A molecular dynamics study of the interprotein interactions in collagen fibrils

Ian Streeter[1,2]**, Nora H. de Leeuw**[1,2]

[1] Department of Chemistry, University College London, 20 Gordon Street, London, United Kingdom WC1H 0AJ.

[2] Insitute of Orthopaedics & Musculoskeletal Science, University College London, Brockley Hill, Stanmore, United Kingdom HA7 4LP

Email: i.streeter@ucl.ac.uk; Fax: +44 (0)20 7679 7463

Abstract

Molecular dynamics simulations of collagen are used to investigate at the atomistic level the nature of the interprotein interactions that are present within a collagen fibril, and which are responsible for the fibril's thermodynamic stability. Simulations both of a collagen fibril and of a fully solvated tropcollagen are compared in order to study the interactions that arise between the proteins upon the process of fibrillogenesis. The interactions studied include direct interprotein hydrogen bonds, water-mediated interprotein hydrogen bonds, and hydrophobic interactions. The simulations are used to quantify the number of interprotein interactions that form; to study which functional groups contribute most towards the interactions; and to study the spatial distribution of interprotein interactions throughout the fibril's D period. The processes of collagen fibrillogenesis and protein folding are then compared with each other, because these two physical processes share many similarities in concept, and the latter has been more widely studied. Molecular dynamics simulations of a bacteriophage T4 lysozyme protein, both in its native state and in and unfolded state, are used as an illustrative example of a typical protein folding process, for direct comparison with the collagen simulations.

1 Introduction

Collagen is a protein that plays an important structural role in the extracellular matrix of all species of vertebrates.[1] Under physiological conditions, solvated collagen proteins spontaneously aggregate to form long thin fibrils with diameters in the range 20–500 nm,[2] in a process known as fibrillogenesis. These collagen fibrils are tough and have a high tensile strength, and they are the fundamental building blocks of structural tissues such as bone, cartilage and tendon. The subject of this paper is the intermolecular interactions that arise upon the formation of a collagen fibril, both between the tightly packed neighbouring proteins, and between the proteins and intrafibrillar water molecules. These molecular interactions direct and control the process of fibrillogenesis; they are responsible for the thermodynamic stability of the fibril and for the supramolecular arrangement of proteins within the fibril. Furthermore, these intermolecular forces act to resist any deformation of the fibril under external stress, and they therefore contribute significantly to collagen's material properties.[3,4]

Figure 1 illustrates the structure of a collagen protein and a collagen fibril. The protein, known as a tropocollagen, consists of three polypeptide strands, which are twisted into a long triple helix.[5] At both ends of the rope-like helix there are short non-helical regions known as telopeptides. Note that the protein illustrated in Figure 1 is not shown to scale; its length, which is approximately 300 nm, is 200 times longer than its diameter, which is approximately 1.5 nm, and the non-helical regions account for only 3.5% of the amino acid sequence.

In a collagen fibril the tropocollagens lie in a staggered parallel arrangement, as shown in Figure 1b. The arrangement has a degree of regularity, with a constant periodicity in the axial direction of approximately 68 nm, referred to as a "D" period. Although there are many other types of protein in nature that interact to form higher-order structures, the collagen fibril is notable as one of the few protein aggregates to display a characteristic periodicity and to have such a regular packing arrangement. Because collagen is truly periodic in the axial direction, one can describe a certain feature (e.g. a telopeptide) as residing at a certain position in the D period, with the implication that this feature also appears in every other D period throughout the fibril. Figure 1b shows that each tropocollagen spans approximately four and a half D periods, which gives rise to both an "overlap" region and a "gap" region within each period. In the cross section of a fibril the tropocollagen proteins are packed together in a quasi-hexagonal arrangement,[6,7] and all remaining spaces between the proteins are filled with intrafibrillar water molecules. As collagen fibrils mature *in vivo*, they are strengthened by the

formation of covalently bonded cross links between neighbouring proteins, but the structures initially formed upon fibrillilogenesis are stabilised only by non-covalent interactions.[8] Interprotein covalent cross links are therefore not considered any further in this paper.

The physical process of collagen fibrillogenesis appears to have many similarities in concept to the process of protein folding, which has been studied and documented in much greater detail.[9–14] In the latter process, a linear polypeptide chain folds to form a protein with a well-defined three-dimensional shape that is determined entirely by its amino acid sequence. Similarly, because collagen fibrillogenesis is a spontaneous process, the information needed to build a fibril must be contained entirely within the tropocollagen's amino acid sequence. For both processes, the driving force is the optimisation of hydrophillic and hydrophobic interactions, which sufficiently overcome the entropic barrier associated with forming a more ordered structure. In the case of fibrillogenesis, the aggregation of collagen proteins causes favourable interactions to form between the amino acid side chains of neighbouring molecules. Indeed, it can be shown by analysis of collagen's amino acid sequence that the number of favourable interprotein interactions between two neighbouring tropocollagens is maximised when they are staggered by an integer multiple of the length of a D period, as they are when packed into a collagen fibril.[15]

Despite the apparent similarities between collagen fibrillogenesis and globular protein folding, we demonstrate in this paper that the analogy is limited, and that there are some notable mechanistic differences between the two processes. We have used molecular dynamics (MD) simulations of collagen both in its fibrillar state and in its fully solvated state, so that comparisons can be drawn between the two. To model the collagen fibril we have used a molecular dynamics approach that was described recently,[16] which accounts for the densely-packed local environment within a fibril by exploiting periodic boundary conditions. This new approach to modelling a fibril differs from the majority of previous computational studies in this area, which typically modelled only short fragments of the collagen protein in a fully solvated state, rather than in a fibrillar environment.[17–21]

In order to compare and contrast the processes of fibrillogenesis and protein folding, we also present MD simulations of a bacteriophage T4 lysozyme protein, both in its globular native state and in an unfolded state.[22] This lysozyme has been the focus of many previous computational studies into protein stability and folding,[23–25] and it is used here as an illustrative example of a typical globular protein containing regions of both α-helix and β-sheet. The T4 lysozyme does not contain any

disulfide bonds, and therefore, analogously to a collagen fibril, its tertiary structure is maintained entirely by non-bonded interactions.

2 Methods

Four different systems were modelled using all-atom molecular dynamics simulations: a collagen fibril comprising type I collagen proteins and intrafibrillar water molecules; the same type I tropocollagen protein but in a fully solvated (non-fibrillar) state; a solvated bacteriophage T4 lysozyme protein in its native folded state; and the same solvated lysozyme in a completely unfolded conformation.

All molecular dyanmics simulations were carried out using SANDER, which is part of the AMBER 9 software.[26] The simulations used the protein-specific force field ff99SB[27] and the proteins were solvated using explicit TIP3P water molecules. The non-bonded interactions in the ff99SB force field are described by pairwise additive Lennard-Jones 6-12 potentials and pairwise additive coulombic potentials, which were calculated using the particle-mesh Ewald summation with a cut-off radius of 8.0 Å.[28]

For all four systems, the MD simulations used a 2 fs time step, and bond lengths involving hydrogen were constrained with the SHAKE algorithm.[29] After an initial minimisation to remove any bad contacts, the systems were heated to 310 K (physiological temperature) at constant volume, and then equilibrated at a constant pressure of 1 atm using the Berendsen barostatic algorithm with a 1.0 ps atm^{-1} relaxation time, except where stated otherwise.[30] For the two collagen simulations and for the unfolded lysozyme simulation, the periodic unit cell was extremely long in one direction, and so it was found that anisotropic coordinate rescaling was more appropriate than isotropic rescaling for maintaining constant pressure, which required a small modification to the AMBER code, as described previously.[16] This anisotropic rescaling involved allowing small variations in length to the unit cell's shortest two edges (in the directions perpendicular to the molecule's long axis), but constraining the longest edge (in the direction parallel to the molecule) to a constant length.

