08123) (The UniProt Consortium, 2014), were used as inputs. Proline residues present in the Yyy position of the triplets were considered to be hydroxyproline in the study and hydroxyl-lysine residues stated in the modified residues of the UniProt entry were also included in the sequence.

To accurately apply the rotation of the collagen molecule, the assumption was made that the collagen molecule must be considered as a straight rod, and thus the straight molecule from the THeBuScr programme was used directly. The next stage in the preparation was to align the principal axis (c-axis) of the collagen molecule to a Cartesian origin axis using the Orient script in VMD (Humphrey, Dalke, & Schulten, 1996). In our case we aligned to the x-axis in such a way, that the backbone atoms had almost zero displacement in the y and z components.

From this aligned straight collagen molecule two different strands are created. First the molecule is replicated along the x-axis, preceded by a 36 nm gap region. The strands are then generated by taking a box of length 360 nm and positioning the box at the beginning of the collagen molecule, containing a strand comprising a full collagen molecule, a gap region and a short 110 residue triple helical and telopeptide section. The second strand is generated by placing the end of the box at the end of the collagen molecule, such that this strand also includes a short triple helical region, telopeptide, a gap region and a full collagen molecule. Thus, we have two models aligned to the x-axis, one of a collagen molecule followed by a short collagen snippet and the second a short collagen-like snippet followed by a collagen molecule.

A script was developed to rotate about a chosen axis a selection or all of a protein by a set number of degrees. Through an input of a PDB file the script uses a rotation matrix, or translation vector to alter the coordinates and generate the modified PDB file. The reference model for our simulations is the 0–0°, two strands not rotated but translated by 17 Å. This orientation is the linear version of the orientation from the 2006 Orgel crystal structure, in which the glycine of the alpha1 chain is above the first residues of the other two chains that lay almost in a horizontal plane parallel to the z axis.

Single point energy scan

A script is used to generate 3600 different models for the two strands, orientated independently at 6° increments. The PDB files for all of the models are then fed through the LeaP part of the AmberTools14 to generate the input files, during which the models are solvated using TIP3P water with a buffer of 8.0 Å and the charge of the system is compensated by the addition of chloride ions. The models then undergo a very short 1000 step conjugate gradient minimisation, during which all of the protein atoms are restrained using a force constant of 1000 kcal/(mol Angstrom^2), which is necessary to remove any high energy fluctuations caused by close contacts with the recently added water and Cl^- ions. A further one-step minimisation was conducted to get a single point total energy for the system, which is then used to direct the search in the second stage of this investigation.

Short MD simulations of orientation model

The structures undergo short molecular dynamics simulations, using a cut-off of 8.0 Å, to allow the models to relax further into their most favourable orientations. The models initially undergo 500 steps steepest descent and 2500 steps of conjugate gradient minimisation, followed by a two-stage heating simulation of 20 ps from 0 to 100 K, and 30 ps from 100 to 310 K. Restraints of 200 kcal/(mol Angstrom^2) were applied on all the protein atoms up until this point in the procedure. Finally, the model undergoes a further 100 ps simulation in the NPT ensemble at 1.0 atm pressure. The results of the 100 ps NPT simulation are used to determine the low energy orientations of the collagen molecules within the fibril in two ways: first, through comparing the energies of the respective orientations, and second, through the use of another script, which calculates the orientations from the trajectories of the simulation, allowing us to monitor and compare the frequency of certain orientations. The 100 ps timescale was chosen, both because the large system size prohibits long timescales, but more importantly because the energies have settled to an acceptable degree within this timescale (SI Figure 2).

Solvent and side chain atoms were removed using ptraj, part of the AmberTools14 package, before a second script utilised vector-based mathematics about the x-axis to determine the relative orientations of each of the heavy atoms within the molecule relative to the 0–0° model. More specifically the script calculates the angle of rotation from the dot product of the vector defined by the new position to a point on the x-axis, relative to the position of the same atom in the 0–0° configuration to a point on the x-axis. The results are then averaged over all the atoms to get the relative orientation of the whole molecule, and this is repeated for each of the time points within the trajectory. The functionality and accuracy of the script were tested on a sample collection ($N = 30$) of known rotated models of the same system, with results reported to within 0.345% accuracy. As the code averages the orientation of each constituent atom, to report a single value for the orientation of the entire collagen molecule it is possible that a portion of the collagen molecule may rotate to a greater extent than the rest of the molecule, resulting in an inaccurate value being recorded for the orientation. To check if twisting was occurring, we verified a random selection ($N = 10$) of the orientations from the short MD runs, to calculate the standard error of the mean for the molecular orientation from its constituent atomic angular displacements. It was found that the standard error of the mean had an average value of ±0.30°, and therefore twisting of the molecule was not occurring to a great enough extent to influence the molecular orientation values reported.

