4

The structure and function of fibrous proteins

Historically a division into globular and fibrous proteins has been made when describing chemical and physical properties. This division was originally made to account for the very different properties of fibrous and globular proteins as well as the different roles each group occupied within cells. Today it is preferable to avoid this distinction and to treat proteins as belonging to families that exhibit structural or sequential homology (see Chapter 6). However, despite very different properties the structures of fibrous proteins were amongst the first to be studied because of their accumulation in bodies such as hair, nails, tendons and ligaments. These proteins demonstrate new aspects of biological design not shown by globular proteins that requires separate description.

Fibrous proteins were named because they were found to make up many of the 'fibres' found in the body. Here, fibrous proteins had a common role in conferring strength and rigidity to these structures as well as physically holding them together. Subsequent studies have shown that these proteins are more widely distributed than previously supposed, being found in cells as well as making up connective tissues such as tendons or ligaments. More importantly, these proteins occupy important biological roles that arise from their wide range of chemical and physical properties. These properties are distinct from globular proteins and arise from the individual amino acid sequences. A common feature of most fibrous proteins is their long, drawn-out, or filamentous structure. Essentially, these proteins tend to occur as 'rod-like' structures extended more in two out of the three possible dimensions and lacking the compactness of globular proteins. As a result fibrous proteins tend to possess architectures based around regular secondary structure with little or no folding resulting from long-range interactions. In other words they lack true tertiary structure.

The second large class of proteins distinct from globular proteins are the membrane proteins. Members of this group of proteins probably make up the vast majority of all proteins found in cells. For many years the purification of these proteins remained very difficult and limited our knowledge of their structure and function. However, slowly membrane proteins have become amenable to biochemical characterization and Chapter 5 will deal with the properties of this important group to highlight in successive sections globular, fibrous and membrane proteins.

The amino acid composition and organization of fibrous proteins

In fibrous proteins at least three different structural plans or designs have been recognized in construction.

Proteins: Structure and Function by David Whitford

These designs include: (i) structure composed of 'coiled-coils' of α helices and represented by the α keratins; (ii) structures made up of extended antiparallel β sheets and exemplified by silk fibroin a collection of proteins made by spiders or silkworms; and (iii) structures based on a triple helical arrangement of polypeptide chains and shown by the collagen family of proteins. The structures of each class of fibrous proteins will be described in the following sections, highlighting how the structure is suitable for its particular role and also emphasizing how defects in fibrous proteins can lead to serious and life threatening conditions.

An analysis of the amino acid composition of typical fibrous proteins (Table 4.1) reveals considerable differences in their constituent amino acids to that described earlier for globular proteins (see Table 2.2, Chapter 2). More significantly the amino acid compositions of fibrous proteins differ with each group. For example, collagen has a proline content in excess of 20 percent whilst in silk fibroin this value is below 1 percent. Similarly in α keratin the cysteine content is 11.2 percent but in collagen and silk fibroin the levels of cysteine are essentially undetectable. In each case the amino acid composition influences the secondary structure formed by fibrous proteins.

Table 4.1 The amino acid composition of three common classes of fibrous proteins in mole percent

Amino acid	Fibroin (silk)	α-keratin	Collagen
Gly	44.6	8.1	32.7
Ala	29.4	5.0	12.0
Ser	12.2	10.2	3.4
Glx	1.0	12.1	7.7
Cys	0	11.2	0
Pro	0.3	7.5	22.1
Arg	0.5	7.2	5.0
Leu	0.5	6.9	2.1
Thr	0.9	6.5	1.6
Asx	1.3	6.0	4.5
Val	2.2	5.1	1.8
Tyr	5.2	4.2	0.4
Ile	0.7	2.8	0.9
Phe	0.5	2.5	1.2
Lys	0.3	2.3	3.7
Trp	0.2	1.2	0
His	0.2	0.7	0.3
Met	0	0.5	0.7

In collagen considerable amounts of hydroxylated lysine and proline are found. Adapted from *Biochemistry*, 3rd edn. Mathews, van Holde & Ahern (eds). Addison Wesley Longman, London.

Keratins

Keratins are the major class of proteins found in hair, feathers, scales, nails or hooves of animals. In general the keratin class of proteins are mechanically strong, designed to be unreactive and resistant to most forms of stress encountered by animals. At least two major groups of keratins can be identified; the α keratins are typically found in mammals and occur as a large number of variants whilst β keratins are found in birds and reptiles as part of feathers and scales containing a significantly higher proportion of β sheet. The β keratins are analogous to the silk fibroin structures produced by spiders and silkworms and described later in this chapter. The α keratins are a subset of a much larger group of filamentous proteins based on coiled-coils called intermediate filaments (IF). The distribution of intermediate filaments is not restricted to mammals but appears to extend to most animal cells as major components of cytoskeletal structures.

In mammals approximately 30 different variants of keratin have been identified with each appearing to be expressed in cells in a tissue-specific manner. In each keratin the 'core' structure is similar and is based around the α helix so that the following discussion of the conformation applies equally well to all proteins within this group. Although the basic unit of keratin is an α helix this structure is slightly distorted as a result of interactions with a second helix that lead to the formation of a left handed coiled-coil. The most common arrangement for keratin is a coiled-coil of two α helices although three helical stranded arrangements are known for extracellular keratins domains whilst in insects four stranded coiled

coils have been found. In 1953 Francis Crick postulated that the stability of α helices would be enhanced if pairs of helices interacted not as straight rods but in a simple coiled-coil arrangement (Figure 4.1). This coiled coil arrangement is sometimes called a super helix with in this instance α keratin found as a left-handed super helix. Detailed structural studies of coiled-coils confirmed this arrangement of helices and diffraction studies showed a periodicity of 1.5 Å and 5.1 Å.

The coiled-coil is formed by each helix interacting with the other and by burying their hydrophobic residues away from the solvent interface. The hydrophobic or non-polar side chains are not randomly located within the primary sequence but occur at regular intervals throughout the chain. This non-random distribution of hydrophobic residues is also accompanied by a preference for residues with charged side chains at positions within helices that are in contact with solvent. As a result of this periodicity a repeating unit of seven residues occurs along each chain or primary sequence. This is called the heptad repeat and the residues within this unit are labelled a, b, c, d, e, f and g. To facilitate identification of each residue within a helix the positions are frequently represented by a helical wheel diagram (Figure 4.2).

Although a regular α helix has 3.6 residues per turn, the coiled-coil arrangement leads to a slight decrease in the number of residues/turn to 3.5. Each helix is inclined at an angle of approximately 18° towards the other helix and this allows for the contacting side chains to make a precise inter-digitating surface. It also leads to a slightly different pitch for the helix of 5.1 Å compared with 5.4 Å in a regular α helix. The hydrophobic residues located at residues a and d of each heptad form a hydrophobic surface interacting with other hydrophobic surfaces. The hydrophobic residues form a seam that twists about each helix. By interacting with neighbouring hydrophobic surfaces helices are forced to coil around each other forming the super helix or coiled-coil. The interleaving of side chains has been known as 'knobs-into-holes' packing and in other proteins as a leucine zipper arrangement, although this last term is slightly misleading. A coiled coil arrangement not only enhances the stability of the intrinsically unstable single α- helix but also confers considerable mechanical strength in a manner analogous to the intertwining of rope or cable. Further aggregation of the coiled-coils occurs and leads to larger aggregates with even greater strength and stability. Besides the high content of hydrophobic residues α keratins also have significant proportions of cysteine (see Table 4.1). The cysteine residues participate in disulfide bridges that cross-link neighbouring coiled-coils to build up a filament or bundle and ultimately the network of protein constituting hair or nail.

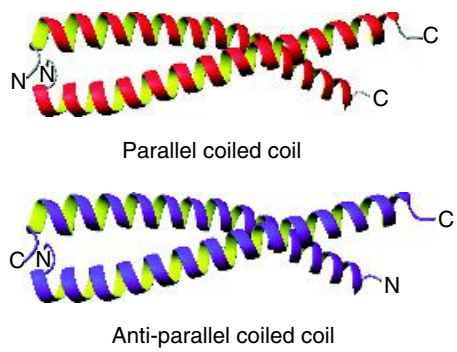

Figure 4.1 Two individual α helices distorted along their longitudinal axes as a result of twisting. The two helices can pack together to form coiled-coils

The coiled-coil containing many heptad repeats extends on average for approximately 300–330 residues and is flanked by amino and carboxy terminal domains. These amino and carboxy domains vary greatly in size. In some keratins very small domains of approximately 10 residues can exist whilst in other homologous proteins, such as nestin, much larger domains in excess of 500 residues are found. More significantly these regions show much greater sequence variability when compared with the coiled-coil regions suggesting that these domains confer specificity on the individual keratins and are tailored towards increasing functional specificity.