2.1 Collagen fibril

The molecular dynamics simulations of the collagen fibril followed a protocol that has recently been described in detail.[16] This protocol achieves the tightly packed arrangement of the collagen proteins in the fibril by using a densely packed unit cell and applying periodic boundary conditions in a manner

that is more commonly associated with simulations of inorganic crystalline solids. An initial system conformation could not be obtained directly from experimental data, because x-ray crystallography of a collagen fibril does not provide atomic resolution. However, a suitable starting conformation was inferred by combining data from high resolution structures of crystallised collagen-mimetic peptides and low resolution structures of the supramolecular fibrillar arrangement.

A possible low-energy conformation of a tropocollagen was first generated using the programme THeBuScr (Triple Helical collagen Building Script), which uses high resolution experimental data to predict and build an all-atom model of a collagen triple helix.[31] However, this programme predicts a perfectly straight conformation of a collagen molecule; it makes no attempt to predict the larger fibrillar structure, or the presence of any bends or kinks in the triple helix. The supramolecular fibrillar structure was therefore taken from a lower resolution x-ray diffraction experiment of a collagen fibril.[32] The latter experiment tells us the overall shape of the collagen molecules, and their arrangement relative to a periodic triclinic unit cell, with edges referred to as *a*, *b* and *c*. It also tells us the approximate conformation of the non-helical telopeptides.[33] The unit cell is extremely long and thin: it measures 677.90 Å in the *c* direction, which represents the fibril's D period, but only 39.97 Å and 26.95 Å in the *a* and *b* directions, respectively, which lie perpendicular to the fibril's long axis. The collagen protein itself is over 3000 Å long, which makes it more than four times longer than the unit cell that defines its periodicity, just like the schematic in Figure 1b. Our all-atom model of the fibril was built so that the local conformation of each tropocollagen was that generated by THeBuScr, but the overall protein shape and fibrillar arrangement was that given by the low resolution x-ray structure.

Water molecules and chloride ions were placed in the periodic unit cell using the programme LEaP, which is part of the Amber Tools software.[26] For each tropocollagen molecule, 11985 intrafibrillar water molecules were added, which equates to a fibrillar water content of 0.75 g water / g collagen. This quantity of water was selected previously using a trial and error approach, such that the crystallographic dimensions of the fibril were conserved during a constant pressure MD simulation; any more water led to an expansion of the crystallographic unit cell, but any less water led to a contraction.[16]

In total, the MD trajectories of the system were calculated for a 60 ns time period, and the atomic coordinates were recorded every 40 ps. The analysis in this paper is based on the final 25 ns of the

trajectory, during which the system energy was stable and the fibrillar conformation was in agreement with experimental measurements, as reported previously.[16] For this system only, the simulations at constant pressure used a barostatic relaxation time of 5 ps atm^{-1}.

2.2 Solvated tropocollagen

The starting conformation of the fully solvated type I tropocollagen was taken directly from the final conformation of the same protein in the collagen fibril simulations described in Section 2.1. For this simulation, the protein was placed in a much larger unit cell containing over 239 thousand water molecules, which prevented any interaction between the protein and its own periodic images. The orthorhombic unit cell was more than 3000 Å long in the direction parallel to the protein, and approximately 50 Å long in each of the other orthogonal directions. To neutralise the charge on the protein, 33 chloride anions were added to low energy positions using the programme LEaP.

The system was heated and equilibrated at constant volume for 20 ps, and then equilibrated at constant pressure for a further 400 ps. For the analysis in this paper, production MD was run for 1 ns, and the atomic coordinates were recorded every 10 ps.

2.3 Native lysozyme

The initial conformation of the bacteriophage T4 lysozyme protein came from the crystal structure in the protein databank (entry 2lzm), which was measured by x-ray diffraction at a resolution of 1.7Å.[22] The unit cell also contained 8500 explicit water molecules and eight chloride ions to neutralise the protein's charge.

The system was heated and equilibrated at constant volume for 80 ps, and equilibrated at constant pressure for a further 800 ps. For the analysis in this paper, production MD was run for 1 ns, and the atomic coordinates were recorded every 10 ps.

2.4 Unfolded lysozyme

The bacteriophage lysozyme protein, with the same amino acid sequence as that described in Section 2.3, was simulated with a completely unfolded conformation. To create the starting structure, a script was used to place backbone carbon and nitrogen atoms such that the polypeptide chain extended in an overall straight line of approximately 580 Å. These atoms all lay in the same plane, and all bond

lengths and bond angles between them were of appropriate values. The programme LEaP was then used to add all other protein atoms, including the amino acid side chains, which pointed alternately in opposite directions in the plane so that neighbouring side chains would not interact. The orthorhombic unit cell was approximately 620 Å long in the direction parallel to the chain, and approximately 35 Å in the orthogonal directions. In addition to the polypeptide, the unit cell contained 27 thousand water molecules and eight chloride ions to neutralise charge.

The system was heated and equilibrated at constant volume for 80 ps, and then equilibrated at constant pressure for 800 ps. For the analysis in this paper, production MD was run for 1 ns, and the atomic coordinates were recorded every 10 ps. In order to maintain the flexible peptide chain in an unfolded conformation, harmonic restraints of 1.0 kcal mol^{-1} $Å^{-1}$ were put on the Cα atoms of the two terminal amino acids in the chain.

3 Results and Disussion

Figure 2 shows typical images of the four different systems that were modelled by molecular dynamics simulations. Figure 2a shows a thin cross sectional slice of the modelled collagen fibril, in which it can be seen how periodic boundary conditions were used to create a continuous array of aligned collagen molecules. The neighbouring tropocollagens were packed in a quasi-hexagonal arrangement, and were sufficiently close together that there were direct interactions between neighbouring proteins. Neighbouring tropocollagens were staggered by integer multiples of the D period, as shown in schematically in Figure 1b, and the unit cell contained both an overlap region and a gap region. The empty intrafibrillar spaces that can be seen between the proteins were filled with water (not shown), with 11985 water molecules per tropocollagen molecule. The supramolecular arrangement of the proteins in the modelled collagen fibril was closely based on x-ray diffraction experiments,[32] and we have previously discussed in detail the unusual aspects of this system in terms of its treatment by MD simulations.[16]

Figure 2b shows a short section of the modelled solvated tropocollagen, which was structurally identical to those in the fibril simulation, but which was surrounded by many more water molecules than in the fibrillar system, such that there were no interprotein interactions. The image shows only a 120 Å section of the rope-like protein, but the complete 3000 Å long molecule was modelled.