Fibrillar collagen simulations detail

MD simulations were performed on all models using SANDER, part of the AMBER12 software package (Case et al., 2012). Periodic boundary conditions were applied to the unit cell in order to simulate the densely packed fibrillar environment. The ff99SB force field was used for the parameterisation of the collagen molecule

with additional terms based on published values for hydroxyproline (Hornak et al., 2006). Water molecules were represented using the TIP3P model (Park, Radmer, Klein, & Pande, 2005). The ff99SB force field was parameterized specifically for biological molecules and describes the non-bonded interactions by pairwise additive Lennard-Jones 6–12 potentials and pairwise additive coulombic potentials. Coulombic potentials were calculated using the Particle Mesh Ewald summation with a cut-off radius of 8.0 Å. An 8.0 Å cut-off was chosen due to the dense fibrillar environment of the collagen molecule within the fibril, which is consistent with many previous modelling studies on fibrillar collagen; the energies converge within the timescale of the simulation (Collier et al., 2015; Klein & Huang, 1999; Marlowe, Singh, & Yingling, 2012; Streeter & de Leeuw, 2010). A time step of 2 fs was adopted for all MD simulations and hydrogen-bond lengths were constrained using the SHAKE algorithm (Ryckaert, Ciccotti, & Berendsen, 1977). Constant temperature and pressure were maintained with the Berendsen algorithm (Berendsen, Postma, van Gunsteren, DiNola, & Haak, 1984) using a barostat time constant of 5.0 ps atm^{-1} and a thermostat time constant of 1.0 ps. As the periodic unit cell has a *c* lattice parameter much larger than *a* and *b*, it is appropriate to use anisotropic coordinate rescaling rather than isotropic rescaling for maintaining constant pressure. This was achieved by making a small modification to the AMBER code, the details of which are discussed in our previous work (Streeter & de Leeuw, 2010). The system density and the potential energy were monitored to determine system convergence.

Supplementary material

The supplementary material for this paper is available online at https://doi.org/10.1080/07391102.2018.1433553.

Acknowledgements

This work made use of the facilities of ARCHER, the UK's national high-performance computing service, which is funded by the Office of Science and Technology through EPSRC's High End Computing Programme.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by a BBSRC grant [grant number BB/K007785]; Via our membership of the UK's HPC Materials Chemistry Consortium EPSRC [grant number EP/L000202].

ORCID

Thomas A. Collier http://orcid.org/0000-0002-6785-9114
Anthony Nash http://orcid.org/0000-0001-8212-0302
Helen L. Birch http://orcid.org/0000-0002-7966-9967
Nora H. de Leeuw http://orcid.org/0000-0002-8271-0545

References

Adams, P. D., Arkin, I. T., Engelman, D. M., & Brünger, A. T. (1995). Computational searching and mutagenesis suggest a structure for the pentameric transmembrane domain of phospholamban. *Nature Structural & Molecular Biology, 2*(2), 154–162. doi:10.1038/nsb0295-154

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403–410. doi:10.1016/S0022-2836(05)80360-2

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research, 28*(1), 45–48. doi:10.1093/nar/28.1.45

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics, 81*(8), 3684–3690.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research, 28*(1), 235–242. doi:10.1093/nar/28.1.235

Berman, H., Henrick, K., & Nakamura, H. (2003). Announcing the worldwide protein data bank. *Nature Structural & Molecular Biology, 10*(12), 980. doi:10.1038/nsb1203-980

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E., Jr, Brice, M. D., Rodgers, J. R., … Tasumi, M. (1977). The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology, 112*(3), 535–542.

Bi, X., Li, G., Doty, S. B., & Camacho, N. P. (2005). A novel method for determination of collagen orientation in cartilage by Fourier transform infrared imaging spectroscopy (FT-IRIS). *Osteoarthritis and Cartilage/OARS, Osteoarthritis Research Society, 13*(12), 1050–1058. doi:10.1016/j.joca.2005.07.008

Brodsky, B. (1999). Hydrogen bonding in the triple-helix. *Proceedings of the Indian Academy of Sciences Chemical Sciences, 111*(1), 13–18.

Brodsky, B., & Ramshaw, J. A. M. (1997). The collagen triple-helix structure. *Matrix Biology, 15*(8–9), 545–554. doi:10.1016/S0945-053X(97)90030-5

Cameron, G. J., Cairns, D. E., & Wess, T. J. (2007). The variability in type I collagen helical pitch is reflected in the D Periodic Fibrillar structure. *Journal of Molecular Biology, 372*(4), 1097–1107. doi:10.1016/j.jmb.2007.05.076

Case, D. A., Darden, T. A., Cheatham, T. E., Simmerling, C. L., Wang, J., Duke, R. E., … Kollman, P. A. (2012). *AMBER 12*. San Fransico: University of California.

