The Molecular Structure of Collagen

Joseph Patrick Rosen O'Dubhthaigh-Orgel

Thesis submitted for the degree of Ph.D.

Department of Biological Sciences
University of Stirling

May 2000

09/01

Acknowledgements

I gratefully acknowledge the financial support provided by the University of Stirling for research during the course of my time at the Department of Biological and Molecular Sciences, University of Stirling. I also wish to thank the Carnegie trust (Scottish Universities) for travel and subsistence support for work carried out at the BioCAT facility, Argonne, Illinois.

The last few years have been hard but worthwhile, a reflection of the patient support of more people than I have space to acknowledge. The completion of a Ph.D. thesis is a land mark achievement in its own right, and this one might not have been possible but for the faith and support of certain people over the last few years. I therefore wish to express my heart felt gratitude to those people, and to the following friends and colleagues:

Dr. Tim Wess and Prof. Andrew Miller, for their supervision, support, and advice throughout this project.

Dr. Tom Irving for being willing to give me some of that much sought after commissioning beam time at the BioCAT facility in 1998, some of the most important work of this thesis might not have happened otherwise.

Dr. Ian Alberts for many robust discussions and advice, and Dr. Andy Hammersly, for less frequent but equally useful robust discussions and advice.

Frank Kelly for advice and assistance on the construction and control of the sample

positioning equipment.

The staff of the ESRF, SRS, and APS, especially on beamlines ESRF ID2 and 22, and SRS 2.1.

Ronnie Balfour and Scott Jackson for being unfailingly cheerful and helpful.

Clare Bird and Dr Linda Wess for kindly reading and checking through the thesis manuscript.

My Office partners (Clare and Sharon), for lots of useful discussions and more importantly their friendship.

Finally, I would like to thank my parents Patrick and Carole, and my father's wife Hilary for their devotion, care and support. Most of all, I thank Yeshua, and Leah for their love.

Abstract

This thesis describes the study of the molecular packing and organisation of collagen molecules within a fibril.

The first two chapters describe the background to the study. In Chapter 1, a review of the extracellular matrix concentrates on the structure and organisation of type I collagen. Chapter 2 summarises the theory of X-ray diffraction by fibres, and Chapter 3 describes X-ray sources and equipment used in data collection. Data treatments and data extraction methods (such as simulated annealing) are also discussed.

Chapters 4 and 5 present the results of the study. Chapter 4 describes the determination of the one-dimensional structure of type I collagen to 0.54 nm resolution using X-ray diffraction and isomorphous derivative phase determination. The significance of the electron density map is interpreted in light of the known amino acid sequence, showing possible variations in the nature of the helix pitch. More importantly, the conformations of the intermolecular crosslink forming non-helical telopeptides were determined.

Chapter 5 provides a detailed background to the current understanding of the three-dimensional packing structure of collagen, and presents the first model-independent phase determined structure of a natural fibre – the lateral packing structure of type I collagen in rat tail tendon. The data extraction methods described in Chapter 3 are employed to calculate an electron density map of anisotropic resolution, from which the

crosslink forming telopeptide segments within the quasi-hexagonal packing structure are identified. Conclusions are drawn concerning the nature of order/disorder within collagen fibrils and the validity of the compressed microfibril model of collagen molecular packing and organisation is discussed.

Chapter 6 summaries the results and evaluates the success of the study. The potential for development of the techniques and results found for further studies are also discussed.

List of abbreviations and symbols

Diffraction terms

a, b, c,	Unit cell dimensions in real space
a*, b*, c*,	Unit cell dimensions in reciprocal space
B	Temperature factor
c	Speed of light
d	Lattice plane spacing
D	Displacement in reciprocal space
D	Collagen repeat period (67/64 nm)
e	Charge of one electron
E	Energy
F	Fourier transform
f	Diffraction structure factor terms
F_p, F_{ph}, F_h	Structure factor terms for (native) protein, derivative, and heavy-atom
h	Planck's constant (6.626×10^{-34} J-s)
h,k,l	Miller indices
i	Square root of -1
L	Distance between origin and diffraction sampling point
O_1,O_2.....	Point scatterers
P	Patterson function
R or r	Radius
r	Vector joining point scatterers

s	X-ray scattering vector
S	Unit vector in direction of scattered beam
S_o	Unit vector in direction of incident beam
t	Time
u	Patterson space vector
V	Volume
Y	Displacement of electromagnetic field in wave of radiation
α, β, γ	Interaxial angles of real space unit cell
$\alpha^*, \beta^*, \gamma^*$	Interaxial angles of reciprocal unit cell
η	Amplitude of scattered radiation
λ	Wavelength
α_s	Scattering phase shift
$\rho(x)$	Electron density distribution
θ_c	Critical angle

Terms used in the context of fibre diffraction

x,y,z	Cartesian coordinates in real space
u,v,w	Lattice coordinates in real space
θ, ϕ	Spherical polar coordinates
R, Z	Cylindrical polar reciprocal space coordinates
m,n	Integer or zero, used in layer-line selection rule
group 1	Refers to the first group of row-lines at 3.9 nm (Figure 4.3)

	row-lines: (1,0,l),(-1,0,l).
group 2	Refers to the second group of row-lines at 2.7 nm (Figure 4.3)
	row-lines: (0,1,l),(0,-1,l),(1,-1,l),(-1,1,l).
group 3	Refers to the third group of row-lines at 1.8 nm (Figure 4.3)
	row-lines: (-1,-1,l),(1,1,l),(2,0,l),(-2,0,l),(-2,1,l),(2,-1,l).
group 4	Refers to the fourth group of row-lines at 1.3 nm (Figure 4.3)
	row-lines: (2,1,l),(-2,-1,l),(-1,2,l),(1,-2,l),(0,2,l),(0,-2,l), (-3,1,l),(3,1,l),(3,0,l),(-3,0,l).

X-ray radiation associated terms and abbreviations

Brilliance	(photons s^{-1} $mrad^{-1}$ mm^{-1} (0.1 BW)$^{-1}$)
BW	Bandwidth
Flux	(photons s^{-1} mm^{-1})
EMS	Electromagnetic spectrum

Collagen associated terms and abbreviations

α-chain	Single polypeptide unit that associates with two other chains to form the collagen triple helix
α1(I),α2(I)	Alpha chain 1 or 2 of collagen type I
ECM	Extracellular matrix
FACIT	Fibril Associated Collagen with Interrupted Triple Helix
Multiplexins	Multiple triple helix domains and interruptions

X or x	The 2nd amino acid position in the collagen triplet repeat
Y or y	The 3rd amino acid position in the collagen triplet repeat
10/3	Ten units (amino acids) within three turns of the helix
7/2	Seven units (amino acids) within two turns of the helix
Triplex	Collagen triple helix

Miscellaneous terms and abbreviations

APS	Advanced Photon Source
DEPC	Diethylpyrocarbonate
ESRF	European Synchrotron Radiation Facility
SRS	Synchrotron Radiation Source
NMR	Nuclear Magnetic Resonance
PTA	Phospho Tungstic Acid
R-factor	Crystallographic/astronomy reliability index
RMS	Root-Mean-Square

Contents

CHAPTER 1

CHAPTER 2

THEORY OF DIFFRACTION 78

2.1 Introduction 79

2.2 Background 80

2.3 Basic principles of X-ray diffraction 81

2.3.1 Wave/particle nature of light 81

2.3.2 Thomson scattering 81

2.3.3 Compton scattering 81

2.3.4 Bragg's law 82

2.4 The scattering of X-rays 86

2.4.1 From one point 86

2.4.2 From two points 90

2.4.3 Scattering from a general number of centres 95

2.5 Diffraction and the Fourier transform 97

2.6 Crystallography 99

2.6.1 Crystals and liquid-crystals 99

2.6.2 The crystal lattice and the reciprocal lattice 101

2.6.3 The Ewald construction 104

2.7 Solutions to the phase problem 106

2.7.1 Patterson function 107

CHAPTER 3

INSTRUMENTATION, DATA COLLECTION, EXTRACTION AND CORRECTION 117

CHAPTER 4

CHAPTER 5

THE THREE-DIMENSIONAL MOLECULAR PACKING STRUCTURE OF FIBRILLAR TYPE I COLLAGEN 191

List of Tables and Figures

Chapter 1

The molecular structure of collagen

1.1 Introduction: collagen form and function

The diverse range of proteins within the collagen family form the major constituent of connective tissues in most vertebrates (Hulmes 1992). Collagen is so prevalent, that it is believed to be the single most abundant protein in the animal kingdom, accounting for about a third of the total protein mass in larger animals. However, its quantitative presence in tissues does not overshadow the qualitative benefits collagens impart to multicellular organisms.

The survival of multicellular life is at a fundamental level, dependent upon sustaining a controlled environment at an inter- as well as at a intra-cellular level. Maintaining surface to surface contact between individual cells is crucial to regulating homeostatic conditions. Multicellular organisms achieve this through a complex system of extracellular matrix proteins and glycoproteins. Within these systems, collagen and collagen-like proteins play a crucial role.

In higher organisms, specialised spatial organisation becomes more crucial, due to the greater complexity of cellular and tissue level interrelationships. Greater physical stresses and strains are placed upon these tissues as a normal parameter of life. This being especially relevant to tissues involved within the circulatory system and those designed to convey mobility to the animal.

Researchers have long sought after the molecular structure of collagens and collagen-like proteins, since attaining this knowledge is of profound importance to

understanding the basis of tissue continuity and strength. It is hoped that the work presented here in this thesis makes a suitable contribution toward this goal.

1.1.1 Structural diversity

The composition of proteins and glycoproteins within the extra cellular matrix of a particular tissue varies according to the structural properties of the tissue in question. Significantly, a wide range of connective tissue types with diversity of function are constructed from a relatively limited repertoire of proteins (Nimni and Harkness 1988, Hulmes 1992, Kielty *et al.,* 1993, Kadler 1995). For instance; tendon, a tissue that needs to be resistant to macroscopic extension, is made up of large, axially parallel, densely packed type I collagen fibrils (Miller 1976), whereas, in tissues where elastic properties are necessary, such as the lungs and blood vessels of vertebrates, collagen fibrils are accompanied by elastic fibres made up of the protein elastin (Barnes 1988). In tissues such as cartilage, the functional role of the tissue requires a structure that is resistant to compression, hence 50% of its make up comprises of highly hydrated proteoglycans (Kielty *et al.,* 1993).

In several tissues, more than one type of collagen is found within the make up of the extracellular matrix, as is the case in cartilage. Types II, VI, IX, and XI are present in varying proportions, providing a dense matrix of stress bearing fibrils (collagen type II being the most predominant form within these fibrils). These fibrils are joined together by associations with type VI and XI collagens, and joined also to the compression resistant proteoglycans through type IX collagen associations (Linsenmayer *et al.,* 1990). Cartilage is a complex tissue with a well designed infrastructure of collagens and other extra cellular matrix (ECM) proteins and is typical in its complexity when compared to that of other connective tissues. This is demonstrated by Table 1.1, which

shows both the range of identified collagens and their distribution amongst connective tissue types. However, connective tissue diversity is not conveyed exclusively by its protein composition. Features such as fibril diameter, packing arrangement and orientation produce significantly different anatomical attributes, even when the predominant collagen type is the same. For instance, the very regular spacing of the fibrils, and uniformly small fibril diameter in corneal stroma, forms the basis of tissue transparency (Nimni and Harkness 1988). Whilst the packing of overlapping collagen fibre 'plates' into concentric rings (surrounding the blood vessel containing central canal) imparts a layered reinforcement to resist breakage in cortical bone (Nimni and Harkness 1988).

The orientation of collagen fibrils in tendon is such that they are aligned approximately parallel to one another; in skin, the collagen fibrils form a coarse feltwork in the plane of the tissue. The tendon fibril arrangement conveys tensile strength when external force is applied parallel to their orientation, whilst the skin arrangement allows a certain degree of flexibility, but is strongly resistant to breakage especially within the plane of the skin surface (Purslow *et al.,* 1998). Fibril diameters range from 50-500 nm in tendon, 40-100 nm in skin, and approximately 25 nm in cornea (all tissue types predominantly contain type I), Kielty *et al.*, (1993).

The genetic variety and diversity of anatomical arrangement of the collagen family found in connective tissues, make understanding the structure, arrangement, and distribution of collagen rewarding across a range of research interests, and of considerable importance. It should be evident that obtaining reasonable estimations of

the packing and orientation parameters of the molecules as well the molecular structure itself, would provide significant insight into how connective tissues are organised and thereby able to fulfil their functions. Of at least equal importance is that this information provides a sound basis from which to explain the fundamental nature of numerous connective tissue diseases (Kadler 1995).