Figure 2c shows a representation of the lysozyme protein as it was modelled in its native state,

with α-helix regions shown in purple and β-sheet regions shown in yellow, and Figure 2d shows a short section of the same protein in a fully unfolded state. The latter image shows only a section of 10 amino acids, but the complete peptide of 164 amino acids was simulated. It can be seen that the unfolded conformation was created such that consecutive sidechains pointed in opposite directions, to prevent any sidechain interactions. We acknowledge that this extremely extended conformation is likely to be an over-exaggeration of an unfolded protein,[34] and so our calculations based on this system, presented in Section 3.2, represent an upper bound of the interactions that arise during the protein folding process.

For all four systems, the unit cell volume and energy stabilised at constant values within the time periods allotted for equilibration, as described in Section 2. It was also confirmed that the quantitative measures of interprotein interactions, reported in Sections 3.1 and 3.2, had stabilised within the equilibration period. The proteins retained their secondary and tertiary structure throughout the production MD simulations.

3.1 Hydrogen bonded interactions in the collagen fibril

The strongest individual noncovalent interaction that can arise between two neighbouring collagen proteins is a hydrogen bond, by which we include hydrogen bonds between oppositely charged amino acids, which are sometimes alternatively categorised as electrostatic attractions. It is therefore likely that any interprotein hydrogen bond could contribute to the stability of the fibrillar structure and could influence the fibril's mechanical properties.[3,4,15] The equivalent interactions are well known in the process of globular protein folding, in which intramolecular hydrogen bonds arise in the folded conformation involving the peptide backbone and polar groups in amino acid sidechains.[10,11,35] The relative strength of intramolecular protein hydrogen bonds versus hydrogen bonds to water, and their tendency to assist the process of protein folding, has been debated in the scientific literature, and general opinion on the subject has often changed.[13] The current prevailing opinion is that intramolecular hydrogen bonds in a globular protein do indeed have a net stabilising effect on the folded conformation, and, in some theories of protein folding, backbone–backbone hydrogen bonding is the dominant driving force for this process.[12]

In the collagen fibril, we are particularly interested in two types of hydrogen bonded interaction, which are shown in Figure 3. The first is a direct interprotein hydrogen bond, which in Figure 3a is

shown between a hydroxyproline hydroxyl group in one tropocollagen and a glycine carbonyl group in its neighbour. The second is a bridging water molecule, which links two adjacent tropocollagens by forming hydrogen bonds to each of them simultaneously, and which in Figure 3a is shown linking the two proteins via a hydroxyproline hydroxyl group in each one. It has previously been postulated, based on experimental observations, that these bridging water molecules could be the driving force for fibrillogenesis and a dominant interaction in stabilising the fibrillar structure.[36,37]

A script was used to search for intermolecular hydrogen bonds and water bridges in each set of recorded coordinates from the collagen simulations, taking care to include any that spanned over the periodic boundary of a unit cell. We only counted water bridges for which the hydrogen bonds were cooperative, which is to say the water donated a hydrogen to one protein and accepted a hydrogen from the other. It should be noted that the ff99SB force field does not explicitly contain an energy term for hydrogen bonds, and so their presence in the modelled systems arose purely due to electrostatic and Van der Waals interactions.[27] For the purpose of this analysis, an intermolecular interaction was designated a hydrogen bond if the two electronegative atoms were closer than 3.2 Å and the angle made with the hydrogen was greater than 150°. This definition of a hydrogen bond is consistent with definitions used in previous MD studies of the collagen structure.[3] The absolute number of hydrogen bonds identified depended on the stringency of these criteria. Therefore, the absolute values of the data presented in this section are not as important as the overall trends and patterns they describe. It was confirmed that these reported trends did not change significantly when a less stringent definition of a hydrogen bond was used.

In the MD simulations of the fully solvated collagen protein, each tropocollagen on average formed 5034 hydrogen bonds with surrounding water molecules at any one time. In the simulated collagen fibril, each tropocollagen on average formed 4625 hydrogen bonds to water and 392 hydrogen bonds to another protein. The tropocollagen therefore formed a similar total number of intermolecular hydrogen bonds in both environments, but within the fibril only 7.8% of these were direct protein–protein interactions. These results contrast with the number of equivalent interactions found in globular protein folding, for which there is a greater propensity for polar functional groups to form intramolecular protein–protein hydrogen bonds. For example, McDonald and Thornton reported that in the central cores of the proteins they studied, only 1.3% of backbone nitrogens and 1.8% of backbone oxygens failed to form any intramolecular hydrogen bond.[35] A key feature of a protein folding

process is that the peptide chain substitutes hydrogen bonds with water molecules for intramolecular hydrogen bonds as water molecules are excluded from the protein core. Within the protein core, it is important to satisfy as many intramolecular hydrogen bonds as possible, because each one stabilises the folded conformation by as much as 3 kcal mol^{-1}, while the net change in free energy of the folding process is typically in the range -5 to -15 kcal mol^{-1}.[13] In comparison to protein folding, in the process of collagen fibrillogenesis a high proportion of the hydrogen bonds with water are retained, because the fibril is highly hydrated and intrafibrillar water molecules permeate the entire structure. These water molecules can ensure that few potential hydrogen bonds are left unsatisfied, and so in the collagen fibril there is less impetus than in globular proteins to maximise the number of protein-protein hydrogen bonds.

Table 1 shows a comparison of the different types of interprotein attraction within the modelled collagen fibril. The values shown correspond to the number of hydrogen bonds or water bridges in a fibril at any one time, as a ratio to the number of tropocollagen molecules in the fibril, and averaged over all recorded sets of coordinates. The table shows whether the bonding hydrogen was donated from an amino acid sidechain to another sidechain (R:R), from a backbone nitrogen to a sidechain (N:R), from a sidechain to a backbone oxygen (R:O) or from a backbone nitrogen to a backbone oxygen (N:O). The final column in Table 1 shows the relative quantities of different intramolecular hydrogen bonds in folded globular proteins, as recorded by Stickle et al., based on a census of the x-ray structures of 42 different proteins.[10]

The data in Table 1 show that in a globular protein intramolecular hydrogen bonds between backbone atoms dominate over any other type of intramolecular hydrogen bond (68.1%). These backbone–backbone interactions have been implicated by some researchers as the dominant driving force of protein folding.[12,13] Conversely, in a collagen fibril there are very few backbone–backbone interprotein hydrogen bonds between neighbouring tropocollagens (2.8%). Furthermore, it is found that of these few backbone–backbone hydrogen bonds in the fibril, more than half of them were made by the proteins' telopeptides; backbone–backbone hydrogen bonds between adjacent triple helices are even more rare. Instead, interprotein hydrogen bonds in the fibril are dominated by sidechain–sidechain (R:R) interactions, with significant contributions from sidechain–backbone (R:O and N:R) interactions. Amino acid side chains in a collagen triple helix are arranged so that they point radially outwards, away from the centre of the molecule (apart from glycine residues), while the protein

backbones lie close to the tropocollagen's central axis.[5] The protein backbones of two neighbouring tropocollagens therefore lie too far apart to form an interprotein hydrogen bond, and they do not approach any closer due to steric hinderance by the sidechains. Conversely, the outward-facing sidechains of neighbouring tropocollagens come into direct contact in the fibril, and so sidechain–sidechain hydrogen bonds abound.