The keratins are often classified as components of the large group of filamentous proteins making up the cytoskeletal system and in particular they are classified as IF (Table 4.2). IF are generally between 8 and 10 nm in diameter and are more common in cells that have to withstand stress or extreme conditions.

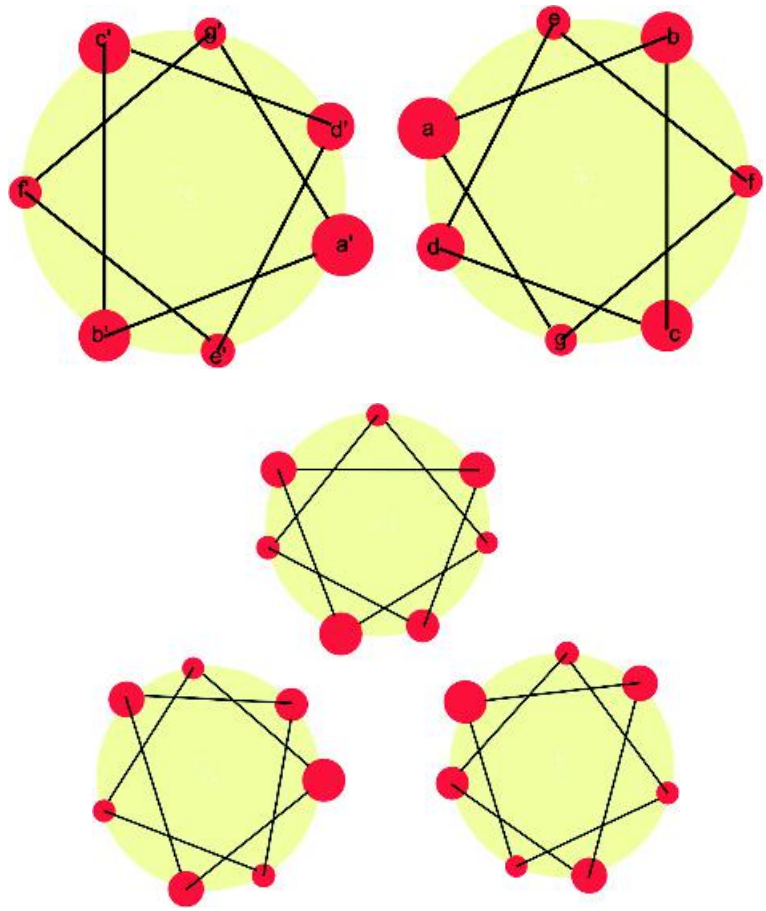

Figure 4.2 A helical wheel representation of the heptad repeat of coiled-coil keratins. The diagram shows a dimer (top) and trimer (bottom) of helices and emphasizes the hydrophobic contact regions between each helix as a result of residues a and d packing together. Leu, Ile and Ala are found frequently at positions a and d. Glu and Gln occur frequently at positions e and g whilst Arg and Lys are found frequently at position g

At least six different IF have been identified with classes I and II represented by acidic and basic keratins. An assignment of the terms 'acidic' and 'basic' to the N and C terminal globular domains of keratin refer to their overall charge. These domains vary in composition from one keratin to another. The acidic domains found at the end of the central α helical regions contain more negatively charged side chains (Glu and Asp) than positively charge ones (Lys, Arg) and consequently have isoelectric points in the pH range from 4 to 6.

Acidic and basic monomers are found within the same cell and the coiled coil or dimer contains one of each type (a heterodimer). Each coiled-coil aligns in a head to tail arrangement and in two staggered rows to form a protofilament. The protofilament dimerizes to form a protofibril with four protofibrils uniting to make a microfibril. The aggregation of protein units is still not finished since further association between microfibrils results in the formation of a macrofibril in reactions that are still poorly understood. The assembly of coiled coils into microfibrils is shown in Figure 4.3.

Table 4.2 The classification of intermediate filaments

Type	Example	Mass (M_r) $\times 10^3$	Location
I	Acidic keratins	40–64	Epithelial
II	Basic keratins	52–68	Epithelial
III	Glial fibrillary acidic protein	51	Astroglia
IV	Vimentin	55	Mesenchymal
VI	Desmin	53	Muscle
	Peripherin	54	Neuronal
	Neurofilaments (L, M, H);	68, 110, 130	Neuronal
	Internexin	66–70	Neuronal
V	Lamins A, B, C	58–70	Most cell types
Unclassified	Septins A, B, C		Some unclassified IFs appear to be found in invertebrates
	Filensin	100	
	Lens	50–60	

At least 6 different classes have been identified from homology profiles although new IF like proteins are causing these schemes to expand. IF are larger than the thin microfilaments (7–9 nm diameter) often made up of actin subunits and smaller than the thick microtubules (~25 nm diameter) made up of tubulin.

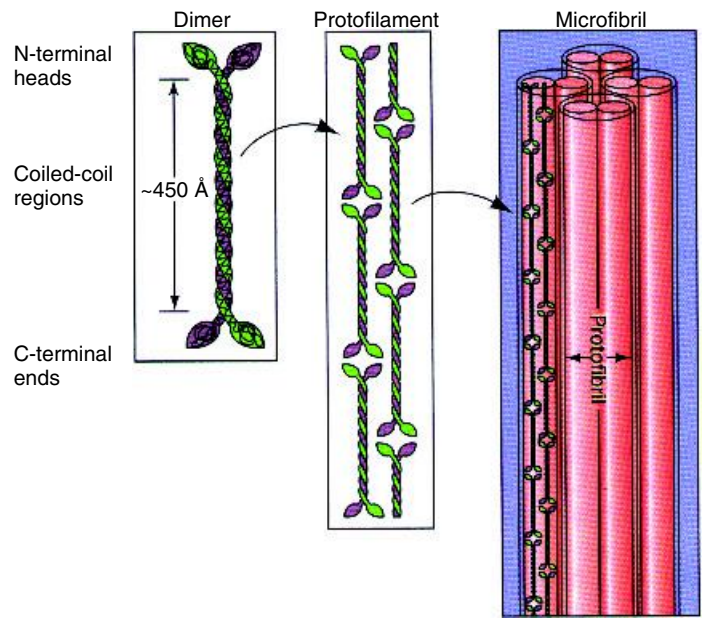

Figure 4.3 Higher order α keratin structure. Left: heterodimeric arrangement of two α helices to form a coiled-coil with both acidic and basic domains. Middle: protofilaments formed by the association of two coiled coils in a head-tail order and in two staggered or offset rows. Right: dimerization of protofilaments to form a protofibril followed by four protofibrils uniting to form a macrofibril (Reproduced from Voet, D, Voet, J.G & Pratt, C.W. *Fundamentals of Biochemistry*, John Wiley & Sons Inc, Chichester, 1999)