Table 1.1 Summary of collagen types by groups

TYPE	MOLECULAR SPECIES	ORGANISATION OF MOLECULAR AGGREGATES	DISTRUBUTION	FUNCTION
I Fib	[α1(I)$_2$ α2(I)] [α1(I)$_3$]	Large diameter banded fibrils staggered by 67nm	Widespread; skin, bone, tendon, ligaments, cornea, lung, etc. Except basement membrane and hyaline cartilage	Supporting fibres; recognised by integrin and syndecan cell surface receptors
II Fib	[α1(II)$_3$]	Small diameter banded fibrils staggered by 67nm	Hyaline cartilage; other cartilaginous tissues, vitreous, and annulus fibrosus	Supporting fibres
III Fib	[α1(III)$_3$]	Small diameter banded fibrils staggered by 67nm	Skin, lungs, aorta, cornea, uterus, nerve, and lymph nodes	Small supporting fibres, form copolymers with type I
V Fib	[α1(V)$_3$] [α1(V)$_2$α2(V)] [α1(V)α2(V) α3(V)]	Probably small diameter banded fibrils staggered by 67nm. Can form molecules with type XI chains	Widespread; skin, bone, tendon, ligament etc., all connective tissues except basement membrane	Small supporting pericellular fibres
XI Fib	[α1(XI)α2(XI) α3(XI)]	Probably small diameter banded fibrils staggered by 67nm. Can form molecules with type V chains	Hyaline cartilage, vitreous humour	Possible core for type II molecules, regulating fibril diameter
IV Net	[α1(IV)$_2$α2(IV)] [α1(IV)$_3$]	Meshwork sheets	Basement membrane	Meshwork scaffolding; contains multiple cell binding sites
VI Fil	[α1(VI)α2(VI) α3(VI)]	Beaded filaments	Wide spread, essentially all connective tissue except basement membranes	Possible interface between major collagen fibrils and cells
VII LCh	[α1(VII)$_3$]	Anchoring fibrils	Sub-basal lamina of skin	Anchoring of dermal epithelial cells to underlying stroma
VIII SCh	[α1(VIII)$_2$ α2(VIII)]	Meshwork sheets	Sclera, subendothelium of large blood vessels, cartilage growth plate	Unknown

X SCh	[α1(X)$_3$]	Meshwork sheets	Hypertrophic cartilage	Endochondral bone development; associates with type II collagen fibrils
XIII SCh	Unknown	Unknown	Cartilage growth plate, perichondria, and uncalcified calvaria	Unknown
IX FAC	[α1(IX)α2(IX) α3(IX)]	Fibril associated collagen with interrupted triple helix (FACIT)	Hyaline cartilage; vitreous humour	Association with type II for; fibril association with cartilage proteglycans
XII FAC	[α1(XII)$_3$]	FACIT	Surface of type I fibrils in; bone, skin, cornea, and type II in cartilage	Association with surface of major types allows further association within ECM
XIV FAC	[α1(XIV)$_3$]	FACIT	Surface of type I fibrils in; bone, skin, cornea, and type II in cartilage	Association with surface of major types allows further association within ECM
XVI FAC	[α1(XVI)$_3$]	FACIT	Heart, lung, pancreas, skeletal muscle, placenta,	Unknown
XIX FAC	Unknown	FACIT	Expressed in human cell lines; rhabdomyosarcoma and fibroblast cells.	Unknown
XV Mul	Unknown	Multiple triple helix domains and interuptions (Multiplexins)	Fibroblasts, endometrium, basement membranes of several tissues	Unknown
XVIII Mul	Unknown	Multiplexins	Lung, liver, kidney	Unknown
XVII Trs	[α1(XVI)$_3$]	Unknown	Basement membranes; skin, cornea, mucous membranes	Linkage of basal cells to stroma

Key

Fib = Fibrillar Net = Network forming Fil = Filamentous LCh = Long Chain
SCh = Short Chain FAC = FACIT Mul = Multiplexins Trs = Transmembrane

Information for this table was derived from several reviews; Hulmes (1992), Kielty *et al.*, (1993), Kadler (1995), and more recent information from more specific papers concerning individual collagen types (see main text).

1.1.2 Collagen, a definition

A protein can be described as a collagen if it contains at least one domain where three polypeptide chains associate to form a triple helix where the sequences of such domains comprise repeating triplets of amino and imino acids. The first position of the triplet repeat sequence is always glycine, with various amino and imino acids filling the second and third positions, the predominant residues at these positions being proline, and hydroxyproline.

1.1.3 The Collagen family

As the number of known collagen types has increased, so the means of classification and identification of these types has changed to accommodate new discoveries to the collagen family. It has been customary to name the collagen types as they have been found sequentially with Roman numerals. For instance, the first three collagen types are labelled types; I, II, and III. Because of their greater abundance accounting for almost 70% of the total collagen found in connective tissues, types I, II, and III have been referred to as the interstitial collagens (historic terminology as discussed by Kielty *et al.*, 1993). This terminology is inappropriate though, since they are not the only collagen types to appear within these tissues. Instead, collagens are now classified on the basis of their primary structure. The presence of Gly-X-Y domains is indicative of the presence of triple helical domains, but several collagens have been isolated that contain large and/or multiple non-helical domains (as indicated by their gene sequence). It is possible on this basis to classify the collagen types into the eight groups listed in Table 1.1, their major differences being schematically illustrated in Figure 1.1.

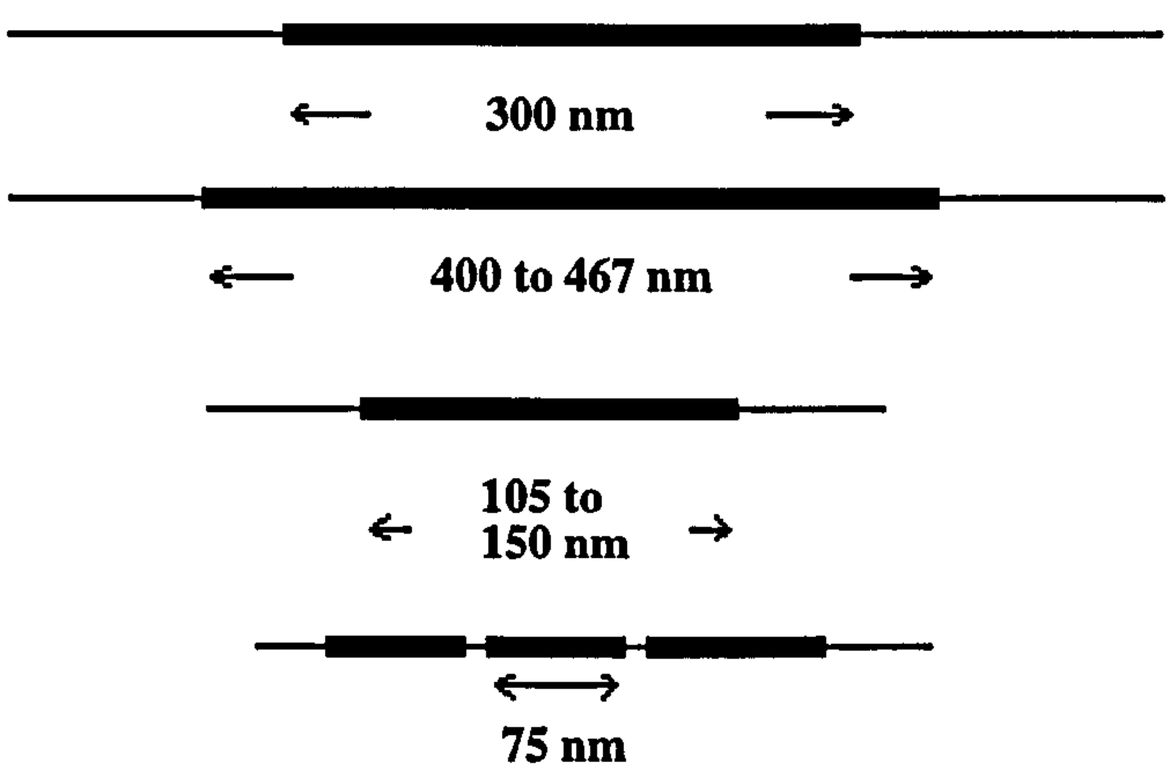


Figure 1.1 The major distinguishing features between the collagen subfamilies

Schematic representation of the basic distinctions made between collagen subfamilies, made on the basis of the collagen types amino–acid sequence. The main classification criteria being on the length of the triple helix domain (represented here by thick lines, non–helical domains being represented by thin lines). Adapted from Kielty *et al.*, (1993).

1.1.4 Collagen subfamilies

Fibril forming collagens are those found to contain an uninterrupted triple helical domain of approximately 300 nm. These collagens are synthesised as three pre-α polypeptide chains which after processing, nucleate to form the characteristic triple helical collagen domain. They are subsequently secreted to the extracellular matrix where they form fibrils and fibres (Chapman 1984). The network forming collagen (type IV) self assembles into a meshwork with multiple cell binding sites, to form a 'scaffolding' structure. It contains 20 short non-helical and possibly flexible domains to facilitate this role (Yurchenco and Schittny 1990), and is found in the basement membrane (Kefalides 1971). The filamentous collagen (type VI) is present in most tissues (Timpl and Engel 1987). It assembles into double beaded microfibrils with 104 nm periodicity, consisting of a short helical collagenous domain of 105 nm, banded by two large non-collagenous terminal domains. Type VII collagen can be classified distinctively as a 'long chain' collagen, and is found in the sub-basal lamina of skin. It has a large helical domain of approximately 400 nm, containing 19 interruptions, and a large non-collagenous N-terminal domain (Burgeson *et al.*, 1985).

Short-chain collagens are those that form comparatively short uninterrupted triple-helical domains approximately 150 nm in length. Types VIII, X, and XIII are all found in cartilage (types VIII and XIII are also found elsewhere, see Table 1.1), where type X is involved in bone development (Juvonen *et al.* 1992), whilst the function of the other two short chain types remains unclear. FACIT collagens (Fibril Associated Collagen with Interrupted Triple Helix) contain short triple-helical domains frequently

interspersed with non-helical collagen sequences. The molecular mass of proteins within the FACIT group vary greatly, due to the presence of large non-collagenous domains at the terminus of some α-chains (Linsenmayer *et al.,* 1991). They associate with the surface of fibrillar collagens, and researchers have suggested that they may regulate fibril assembly (Linsenmayer *et al.,* 1991). FACIT also associate between the surface of fibrillar collagens and the extra-cellular matrix (ECM) (e.g. with proteoglycans), (Linsenmayer *et al.,* 1990). The multiplexins (Multiple triple helix domains and interruptions), types XV and XVIII have been identified at the genomic level but have not yet been characterised. Type XV is found in the basement membrane of numerous tissues, whilst type XVIII is found in the basement membrane of lung, liver and kidney (Weckmann and Cabral 1996, Halfter *et al.,* 1998). Collagen XVII appears to be a homotrimeric transmembrane molecule of three 180-kDa α1(XVII) chains with a globular intracellular domain and a N-glycosylated extracellular domain of three 120-kDa polypeptides which form a triple helix at physiological temperatures. It is found in the basement membrane of skin, and cornea, and has been characterised via an artificially-expressed gene product (Schacke *et al.,* 1998).

1.2 Collagen type I: a fibril forming type

The subject of this thesis is the molecular structure of type I collagen, therefore it is necessary to focus on this single type in further discussion unless otherwise stated. However many of the structural features of type I collagen are common to all fibril-forming types, and important insights into the regions of helical and non-helical organisation of type I collagen may very well have direct relevance to the non-fibril forming types as well.

1.2.1 Physical properties

A collagen fibril can be described as crystalline due to its regular arrangement of identical subunits in the axial direction, and semi-crystalline in the lateral direction corresponding to disorder in the packing arrangement of the collagen fibrils (Fraser *et al.*, 1979). The axial arrangement becomes obvious under the electron microscope after appropriate staining and fixing. Under suitable conditions collagen fibres can clearly be seen to be composed of bundles of fibrils, each of which possesses the characteristic fibrillar banding pattern (Figure 1.2).

It was thought that the individual collagen molecules themselves were simply like stiff rods (in solution, see review by Chapman 1984). However from more recent studies and sequence information, although imino rich regions are 'stiffer' due to steric conformational restrictions, it seems that there are more flexible regions (Gelman and Piez 1980, Hofmann *et al.*, 1980). Fan *et al.*, (1993) demonstrated with model peptides of low imino content, an increased flexibility in the triplex backbone (NMR relaxation

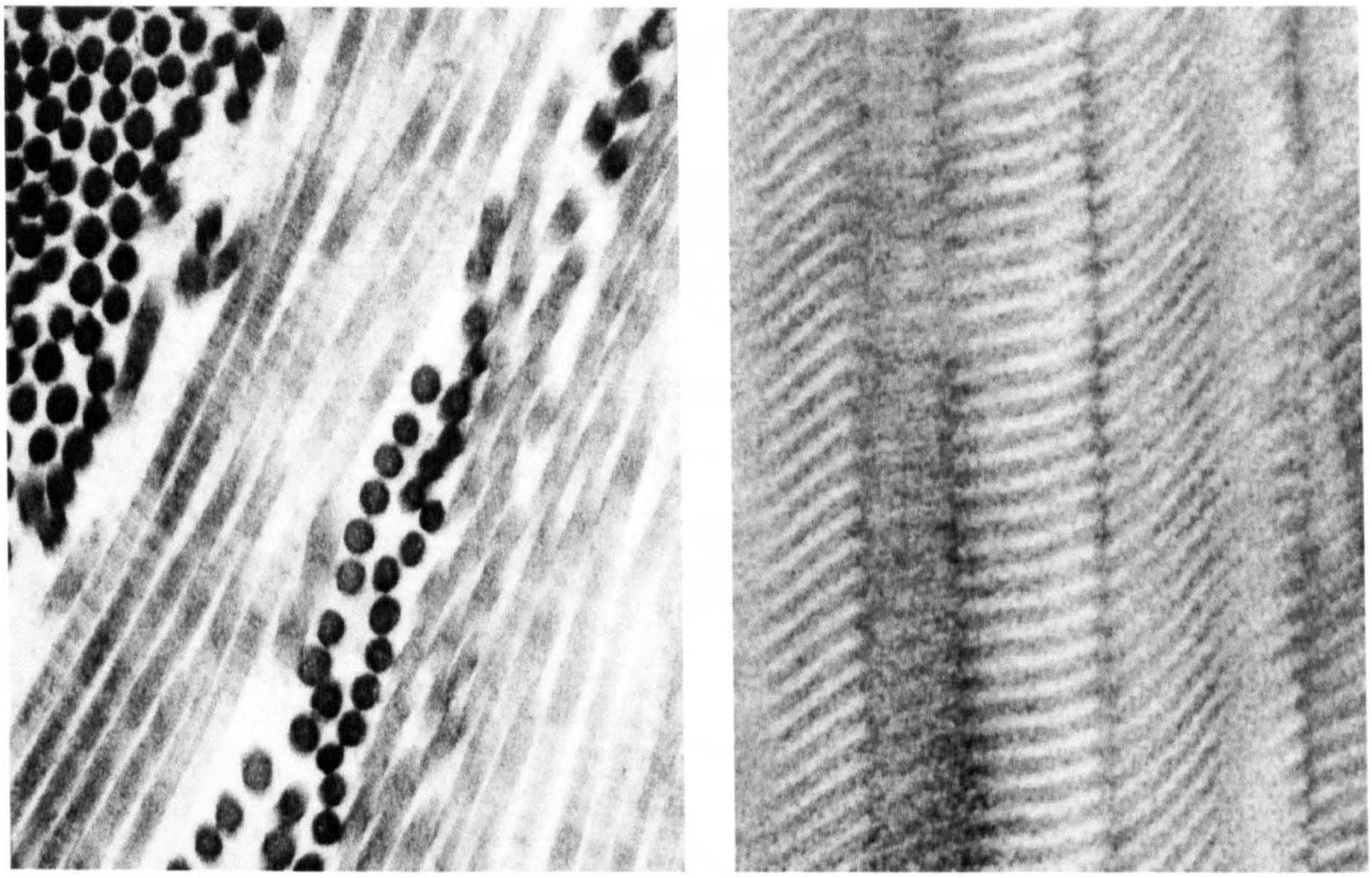
studies), whilst imino-rich regions were comparatively rigid. Overall, the collagen molecule in solution is neither rigid nor randomly flexible but appears to possess an intermediate level of semi-flexibility (Chapman 1984).