It can be seen from Table 1 that the number of direct interprotein hydrogen bonds in the collagen fibril exceeded the number of bridging water molecules, but not excessively. Our interest in bridging water molecules arose because it had previously been suggested that they are the dominant interprotein attraction within the collagen fibril.[36,37] The data in Table 1 support the claims that water-mediated interprotein bonds exist, but in terms of their quantity they do not appear to be more important than direct hydrogen bonds. Compared to direct hydrogen bonds, a higher fraction of these water-mediated interactions were backbone–backbone or backbone–sidechain interactions. This is presumably because water bridges can span a greater distance than direct hydrogen bonds, so interactions with the central protein backbone were more feasible.

It is interesting that between neighbouring collagen proteins the sidechain to backbone (R:O) interactions outnumber the backbone to sidechain (N:R) interactions in terms of both direct hydrogen bonds and bridging water molecules, whereas the equivalent intramolecular interactions in folded proteins are approximately equal in number. One possible reason is that proline and hydroxyproline residues, which together account for 21% of the amino acids in collagen, do not have backbone hydrogen atoms to donate, and glycine residues in the collagen triple helix, which account for 32% of the entire protein, cannot donate their backbone hydrogens because they are already involved in intramolecular interactions that serve to stabilise the protein's triple helix conformation.[5]

Table 2 shows the relative contributions of different hydrophillic amino acid sidechains to the interprotein hydrogen bonds and interprotein water bridges in the modelled collagen fibril, averaged over all recorded sets of coordinates. The amino acids are listed in the order of the number of times they appear in the primary structure of a tropocollagen, *n*. Charged amino acids are especially prolific at forming interprotein hydrogen bonds, with arginine (Arg) and lysine (Lys) contributing the most as hydrogen bond donors, and with glutamate (Glu) and aspartate (Asp) contributing the most as hydrogen bond acceptors. This complements the experimental findings of Stickle et al. for the intramolecular hydrogen bonds in folded globular proteins, in which it was reported that 41% of all

sidechain donators were either Lys or Arg, and 59% of all sidechain acceptors were either Asp or Glu.[10] From the data in Table 2, charged amino acids are also more likely than neutral amino acids to form a water-bridged interprotein interaction. However, this tendency is less pronounced than for direct interprotein hydrogen bonds; i.e. the relative contributions of Arg, Lys, Glu and Asp are less in the last two columns of Table 2 than in the preceding two columns.

Figure 4a shows the distribution of direct interprotein hydrogen bonds throughout the D period, and Figure 4b shows the equivalent distribution for interprotein water bridges. These plots show that both types of interprotein attraction tend to form in clusters, rather than uniformly throughout the D period. For example, both types of interaction have a peak at approximately 200 Å with local minima on either side at approximately 170 Å and at 235 Å. This can be explained by the spatial distribution of charged amino acids in the fibril, which we have established have the strongest propensity to form interprotein interactions. Figure 4c shows the distribution of the charged residues throughout the D period. Note that because this plot is the sum of both positive and negative amino acids, it does not represent the net local charge through the fibril. It has been described previously in the literature that charged residues tend to occur in clusters in the D period, and it is this periodic distribution that guides tropocollagens into their correct axial positions during the process of fibrillogenesis.[15] Comparing the plots in Figure 4, it can be seen that the regions of increased interprotein hydrogen bonded interaction correspond to regions of the fibril with a large number of charged amino acids. It is interesting that the clustering of charged residues, which is necessary for maintaining the fibril's regular structure, results in some parts of the fibril that are held together tightly by hydrogen bonds and water bridges, and other regions where these features are largely absent.

It can be seen from Figure 4 that the gap region of the fibril has a lower average density of both types of interprotein attraction than the overlap region: the average density of interprotein hydrogen bonds is 44% lower in the gap region and the average density of interprotein water bridges is 34% lower. These reductions in the number of interactions exceed the value of 20% that might be expected based on the relative number of proteins in the two regions (see Figure 1b), and moreover they exceed the value of 11% that might be expected based on the relative number of charged amino acids in either region. This is presumably because the tropocollagens in the overlap region are packed together more tightly, and so interprotein attractions can form more easily.

3.2 Hydrophobic interactions in the collagen fibril

The expression 'hydrophobic interaction' in this context refers to the tendency for nonpolar groups in organic molecules to cluster together in aqueous solvent, thereby reducing the surface area of the interface between hydrophobic groups and water molecules. The common interpretation of the hydrophobic effect is that water molecules at room temperature form ordered cage-like structures around hydrophobic molecules, because this maximises the number of water–water hydrogen bonds, but it does so at the expense of a reduction in entropy.[9,38] The nonpolar regions of a solvated protein tend to cluster together to form a central hydrophobic core, because this reduces the size of the hydrophobic interface exposed to water, and so there is a corresponding gain in entropy from the surrounding water molecules.

In this section, we study the degree to which interprotein hydrophobic interactions were present in the collagen fibril MD simulations, because this could potentially be a source of stability of the fibrillar structure, and a driving force for fibrillogenesis. Tropocollagen proteins contain many hydrophobic amino acids, and there is evidence that the packing of these proteins into a fibril acts to optimise the alignment of hydrophobic sidechains between neighbouring tropocollagens.[15] Furthermore, the process of fibrillogenesis is thermodynamically driven by a favourable change in entropy,[39] which could implicate hydrophobic interactions as a major driving force.

There are differing opinions in the scientific literature concerning the relative importance of the hydrophobic effect compared with other stabilising factors in driving the process of protein folding.[9,13] There are also differing opinions concerning the precise mechanistic details that underlie the hydrophobic effect.[38] Nonetheless, hydrophobic interactions are likely to at least contribute towards the thermodynamic stability of a folded protein or a collagen fibril,[40] and we assume for the calculations in this section that these stabilising interactions can be identified by a reduction in size of the protein's hydrophobic interface. The hydrophobic interfaces of the modelled systems are defined here as all sidechains of the seven most highly ranked amino acids in the hydrophobicity scale of Kyte and Doolittle, which all have a positive free energy change of solvation.[41] These seven hydrophobic residues are Ile, Val, Leu, Phe, Cys, Met, and Ala.

For all four of the modelled systems shown in Figure 2, we calculated the distance of each water molecule in the system to its nearest hydrophobic amino acid sidechain. More specifically, the distance measured was from the oxygen atom of the water molecule to the nearest hydrophobic sidechain

heavy atom (non-hydrogen). Figure 5 shows the distributions of these measured distances, averaged over all recorded sets of coordinates. Figure 5a shows the distribution of water molecules for the collagen fibril and for the fully solvated tropocollagen, normalised per collagen molecule, and Figure 5b shows the equivalent distributions for the lysozyme protein in its native state and in its denatured state. For all four systems, the peak centred at approximately 3.7 Å corresponds to water molecules that occupied the first hydration shell around the hydrophobic sidechains, and which, according to the theory of the hydrophobic effect, were therefore in an entropically unfavourable state. The fibrillar system differs from the others in that the distribution decreases towards zero at distances greater than 6.3 Å. This is because the fibril contained only a limited amount of water interspersed between densely packed collagen proteins, and so no water molecule was ever further than 12.5 Å from a hydrophobic sidechain. The other three distributions trend upwards in this region because of the increasing volume of the solvation shell at larger radial distances from the protein.