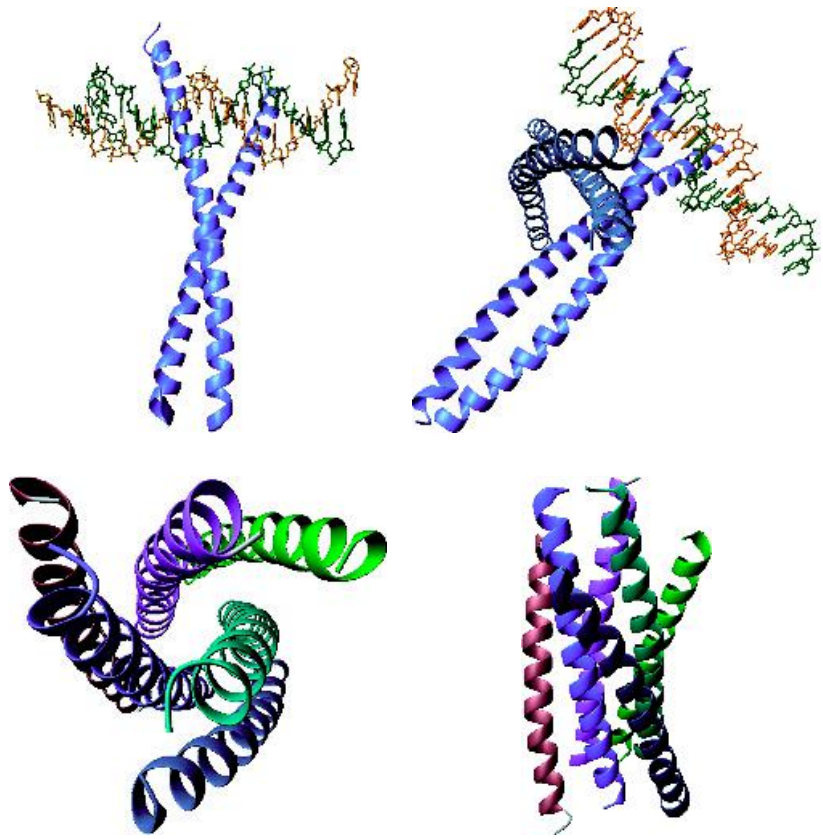

Figure 4.4 Examples of coiled coil motifs occurring in non-keratin based proteins. In a clockwise direction from the top left the diagrams show part of the leucine zipper DNA binding protein GCN4 (PDB:1YSA); a heterodimer of the c-jun proto-oncogene (the transcription factor Ap-1) dimerized with c-fos and complexed with DNA (PDB:1FOS); and two views of the gp41 core domain of the simian immunodeficiency virus showing the presence of six α helices coiled together as two trimeric units – an inner trimer often called N36 and an outer trimer called C34 (PDB:2SIV)

Crystallographic studies have shown that the coiled-coil motif occurs in many other proteins as a recognizable motif. It occurs in viral membrane-fusion proteins including the gp41 domain found as part of the human and simian immunodeficiency virus (HIV/SIV), in the haemagglutinin component of the influenza virus as well as transcription factors such as the leucine zipper protein GCN4. It is also found in muscle proteins such as tropomyosin and is being identified with increasing regularity in proteins of diverse function and cellular location. Although first recognized as part of long fibrous proteins the coiled-coil is now established as a structural element of many other proteins with widely differing folds (Figure 4.4). In the coiled-coil structures the α helices do not have to run in the same direction for hydrophobic interactions to occur. Although a parallel conformation is the most common arrangement antiparallel orientations, where the chains run in opposite directions, occur in dimers but are very rare in higher order aggregates.

Mutations in the genes coding for keratin lead to impaired protein function. In view of the almost ubiquitous distribution of keratin these genetic defects can have severe consequences on individuals. Defects prove particularly deleterious to the integrity of skin and several inherited disorders are known where

cell adhesion, motility and proliferation are severely disturbed. Since many human cancers arise in epithelial tissues where keratins are prevalent such defects may predispose individuals to more rapid tumour development.

Keratins, the most abundant proteins in epithelial cells, are encoded by two groups of genes designated type I and type II. There are >20 type I and >15 type II keratin genes occurring in clusters at separate loci in the human genome. A distinction is often made on the type of cell from which the keratin is derived or linked. This leads to type II keratin proteins from soft epithelia labelled as K1–K8 whilst those derived from hard epithelia (such as hair, nail, and parts of the tongue) are designated Hb1–Hb8. Similarly type I keratins are comprised of K9–K20 in soft epithelia and Ha1–Ha10 in hard epithelia. All α keratins are rich in cysteine residues that form disulfide bonds linking adjacent polypeptide chains. The term 'hard' or 'soft' refers to the sulfur content of keratins. A high cysteine (i.e. sulfur) content leads to hard keratins typical of nails and hair and is resistant to deformation whilst a low sulfur content due to a lower number of cysteine residues will be mechanically less resistant to stress.

In vitro the combination of any type I and type II keratins will produce a fibrous polymer when mixed together but *in vivo* the pairwise regulation of type I and II keratin genes in a tissue specific manner gives rise to 'patterns' that are very useful in the study of epithelial growth as well as disease diagnosis. The distribution of some of these keratins is described in Table 4.3 and with over 30 different types of keratin identified there is a clear preferential location for certain pairs of keratins.

In all complex epithelia a common set of keratin genes are transcribed consisting of the type II K5 and the type I K14 genes (along with variable amounts of K15 or K19, two additional type I keratins). Post-mitotic, suprabasal cells in these epithelia transcribe other pairs of keratin genes, the identity of which depends on the differentiation route of these cells. Thus the K1 and K10 pair is characteristic of cornifying epithelia such as the epidermis, whilst the K4 and K13 pair is expressed in epithelia found lining the oral cavity, the tongue and the oesophagus, and the K3 and K12 pair is found in the cornea of the eye.

Table 4.3 Distribution of type I and II keratin in different cells

Type I (acidic)	Type II (basic)	Location
K10	K1	Suprabasal epidermal keratinocytes
K9	K1	Suprabasal epidermal keratinocytes
K10	K2e	Granular layer of epidermis
K12	K3	Cornea of eye
K13	K4	Squamous epithelial layers
K14	K5	Basal layer keratinocytes
K15	K5	Basal layer of non-keratinizing epithelia
K16	K6a	Outer sheath of hair root, oral epithelial cells, hyperproliferative keratinocytes
K17	K6b	Nails
	K7	Seen in transformed cells
K18	K8	Simple epithelia
K19		Follicles, simple epithelia
K21		Intestinal epithelia

In skin diseases such as psoriasis and atopic dermatitis it is noticed that keratin 6 and 16 predominate whilst a congenital blistering disease, epidermolysis bullosa simplex, arises from gene defects altering the structure of keratins 5 and 14 at the basal layer.

A property of disulfide bridges between cysteine residues is their relative ease of reduction with reducing agents such as dithiothreitol, mecaptoethanol or thioglycolate. In general these reagents are called mercaptans, and for hair the use of thioglycolate allows the reduction of disulfide bridges and the relaxing of hair from a curled state to a straightened form. Removal of the reducing agent and the oxidation of the thiol groups allows the formation of new disulfide bridges and in this way hair may be reformed in a new 'curled', 'straightened' or 'permed' conformation. The springiness of hair is a result of the extensive number of coiled-coils and their tendency in common with a

conventional spring to regain conformation after initial stretching. The reduction of disulfide bridges in hair allows a keratin fibre to stretch to over twice their original length and in this very extended 'reduced' conformation the structure of the polypeptide chains shift towards the β sheet conformations found in feathers or in the silk-like sheets of fibroin. In the 1930s W.T. Astbury showed that a human hair gave a characteristic X-ray diffraction pattern that changed upon stretching the hair, and it was these two forms that were designated α and β.

Fibroin

A variation in the structure of fibrous proteins is seen in the silk fibroin class made up of an extended array of β strands assembled into a β sheet. Insect and spiders produce a variety of silks to assist in the production of webs, cocoons, and nests. Fibroin is produced by cultivated silkworm larvae of the moth *Bombyx mori* and has been widely characterized. Silk consists of a collection of antiparallel β strands with the direction of the polypeptide backbone extending along the fibre axis. The high content of β strands leads to a microcrystalline array of fibres in a highly ordered structure. The polypeptide backbone has the extended structure typical of β strands with the side chains projecting above and below the plane of the backbone. Of great significance in silk fibroin are the long stretches of repeating composition. A six residue repeat of $(Gly\text{-}Ser\text{-}Gly\text{-}Ala\text{-}Gly\text{-}Ala)_n$ is observed to occur frequently and it is immediately apparent that this motif lacks large side chains. These three residues appear to represent over 85 percent of the total amino acid composition with approximate values for the individual fractions being 45 percent Gly, 30 percent Ala and 15 percent Ser in silk fibroin (see Table 4.1). The sequence of six residues is part of a larger repeating unit

$$\text{-}(Gly\text{-}Ala)_2\text{-}Gly\text{-}Ser\text{-}Gly\text{-}Ala\text{-}Ala\text{-}Gly\text{-}$$
$$(Ser\text{-}Gly\text{-}Ala\text{-}Gly\text{-}Ala)_8\text{-}Tyr\text{-}$$

that may be repeated up to 50 times leading to masses for silk polypeptides between 300 000 and 400 000 (see Figure 4.5). Glycine, alanine and serine are the three smallest side chains in terms of their molecular volumes (see Table 2.3) and this is significant in terms of packing antiparallel strands. More importantly the order of residues in this repeating sequence places the glycine side chain (simply a hydrogen) on one side of the strand whilst the Ser and Ala side chains project to the other side. This arrangement of side chains leads to a characteristic spacing between strands that represents the interaction of Gly residues on one surface and the interaction of Ala/Ser side chains on the other. The interaction between Gly surfaces yields an inter-sheet spacing or regular periodicity of 0.35 nm whilst the interaction of the Ser/Ala rich surfaces gives a spacing of 0.57 nm between the strands in silk fibroin. Larger amino acid side chains would tend to disrupt the regular periodicity in the spacing of strands and they tend to be located in regions forming the links between the antiparallel β strands. The structure of these linker regions has not been clearly defined.