1.2.2 Type I collagen structure and organisation

The structure of the fibrillar collagens is not as straight-forward as one might think (see Figure 1.3, section 1.7.2, and Chapters 4 and 5). The complexity of the molecular structure and organisation of collagen fibrils is indicated in their biosynthesis, a multi-step process with significant post-translational modification (see Figure 1.4 for summary).

Because of the hierarchical structure of collagen (peptide chains to collagen molecules to fibrils and tissues) it is appropriate to describe the nature of the constituent parts of the structure before discussing their interconnectivity in forming collagen fibrils. In appreciating this description, it is useful to have a grasp of the organisation of the collagen molecule, which is summarised below:

Type I collagen is a heteropolymer composed of two $\alpha 1$ and one $\alpha 2$ peptide chains. (Piez *et al.,* 1961,1963). These form a triple helix region which extends for a little less than 300 nm, is 1.4 nm in diameter and approximately 285 Kdal in mass (Nold *et al.,* 1970). The $\alpha 1$ chains are some 26 residues longer than the $\alpha 2$ chain, which results in short regions at the N and C termini of the collagen chains that do not adopt a helical conformation.



a **b**

Figure 1.2 Electron micrographs of positive and negatively stained collagen fibrils

a) Positively stained collagen fibres of mammalian colonic submucosa. The fibrils are in lateral and longitudinal cross–section due to the crossply organisation of the tissue.

b) Negatively stained collagen fibrils of rat tail tendon (vertical alignment). Negative stain fills the gap region (dark bands) and binds to residues within the overlap region in much smaller quantities (light bands.).

Images courtesy of Linda and Tim Wess respectively

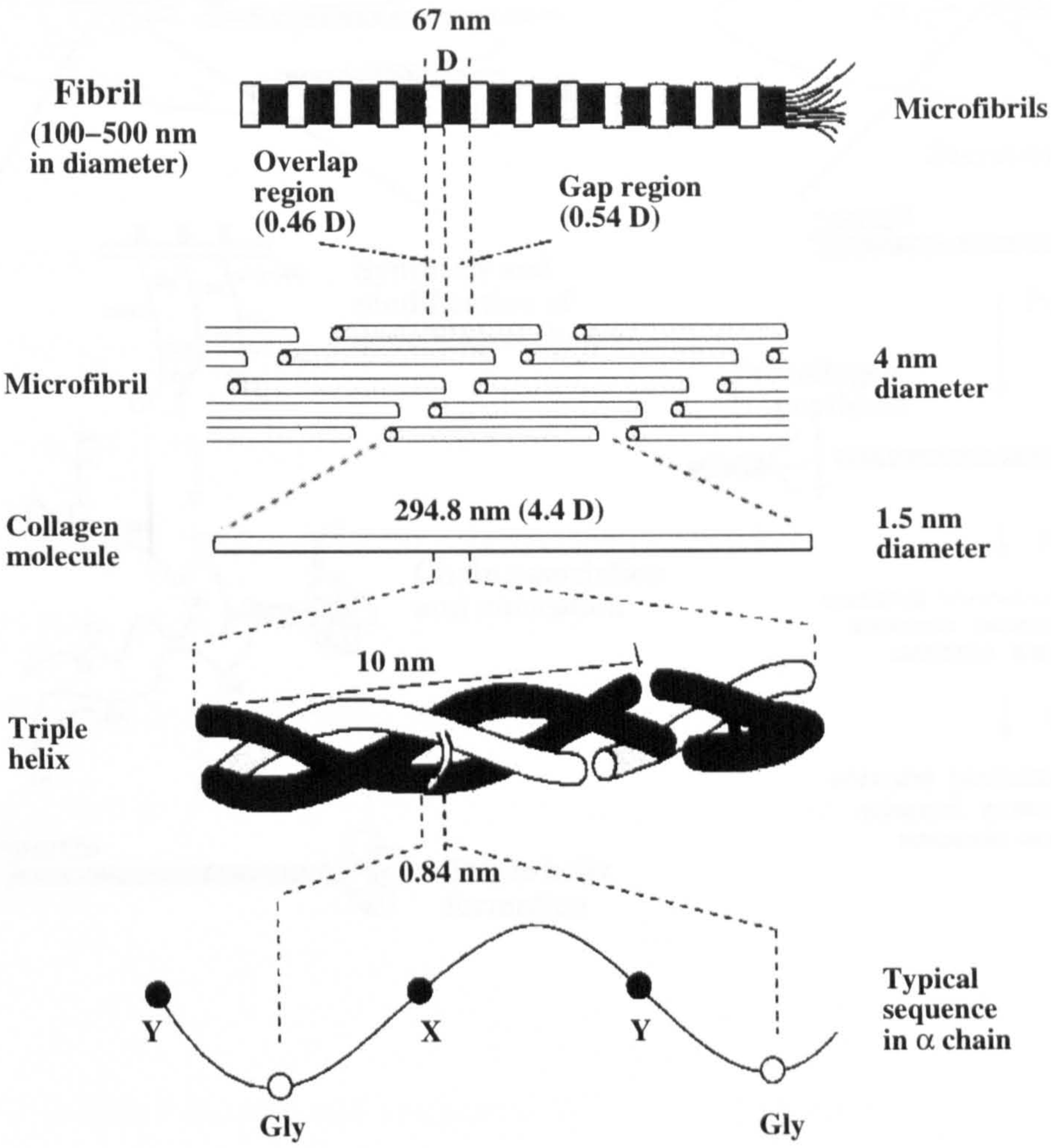


Figure 1.3 The structural hierachy of the collagen fibril

The principal component of rat tail tendon is the collagen fibril. The fibril (shown here with schematic negatively stained banding) is made up of collagen triple helices that are organised within a quasi–hexagonal 3D packing scheme with a regular D–staggered axial packing arrangement (see Chapter 5). The dimensions used here are approximate. This figure is based upon the information contained in several reviews (Engel and Prockop 1991, Hulmes 1992, Kielty *et al.,* 1993, Kadler 1995).

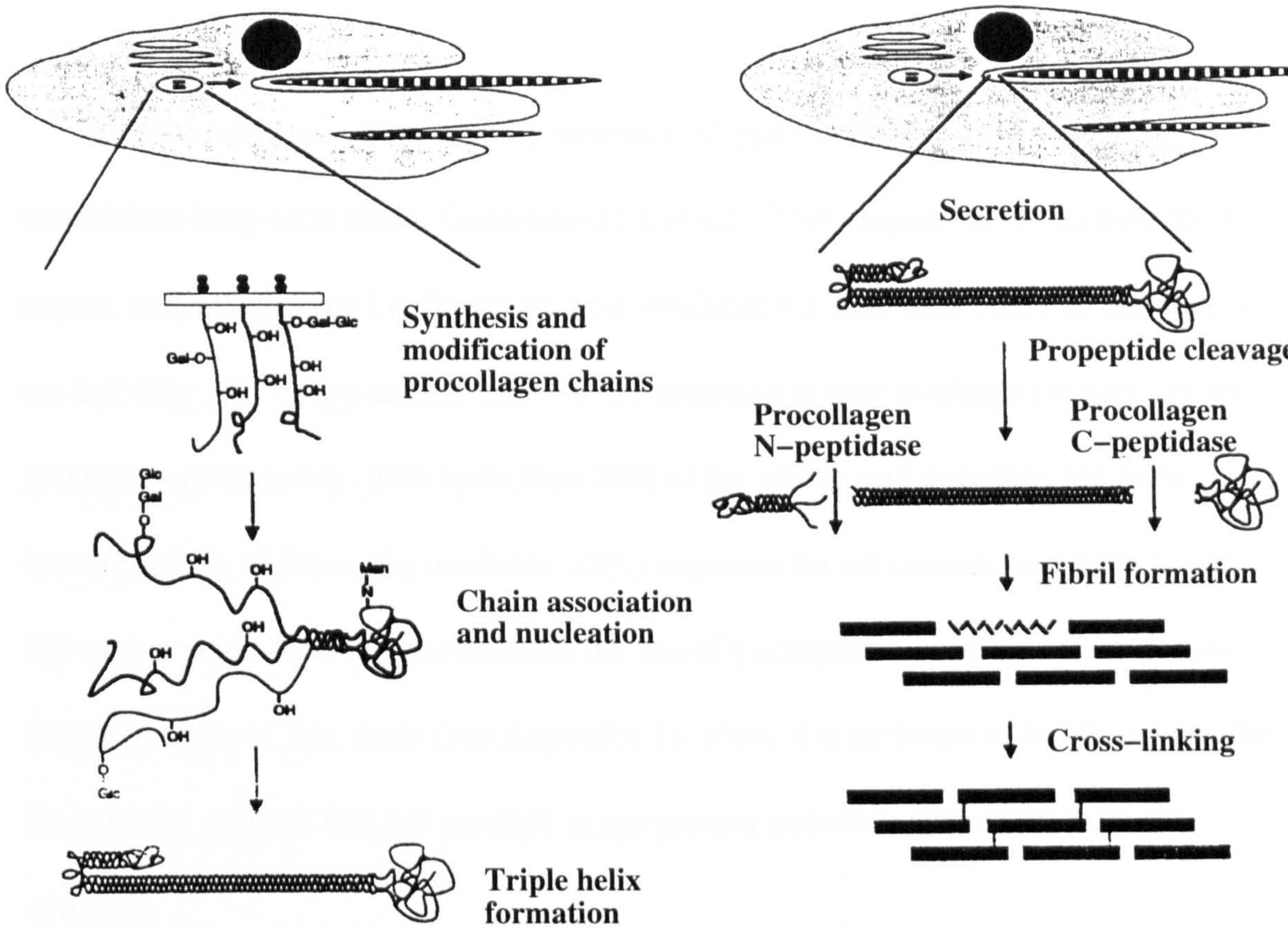


Figure 1.4 Collagen biosynthesis

Schematic summary of the steps (intracellular and extracellular) involved in the synthesis and assembly of type I collagen. Collagen polypeptide chains are synthesised by the ribosomes of the rough endoplasmic reticulum and secreted into the lumen, where they undergo enzymatic modification (Prockop *et al.,* 1976; Kivirikko and Myllyla 1989). Chain association and the formation of the triple helix follows, and the newly formed procollagen molecules are secreted (via Golgi secretory vesicles) into the extracellular space (Prockop *et al.,* 1976). Specific proteinases cleave off the N– and C– propeptides and the tropocollagen molecules assemble into the characteristic staggered array, stabilised by the (enzymatically–initiated) formation of intermolecular crosslinks (Hulmes *et al.,* 1989).

1.3 Primary structure of type I collagen

The determination of the primary structure of type I collagen is the result of tremendous long-term effort. Complete α1 and α2 cDNA sequences for human, chick, mouse, and bovine type I collagen are now available via such data banks as that held at the ExPASy site, (www.expasy.ch). The α1 sequence is now available (in part) for rat, although unfortunately, little more than 50% of the amino acid sequence has been determined. In addition, the available α2(1) sequence for rat constitutes a little less than 200 amino acids. This has necessitated the use of a composite cross-species sequence for the purpose of this study (see Appendix 1). First, it is pertinent to briefly review the background of work that has resulted in our present understanding of the primary structure.

1.3.1 Residue content and sequence

The unique amino acid content of type I collagen, its low levels of tryptophan and tyrosine, absence of cysteine, and its high glycine and imino acid content has set the protein apart as significantly different at every level of structure from that of globular proteins, a fact observed at the very outset of collagen research by Bowes and Kenten (1948). The predominant form of collagen type I is a heterotrimer made of two α1 and one α2 polypeptide chains (Piez *et al.*, 1963). Initially it was thought that this ratio is persistent throughout all vertebrate collagens, but with three decades of study this has been shown to be incorrect; for instance, types II and III are homotrimers (Kadler 1995).

Sequence data available firstly as composite sequences from calf and rat skin (Hulmes *et al.,* 1973, Fietzek and Rexrodt 1975) and later through an increasing number of completed sequences (Gene bank, Bio-rad, ExPASy), shows a great deal of similarity between the α1 and α2 sequences. Although both polypeptides are coded for on separate genes and are separable by ion-exchange chromatography and SDS-polyacrylamide gel electrophoresis (Kielty *et al.,* 1993), the α1 chains of types I and III collagen have more similarities with one another than the α1 and α2 chains of type I collagen do (Hofmann *et al.,* 1980). Both the α1 and α2 chains have three distinctive regions, the longest, a central region of 1014 amino acid residues, contains the tripeptide repeat unit Gly-X-Y and hence adopts the triple helix conformation. At both ends of this region there are two short peptide sequences within which the tripeptide repeat sequence does not occur. These two regions are located at the N and C termini of the molecule and are termed the non-helical telopeptides - since they do not adopt the triple helical conformation (Hulmes *et al.,* 1980). The length of the C-α1 and the C-α2 chains in the C-terminal region are 25 and 6 residues in length respectively. Whilst in the N terminal region the α1 and α2 chain are 16 and 9 residues (See Appendix 1).

1.3.2 Sequence to structure

The Gly-X-Y repeat unit is an absolute requirement for the formation of the triple helical structure (Rich and Crick 1955). Within this sequence the strategic location of imino acids play a crucial role in conveying rigidity or rather imposing conformational restrictions upon a particular region of the helix. Such regions are particularly prominent at the end of the central helix region before the interface with the non-helical telopeptides.

The conformational restrictions of imino-acid rich regions on the triple helix structure can be measured as a function of the thermal stability of the triple helix (for a review see Engel and Prockop, 1991). A number of recent spectroscopic, computational, and crystallographic studies based on this premise have investigated the stability of different imino-rich, -poor, and -absent regions (Li *et al.*, 1993, Fan *et al.*, 1993, Bella *et al.*, 1994, Lee *et al.*, 1996, Chan *et al.*, 1997, Kramer *et al.*, 1998, Ramshaw *et al.*, 1998, Nagarajan *et al.*, 1998), as well as point substitutions of Gly for other amino acids (Bella *et al.*, 1994, Kuivaniemi *et al.*, 1991, Bella *et al.*, 1996, Yang *et al.*, 1997). Notably, the presence of the repeat Gly-Pro-Hyp conveys a great deal of stability to the helix, whilst other regions allow the molecule greater flexibility (Li *et al.*, 1993, Paterlin 1995).