It can be seen from Figure 5a that in the collagen fibril there were fewer water molecules (per collagen molecule) in close proximity to hydrophobic groups, compared to the fully solvated tropocollagen. To quantify this, the number of water molecules in the first hydration shell around the hydrophobic sidechains, defined here as being within 5 Å, was 6662 for the fully solvated tropocollagen, and 4809 per tropocollagen in the fibril. This is equivalent to a decrease of 27.8% water molecules at a hydrophobic interface upon formation of a collagen fibril from the fully solvated proteins. From Figure 5b it can be seen that there is a much more significant difference in the number of water molecules at a hydrophobic interface between the folded and unfolded conformations of lysozyme. To quantify this, the number of water molecules in the first hydration shell around the lysozyme's hydrophobic sidechains was 1062 for the unfolded protein, and 247 for the native protein, which is a decrease of 76.7%. This decrease occurs because the majority of lysozyme's hydrophobic sidechains were buried in the central hydrophobic core of the folded protein, whereas they were fully exposed to water in the unfolded conformation. This is a typical example the hydrophobic effect in stabilising the native conformation of a globular protein, with an associated decrease in surface area of the exposed hydrophobic interface. It is interesting that we observe similarities between the processes of protein folding and fibrillogenesis in that both processes result in fewer water molecules at a hydrophobic interface, although this effect is much smaller in magnitude for fibrillogenesis. This supports the notion that the hydrophobic effect could contribute towards the stability of a collagen fibril and towards

driving the process of fibrillogenesis.

We identified two main contributions to the decrease in number of waters at the hydrophobic interface in the collagen fibril compared to solvated tropocollagens. Firstly, because of the dense fibrillar packing arrangement visible in Figure 2a, many of the outward-facing hydrophobic sidechains from one tropocollagen were in direct contact with other neighbouring tropocollagens. This effectually reduced the surface area of the exposed hydrophobic interface, and is conceptually equivalent to the aforementioned packing of hydrophobic sidechains of the lysozyme protein into a central hydrophobic core.

The second contribution was due to intrafibrillar water molecules that were sandwiched in between hydrophobic groups from two (or more) different tropocollagens. These sandwiched water molecules were present when the outward-facing hydrophobic sidechains from two neighbouring proteins did not extend far enough to make contact, but instead left a gap that was filled by a water molecule. Figure 6 shows an example of such a water molecule, in this case sandwiched between a phenylalanine and an isoleucine sidechain from two neighbouring tropocollagens. Both of the highlighted hydrophobic sidechains are approximately 3.6 Å from the water molecule, so the hydrophobic interface in this image is no smaller than if the tropocollagens had been fully solvated. However, because the sandwiched water molecule constitutes the first hydration shell of both hydrophobic sidechains, there is an overall decrease in the number of water molecules at the hydrophobic interface, compared to fully solvated tropocollagens. To quantify the magnitude of this effect in the fibril, we identified 689 water molecules per collagen tropocollagen on average that were within 5 Å of two (or more) hydrophobic sidechains from different tropocollagens. These sandwiched water molecules therefore accounted for 14.3% of all water molecules at a hydrophobic interface within the fibril. The common interpretation of the hydrophobic effect is that a water molecule suffers a loss of entropy when it is in close proximity to a hydrophobic interface, because it may no longer freely rotate as it does in the bulk phase.[9,38] There is likely to be a subsequent loss of entropy if that same water molecule comes into proximity to a second hydrophobic interface simultaneously, as is the case in Figure 6. If the loss of entropy due to the second hydrophobic interface is less than the loss of entropy associated with the first hydrophobic interface, then the presence of sandwiched water molecules is likely to have a net stabilising effect on the collagen fibril.

To further investigate the extent to which hydrophobic regions of the modelled protein systems

were shielded from water, we counted the number of water molecules in the first hydration shell around each hydrophobic sidechain, defined as being within 5 Å of a hydrophobic heavy atom. More specifically, for each integer number of water molecules, we counted the number of hydrophobic sidechains in the system that were observed to contain that many water molecules in its first hydration shell, and then averaged the distribution over all recorded timesteps. Figure 7a shows the recorded distributions for the fully solvated tropocollagen and for the collagen fibril, normalised per collagen molecule, and Figure 7b shows the equivalent distributions for the lysozyme protein in its folded and unfolded conformations. The distributions span a wide range of integer numbers of water molecules along the abscissa, because our method of counting did not discriminate between large sidechains (e.g. phenylalanine) and small sidechains (e.g. alanine).

For the collagen systems, the distribution in Figure 7a shifts along the abscissa to lower values for the collagen fibril compared to the fully solvated tropocollagen, which indicates that hydrophobic groups in the fibril were in contact with fewer water molecules on average, because they were somewhat shielded by neighbouring proteins. The lysozyme protein shows an equivalent shift of the distribution in Figure 7b towards fewer water molecules for the folded conformation compared to the unfolded conformation, and the magnitude of this shift is much greater than that observed between the two collagen systems. For the folded lysozyme, the distribution in Figure 7b has its highest value at zero water molecules, which indicates that many of the hydrophobic sidechains were completely buried in the central hydrophobic core of this protein, and therefore were not in contact with any water molecules at all. For the lysozyme protein, the average number of water molecules in contact with a hydrophobic sidechain decreased from 19.4 to 5.0 upon folding of the protein. For a tropocollagen protein, the average number of water molecules in contact with its hydrophobic sidechains decreased from 12.9 to 10.9 upon the formation of a fibril. From these data it can be seen that the formation of a collagen fibril tends to reduce the surface area of the exposed hydrophobic interface, but the effect is small in magnitude when compared to the hydrophobic effects that are associated with the folding of globular proteins.

4 Conclusions

The interprotein interactions that arise between tropocollagens have a number of direct implications for the physical properties of a collagen fibril. For example, they determine the thermodynamic sta-

bility of the fibril, relative to fully solvated tropocollagens, and they therefore drive the process of fibrillogenesis. They act to oppose any deformation to the collagen fibril, and so they are ultimately responsible for the strength and toughness of collagenous tissues.[3,4] Finally, because the tropocollagen contains a spatial distribution of both hydrophobic and hydrophillic regions, the interprotein interactions determine the periodic ordered supramolecular arrangement of proteins within the fibril.[15]

The simulations reported in this paper have allowed a study of the nature of interprotein interactions at the atomistic level in a collagen fibril. In particular, the simulations have allowed us to make predictions of the number of intermolecular hydrogen bonds in a fibril; of the functional groups responsible for forming interprotein hydrogen bonds; of the behaviour of water molecules in mediating interprotein interactions between polar groups; and of the extent to which hydrophobic sidechains in the fibril are buried from water molecules. The simulations did not tell us the extent to which any individual interaction stabilises the fibrillar structure, but we anticipate that experimental data on collagen fibrils can now be interpreted in terms of the range of interactions that we have identified.

The process of fibrillogenesis has similarities in concept with the process of protein folding, in that both processes are driven by the optimisation of hydophillic and hydrophobic interactions between the constituent amino acids. The simulations reported here have highlighted both similarities and differences between the interactions that drive the two processes. For example, when a protein folds, polar functional groups buried at the centre of the protein have a strong impetus to form intramolecular hydrogen bonds; conversely, in a collagen fibril, polar groups at the centre of the fibril are in contact with relatively large quantities of intrafibrillar water, and so they exhibit less tendency to form protein–protein hydrogen bonds. The majority of intramolecular hydrogen bonds in a folded proteins are backbone–backbone interactions, but the majority of interprotein hydrogen bonds in a collagen fibril are sidechain–sidechain interactions. The processes of protein folding and collagen fibrillogenesis both act to shield hydrophobic sidechains from water molecules, but this effect is far smaller in magnitude for fibrillogenesis.