Silk has many remarkable properties. Weight-for-weight it is stronger than metal alloys such as steel, it is more resilient than synthetic polymers such as Kevlar©, yet is finer than a human hair. It is no exaggeration to say that silk is nature's high-performance polymer fine-tuned by evolution over several hundred million years. As a result of these desirable properties there have been many attempts to mimic the properties of silk with the development of new materials in the area of biomimetic chemistry. Silk is extremely strong because in the fully extended conformation of β strands any further extension would require the breakage of strong covalent bonds. However, this strength is coupled with surprisingly flexibility that arises as a result of the weaker van der Waals interactions that exist between the antiparallel β strands. These desirable physical properties are very difficult to reproduce in most synthetic polymers.

Collagen

Collagen is a major component of skin, tendons, ligaments, teeth and bones where it performs a wide variety of structural roles. Collagen provides the framework that holds most multicellular animals together and constitutes a major component of connective tissue. Connective tissue performs many functions including

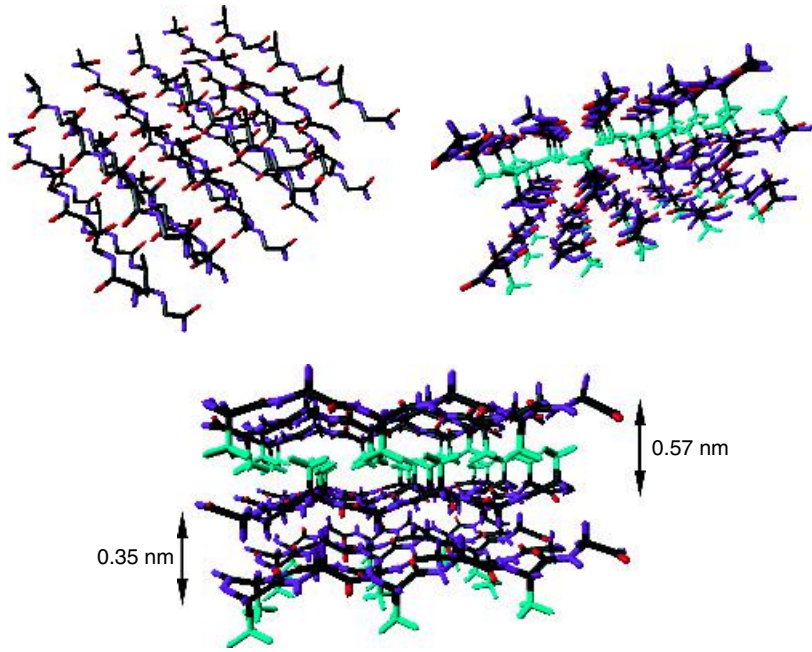

Figure 4.5 The interaction of alternate Gly and Ala/Ser rich surfaces in antiparallel β strands in the silk fibroin structure from *Bombyx mori*. The spacing between strands alternates between 0.35 and 0.57 nm as shown in the bottom figure. The structures of silk were generated for an alternating series of Gly/Ala polymers (PDB: 2SLK). Top left: the arrangement of strands in the antiparallel β sheet structure showing backbone only. Top right: The antiparallel β strands showing the interaction between Ala residue (shown by cyan colour) in an end on view where the backbone is running in a direction proceeding into and out of the page. In the bottom diagram the strands are running laterally and the Ala/Ser rich interface shown by the cyan side chains is wider than the smaller interface formed between glycine residues

binding together body structures and providing support and protection. Connective tissue is the most abundant tissue in vertebrates and depends for its structural integrity primarily on collagen. In vertebrates connective tissue appears to account for approximately 30 percent of the total mass. Although collagen is often described as the single most abundant protein it has many diverse biological roles and to date at least 30 distinct types of collagen have been identified from the respective genes with each showing subtle differences in the amino acid sequence along their polypeptide chains. Collagen is usually thought of as a protein characteristic of vertebrates such as mammals but it is known to occur in all multicellular animals. Sequencing of the nematode (flatworm) genome of *Caenorhabditis elegans* has revealed over 160 collagen genes. In this nematode collagen proteins are the major structural component of the exoskeleton with the collagen genes falling into one of three major gene families.

All collagens have the structure of a triple helix described in detail below and are assembled from three polypeptide chains. Since these chains can be combined in more than one combination a great many distinct collagens can exist and several have been identified as occurring predominantly in one group of vertebrate tissues. In humans at least 19 different collagens are assembled from the gene products of approximately 30 distinct and identified genes. Within these 19 structural types four major classes or groupings are generally identified. These groupings are summarized in Table 4.4. Type I collagen consists of two identical chains called α_1(I) chains and a third chain called α_2.

Table 4.4 The major collagen groups

Type	Function and location
Type I	The chief component of tendons, ligaments, and bones
Type II	Represents over 50 % of the protein in cartilage. It is also used to build the notochord of vertebrate embryos
Type III	Strengthens the walls of hollow structures like arteries, the intestine, and the uterus
Type IV	Forms the basal lamina (sometimes called a basement membrane) of epithelia. For example, a network of type IV collagens provides the filter for the blood capillaries and the glomeruli of the kidneys

In contrast Type II collagen contain three identical α_1 chains.

In a mature adult collagen fibres are extremely robust and insoluble. The insolubility of collagen was for many years a barrier to its chemical characterization until it was realized that the tissues of younger animals contained a greater proportion of collagen with higher solubility. This occurred because the extensive cross-linking of collagen characteristic of adults is lacking in young animals and it was possible to extract the fundamental structural unit called tropocollagen.

The structure and function of collagen

Tropocollagen is a triple helix of three similarly sized polypeptide chains each on average about 1000 amino acid residues in length. This leads to an approximate M_r of 285 000, an average length of ~300 nm and a diameter of ~1.4 nm. A comparison of the dimensions of tropocollagen with an average globular protein is useful and instructive. The length of collagen is approximately 100 times greater than myoglobin yet its diameter is only half that of the globular protein emphasizing its extremely elongated or filamentous nature. The polypeptides of tropocollagen are unusual in their amino acid composition and are defined by high proportions of glycine residues (see Table 4.2) as well as elevated amounts of proline. Collagen has a repetitive primary sequence in which every third residue is glycine. The sequence of the polypeptide chain can therefore be written as

-Gly-Xaa-Yaa−Gly-Xaa-Yaa-Gly-Xaa-Yaa-

where Xaa and Yaa are any other amino acid residue. However, further analysis of collagen sequences reveals that Xaa and Yaa are often found to be the amino acids proline or lysine. Many of the proline and lysine residues are hydroxylated via post-translational enzymatic modification to yield either hydroxyproline (Hyp) or hydroxylysine (Hyl) (see Chapter 8). The sequence Gly-Pro-Hyp occurs frequently in collagen. The existence of repetitive sequences is a feature of collagen, keratin and fibroin proteins and is in marked contrast to globular proteins where repetitive sequences are the exception.

Each polypeptide chain intertwines with the remaining two chains to form a triple helix (Figure 4.6). The helix arrangement is very different to the α helix and shows most similarity with the poly-proline II helices described briefly in Chapter 3. Each chain has the sequence Gly-Xaa-Yaa and forms a left-handed super helix with the other two chains. This leads to the triple helix shown in Figure 4.6. When viewed 'end-on' the super helix can be seen to consist of left handed polypeptide chain supercoiled in a right handed manner about a common axis.