The non-random occupancy of positions X and Y by amino acids rather than imino acids (Chan *et al.*, 1997, Yang *et al.*, 1997, Ackerman *et al.*, 1999) also plays an important although not well understood role in conveying stability to the triplex and

packing arrangement of collagen chains through electrostatic and hydrophobic associations (Venugopal *et al.,* 1994, Vitagliano *et al.,* 1993) and covalent crosslinking (Piez *et al.,* 1961, Nakamura 1987). These associations, together are believed to convey a great deal of stability to the overall structure of the collagen fibril.

The tripeptide repeat sequence of collagen is an essential prerequisite to the structural integrity of the triple helix. For instance, when glycine (naturally occurring as every third residue in the amino acid sequence) is replaced by another amino acid residue, the stability of the helix is greatly disrupted (Paterlin *et al.,* 1995, Kuivaniemu *et al.,* 1991, Ryhänen *et al.,* 1983, Long *et al.,* 1993). The substitution of glycine at the beginning of the C-telopeptide also disrupts the molecular arrangement significantly enough so as to cause mild osteogenesis imperfecta (Cohn *et al.,* 1988). The prevalence of the imino acids also plays a major role in conveying a unique structure to the collagen triple helix through their limited range of conformational alternatives, and through the hydrogen bonding system of hydroxyproline (Brodsky 1999).

1.4 The triple helix structure of collagen as determined by X-ray fibre diffraction

Our current understanding of the conformation of the collagen triple helix has come about through careful consideration of the unique amino acid make-up of collagen, and by reconciling this with the available X-ray data. The conclusions of these studies, essentially of a model building nature, have stood the test of time well, and have recently been confirmed by various NMR and X-ray crystallography studies discussed earlier (a review of the earliest work can be found in Astbury 1933).

Significant improvements in the quality of available X-ray data came about once it was demonstrated that the quality of data could be enhanced through limited stretching of the mounted sample before data collection (Cowan *et al.*, 1953; Fraser *et al.*, 1979). However, it still proved impossible to determine directly the helical parameters from the diffraction data. For this reason investigators moved to an approach based on modelling the observed data, assuming a helical model limited by the torsion angles of the constituent residues within the polypeptide backbone (Cochran *et al.*, 1952). The fibre diagram was then indexed accordingly, that is, a structure containing ten equivalent units within three turns of a helix (Cohen and Bear 1953). However, it was not until Ramachandran and Kartha (1954) postulated that collagen was of a triple helical nature, that the triple helical model and the quantitative data were reconciled. In their model, each α chain formed a twisted, left handed helix, which together with the other two chains form a triple helical right handed coiled coil (shown schematically in Figure 1.5 part a). This model explains the triple helical nature of the collagen molecule whilst also accounting for the persistent presence of glycine in every third position in

the collagen amino acid sequence. The close proximity of the chains at the third repeat position of the tripeptide repeat leaves little room for glycine, and insufficient space for amino-acids with larger side groups, a fact made graphically clear by the prevalence of life threatening conditions where base substitutions have resulted in alternative amino acids occurring at the position of glycine in the triple helix (e.g. Cohn *et al.,* 1988, Kuivaniemi *et al.,* 1991, Bella *et al.,* 1996, Yang *et al.,* 1997).

The basis of the helical twist is believed to be due to the presence of proline, since the torsion angles, ϕ, and φ, available to the imino acid are more restricted than that of other amino acids. This effectively limits the conformational possibilities resulting in the unique collagen arrangement. Similar helical paths are exhibited by the synthetic peptides poly(Pro)II and poly(Gly)II (Ramachandran 1968), as confirmed by Yonath and Traub (1969). X-ray diffraction of a synthetic peptide, poly-(Gly-Pro-Pro), proved to be similar to that of collagen X-ray diffraction data. The other imino acid, hydroxyproline also plays a crucial role in conveying stability to the helix through its hydrogen bonding system (Fan *et al.,* 1993, Li *et al.,* 1993). It was the hydrogen bonding system of the collagen II model (Rich and Crick 1955) (itself a development of the Ramachandran model) that proved to be the most consistent with the data available to that date. The major weakness of the Ramachandran model being the stochastic consequence of the Ramachandran model triplet repeat (Gly-nonPro-Pro/Hyp). This model in turn was then refined by a linked-atom least squares procedure using quantitative X-ray fibre diffraction data (Fraser *et al.,* 1979). This resulted in the development of the molecular model, to include a supercoiled conformation. The refinement of the model to define the collagen molecule as a three start helix in turn

accounts for a hydrogen bonding system that stabilises the conformation further. The stabilising effect would come from the inter-chain hydrogen bonding between the peptide amino group of the glycine residue in one chain and the peptide carboxyl group of the residue in the X position of the tripeptide repeat of another chain. Leaving the residue in position Y free to hydrogen bond with water, as an important source of stabilisation in this structural protein as is the case with globular proteins (Fraser *et al.*, 1987, Bella *et al.*, 1994, Brodsky 1999).

1.4.1 Collagen helical parameters

The observed unit height (translation per amino acid) of the helix has been determined to be 0.2983 nm in hydrated stretched collagen (Fraser *et al.*, 1979), the displacement of each amino acid parallel to the helix axis. The unit twist of the superhelix being 107.1 degrees +/- 0.6 degrees. This corresponds to a 10/3 turn helix (ten units within three turns of the helix). It must be remembered that these parameters are averages, since local variations within the helix (a low imino acid concentration for instance) could significantly alter the bond length and torsion angles as indicated by Bella *et al.*, (1994) and Kramer *et al.*, (1999) in their X-ray crystallography studies of collagen-like peptides, as discussed in the following sections.

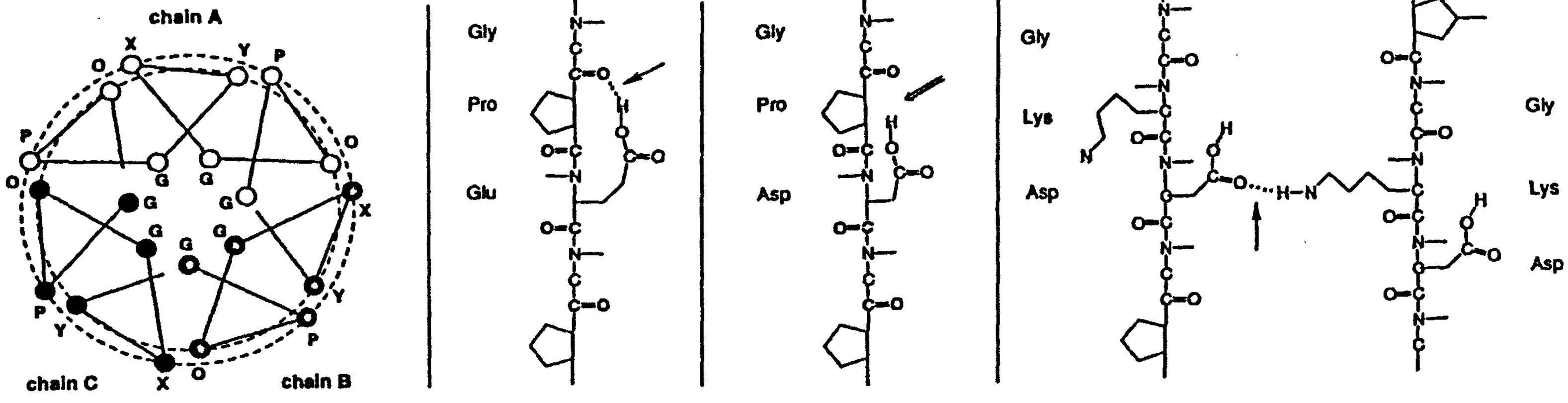


Figure 1.5 Schematic diagrams of the steric positions of ionizable residues in the collagen–like peptide triple helix

a) Cross–section of the 10/3 triple helix of the collagen–like peptide, showing the increased solvent exposure of residues occupying position X of the GXY triplet repeat relative to residues occupying position Y. Residues in position Y of chain A are at the same axial level and sterically close to residues in position X of chain B.

b) A single chain of a GPE–containing peptide. The abilty of the glutamic acid side chain to form an intrachain hydrogen bond (marked) may in part explain the stabilising effect of Glu in position Y.

c) A single chain of a GKD–containing peptide. The shorter side chain of the Asp residue is unable to form the intrachain hydrogen bond (marked).

d) Two staggered GKD–containing peptide chains (of the triplex). Favourable electrostatic interactions may occur between the Lys and Asp residues of neighbouring chains (marked). When the Lys is in position X and the Asp residue in the Y position, the residues are at the same axial level, and sterically close to one another. Adapted from Chan *et al.,* (1996).

1.5 Crystallographic and NMR studies of short collagen-like peptides

The crystallographic study of Bella *et al.*, (1994) of a collagen-like peptide ($[GPO]_{10}$), provided the first visualisation of the conformation of individual residues in the collagen helix, and identification of specific interactions with bound water molecules. Therefore a review of this, and the closely related studies that followed it is necessary here.

The structural solution of Bella *et al.*, was based upon the fibre diffraction model for collagen in kangaroo tail tendon (Bella *et al.*, 1994, Fraser *et al.*, 1979). It showed reasonable agreement with the X-ray data and the stereochemical restraints of the constituent residues. The three peptides coiled into polyproline II helices which supercoiled into a stable triple helix with a slight distortion in the locality of an alanine for glycine substitution, which did not disrupt the overall helix structure. The three helices were shown to be held in specific register by the formation of interchain X=O——H–N hydrogen bonds (see Figure 1.5), verifying earlier fibre diffraction studies (Rich and Crick 1961, Fraser *et al.*, 1979).

Helical parameters were determined from the Gly-Pro-Hyp regions of the peptide, and compared to those measured from the X-ray fibre diagram. The unit height over this Gly-Pro-Hyp region was measured to be 0.286 nm, consistent with that determined for unstretched collagen or model peptides (Rich and Crick 1961), but is shorter than that of stretched collagen (0.2983 nm, Fraser *et al.*, 1979). It should be noted, that the fibre diffraction-determined values are averages over the length of the collagen helix, whilst

those determined from the crystal structure are that of local imino rich regions. Bella *et al.*, (1994) observed that in the collagen like-peptide, the imino content constitutes 66% of the total residues, whereas the proportion in native collagen is approximately 20%. The average twist of the basic helix (terminology as used by Bella *et al.*, meaning the twist of the individual peptide chains) was shown to be 60°, and -100° for the super helix (triplex) although both twist angles showed a broad distribution from the average, from residue to residue. The unit heights of the basic helix and superhelix in contrast remained consistent, these being 0.84 nm and 0.28 nm respectively (0.84 nm / 3 = 0.28 nm, the projected spacing of residues on the Z-axis of the triple helix). These features are shown schematically in Figure 1.5. In addition to confirming the overall structural details of the collagen triple helix, Bella *et al.*, made five significant observations:

1. Triple helices are able to change their twist through small variations of main chain torsion angles, without apparent changes in the unit height or interchain hydrogen bonding pattern.

2. The Gly-Ala substitution resulted in a small local distortion to the triple helix, but the triplex was still stable and able to overcome the steric problems caused by the addition of methyl groups to the Gly site in two ways:

i) A twist relaxation, that is the triple helix 'unscrews' in the region of the Ala substitution.

ii) The presence of four interstitial water molecules that provided additional hydrogen bonding and helped mediate the triple helical assembly.

3. The triple helices were surrounded by a highly structured cylinder of hydration mediated largely by the hydroxyproline residues. This cylinder of hydration may play a crucial role in determining the lateral packing of the collagen molecules. A 1.4 nm reflection was observed from the crystal that seems to correspond to the inter-molecular lateral spacing observed in (hydrated) rat tail tendon (Bella *et al.*, 1994, Lee *et al.*, 1996, Kramer *et al.*, 1998).

4. The hydrogen bonded water molecules are bound to the Gly and Hyp carbonyl groups as well as the 4'OH groups of Hyp (Bella *et al.*, 1994, Nagarajan *et al.*, 1999, Brodsky 1999). This indicates that ordered water is in direct contact with the peptide acceptor groups, conveying a degree of protection to the peptide from the bulk solution.

5. The subtle defects revealed by the Gly-Ala substitution in the triple helix may shed light on the nature of interstitial diseases such as osteogenesis imperfecta. It has been demonstrated that the Gly-Ala substitution significantly reduces the thermal stability of the triplex (Yang *et al.*, 1997).

1.5.1 Host-guest model peptide studies

By demonstrating the validity of the collagen-like peptide as a model for the collagen triple helix, a number of studies investigating the role of different residue types in the triplet repeat followed. These have been centred around the host-guest peptide approach, where a host peptide chiefly consisting of the basic sequence Gly-Pro-Hyp, contains a point change from the Gly-Pro-Hyp norm (e.g. Gly-Pro-Arg, or Ala-Pro-Hyp etc.). The importance of this (non-random) distribution of ionizable and polar residues has been indicated in a number of studies (Li *et al.*, 1993, Fan *et al.*, 1993, Bella *et al.*, 1994, Lee *et al.*, 1996, Chan *et al.*, 1996, Chan *et al.*, 1997, Kramer *et al.*, 1998, Ramshaw *et al.*, 1998, Nagarajan *et al.*, 1998).

Amongst the observations made, Chan *et al.*, (1996,1997) demonstrated the equivalence of Arg to that of Hyp in terms of conveying thermal stability to the triplex when substituting Hyp in the Y position of the triplet repeat. Yang *et al.*, (1997) went on further to synthesise a GPR8 peptide (Gly-Pro-R, where R = any amino/imino acid, 8 referring to the length of the peptide) differing from the host peptide (Gly-Pro-Hyp), in that Hyp had been completely replaced throughout the three peptides in the helix by Arg. The GPR8 peptide was slow to polymerise into the helix structure, and its thermal stability was some 13°C lower than that of the host peptide, indicating that Arg only conveys equivalent stability to that of Hyp when the replacement is made in an imino-rich environment. Replacement studies with guest peptides with other charged residues in the Y position seemed to indicate that residues with relatively long side chains stabilise the triplex, through same chain, and inter-chain hydrogen bonding. Whilst the

Asp guest peptide, with the comparatively short side chain of the Asp residue, showed a significant reduction in the thermal stability of the triple helix. Chan *et al.,* also indicated the non-equivalence of the X and Y position in terms of residue replacement verses thermal stability. For instance, the replacement of Hyp in the Y position resulted in a spread of thermal stabilities in the guest peptides across 14°C, whilst replacements made at the X position resulted in a spread of thermal stabilities within a 5°C range, between that of Gly-Pro-Hyp, and Gly-Ala-Hyp (see Figure 1.6). This could be due to greater solvent exposure for the residue in the X position relative to that of the residue in the Y position; the charged residue in the Y position causes a greater charge repulsion in the triplet than would a charged residue at position X (see Figure 1.5).