There has been much discussion in the scientific literature about whether the dominant stabilising factor of a protein's folded conformation is the hydrophobic effect or intramolecular hydrogen bonds.[9,12,13,40] The MD simulations in this paper suggest that the interprotein interactions in a collagen fibril, both hydrophobic and hydrophillic, are generally less pronounced than the equivalent

intramolecular interactions that are found in globular proteins. It is therefore very possible that the processes of protein folding and collagen fibrillogenesis are not primarily driven by the same types of interaction.

The collagen fibril simulations in this paper made use of a modelling procedure that accounted for the local environment within a collagen fibril by using periodic boundary conditions.[16] The data reported here could not have been collected if our simulations had followed the methods used in most other reported MD simulations of collagen, which have generally modelled only small fragments of collagen molecules in a fully solvated state.[17–21] Collagen fibrils can vary in their quantity of intrafibrillar water under different experimental or physiological conditions,[42] but the simulations in this paper used just a single water content of 0.75 g water / g collagen. In future work we intend to use similar MD simulations to study how the relative quantities of the different interprotein interactions differ with varying water content.

References

[1] Ottani, V.; Martini, D.; Franchi, M.; Ruggeri, A.; Raspanti, M. *Micron* **2002,** *33,* 587–596.

[2] Parry, D. A. D.; Craig, A. S. *Nature* **1979,** *282,* 213–215.

[3] Zhang, D.; Chippada, U.; Jordan, K. *Ann. Biomed. Eng.* **2007,** *35,* 1216–1230.

[4] Gautieri, A.; Vesentini, S.; Redaelli, A.; Buehler, M. J. *Int. J. Mater. Res.* **2009,** *100,* 921–925.

[5] Bhattacharjee, A.; Bansal, M. *IUBMB Life* **2005,** *57,* 161–172.

[6] Hulmes, D. J. S.; Miller, A. *Nature* **1979,** *282,* 878–880.

[7] Wess, T. J.; Hammersley, A. P.; Wess, L.; Miller, A. *J. Mol. Biol.* **1998,** *275,* 255–267.

[8] Bailey, A. J.; Paul, R. G.; Knott, L. *Mech. Ageing Dev.* **1998,** *106,* 1–56.

[9] Dill, K. A. *Biochemistry* **1990,** *29,* 7133–7155.

[10] Stickle, D. F.; Presta, L. G.; Dill, K. A.; Rose, G. D. *J. Mol. Biol.* **1992,** *226,* 1143–1159.

[11] Eswar, N.; Ramakrishnan, C. *Protein Eng.* **2000,** *13,* 227–238.

[12] Rose, G. D.; Fleming, P. J.; Banavar, J. R.; Maritan, A. *Proc. Natl. Acad. Sci. USA* **2006,** *103,* 16623–16633.

[13] Bolen, D. W.; Rose, G. D. *Annu. Rev. Biochem.* **2008,** *77,* 339–362.

[14] Zhang, J.; Li, W.; Wang, J.; Qin, M.; Wu, L.; Yan, Z.; Xu, W.; Zuo, G.; Wang, W. *IUBMB Life* **2009,** *61,* 627–643.

[15] Hulmes, D. J. S.; Miller, A.; a. D. Parry, D.; Piez, K. A.; Woohead-Galloway, J. *J. Mol. Biol.* **1973,** *79,* 137–148.

[16] Streeter, I.; de Leeuw, N. H. *J. Phys. Chem. B* **2010,** (in press).

[17] Klein, T. E.; Huang, C. C. *Biopolymers* **1999,** *49,* 167–183.

[18] Mooney, S. D.; Kollman, P. A.; Klein, T. E. *Biopolymers* **2002,** *64,* 63–71.

[19] Lorenzo, A. C.; Caffarena, E. R. *J. Biomech.* **2005,** *38,* 1527–1533.

[20] Ravikumar, K. M.; Hwang, W. *Proteins: Struct., Funct., Bioinf.* **2008,** *72,* 1320–1332.

[21] Gurry, T.; Nereberg, P. S.; Stultz, C. M. *Biophys. J.* **2010,** *98,* 2634–2643.

[22] Weaver, L. H.; Matthews, B. W. *J. Mol. Biol.* **1987,** *193,* 189–199.

[23] Veenstra, D. L.; Kollman, P. A. *Protein Eng.* **1997,** *10,* 789–807.

[24] Dong, F.; Zhou, H. X. *Biophys. J.* **2002,** *83,* 1341–1347.

[25] Ghosh, A.; Brinda, K. V.; Vishveshwara, S. *Biophys. J.* **2007,** *92,* 2523–2535.

[26] Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005,** *26,* 1668–1688.

[27] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct. Funct. Bioinf.* **2006,** *65,* 712–725.

[28] Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993,** *98,* 10089–10092.

[29] Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977,** *23,* 327–341.

[30] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984,** *81,* 3684–3690.

[31] Rainey, J. K.; Goh, M. C. *Bioinformatics* **2004,** *20,* 2458–2459.

[32] Orgel, J. P. R. O.; Irving, T. C.; Miller, A.; Wess, T. J. *Proc. Natl. Acad. Sci. USA* **2006,** *103,* 9001–9005.

[33] Orgel, J. P.; Wess, T. J.; Miller, A. *Structure* **2000,** *8,* 137–142.

[34] Shortle, D. *Curr. Opin. Struct. Biol.* **1993,** *3,* 66–74.

[35] McDonald, I. K.; Thornton, J. M. *J. Mol. Biol.* **1994,** *238,* 777–793.

[36] Leikin, S.; Rau, D. C.; Parsegian, V. A. *Nat. Struct. Mol. Biol.* **1995,** *2,* 205–210.

[37] Kuznetsova, N.; Chi, S. L.; Leikin, S. *Biochemistry* **1998,** *37,* 11888–11895.

[38] Southall, N. T.; Dill, K. A.; Haymet, A. D. J. *J. Phys. Chem. B* **2002,** *106,* 521.

[39] Kadler, K. E.; Hojima, Y.; Prockop, D. J. *J. Biol. Chem.* **1987,** *260,* 15696–15701.

[40] Pace, C. N.; Shirley, B. A.; McNutt, M.; Gajiwala, K. *FASEB J.* **1996,** *10,* 75–83.

[41] Kyte, J.; Doolittle, R. F. *J. Mol. Biol.* **1982,** *157,* 105–132.

[42] Sivan, S.; Merkher, Y.; Wachtel, E.; Ehrlich, S.; Maroudas, A. *J. Orthop. Res.* **2006,** *24,* 1292–1298.