Collier, T. A., Nash, A., Birch, H. L., & de Leeuw, N. H. (2015). Preferential sites for intramolecular glucosepane cross-link formation in type I collagen: A thermodynamic study. *Matrix Biology: Journal of the International Society for Matrix Biology, 48*, 78–88. doi:10.1016/j.matbio.2015.06.001

De Simone, A., Vitagliano, L., & Berisio, R. (2008). Role of hydration in collagen triple helix stabilization. *Biochemical and Biophysical Research Communications, 372*(1), 121–125. doi:10.1016/j.bbrc.2008.04.190

Fraser, R. D. B., MacRae, T. P., Miller, A., & Suzuki, E. (1983). Molecular conformation and packing in collagen fibrils. *Journal of Molecular Biology, 167*(2), 497–521. doi:10.1016/S0022-2836(83)80347-7

Galvis, L., Dunlop, J. W. C., Duda, G., Fratzl, P., & Masic, A. (2013). Polarized raman anisotropic response of collagen in tendon: Towards 3D orientation mapping of collagen in tissues. *PLoS ONE, 8*(5), e63518. doi:10.1371/journal.pone.0063518

Gautieri, A., Vesentini, S., Redaelli, A., & Buehler, M. J. (2011). Hierarchical structure and nanomechanics of collagen microfibrils from the atomistic scale up. *Nano Letters, 11*(2), 757–766. doi:10.1021/nl103943u

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., & Simmerling, C. (2006). Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics, 65*, 712–725. doi:10.1002/prot

Hulmes, D. J. S., & Miller, A. (1979). Quasi-hexagonal molecular packing in collagen fibrils. *Nature, 282*(5741), 878–880. doi:10.1038/282878a0

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics, 14*, 33–38.

Kadler, K. E., Baldock, C., Bella, J., & Boot-Handford, R. P. (2007). Collagens at a glance. *Journal of Cell Science, 120*(12), 1955–1958. doi:10.1242/jcs.03453

Kannus, P. (2000). Structure of the tendon connective tissue. *Scandinavian Journal of Medicine & Science in Sports, 10*(6), 312–320. doi:10.1034/j.1600-0838.2000.010006312.x

Keshwani, N., Banerjee, S., Brodsky, B., & Makhatadze, G. I. (2013). The role of cross-chain ionic interactions for the stability of collagen model peptides. *Biophysical Journal, 105*(7), 1681–1688. doi:10.1016/j.bpj.2013.08.018

Klein, T. E., & Huang, C. C. (1999). Computational investigations of structural changes resulting from point mutations in a collagen-like peptide. *Biopolymers, 49*(2), 167–183. doi:10.1002/(SICI)1097-0282(199902)49:2<167::AID-BIP5>3.0.CO;2-5

Kukreti, U., & Belko, S. M. (2000). Collagen fibril D-period may change as a function of strain and location in ligament. *Journal of Biomechanics, 33*(12), 1569–1574.

Kuznetsova, N., Rau, D. C., Parsegian, V. A., & Leikin, S. (1997). Solvent hydrogen-bond network in protein self-assembly: Solvation of collagen triple helices in nonaqueous solvents. *Biophysical Journal, 72*(1), 353–362. doi:10.1016/S0006-3495(97)78674-0

Liao, J., Yang, L., Grashow, J., & Sacks, M. S. (2005). Molecular orientation of collagen in intact planar connective tissues under biaxial stretch. *Acta Biomaterialia, 1*(1), 45–54. doi:10.1016/j.actbio.2004.09.007

Marlowe, A. E., Singh, A., & Yingling, Y. G. (2012). The effect of point mutations on structure and mechanical properties of collagen-like fibril: A molecular dynamics study. *Materials Science and Engineering: C, 32*(8), 2583–2588. doi:10.1016/j.msec.2012.07.044

Masic, A., Bertinetti, L., Schuetz, R., Galvis, L., Timofeeva, N., Dunlop, J. W. C., … Fratzl, P. (2011). Observations of multiscale, stress-induced changes of collagen orientation in tendon by polarized raman spectroscopy. *Biomacromolecules, 12*(11), 3989–3996. doi:10.1021/bm201008b

Moger, C. J., Barrett, R., Bleuet, P., Bradley, D. A., Ellis, R. E., Green, E. M., … Winlove, C. P. (2007). Regional variations of collagen orientation in normal and diseased articular cartilage and subchondral bone determined using small angle X-ray scattering (SAXS). *Osteoarthritis and Cartilage/OARS, Osteoarthritis Research Society, 15*(6), 682–687. doi:10.1016/j.joca.2006.12.006