About 8% of triplets found in fibrils from collagen are those where the Pro-Hyp motif in the X-Y position have been replaced with oppositely charged residues. The majority of these have a basic residue at X, and an acidic residue at Y (Hofmann *et al.*, 1980, Doyle *et al.,* 1974a). The work of Chan *et al.,* (1997), demonstrated that this arrangement is stabilising, where each residue individually contributes to the (cumulative) stabilising effect. Although this does not rule out the possibility of ion pairs forming since the residue in position Y of one chain is at the same axial height and sterically near the residue in the X position of a neighbouring chain.

Electron microscopy studies (Doyle *et al.,* 1975, Chapman and Hulmes 1984) reveal the distribution of bands of charged residues along the length of fixed collagen fibrils. cDNA sequences confirm these observations. The fact that the charged residue distribution is banded like this, also suggest that their distribution is not random, as

confirmed from the available amino acid sequence data (see Appendix 1). This cDNA sequence information also reveals the presence and location of a small number of residues able to selectively bind to heavy metal stain. Methionine, histidine, and tyrosine residues are present in small quantities and when bound to heavy atoms, produce a limited number of stain vectors in diffraction experiments (Bradshaw *et al.*, 1989). Tyrosine is of particular interest since it is present only at the start and end of the C-telopeptide, and at the end of the N-terminal telopeptide, the structural significance of which has not been yet made clear.

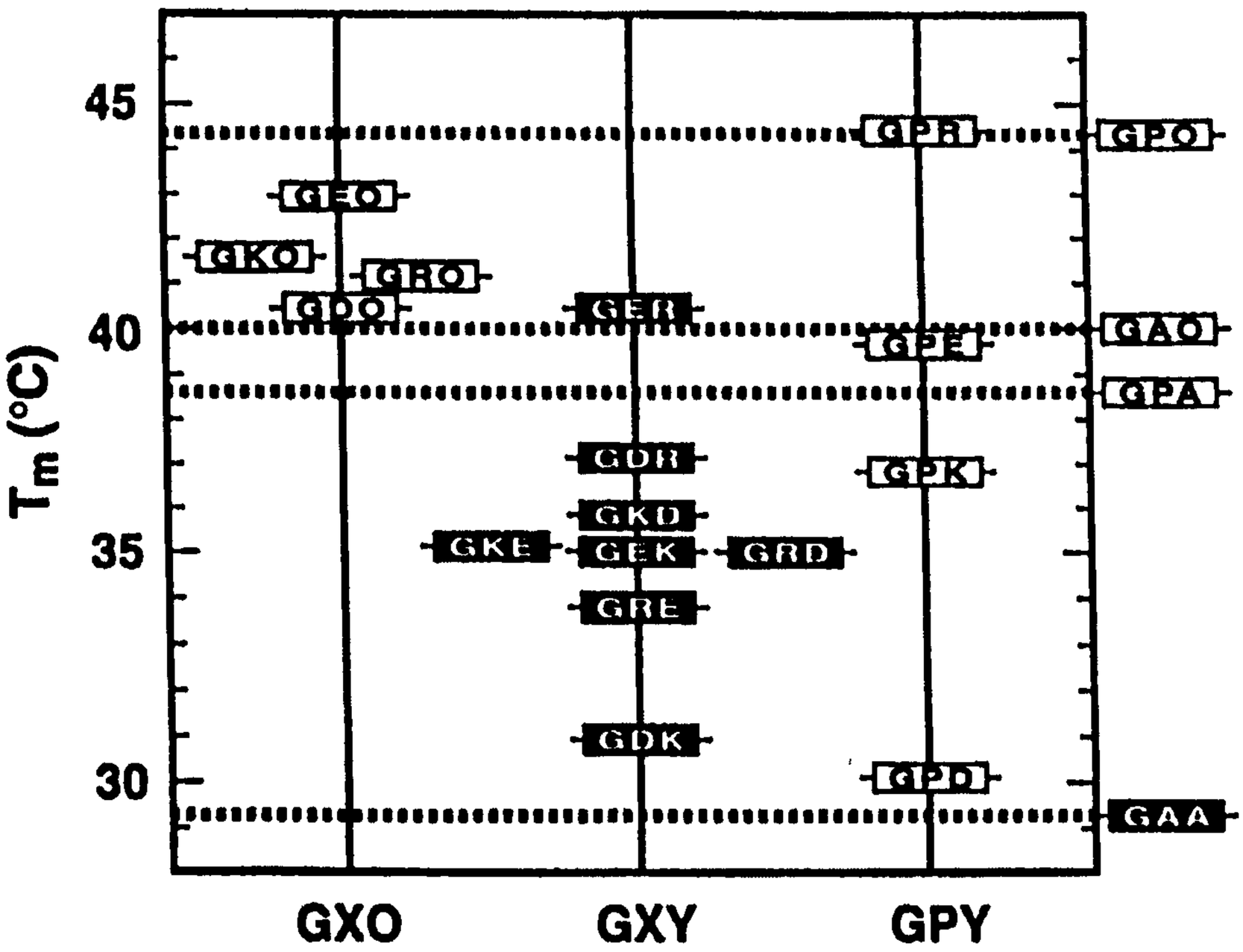


Figure 1.6 Triple helix thermal stabilty of short collagen–like peptides

The host–guest system (descibed in the main text) of Chan *et al.*, (1997) involves the inclusion of a non–imino acid triplet into an imino acid rich peptide 'host'. The thermal stabilty of each of these peptides are plotted here against the position (X or Y) in the triplet repeat of specific amino acid residues (see main text for discussion). Adapted from Chan *et al.*, (1997).

1.6 Telopeptides

The Gly-X-Y repeating sequence of amino-acids characteristic of the collagen triple helix does not continue into the N and C termini of the molecule. Hence, the telopeptides are incapable of forming a collagen triple helix, and therefore do not necessarily conform to the same axial translation as that of the rest of the molecule. The residue to residue spacing of the non-helical telopeptide regions has been demonstrated (by modelling studies) to be less than that of the main chain collagen helix; Hulmes *et al.*, (1980) reported a value of 0.282 nm for the main chain, 0.241 nm in the N-terminal telopeptide and 0.2 nm in the C-terminal telopeptide. This does not mean that the non-helical regions make any less significant contribution to the overall conformation of collagen. In the self-assembly process of collagen fibrils, the short non-helical end regions of the molecules are required to ensure correct molecular registration and crosslink formation (Helseith and Veis 1981), and they ensure the development of structural strength and integrity. Where crosslinking has been found to be inhibited, or hyperstabilsed (additional crosslinking), numerous connective tissue diseases have been cited to occur (Eyre *et al.*, 1984). Hence, the conformation of the telopeptide regions has been sought in order to determine their key role in fibrillogenesis, and their effect on packing organisation (Wess *et al.* 1995,1998).

In a model-dependent X-ray diffraction study (Hulmes *et al.*, 1977), it was found that localised amino-acid sequence and conformation changes within the telopeptide regions significantly affected the intensity distribution of the low-angle meridional orders. This demonstrated that the conformation of the non-helical telopeptides

significantly affects the overall structure of the collagen molecule. A number of other studies based upon the primary structure of type I collagen (Hulmes *et al.,* 1980, Jones and Miller 1987,1991, Vitagliano *et al.,* 1995, Wess *et al.,* 1995, 1998a), have indicated that the telopeptides are theoretically capable of adopting a range of contracted, extended, folded or mixed combinations of these conformations.

Otter *et al.,* (1988) working with a single synthetic α1 chain C-telopeptide and NMR spectroscopy, were able to demonstrate that the isolated chain was axially condensed, with a possible disposition to folding. Scott (1986), proposed a folded C-telopeptide conformation to account for the spectroscopic data obtained in the analysis of a single α1 chain C-telopeptide isolated from calf skin, and Bradshaw *et al.,* (1989) from an X-ray study of the native protein demonstrated that the telopeptide regions are axially contracted. Although in the study of Bradshaw *et al.,* (1989), they were able to demonstrate the contraction of the telopeptide regions compared with the triple helical region. Their study was not of sufficiently high resolution to determine what conformations the N or C telopeptides might adopt, whether folded or simply contracted.

1.7 Fibrillar structure

The fibrillar nature of type I collagen has been investigated in rat and kangaroo tail tendon where it was found to possess significantly greater crystallinity in both the axial direction and the three dimensional packing structure than other tissues (Jesior *et al.*, 1980, Wess *et al.*, 1995, 1998a). Even here, differences in the crystallinity and complexity of the axial and lateral organisation of the fibril have resulted in a separation of these structural aspects (for the purposes of simplifying investigation) in the vast majority of studies.

The term 'axial structure' refers to the projection of the molecular array onto the single (fibril) axis. By disregarding the lateral organisation in the other two directions some aspects of the structure determination are simplified. However, the main reason for this consideration is due to the more readily available and interpretable data in the axial directions. This is demonstrated by the fact that type I collagen possesses a high degree of crystallinity in the axial orientation (Wess *et al.*, 1998a, Orgel *et al.*, 2000). However, numerous attempts have been made to elucidate the nature of the lateral packing arrangement of the collagen fibres, chiefly through the use of the available X-ray diffraction data (see sections 1.7.2 and 5.1). These approaches have been model based, and there have been no attempts to produce a structure based on phases solved via non-model dependent means, due in part to the complexity involved in measuring the overlapping Bragg reflections in the equatorial direction of the fibre diagram.

1.7.1 Axial organisation

The first X-ray diffraction evidence that type I collagen fibres possess long-range periodic order in the axial direction was presented over 60 years ago (Clark *et al.*, 1935, Wyckoff *et al.*, 1935, Clark & Schaad, 1936, Corey & Wyckoff, 1936). Since then, data from electron microscopy, X-ray and neutron diffraction studies have led to the generally accepted understanding of the axial or one dimensional structure, and molecular arrangement. Schmitt *et al.*, (1942) were the first to visualise the axial fibrillar scheme. This electron microscopy study revealed that the fibril pattern has a repeat periodicity of 64 nm (in dry samples). This is very much less than the length of a single collagen molecule (L), approximately 300 nm. In 1955, Schmitt *et al.*, proposed that L=4D and put forward the quarter stagger model to account for the discrepancy between D and L. This model proposed that the collagen molecules are organised parallel to one another, but staggered by 64 nm and, or multiplies thereof.

The addition of electron dense atoms to collagen fibrils results in a dark and lighter banding pattern within the D-repeat of the fibril repeat. It was thought that the dark regions correspond to regions where a molecular segment is missing in the gap region. This accounts for why these areas are darker when negatively stained, since more electron dense atoms are found in these areas due to the greater available space to fill. Hodge and Petruska (1962,1963) proposed a non-integer value for L (4.4D) and a ratio of 0.4D for the overlap and 0.6D for the gap regions that resulted from their axial stagger model of dry collagen and accounted for the banding pattern of negatively stained samples.

The Hodge-Petruska scheme divides the collagen molecule into 5 units, all of regular length (1D) with the exception of unit 5, which has a length of approximately 0.4D. Applying this model to the electron micrograph of negativity stained dry collagen, the dark regions correspond to the region beginning at the end of the short segment, unit 5, whilst the lighter bands are the areas where all five segments are present, (shown in Figure 1.2b).Positive staining differs from negative staining in that most of the stain is washed out, leaving only sites where the stain has electrostatic interaction with the amino acid residues of that locus (Nemetschek *et al.,* 1955). It has been shown that the non-periodic banding pattern that results (the pattern resulting from bands of charged residues) is closely related to the amino acid sequence (Doyle *et al.,* 1974a; Chapman, 1974; Chapman & Hardcastle, 1974; Bruns & Gross 1973, 1974; Doyle *et al.,* 1975). Meek *et al.,* (1979) and Doyle *et al.,* (1974b) were both in agreement that the number of residues within a D repeat is 234.2 +/- 0.5.

X-ray diffraction work confirmed the 64 nm periodicity in dry collagen (Bear 1942,1944). But it was not until 1956 (Tomlin and Worthington) that the band and interband scheme (now known as the gap and overlap) was confirmed to occur similarly in the native wet state. The inherent strength of X-ray diffraction studies, such as those made by Tomlin and Worthington, is that they do not require a sample preparation, a process that significantly alters the native wet state of the collagen fibrils. Subsequent X-ray diffraction studies have shown that it is possible to make detailed investigations of collagen structure *in situ*, which is simply impossible for studies reliant on conventional transmission electron microscopy. In addition, diffraction techniques

allow for a visualisation of the whole structure at high resolution, not just the surface of molecular arrangements, as is the limitation of even advanced electron microscopy techniques currently available (with the exception of TEM – by definition a *transmission* technique). There is however a major obstacle to overcome in any diffraction study, known as the 'phase-problem'.

In several studies, investigators have elected to avoid solving the phase component of the structure factors altogether (Hulmes *et al.,* 1977,1980). Hulmes *et al.,* using neutron and X-ray diffraction data were able to confirm a number of previously determined parameters for axial collagen studies. However, significant differences were found in the gap, overlap ratio and the size of the D-period, but explainable since the study was on native state (wet) collagen fibres, and the results compared to dry state parameters. Hulmes *et al.,* were able to make a significant contribution however, confirming the number of amino acids within the axial unit cell to be 234. They also presented compelling evidence for the axial contraction of the non-helical telopeptides, which themselves determine the boundaries of the gap:overlap region. This has been confirmed by successful attempts to phase the meridional section of the collagen X-ray diffraction pattern (Bradshaw *et al.,* 1989, Orgel *et al.,* 2000). In the study of Bradshaw *et al.,* 52 meridional orders of diffraction for the native protein and three isomorphous derivatives were used to calculate a map of a maximum axial resolution of 1.29 nm. In this study, Bradshaw *et al.,* were able to show good correlation between the estimated and observed labelling positions of the heavy atoms used, based on an understanding of the ligand/metal chemistry, amino-acid sequence, and the experimental data. They also confirmed the earlier work of Hulmes *et al.,* (1977, 1980). The similar although much

higher resolution study of Orgel *et al.*, (2000) is described and discussed in Chapter 4 (the work of Orgel *et al.*, being based upon the thesis work presented in Chapter 4).