The rise or translation distance per residue for each chain in the triple helix is 0.286 nm whilst the number of residues per turn is 3.3. Combining these two figures yields a value of ~0.95 nm for the helix pitch and reveals a more extended conformation especially when compared with α helices or 3_{10} helices (pitch ~0.5–0.54 nm).

Glycine lacking a chiral centre and possessing considerable conformational flexibility presents a significant contrast to proline. In proline conformational restraint exists as a result of the limited variation in the torsion angle (ϕ) permitted by a cyclic pyrrolidone ring. The presence of large amounts of proline in tropocollagen is also significant because the absence

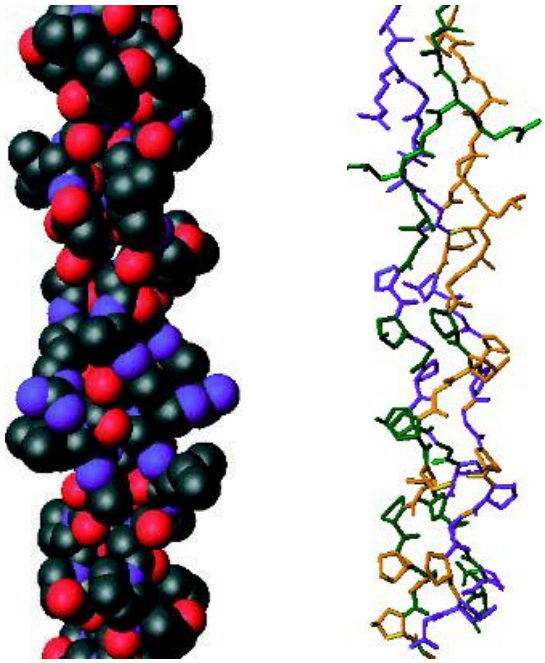

Figure 4.6 The basic structure of the triple helix of collagen. Space filling representation of the triple helix and a simpler wireframe view of the different chains (orange, dark green and blue chains). Careful analysis of the wireframe view reveals regular repeating Pro residues in all three chains

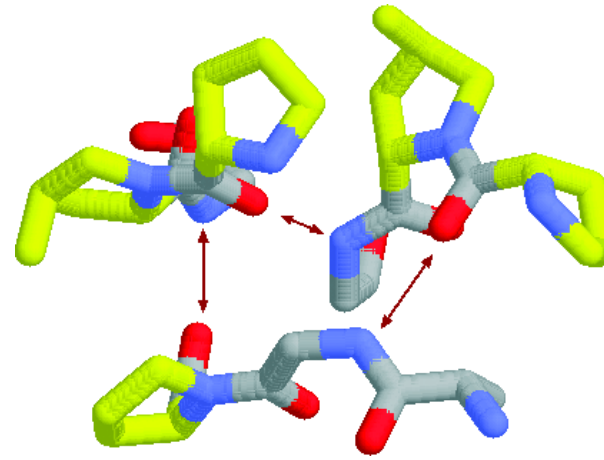

Figure 4.7 The interaction of glycine residues at the centre of the collagen helix. Note the side chains of residues Xaa and Yaa are located towards the outside of the helix on each chain where they remain sterically unhindered. In each of the three chains proline and hydroxyproline side chains are shown in yellow with the remaining atoms shown in their usual CPK colour scheme. Only heavy atoms are shown in this representation. The arrows indicate the hydrogen bonds from the glycine NH to the CO of residue Xaa in the neighbouring chain. Each chain is staggered so that Gly and Xaa and Yaa occur at approximately the same level along the axis of the triple helix (derived from PDB:1BKV)

of the amide hydrogen (HN) eliminates any potential hydrogen bonding with suitable acceptor groups. As a result of the presence of both glycine and proline in high frequency in the collagen sequences the triple helix is forced to adopt a different strategy in packing polypeptide chains. Since the glycine residues are located at every third position and make contact with the two remaining polypeptide chains it is clear that only a very small side chain (i.e. glycine) can be accommodated at this position. Any side chain bigger than hydrogen would disrupt the conformation of the triple helix. As a result there is very little space along the helix axis of collagen and glycine is always the residue closest to the helix axis (Figure 4.7). The side chains of proline residues along with lysine and other residues are on the outside of the helix.

The close packing of chains clearly stabilizes the triple helix through van der Waals interactions, but in addition extensive hydrogen bonding occurs between polypeptide chains. The hydrogen bonds form between the amide (NH) group of one glycine residue and the backbone carbonyl (C=O) group of residue Xaa on adjacent chains. The direction of the hydrogen bonds are transverse or across the long axis of the helix. Interactions within the triple helix are further enhanced by hydrogen bonding between amide groups and the hydroxyl group of Hyp residues. An indication of the importance of hydrogen bonding interactions in collagen helices has been obtained through constructing synthetic peptides and determining the melting temperature of the collagen triple helix. The melting temperature is the temperature

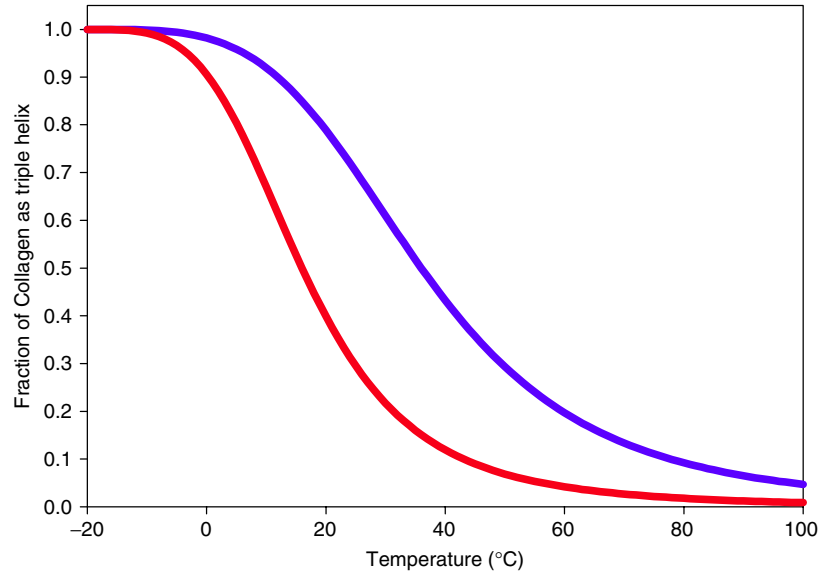

Figure 4.8 Thermal denaturation curve for collagen. In normal collagens the transition midpoint temperature or T_m is related to the normal body temperature of the organism and for mammals is above 40 °C as shown by the blue line in the above graph

at which half the helical structure has been lost and is characterized by a curve showing a sharp transition at a certain temperature reflecting the loss of ordered structure. The loss of structural integrity reflects denaturation of the triple helix and is accompanied by a progressive loss of function.

The importance of hydroxyproline to the transition temperature is shown by synthetic peptides of (Gly-Pro-Pro)$_n$ and (Gly-Pro-Hyp)$_n$. The former has a transition mid point temperature (T_m) of 24 °C whilst the latter exhibits a much higher T_m of ~60 °C (Figure 4.8). This experiment strongly supports the idea of triple helix stabilization through hydroxylation of proline (Hyp) and the formation of hydrogen bonds with neighbouring chains. Heating of collagen forms gelatin a disordered state in which the triple helix has dissociated. Although cooling partially regenerates the triple helix structure much of the collagen remains disordered. The reasons underlying this observation were not immediately apparent until the route of collagen synthesis was studied in more detail. Collagen is synthesized as a precursor termed procollagen in which additional domains at the N and C terminal specifically modulate the folding process. Mature collagen lacks these domains so any unfolding or disordering that occurs remains difficult to reverse.

Although hydrogen bonds and van der Waals interactions impart considerable stability to the tropocollagen triple helix and underpin its use as a structural component of many cells further strength arises from the association of tropocollagen molecules together as part of a collagen fibre. Each tropocollagen molecule is approximately 300 nm in length and packs together with neighbouring molecules to produce a characteristic banded appearance of fibres in electron micrographs. The banded appearance arises from the overlapping of each triple helix by approximately 64 nm thereby producing the striated appearance of collagen fibrils. This pattern of association relies on further cross-linking both within individual helices, known as *intra*molecular cross-links, as well as bonds between helices where they are called *inter*molecular cross-links. Both cross-links are the result of covalent bond formation.