Tables

	collagen fibril interprotein hydrogen bonds	collagen fibril interprotein water bridges	folded globular proteins intramolecular hydrogen bonds[10]
R:R	122.6 (62.6%)	53.9 (33.0%)	10.6%
N:R	19.4 (9.9%)	30.4 (18.6%)	10.4%
R:O	48.4 (24.7%)	62.4 (38.3%)	10.9%
N:O	5.6 (2.8%)	16.5 (10.1%)	68.1%
total	196.0	163.1	

Table 1: A comparison of the different types of hydrogen bonded interactions in the collagen fibril. The first two columns of data refer to the average number of interprotein attractions per collagen protein in the simulated fibril at any point in time. The final column shows experimental data for equivalent interactions in folded globular proteins recorded by Stickle et al.[10] Percentages are shown such that each column sums to 100%.

sidechain	*n*	hydrogen bond donor	hydrogen bond acceptor	water bridge donor	water bridge acceptor
Hyp	328	11.2%	7.0%	21.2%	11.1%
Arg (+)	158	47.6%	0.0%	39.7%	0.0%
Glu (−)	145	0.0%	54.1%	0.0%	45.6%
Ser	131	12.2%	3.5%	8.5%	4.9%
Lys (+)	110	15.9%	0.0%	14.1%	0.0%
Asp (−)	90	0.0%	26.7%	0.0%	23.5%
Gln	83	5.9%	5.8%	8.0%	8.9%
Thr	64	3.1%	0.7%	3.3%	2.5%
Asn	48	1.8%	1.5%	4.1%	2.4%
Tyr	14	1.1%	0.5%	0.4%	0.8%
His	13	1.1%	0.2%	0.7%	0.4%

Table 2: Relative contributions of different amino acid sidechains to the interprotein attractions. Hyp is hydroxyproline; all other residues are standard. Amino acids are listed in the order of the number of times they appear in a collagen protein, *n*, and charged amino acids are indicated in parentheses. The data are presented so that each column sums to 100%

Figures

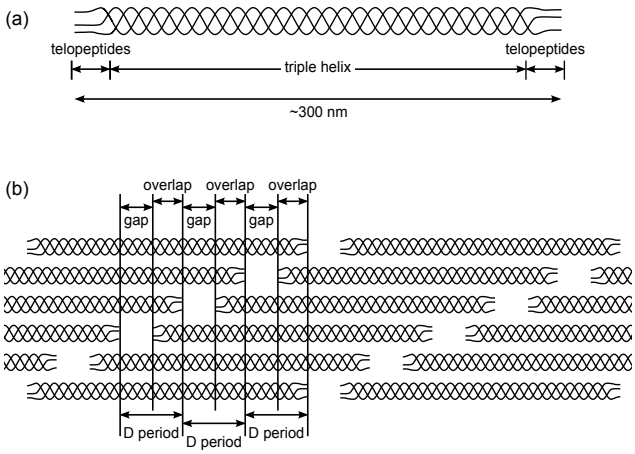

Figure 1: Schematic of (a) a single tropocollagen protein; (b) the supramolecular arrangement of tropocollagens in a fibril. Tropocollagens are illustrated as a triple helix with nonhelical telopeptides, not shown to scale.

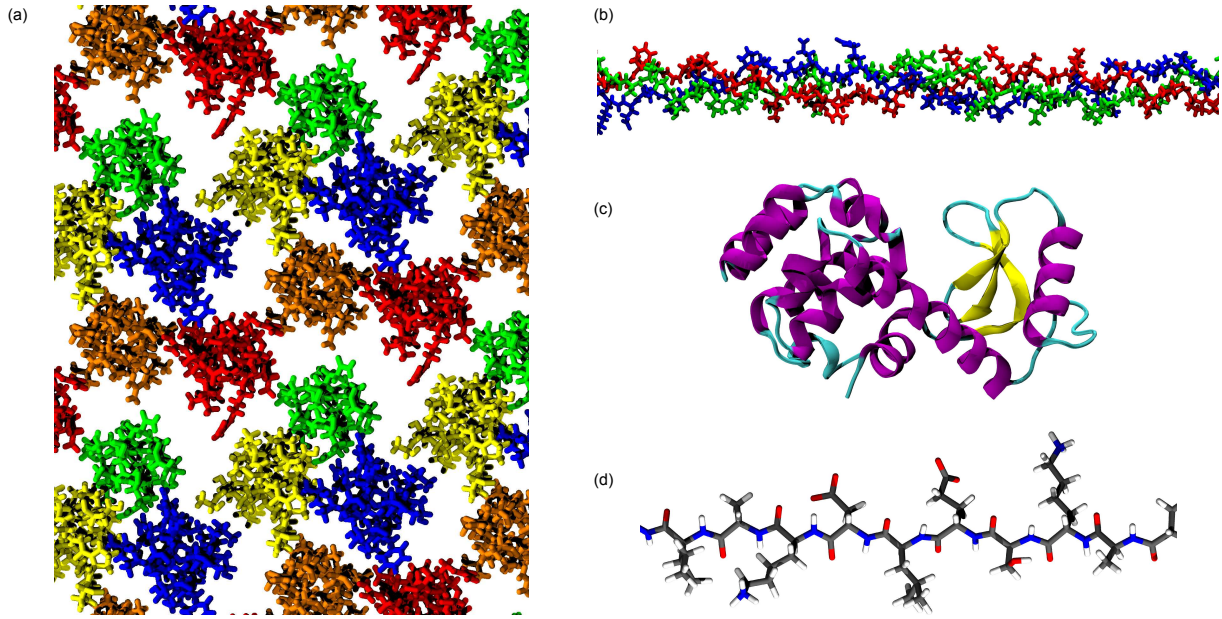

Figure 2: Images of the four modelled systems at the completion of the MD simulations: (a) the collagen fibril, shown as a thin cross sectional slice with each tropocollagen in a different colour, (b) a short section of the fully solvated tropocollagen, with each peptide chain shown in a different colour, (c) the lysozyme protein in its native state, (d) a short section of the lysozyme protein in an unfolded state. Water molecules have been omitted from all images for clarity.

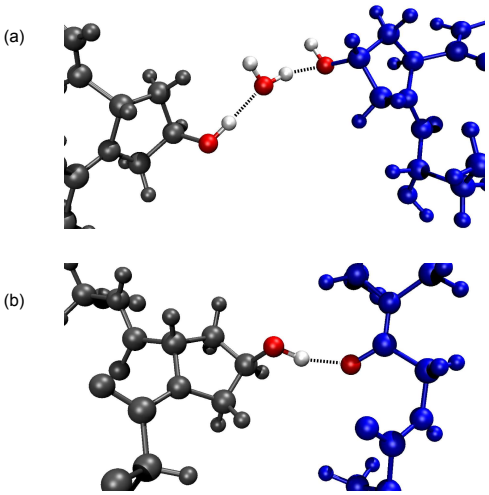

Figure 3: Two types of hydrogen bonded interprotein interaction. (a) A direct interprotein hydrogen bond. (b) A bridging water molecule. The two tropocollagens are shown as grey and blue, and the relevant polar groups are highlighted in red and white.