1.7.2 Lateral packing arrangement

Although not all collagen-containing tissues are observed to possess lateral crystallinity and examples are limited to a number of specific tissues (Jesior *et al.*, 1980, Eikenberry *et al.*, 1984, Miller and Wray 1971), it is believed that the distribution of collagen fibres is not completely random in tissues that do not show crystallinity via X-ray diffraction (Hulmes *et al.*, 1995, Wess *et al.*,1998a).

In contrast to investigations of the axial structure, there are significantly fewer publications concerning the lateral organisation (packing structure) of collagen. This is due principally to difficulties in obtaining data of at least reasonable quality so as to ensure unambiguous interpretation. Studies that rely solely on the electron microscope are unable to answer the question of how collagen packs in three dimensions, due to limited resolution and possible disruption of the native state during sample preparation and data collection. Hulmes *et al.*, (1981,1985) were able to observe preferential lattice orientation of spacings, lattice curvature and discontinuities between crystalline domains in their freeze-fracture electron microscopy studies. In contrast, X-ray diffraction reveals lateral crystallinity (North *et al.*, 1954), and provides detailed information. However, the reflections are often diffuse and difficult to interpret due to the 'liquid crystal' nature of the lateral packing (Hulmes *et al.*, 1995). The diffuse 'equatorial fan' of background intensity to the discrete equatorial Bragg reflections is,

although significant as a function of the inherent disorder within the tendon, problematic to measuring the intensity of the discrete reflections.

Analysis of the intensity peaks on the equator resulted in a number of (mostly) conflicting early models to account for the observations (for a review of the earliest models see Miller 1976, and Bornstein and Traub 1979). Of these, the most widely accepted is the microfibril model first put forward by Smith (1968). He proposed that the basic unit of packing is a group of five molecular chains that pack with helical symmetry. These basic units are referred to as microfibrils; they are discrete, rope-like, and pack together on a quasi-hexagonal lattice (not the pentagonal arrangement of Smith 1968) to form fibrils (Fraser and MacRae 1981, Piez and Trus 1981, Fraser *et al.*, 1983, Fraser *et al.*, 1987, Wess *et al.*, 1998a).

The model-based approach to the 3D collagen structure elucidation received a much needed boost as the quality of data improved significantly over the last three decades. As a result, Hulmes and Miller (1979) were able to propose that the molecules packed in a quasi-hexagonal lattice. Later, Fraser *et al.*, (1983), utilising phospho tungstic acid (PTA) stain to further improve the contrast of the discrete intensity over the diffuse, concluded that the unit cell is triclinic, later confirmed by Wess *et al.*, (1995). Wess *et al.*, (1995) demonstrated that all the row-lines predicted by the unit cell were sampled by Bragg peaks to an equatorial resolution of 1.0 nm^{-1}. The parameters of the lateral packing model of Fraser *et al.*, (1983,1987) and Wess *et al.*, (1995,1998) are shown in Figures 5.4-5.7 (Chapter 5), the model structure of Wess *et al.*, being a 1D staggered left-handed, 5-stranded compressed microfibril. Fraser *et al.*, (1987) observed that the

lack of high angle reflections in the equatorial plane was due to greater flexibility of the four molecular segments in the gap regions.

Although a general consensus is being reached on the packing structure of collagen (Wess *et al.,* 1998b), the exact arrangement of the collagen chains within the microfibril has still to be determined. There are some forty possible conformations outstanding, which are reduced down to 6 should the biochemical evidence of crosslinking sites prove accurate and the cyclic microfibril model valid (Fraser *et al.,* 1987). This topic is covered in greater detail in section 5.1, as a background to discussing the significance of the work presented in Chapter 5.

1.7.3 Crosslinking

The identification of intermolecular crosslinking sites (Piez and Trus 1981, Nakamura 1987, Wess *et al.*, 1990) provides significant insight into the mechanism by which collagen chains are held in position, and the overall structure of the microfibril and fibril is maintained. Their importance is of a critical nature; where crosslinking is impaired or hyperstabilsed, numerous connective tissue disorders are known to occur (hyperstabilsation being associated with the process of normal ageing) (Eyre *et al.*, 1984, Helseth and Veis 1981).

The covalent crosslinking of molecules within a fibril is effectively the final step in the formation of functional fibrils (Henkel and Glanville 1982). Formation of crosslinks is initiated by the post-translational enzymatic oxidative deamination of lysine or hydroxylysine residues by lysyl oxidase. The products, allysine and hydroxyallysine are then able to spontaneously form intra and intermolecular crosslinks with other lysine or hydroxylysine residues. (see Figure 1.7 and Table 1.2).

Lysyl oxidase activity in the oxidative deamination of lysyl and hydroxylysyl residues is regulated by strict steric requirements (the quarter staggering of collagen molecules) and by the amino acid sequence directly surrounding the target lysyl/hydroxylysyl residues (Eyre, 1987; Last *et al.*, 1990). This is strong evidence of a highly regularised process in the formation of normal crosslinks.

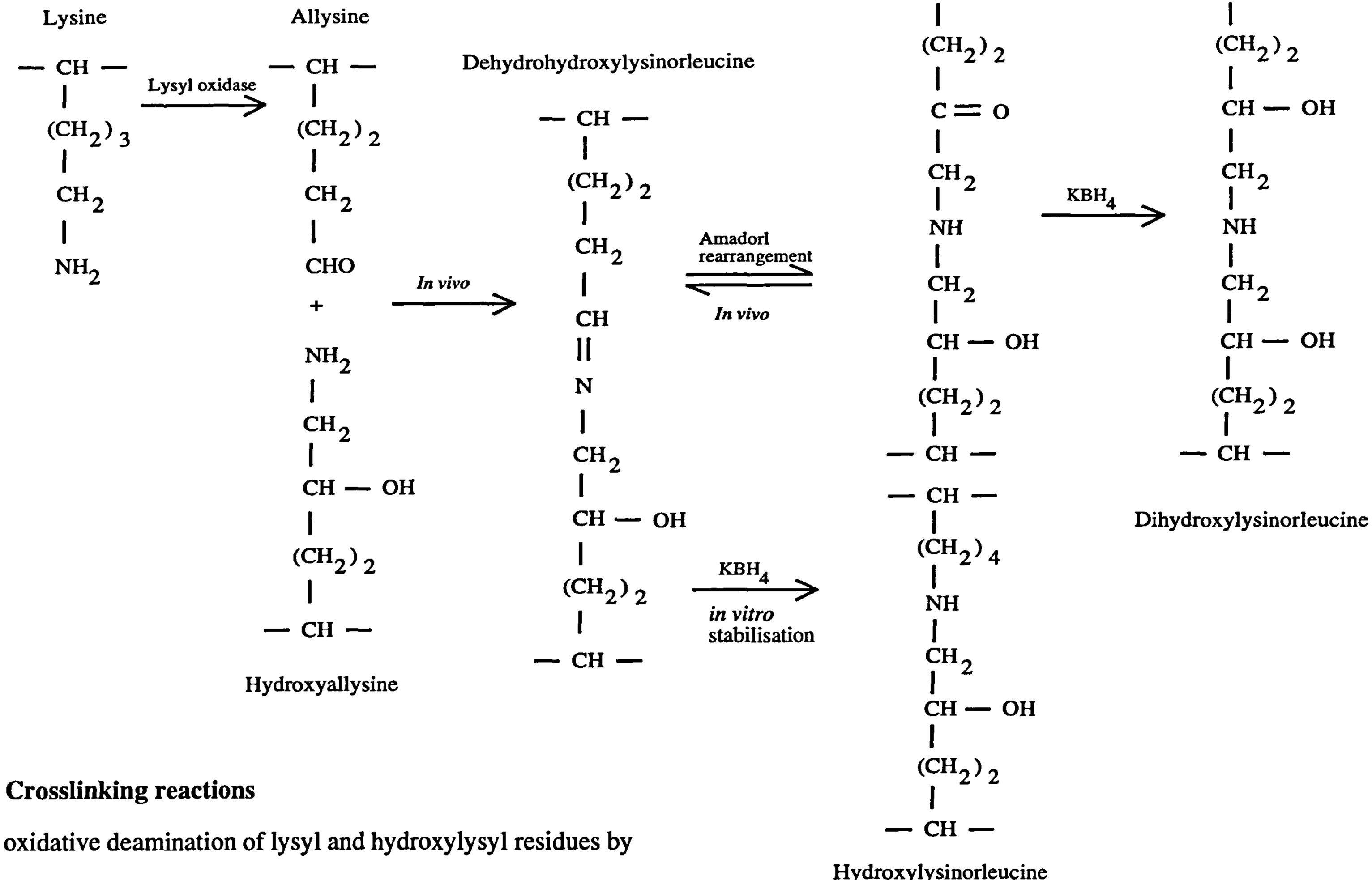


Figure 1.7 Crosslinking reactions

Enzymatic oxidative deamination of lysyl and hydroxylysyl residues by lysyl oxidase, and the spontaneous formation of reducible crosslinks.

Adapted from Kadler (1995).

STAGGER	AMINO ACID RESIDUE 1	AMINO ACID RESIDUE 2	AMINO ACID RESIDUE 3
2D	219Hly	681Hyl	*
3D	219Hyl	927Hly	929His
4D	9NLys	927Hyl	*
4D	87Hyl	17CLys	*
1D	9NLys	219Hyl	*

Table 1.2 The location of possible crosslinks as determined biochemically

The possible amino acid locations of intermolecular crosslinks as determined by Nakamura (1987). Amino acids believed to be crosslinked to one another are listed by row. **N** and **C** denote that the residue occurs within the N- or C-terminal telopeptide respectively, the remaining residues occur within the main chain.

1.7.3.1 Crosslinks at the telopeptides

Both telopeptides (N and C termini) of the collagen molecule contain lysine residues which are involved as aldehyde donors in intermolecular crosslink formation (Helseth and Veis 1981). Proteolytic cleavage of the collagen fibril produces peptide fragments containing the telopeptides covalently bound with main chain fragments (Piez and Trus 1981, Nakamura 1987), the residues believed to be specifically involved being N-terminal Lys 9 with Hyl 927, and C-terminal Lys 17 with Hyl 87. It is clear therefore, that they are directly involved in fibril stabilisation.

Supporting evidence for the *in situ* location of crosslinks at the telopeptides came from the neutron diffraction study of Wess *et al.*, (1990). Two deuterated rat tail tendons (6 and 30 minutes incubation time in deuterated KBH_4), and the native data sets were used in a multiple isomorphic phase determination of the neutron diffraction structure factors. The difference Fourier maps of the derivative electron densities showed the positions of reducible crosslinks in the location of the telopeptides and indicated main chain lysine/hydroxylsyine residues, although the resolution of the study (2.91 nm) was insufficient to make direct allocations. Interestingly, Wess *et al.*, observed that the labelling at the C-terminal telopeptide was significantly less than expected, and they speculated on the possibility that this could be due to the label encountering a steric restriction caused by the conformation (the axial contraction) of the C-telopeptide (See Chapter 4).

1.7.4 Glycation

In the same study as that mentioned above, Wess *et al.,* (1990) observed a small number of peaks in the difference Fourier maps that could not be directly related to intra- or inter-molecular crosslink sites. They suggested that this observation was related to the non-enzymatic glycosylation (gylcation) previously observed in collagen (Rosenberg *et al.,* 1979). These peak positions being the location of non-enzymatic carbohydrate attachment sites.

The presence of a large peak in the gap region in the axial location of lysine 434 of the α1 chain, and its discrete nature led to a speculative suggestion by Wess *et al.,*(1990); should this peak be due to the deuteration of a sugar linkage, then the glycosylation process may be more specific than previously thought.

Collagen type I is known to contain a number of potential glycosylation sites, hydroxylysine residues being the likely attachment sites for carbohydrate (Spiro 1969) due to the hydroxy group of hydroxylysine facilitating the attachment of the sugar through the enhancement of the ε-NH_2 nucleophility (Perejda *et al.,* 1984, Le Pape *et al.,* 1984).

Generally it is believed that the deposition of sugar attachment sites through glycation is related to the ageing of tendon and the raised interstitial sugar concentrations found in sufferers of diabetes. It has been suggested that the presence of a bulky carbohydrate side group (an enzymatically determined glycosylation site) may play a significant role in the packing of collagen chains (Morgan *et al.,* 1970).

1.8 Tendon structure and function

The basic function of tendon is to act as an extension to muscle for attachment to bone and conducting the force exerted by the muscles to move appendages. The tendons are composed of mostly type I collagen fibres oriented parallel to the longitudinal axis of the tendon, although strictly speaking the fibres are not exactly parallel to the longitudinal axis, due to 200 μm supermolecular "crimp" (Rowe 1985). This is the "zig-zag" path that the fibres take along the length of the tendon, which allows extensive intramolecular and intermolecular cross-linking for increased tensile strength. The crimping is visible when viewed with the light microscope, but disappears when the tendon is stretched by approximately 4% (Kastelic *et al.*, 1978) as the fibres are brought into parallel alignment. The fibres are arranged in bundles and surrounded by loose connective tissue, which contains blood vessels, nerves and lymphatics. Tenocytes (fibrocytes), the cells responsible for collagen production and secretion, are relatively inactive metabolically, and are located between collagen fibers, embedded in an amorphous polysaccharide ground substance composed of glycosaminoglycans and proteoglycan (Kastelic *et al.*, 1978, Kastelic and Baer 1980).