The covalent cross-links among collagen molecules are derived from lysine or hydroxylysine and involve

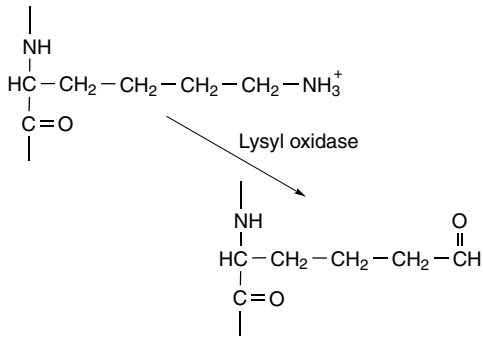

Figure 4.9 Oxidation of ε-amino group of lysine to form aldehyde called allysine

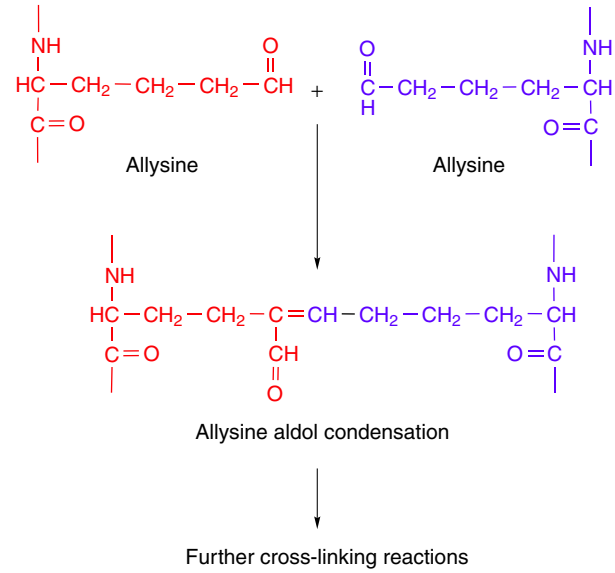

Figure 4.10 Outline of the reaction pathway leading to the formation of lysine crosslinks in collagen fibres. The first oxidative step results in the deamination of the ε-amino group and the formation of allysine (or hydroxyallysine). Two allysine residues condense to form a stable cross-link which can undergo further reactions that heighten the complexity of the cross-link. The allysine route predominates in skin whilst the hydroxyallysine route occurs in bone and cartilage

the action of an enzyme called lysyl oxidase. This copper dependent enzyme oxidizes the ε-amino group of a lysine side chain and facilitates the formation of a cross link with neighbouring lysine residues (Figure 4.9). Lysine sidechains are oxidized to an aldehyde called allysine and this promotes a condensation reaction between two chains forming strong covalent cross-links (Figure 4.10). A reaction between lysine residues in the same collagen fibre results in an intramolecular cross-link whilst reaction between different triple helices results in intermolecular bridges. In view of the presence of hydroxylysine and lysine in collagen these reactions can occur between two lysine, two hydroxylysines or between one hydroxylysine and one lysine. The products are called hydroxylysinonorleucine or lysinonorleucine. Further cross-linking can form trifunctional cross-links and a hydroxypyridinoline structure.

Cross-linking of collagen is a progressive process but does not occur in all tissues to the same extent. In general, younger cells have less cross-linking of their collagen than older cells, with a visible manifestation of this process being the increase in the appearance of wrinkled skin in the elderly, especially when compared with that found in a newborn baby. It is also the reason why meat from older animals is tougher than that derived from younger individuals.

Collagen biosynthesis

Collagen is a protein that undergoes significant post-translational modification and serves to introduce a subject that is described in more detail in Chapter 8. The initial translation product synthesized at the ribosome is very different to the final product and without these subsequent modifications it is extremely unlikely that the initial translation product could perform the same biological role as mature collagen. It is also true to say that any process interfering with modification of collagen tends to result in severe forms of disease.

The biosynthesis of collagen is divided into discrete reactions that differ not only in the nature of the modification but their cellular location. Step 1 is the initial formation of preprocollagen, the initial translation product formed at the ribosome. In this state the collagen precursor contains a signal sequence that directs the protein to the endoplasmic reticulum membrane and

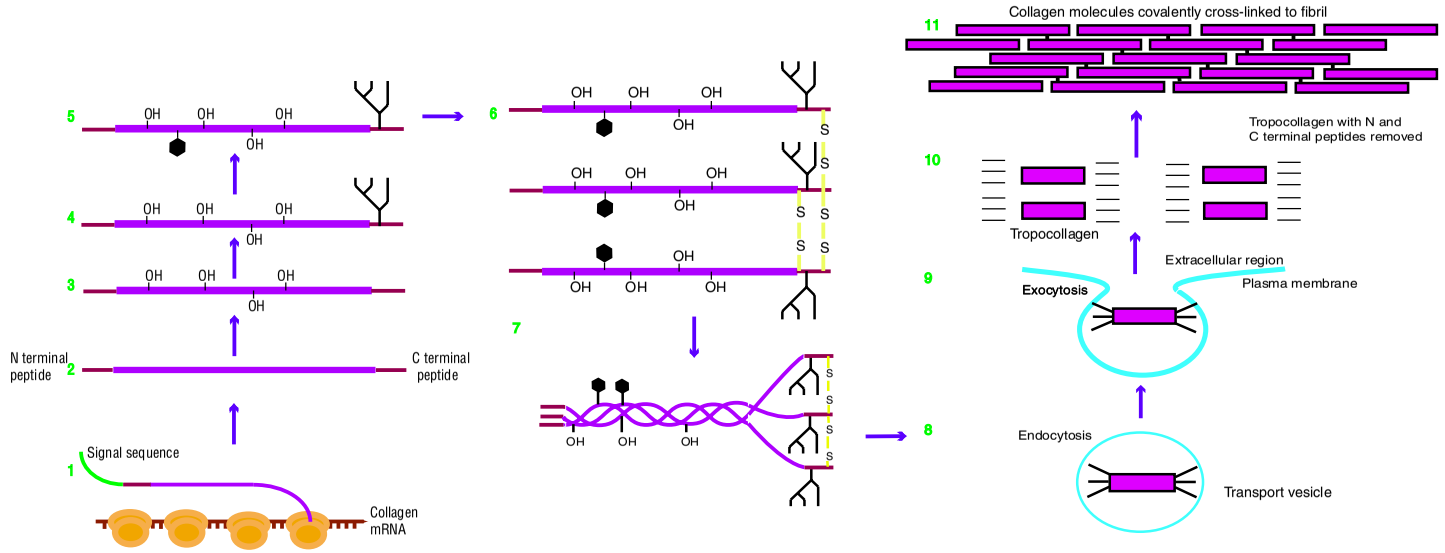

Figure 4.11 Processing of procollagen removes N and C terminal extensions and allows association into staggered, cross-linked fibrils.

1. Synthesis on ribosomes. Entry of chains into lumen of endoplasmic reticulum occurs with first processing reaction removing signal peptide
2. Collagen precursor with N and C terminal extensions.
3. Hydroxylation of selected proline and lysines
4. Addition of Asn-linked oligosaccharides to collagen
5. Initial glycosylation of hydroxylysine residues
6. Alignment of three polypeptide chains and formation of inter-chain disulfide bridges
7. Formation of triple helical procollagen
8. Transfer by endocytosis to transport vesicle
9. Exocytosis transfers triple helix to extracellular phase
10. Removal of N and C terminal propeptides by specific peptidases
11. Lateral association of collagen molecules coupled to covalent cross linking creates fibril

facilitates its passage into the lumen where the signal peptide is cleaved by the action of specific proteases. However, while still associated with the ribosome, this polypeptide is hydroxylated from the action of prolyl hydroxylase and lysyl hydroxylase, resulting in the formation of hydroxyproline and hydroxylysine, and is followed by transfer of the polypeptide into the lumen of the endoplasmic reticulum. Here, a third step involves glycosylation of the collagen precursor and the attachment of sugars, chiefly glucose and galactose, occurs via the hydroxyl group of Hyl. Frequently, the sugars are added as disaccharide units. In this state the pro-α-chains join forming procollagen whilst the N- and C-terminal regions form inter-chain disulfide bonds and the central regions pack into a triple helix. In this state the collagen is termed procollagen and it is transported to the Golgi system prior to secretion from the cell. Procollagen peptidases remove the disulfide-rich N and C terminal extensions leaving the triple helical collagen in the extracellular matrix (Figure 4.11) where it can then associate with other collagen molecules to form staggered, parallel arrays (Figure 4.12). These arrays undergo further modification by the formation of cross-links through the action of lysyl oxidase, as described above.