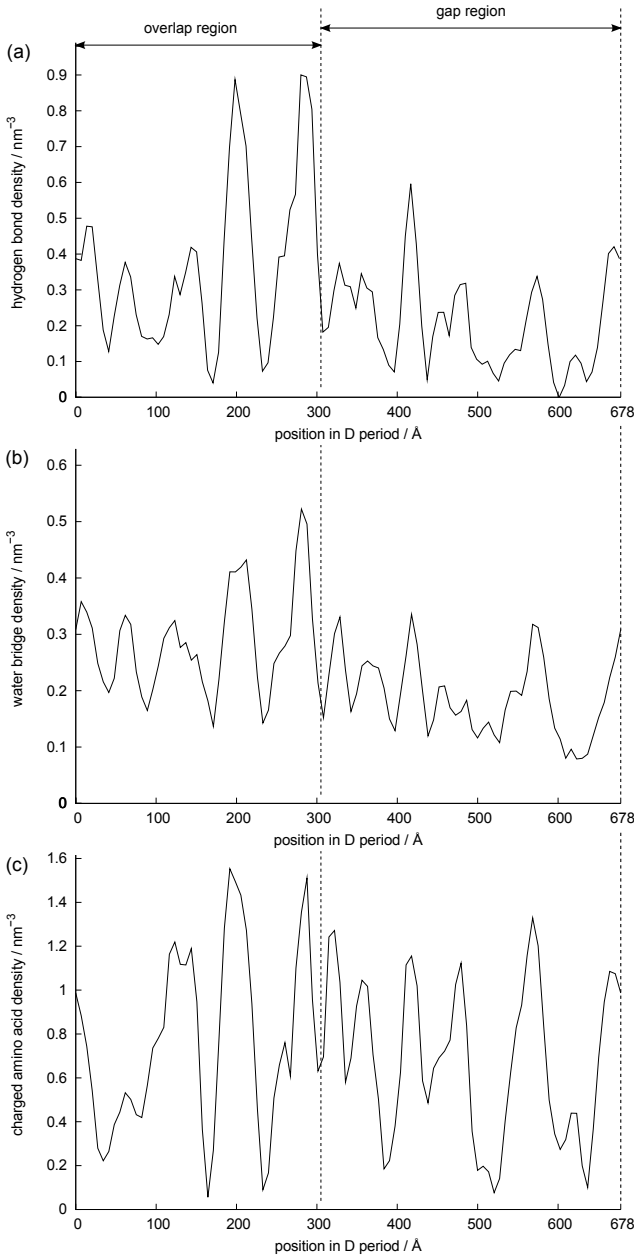

Figure 4: (a) The density of interprotein hydrogen bonds throughout a D period. (b) The density of bridging water molecules throughout the fibril. (c) The density of charged amino acids throughout a D period. All plots were calculated from the MD simulations and averaged over time. The units of the ordinates are number per nm^3 of fibril.

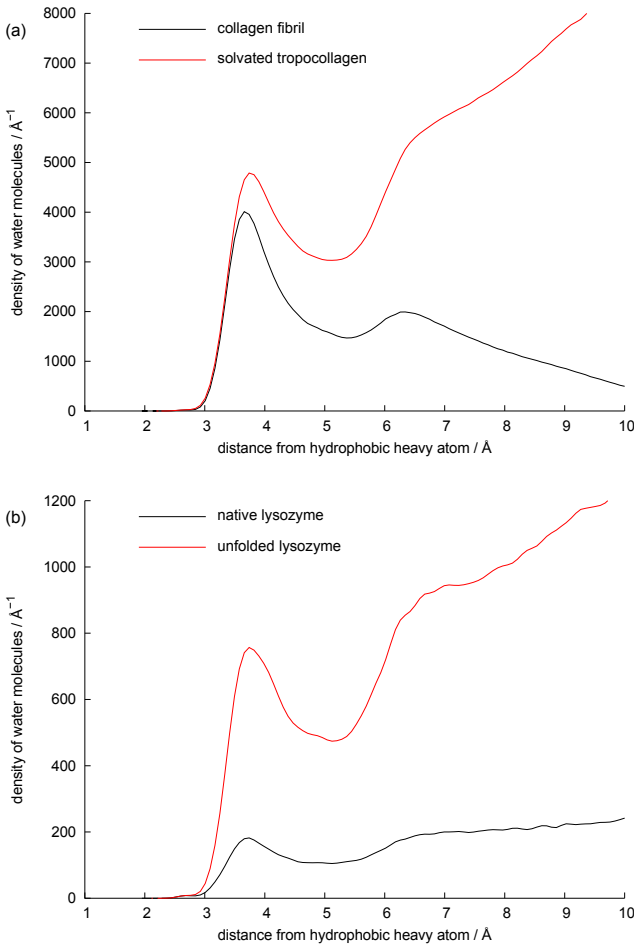

Figure 5: Distributions showing the distance of each water molecule to the nearest hydrophobic sidechain. (a) The collagen fibril and fully solvated tropocollagen. (b) The lysozyme protein in its native and unfolded conformations. The distributions have been averaged over all recorded time steps.

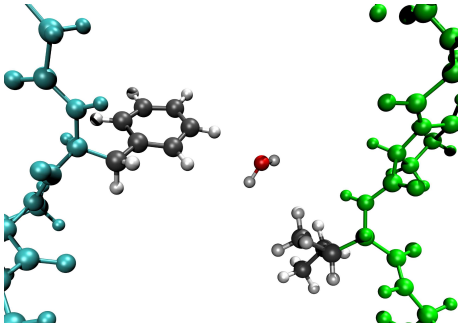

Figure 6: A water molecule in the first hydration shell of two different hydrophobic sidechains. The two neighbouring tropocollagens are shown as blue and green, and the relevant hydrophobic sidechains are highlighted in black and white.

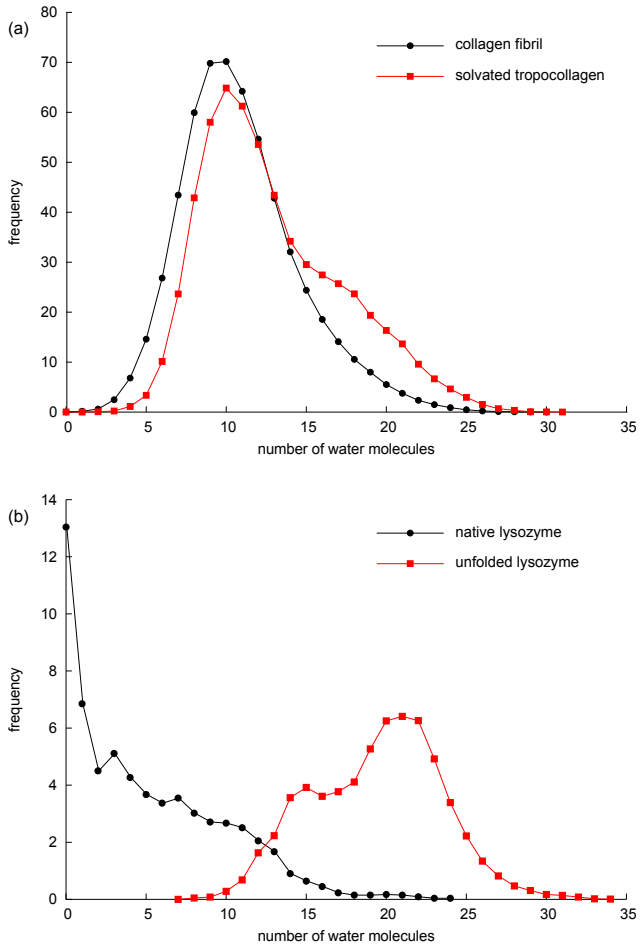

Figure 7: Distributions showing the number of hydrophobic sidechains to be surrounded by different numbers of water molecules. For each integer number of water molecules, the distributions show how many hydrophobic sidechains had that many water molecules within 5 Å. (a) The collagen fibril and fully solvated tropocollagen. (b) The lysozyme protein in its native and unfolded conformations. The distributions have been averaged over all recorded time steps.