Monnier, V. M., Sun, W., Sell, D. R., Fan, X., Nemet, I., & Genuth, S. (2014). Glucosepane: A poorly understood advanced glycation end product of growing importance for diabetes and its complications. *Clinical Chemistry and Laboratory Medicine: CCLM/FESCC, 52*(1), 21–32. doi:10.1515/cclm-2013-0174

Orgel, J. P. R. O., Irving, T. C. C., Miller, A., & Wess, T. J. J. (2006). Microfibrillar structure of type I collagen *in situ*. *Proceedings of the National Academy of Sciences, 103*(24), 9001–9005. doi:10.1073/pnas.0502718103

Ottani, V., Raspanti, M., & Ruggeri, A. (2001). Collagen structure and functional implications. *Micron, 32*(3), 251–260. doi:10.1016/S0968-4328(00)00042-1

Pannarale, L., Braidotti, P., D'Alba, L., & Gaudio, E. (1994). Scanning electron microscopy of collagen fiber orientation in the bone lamellar system in non-decalcified human samples. *Cells Tissues Organs, 151*(1), 36–42. doi:10.1159/000147640

Park, S., Radmer, R. J., Klein, T. E., & Pande, V. S. (2005). A new set of molecular mechanics parameters for hydroxyproline and its use in molecular dynamics simulations of collagen-like peptides. *Journal of Computational Chemistry, 26*(15), 1612–1616. doi:10.1002/jcc.20301

Persikov, A. V., Ramshaw, J. A. M., Kirkpatrick, A., & Brodsky, B. (2002). Peptide investigations of pairwise interactions in the collagen triple-helix. *Journal of Molecular Biology, 316*(2), 385–394. doi:10.1006/jmbi.2001.5342

Persikov, A. V., Ramshaw, J. A. M., Kirkpatrick, A., & Brodsky, B. (2005). Electrostatic interactions involving lysine make major contributions to collagen triple-helix stability. *Biochemistry, 44*(5), 1414–1422. doi:10.1021/bi048216r

Petruska, J. A., & Hodge, A. J. (1964). A subunit model for the tropocollagen macromolecule. *Proceedings of the National Academy of Sciences, 51*(5), 871–876.

Pingel, J., Lu, Y., Starborg, T., Fredberg, U., Langberg, H., Nedergaard, A., … Kadler, K. E. (2014). 3-D ultrastructure and collagen composition of healthy and overloaded human tendon: Evidence of tenocyte and matrix buckling. *Journal of Anatomy, 224*(5), 548–555. doi:10.1111/joa.12164

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research, 35*(SUPPL. 1), 501–504. doi:10.1093/nar/gkl842

Rainey, J. K., & Goh, M. C. (2004). An interactive triple-helical collagen builder. *Bioinformatics, 20*(15), 2458–2459. doi:10.1093/bioinformatics/bth247

Ryckaert, J.-P., Ciccotti, G., & Berendsen, H. J. C. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *Journal of Computational Physics, 23*(3), 327–341. doi:10.1016/0021-9991(77)90098-5

Schrof, S., Varga, P., Galvis, L., Raum, K., & Masic, A. (2014). 3D Raman mapping of the collagen fibril orientation in human osteonal lamellae. *Journal of Structural Biology, 187*(3), 266–275. doi:10.1016/j.jsb.2014.07.001

Silver, F. H., Freeman, J. W., & Seehra, G. P. (2003). Collagen self-assembly and the development of tendon mechanical

properties. *Journal of Biomechanics, 36*(10), 1529–1553. doi:10.1016/S0021-9290(03)00135-0

Streeter, I., & de Leeuw, N. H. (2010). Atomistic modeling of collagen proteins in their fibrillar environment. *The Journal of Physical Chemistry B, 114*(41), 13263–13270. doi:10.1021/jp1059984

Streeter, I., & de Leeuw, N. H. (2011). A molecular dynamics study of the interprotein interactions in collagen fibrils. *Soft Matter, 7*(7), 3373–3382. doi:10.1039/c0sm01192d

The UniProt Consortium (2014). Activities at the universal protein resource (UniProt). *Nucleic Acids Research, 42*, D191–D198.

Ugryumova, N., Jacobs, J., Bonesi, M., & Matcher, S. J. (2009). Novel optical imaging technique to determine the 3-D orientation of collagen fibers in cartilage: Variable-incidence angle polarization-sensitive optical coherence tomography. *Osteoarthritis and Cartilage/OARS, Osteoarthritis Research Society, 17*(1), 33–42. doi:10.1016/j.joca.2008.05.005

Yang, W., Chan, V. C., Kirkpatrick, A., Ramshaw, J. A. M., & Brodsky, B. (1997). Gly-pro-arg confers stability similar to gly-pro-hyp in the collagen triple-helix of host-guest peptides. *Journal of Biological Chemistry, 272*(46), 28837–28840. doi:10.1074/jbc.272.46.28837