Collagenases degrade collagen and have been shown to be one member of a large group of enzymes called matrix metalloproteinases (MMPs). These enzymes degrade the extracellular matrix. Abnormal metalloproteinase expression leads to premature degradation of the extracellular matrix and is implicated in diseases such as atherosclerosis, tumour invasion and rheumatoid arthritis.

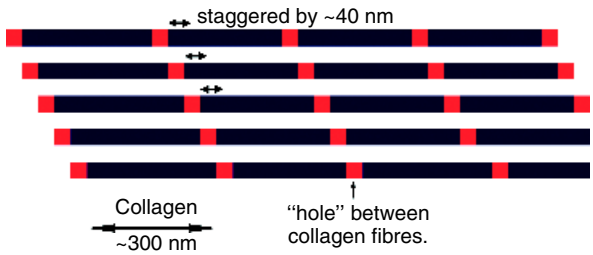

Figure 4.12 Association of procollagen into staggered parallel arrays making up collagen fibres. Each collagen fibre is approximately 300 nm in length and is staggered by ~40 nm from adjacent parallel fibres. This 'hole' can represent the site of further attachment of extracellular proteins and can become filled with calcium phosphate

Disease states associated with collagen defects

The widespread involvement of collagen in not only tendons and ligaments, but also the skin and blood vessels, means that mutations in collagen genes often result in impaired protein function and severe disorders affecting many organ systems. Mutations in the 30 collagen genes discovered to date give rise to a large variety of defects in the protein. In addition defects have also been found in the enzymes responsible for the assembly and maturation of collagen creating a further group of disease states. In humans defects in collagen as a result of gene mutation lead to osteogenesis imperfecta, hereditary osteoporosis and familial aortic aneurysm.

Several hereditary connective tissue diseases have been identified as arising from mutations in genes encoding collagen chains. Most common are single base mutations that result in the substitution of glycine by a different residue thereby destroying the characteristic repeating sequence of Gly-Xaa-Yaa. A further consequence of these mutations is the incorrect assembly or folding of collagen. Two particularly serious diseases attributable to defective collagen are osteogenesis imperfecta and Ehlers–Danlos syndrome. The molecular basis for these diseases in relation to the structure of collagen will be described.

Osteogenesis imperfecta is a genetic disorder characterized by bones that break comparatively easily often without obvious cause. It is sometimes called brittle bone disease. At least four different types of osteogenesis imperfecta are recognized by clinicians (Table 4.5) although all appear to arise from mutation of the collagen genes coding for either the α_1 or α_2 chains of type I collagen.

Osteogenesis imperfecta is caused by a mutation in one allele of either the α_1 or α_2 chains of the major collagen in bone, type I collagen. Type I collagen contains two α_1 chains together with a single

Table 4.5 Classification of osteogenesis imperfecta. Currently, diagnosis of type II form of osteogenesis imperfecta is determined *in utero* whilst diagnosis of other forms of the disease can only be made antenatal

Type	Inheritance	Description
I	Dominant	Mild fragility, slight deformity, short stature. Presentation at young age (2–6)
II	Recessive	Lethal: death at pre- or perinatal stage
III	Dominant	Severe, progressive deformity of limbs and spine
IV	Dominant	Skeletal fragility and osteoporosis, bowing of limbs. Most mild form of disease

α_2 chain and the presence of one mutant collagen allele is sufficient to produce a defective collagen fibre. The incidence of all forms of osteogenesis imperfecta is estimated to be approximately 1 in 20 000; from screening individuals it was found that glycine substitution was a frequently observed event with cysteine, aspartate and arginine being the most frequent replacement residues. The disruption of the Gly-Xaa-Yaa repeat has pathological consequences that vary considerably, as shown by Table 4.5 where four different types of osteogenesis imperfecta ranging in severity from lethal to mild are observed. One consequence of glycine substitution is defective folding of the collagen helix and this is accompanied by increased hydroxylation of lysine residues N-terminal to the mutation site. Since hydroxylation occurs on unfolded chains one possible effect of mutation is to inhibit the rate of triple helix formation rendering the proteins susceptible to further modification or interaction with molecular chaperones or enzymes of the endoplasmic reticulum that alter their processing or secretion. A further sequela of mutant collagen fibres is incorrect processing by *N*-propeptide peptidases (see section on collagen biosynthesis) and where mutant collagen molecules are incorporated into fibrils there is evidence of poor mineralization. It seems very likely that the exact site of the mutation of α_1 or α_2 chains will influence the overall severity of the disease with some evidence suggesting mutations towards the C-terminal produce more severe phenotypes. However, the precise relationship between disease state, mutation site and perturbation of collagen structure remain to be elucidated.

A second important disease arising from mutations in a single collagen gene is Ehlers–Danlos syndrome. This syndrome results in widely variable phenotypes, some relatively benign whilst others are life threatening. Classically the syndrome has been recognized by physicians from the presentation of patients with joint hypermobility as well as extreme skin extensibility, although many other diagnostic traits are now recognized including vascular fragility. Many different types of the disease are recognized both medically and at the level of the protein. The variability arises from mutations at different sites and in different types of collagens leading to the variety of phenotypes. These mutations can lead to changes in the levels of collagen molecules, changes in the cross-linking of fibres, a decreased hydroxylysine content and a failure to process collagen correctly by removal of the N-terminal regions. The common effect of all mutations is to create a structural weakness in connective tissue as a result of a molecular defect in collagen.

Related disorders characterized at a molecular level

Marfan's syndrome is an inherited disorder of connective tissue affecting multiple organ systems including the skeleton, lungs, eyes, heart and blood vessels. For a long time Marfan's syndrome was believed to be caused by a defect in collagen but this defect is now known to reside in a related protein that forms part of the microfibrils making up the extracellular matrix that includes collagen. The condition affects both men and women of all ethnic groups with an estimated incidence of ~1 per 20 000 individuals throughout the world.

Before identification of Marfan's syndrome became routine and surgical management of this disease was common most patients died of cardiovascular complications at a very early age and usually well before the age of 50. In 1972 the average life expectancy for a Marfan sufferer was ~32 years but today, with increased research facilitating recognition of the disease and surgical intervention alleviating many of the health problems, the expected life span had increased to over 65 years.

Marfan's syndrome is caused by a molecular defect in the gene coding for fibrillin, an extracellular protein found in connective tissue, where it is an integral component of extended fibrils. Microfibrils are particularly abundant in skin, blood vessels, perichondrium, tendons, and the ciliary zonules of the eye. The elastin-based fibres form part of an extracellular matrix structure that provides the elastic properties to tissues. Both morphological and biochemical characterization of fibres reveal an internal core made up primarily of the protein elastin together with a peripheral layer of microfibrils composed primarily of fibrillin. Humans have two highly homologous fibrillins, fibrillin-1 (Figure 4.13) and fibrillin-2, mapping to chromosomes 15 and 5 respectively. In 1991, the first mutation in fibrillin 1 was reported and subsequently over 50 mutations in individuals with Marfan syndrome have been described. A characteristic feature of both fibrillins is their mosaic composition where numerous small modules combine to produce the complete, very large, protein of 350 kDa.

The majority of fibrillin consists of epidermal growth factor-like subunits (47 epidermal growth factor (EGF)-like modules) of which 43 have a consensus sequence for calcium binding. Each of these domains is characterized by six cysteine residues three disulfide bridges and a calcium binding consensus sequence of D/N-x-D/N-E/Q-x_m-D/N*-x_n-Y/F (where m and n are variable and * indicates possible post-translational modification by hydroxylation). Other modules found in fibrillin-1 including motifs containing eight Cys residues, hybrid modules (two) along with sequences unique to fibrillin (three). These domains are interspersed throughout the molecule with the major differences between fibrillins residing in a proline-rich region close to the N-terminus in fibrillin-1 that is replaced by a glycine-rich region in fibrillin-2.

The EGF domain occurs in many other proteins including blood coagulation proteins such as factors X, VII, IX, and the low density lipoprotein receptors.