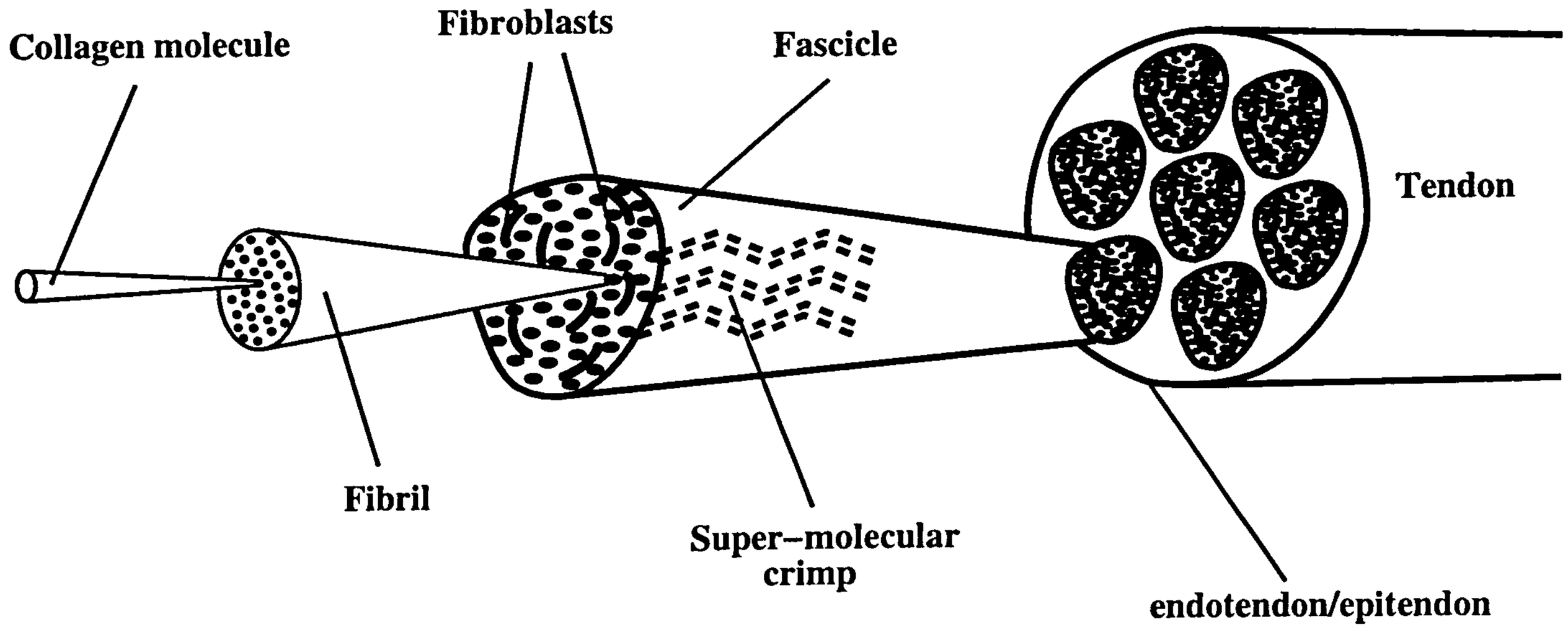


Figure 1.8 Stylised representation of the structual hierachy in tendon

Adapted from Kastelic and Baer (1980).

1.8.1 Tendon structural hierarchy

Tendon structure, like that of a single fibril, is of a hierarchical nature. The fundamental unit of that structure being the fibril (see Figure 1.8).

Fibrils are associated together in a proteoglycan matrix as a primary bundle, each primary bundle being separated from its neighbours by a thin layer of tenocytes (fibroblasts). Several primary bundles are surrounded by a sheath-like structure to form a single unit (the paratendon) defined as the fascicle. Groups of fascilles in turn are encased in the epitendon, which is thought to contain a system of interlacing collagen fibrils continuous with the endotendon. The endotendon delineates the overall structural unit - the tendon (Rowe 1985).

1.9 Conclusion

This chapter has presented a background to the research of connective tissues, more specifically that of type I collagen. The currently held views of the structure of tendon collagen at the molecular level have been given close attention.

The extensive efforts of researchers during the greater part of the last century have resulted in a basic understanding of the structure of collagen. Although this has been greatly advanced in the last twenty years or so with improved techniques and technology, the relation of studies on collagen-like peptides (for instance) to the *in situ* structure of the much longer collagen molecule is unknown. Similarly, model studies that have speculated on the possible lateral packing arrangement of collagen molecules are, although eloquently conducted, still prone to not necessarily giving rise to a single unambiguous solution. Inevitably, since the nature of the problem (determining the molecular structure of collagen *in vitro*) is inherently complicated, the full structure of collagen at the molecular level has still to be determined.

Answers to important questions such as the location of the intermolecular crosslinking sites, the conformation of the non-helical telopeptides (and consequential effect on the stability of the fibril), and lateral packing arrangement of collagen molecules could produce a picture of type I collagen structure that nears completion. The work completed and presented within this thesis has contributed significantly to completing this picture.

Each step in the process of structural elucidation of collagen via X-ray fibre diffraction, from the collection of high quality data, data extraction, processing and interpretation is difficult. This would explain (in part) why previous studies have relied heavily on model based approaches, a habit that needs to be overcome if at least some of the answers to questions expressed above, are to be found for the protein whilst *in situ*.

The purpose of this thesis has been to investigate the possibility of collecting and processing high-angle axial diffraction data in excess of 0.67 nm resolution, in the attempt to define the axial location of the crosslinking sites and the telopeptide structure from the one dimensional profile. Additionally, it was proposed that it might be possible to quantify the equatorial diffraction pattern and produce a structural solution for the lateral packing arrangement. The results of these experiments are reported in the following chapters. The significant results from which have been:

1. The production of a high angle 1D electron density map to 0.54 nm resolution, from which the conformation of the telopeptides has been deduced (Chapter 4).

2. The calculation of the 3D electron density map of anisotropic resolution that reveals for the first time, the molecular packing arrangement within the overlap region, and the assignment of coordinate positions of the crosslink forming molecular segments (Chapter 5).

Chapter 2

Theory of diffraction

2.1 Introduction

The focus of this thesis is the determination of the structure of collagen at the molecular level. The means of investigation used (X-ray fibre diffraction) makes it theoretically possible to elucidate the native (hydrated) conformation and packing arrangement of collagen molecules.

The inherent difficulties of X-ray diffraction techniques make this objective far from easy to achieve. By far the largest technical problem is that of not being able to record the phase component of diffracted X-rays. Because of this, researchers rely heavily on computational techniques to correct and then interpret their results and overcome the 'phase problem'. This in turn means that a sound understanding of the theory of X-ray fibre diffraction, and techniques available for overcoming the phase problem are needed in the elucidation of collagen structure. It is therefore appropriate to review the fundamentals of X-ray fibre diffraction and phase solution strategies here.

2.2 Background

The electromagnetic nature of X-rays and visible light allows an analogy to be drawn between them in terms of the formation of images; An object is illuminated, and that object (diffraction grating) scatters the incident light rays. These rays can be recombined by using a focusing lens (a 'reverse' diffraction event). Focused rays, once past the back focal plan of the lens, reconverge and combine to form an image of the object at the focal point, as is the case in light microscopy.

The first part of this process is analogous for X-rays in that when they illuminate an object (crystal or other sample), the X-rays are scattered. However, refocusing the scattered X-rays satisfactorily, although theoretically possible, is in practice very difficult since the refractive index of X-rays is very close to unity. It is relatively easy to collect images of the diffracted X-rays using film or now more commonly electronic devices. The recorded X-ray scatter can then be recombined analytically with the use of computers, which is the basis of X-ray crystallography. This is also the major inherent problem with this technique, in that to analytically recombine the scattered rays, two components known mathematically as phase and amplitude (describing a wave function) are needed. Collecting images of diffracted X-rays only provides one of those components, the amplitude. This is known as the phase problem and various solutions to it are discussed later in this chapter.

2.3 Basic principles of X-ray diffraction

2.3.1 Wave/particle nature of light

The scattering process falls into two categories of process known as Thomson or coherent scatter, and Compton or incoherent scattering.

2.3.2 Thomson scattering

Electromagnetic radiation that is incident to a free electron forces it to oscillate at the same frequency as the incident wave. The oscillating electron's charge becomes the point of origin of a secondary, scattered ray of the same wavelength as the incident radiation but 180° out of phase with it. Since further rays scattered from the same electron have the same phase relationship to the incident beam, the scattering is coherent. Scattering is weak, and in a 1 mm thick theoretical 'crystal' composed only of free electrons, only 2% of the incident radiation is scattered (Woolfson 1970).

2.3.3 Compton scattering

Whereas Thomson scattering demonstrates the wave nature of electromagnetic radiation, Compton scattering demonstrates its particle nature in the context of X-ray scattering. Incident photons collide with relatively loosely bound electrons and are deflected with a loss of energy. Because of the energy change, the deflected wavefront also experiences a change in wavelength.

Compton scattering is particularly strong in comparison to Thomson scattering at high angles, or in non-crystalline samples. But in the case of X-ray diffraction of a crystal or crystallite, the co-operative effect of the coherent scattering from many equivalent atoms becomes cumulative, and is greater than the sum of incoherent scattering contributions and therefore becomes the dominant feature of the diffraction pattern.

2.3.4 Bragg's law

Bragg's law visualises scattering by a crystal in terms of reflections from planes of atoms (Bragg 1913). When these planes are illuminated at an oblique angle (Figure 2.1), the scattered X-rays are reflected at an equivalent angle in the same plane (angle of reflections is equal to angle of incidence). Constructive interference occurs between rays reflected from different planes, if the path difference between the reflected rays is equal to an integral number multiplied by the radiation's wavelength. This is represented by equation 2.1:

$$2d \sin \theta = n \lambda \qquad 2.1$$

Where d equals the distance of separation between successive planes of atoms, λ the wavelength of the incident radiation, and θ the angle of incidence/reflection.

This equation predicts the position in space of any diffracted ray, hence the first order of diffraction from the given atomic planes will occur at n=1, the third n=3, the

tenth n=10, and so on.

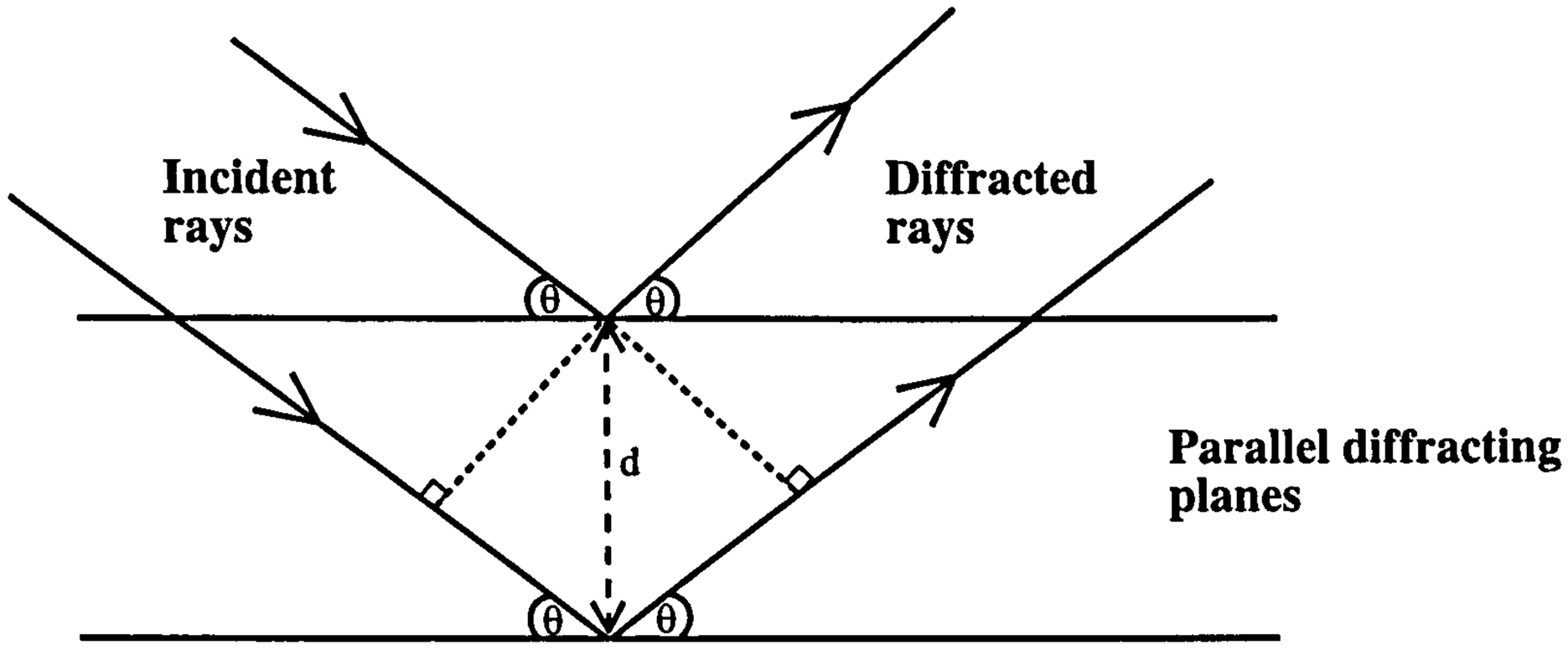


Figure 2.1 Braggs law

X–rays are reflected from parallel planes of atoms so that the angle of reflection is equal to the angle of incidence (θ) and satisfy the equation; $2d \sin \theta = n\lambda$ (see main text). Parallel rays reflected from points on neighbouring partially reflecting planes are in phase when Braggs law is obeyed.