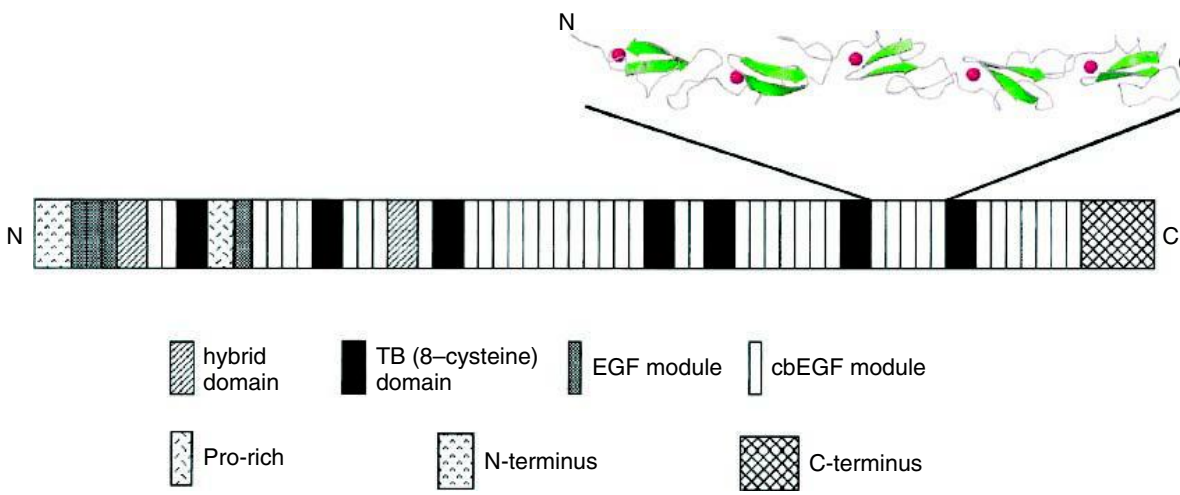

Figure 4.13 The modular organization of fibrillin-1 (reproduced with permission from Handford, P.A. *Biochim. Biophys. Acta* 2000, **1498** 84–90. Elsevier)

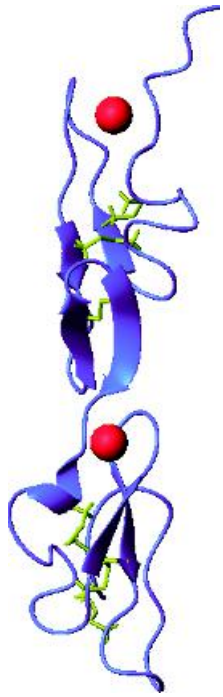

Figure 4.14 The structure of a pair of calcium binding EGF domains from fibrillin-1 (PDB:1EMN)

The structure of a single EGF like domain and a pair of calcium-binding domains confirmed a rigid rod-like arrangement stabilized by calcium binding and hydrophobic interactions (Figure 4.14). Mutations known to result in Marfan's syndrome lead to decreased calcium binding to fibrillin, and this seems to play an important physiological role. The importance of the structure determined for the modules of fibrillin was that it offered an immediate explanation of the defect at a molecular level. The structural basis for Marfan's syndrome resides in the disruption of calcium binding within fibrillin as a result of single site mutation.

Marfan's syndrome is an autosomal dominant disorder affecting the cardiovascular, skeletal, and ocular systems. The major clinical manifestations are progressive dilatation of the aorta (aortic dissection), mitral valve disorder, a tall stature frequently associated with long extremities, spinal curvature, myopia and a characteristic thoracic deformity leading to a 'pigeon-chest' appearance. The majority of the mutations known today are unique (i.e. found only in one family) but at a molecular level it results in the substitution of a single amino acid residue that disrupts the structural organization of individual EGF-like motifs.

Flo Hyman, a famous American Olympian volleyball player, was a victim of Marfan's syndrome and a newspaper extract testifies to the sudden onset of the condition in apparently healthy or very fit athletes. "During the third game, Hyman was taken out in a routine substitution. She sat down on the bench. Seconds later she slid silently to the floor and lay there, still. She was dead". Many victims of Marfan's disease are taller than average. As a result basketball and volleyball players are routinely screened for the genetic defect which despite phenotypic characteristics such as pigeon chest, enlarged breastbone, elongated fingers and tall stature often remains undiagnosed until a sudden and early death. It has also been speculated largely on the basis of their physical appearances that Abraham Lincoln (16th president of the United States, 1809–1865) and the virtuoso violinist Niccolò Paganini (1782–1840) were suffering from connective tissue disorders.

Summary

Fibrous proteins represent a contrast to the normal topology of globular domains where compact folded tertiary structures exist as a result of long-range interactions. Fibrous proteins lack true tertiary structure, showing elongated structures and interactions confined to local residues.

The amino acid composition of fibrous proteins departs considerably from globular proteins but also varies widely within this group. This variation reflects the different roles of fibrous proteins.

Three prominent groups of fibrous proteins are the collagens, silk fibroin and keratins and all occupy pivotal roles within cells.

Collagen, in particular, is very abundant in vertebrates and invertebrates where the triple helix provides a platform for a wide range of structural roles in the extracellular matrix delivering strength and rigidity to a wide range of tissues.

The triple helix is a repetitive structure containing the motif (Gly-Xaa-Yaa) in high frequency with Xaa

and Yaa often found as proline and lysine residues. Repeating sequences of amino acids are a feature of many fibrous proteins and help to establish the topology of each protein. In collagen the presence of glycine at every third residue is critical because its small side chain allows it to fit precisely into a region that forms from the close contact of three polypeptide chains. Larger side chains would effectively disrupt this region and perturb the triple helical structure.

Although based around a helical design collagen differs considerably in dimensions to the typical α helix. The triple helix of collagen undergoes considerable post-translational modification to increase strength and rigidity.

Keratins make up a considerable proportion of hair and nails and contain polypeptide chains arranged in an α helical conformation. The helices interact via supercoiling to form coiled-coils. Of importance to coiled-coils are specific interactions between residues in different helices via non-polar interactions that confer significant stability. The basis of this interaction is a heptad of repeating residues along the primary sequence.

A heptad repeat possesses leucine or other residues with hydrophobic side chains arranged periodically to favour inter-helix interactions. Helices are usually arranged in pairs. Significantly, this mode of organization is found in other intracellular proteins as well as in viral proteins. Several DNA binding proteins contain coiled-coil regions and the hydrophobic-rich domains are often called leucine zippers.

In view of their widespread distribution in all animal cells mutations in fibrous proteins such as keratins or collagens lead to serious medical conditions. Many disease states are now known to arise from inherited disorders that lead to impaired structural integrity in these groups of proteins.

Problems

1. Describe the different amino acid compositions of collagen, silk and keratins?
2. Explain why glycine and proline are found in high frequency in the triple helix of collagen but are not found frequently within helical regions of globular proteins?
3. Why is silk both strong and flexible?
4. Why is wool easily stretched or shrunk? Why is silk more resistant to these deformations.
5. Poly-L-proline is a synthetic polypeptide that adopts a helical conformation with dimensions comparable to those of a single helix in collagen. Why does poly-L–proline fail to form a triple helix. Will the sequence poly-(Gly-Pro-Pro)$_n$ form a triple helix and how does the stability of this helix compare with native collagen. Does poly-(Gly-Pro-Gly-Pro)$_n$ form a collagen like triple helix?
6. A mutation is detected in mRNA in a region in frame with the start codon (AUG) CCCUAAAUG. . . . . .
   GGACCCAAAGGACCUAAG**U**GUCCAUCUGGU
   CCGAAGGGGUCCAACGGACCCAAGGGU. . . . . .
   Establish the identity of the peptide and describe the possible consequences of this mutation.
7. Describe four post-translational modifications of collagen and list these modifications in their order of occurrence?
8. Gelatin is primarily derived from collagen, the protein responsible for most of the remarkable strength of connective tissues in tendons and other tissues. Gelatin is usually soft and floppy and generally lacking in strength. Explain this observation.
9. Ehlers–Danlos syndromes are a group of collagen based diseases characterized by hyperextensible joints and skin. What is the most probable cause of this disorder in terms of the structure of collagen?
10. Keratin is based on a seven residue or heptad repeat. Describe the properties of this sequence that favour coiled-coil structures.