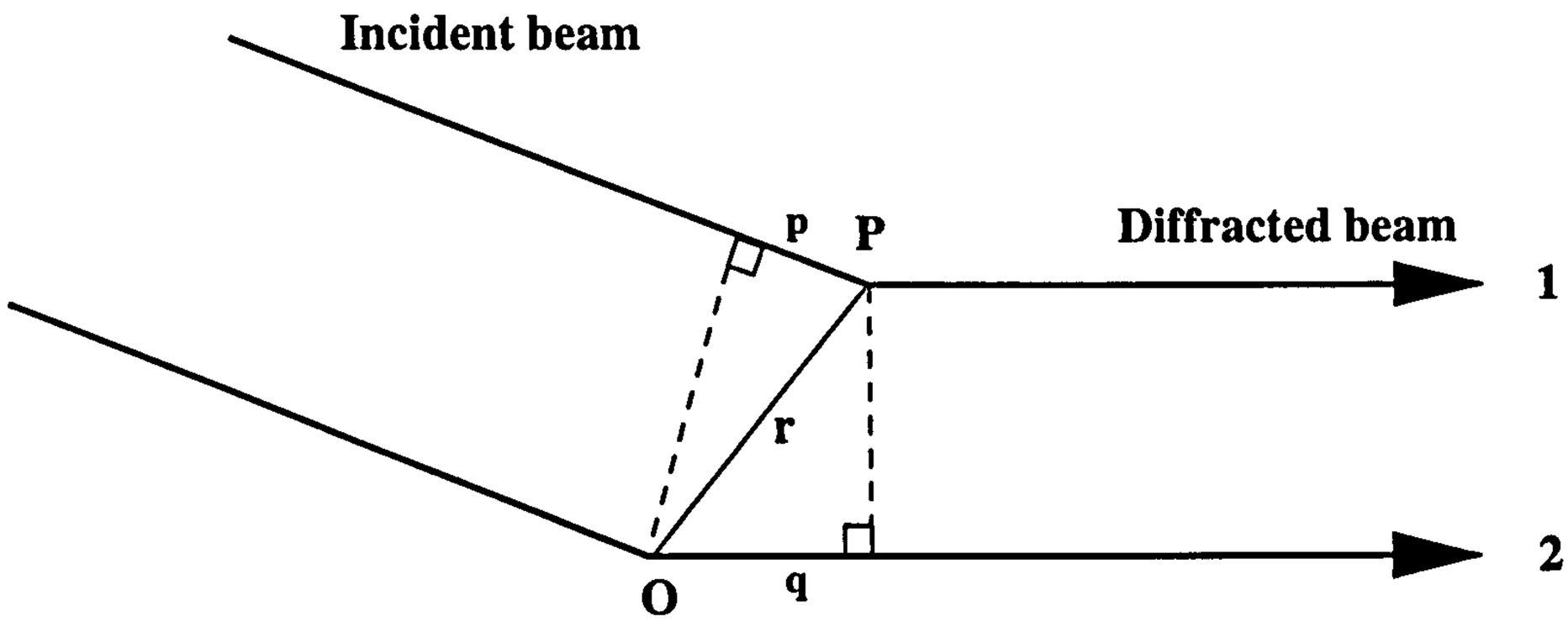


Figure 2.2 Scattering from a single centre (P) relative to an origin (O)

Incident radiation is scattered at a point O. This scattering event introduces a phase shift in the diffracted beam. The path difference between the through beam and the diffracted beam is; $-2\pi L\lambda$, L being the path difference between rays 1 and 2 (p–q).

2.4 The scattering of X-rays

2.4.1 From one point

If a point, O, represents a scattering centre, and the incident radiation is monochromatic, then the expression 2.2 can be used to describe the displacement of the incident rays approaching O, in relation to time.

$$y = A \cos 2\pi vt \qquad 2.2$$

Where v is the frequency of the radiation, A is the maximum amplitude of the wave, t is time, and y is the displacement.

At a point along the propagation path of the scattered beam, P, the wave will have a displacement, which is dependent upon a number of factors (see Figure 2.2):

a) The phase difference, dependent upon the length of OP (=L) between the wave at point O, and the scattered wave at point P. The phase difference can be described as; $-2\pi L/\lambda$, where λ is the wavelength of the radiation. An alternative to this formulation is; $-2\pi Lv/c$, v being the frequency, and c being the velocity of the propagating wave.

b) The scattering event itself may create a phase difference making the scattered wave at point O retarded in relation to the incident wave at point O. This quantity may be referred to as the 'scattering phase shift' (α_s).

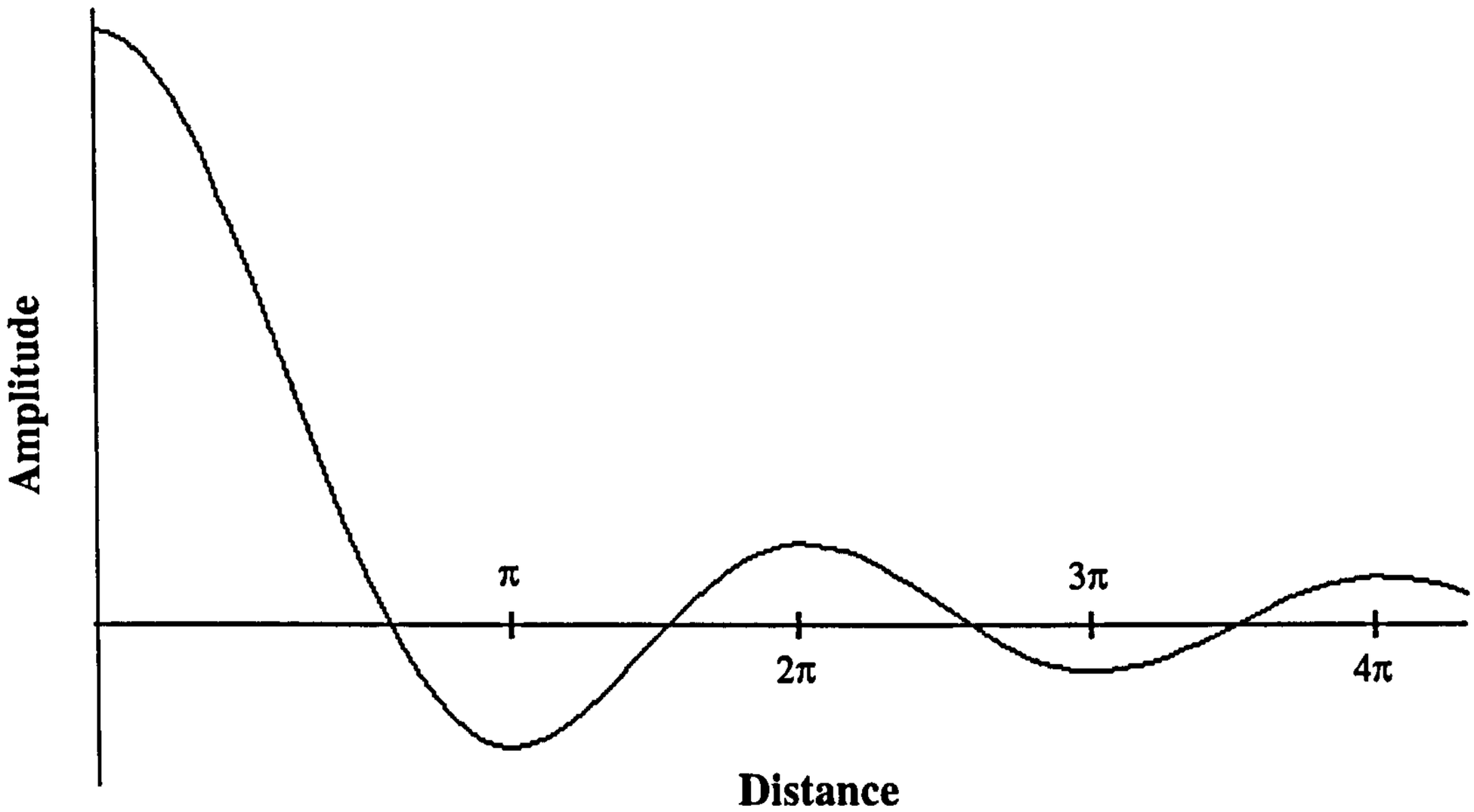


Figure 2.3 The inverse square law

The amplitude of scattered radiation diminishes inverse proportionally to the distance travelled (Woolfson 1970).

c) The fall off of amplitude of the scattered radiation. This obeys the inverse-square law or reduction of intensity with respect to distance (see Figure 2.3).

The displacement (y) of the scattered wave at point P can now be determined using expression 2.3

$$y(2\theta,L,t) = f_{2\theta}\, A/L \cos [\, 2\pi v(t-L/c) - \alpha_s \,] \qquad 2.3$$

$f_{2\theta}$ is a constant of proportionality (scattering length) associated with the scattering centre and a function of the scattering angle 2θ.

Expression 2.2 can be expressed in complex form:

$$Y = Y_0 \cos 2\pi v(t-x/c) + iY_0 \sin 2\pi v(t-x/c) \qquad 2.4$$

Here Y_0 is the wave amplitude, the real part the displacement of the wave, and the ratio of imaginary part/real part is the tangent of the phase at (x,t) relative to the origin (0,0). This enables a description of the time dependence of the disturbance at P as:

$$y(2\theta,L,t) = f_{2\theta}(A/L)\exp[2\pi iv(t-(L/c))-i\alpha_s\ ]$$

2.5

Hence the amplitude at P, of a wave scattered from a single centre is given by:

$$\eta(2\theta,L) = f_{2\theta}(A/L)$$

2.6

And the phase shift at point P, of the scattered wave relative to the incident radiation at O is:

$$\alpha_{OP} = 2\pi vL/c + \alpha_s$$

2.7

2.4.2 From two points

In the situation where radiation is incident on two identical centres (O_1 and O_2, Figure 2.4), the resultant scattered radiation has been scattered effectively through the same angle (2θ). This holds true in the case where the distance between O_1 and a point P is very large relative to the distance O_1O_2.

The phase shift associated with the scattered radiation can be expressed as:

$$\alpha_{o1o2} = -2\pi/\lambda\ (CO_2 + O_2D) \qquad 2.8$$

Note that since the scattering centres are identical, the scattering phase shift (α_s) will be the same for both centres.

If $\bar{S}_o$ and $\bar{S}$ are two unit vectors on the path of the incident and scattered beams respectively and r is the vector between O_1 and O_2. Then the path difference δ_s for two incident beams at O_1 and O_2 and scattered in the direction 2θ can be described as:

$$\delta_s = CO_2 + O_2D = r.\bar{S}_o - r.\bar{S} \qquad 2.9$$

Combining 2.8 and 2.9, the phase difference is:

$$\alpha_{o_1o_2} = 2\pi/\lambda \; .r.(\bar{S} - \bar{S}_o) \qquad 2.10$$

The quantity $(\bar{S} - \bar{S}_o)/\lambda$ can be defined as 's' (see Figure 2.5), so that;

$$\alpha_{o_1o_2} = 2\pi r.s \qquad 2.11$$

$\bar{S}_o/\lambda$ and $\bar{S}/\lambda$ in both the incident and scattered directions are of equal magnitude $(1/\lambda)$.

Figure 2.5 shows that s is proportional to the bisector of the angle between $\bar{S}_o$ and $\bar{S}$, hence its magnitude is given by:

$$|s| = (2 \sin\theta) / \lambda \qquad 2.12$$

Therefore the displacement of the radiation at a point P at a distance L from O_1 will be:

$$y(2\theta,L,t) = f_{2\theta}\,(A/L)\,\exp[2\pi iv(t-(L/c))-i\alpha_s](1+\exp 2\pi ir.s) \qquad 2.13$$

The amplitude of this can be expressed as:

$$\eta_2(2\theta,L) = f_{2\theta}\,(A/L)\,(1+\exp 2\pi ir.s) \qquad 2.14$$

Which with reference to 2.6 can be expressed in terms of the amplitude of scattering from a single unit:

$$\eta_2(2\theta,L) = \eta(2\theta,L)\,(1+\exp 2\pi ir.s) \qquad 2.15$$

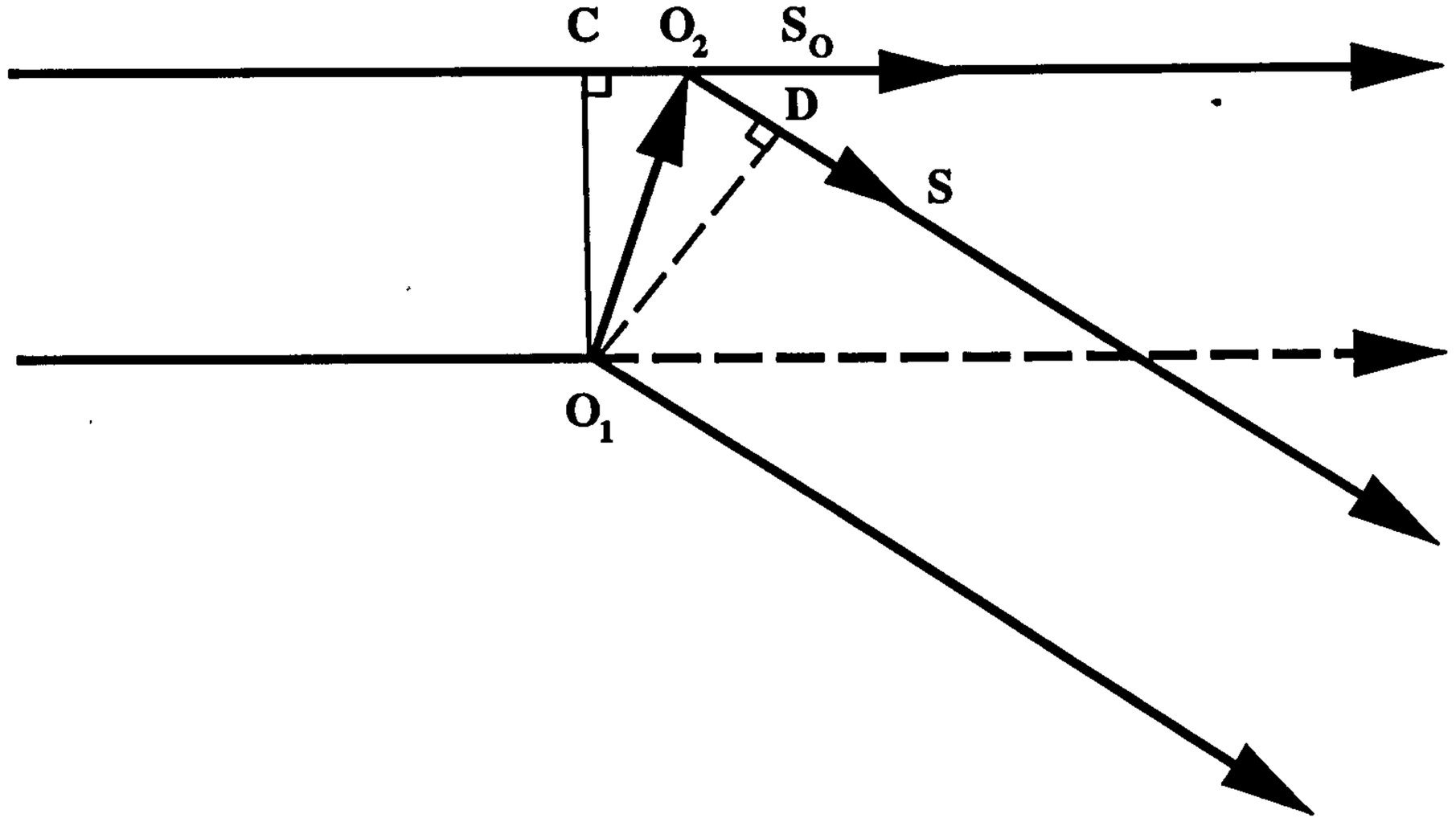


Figure 2.4 Scattering from two identical centres

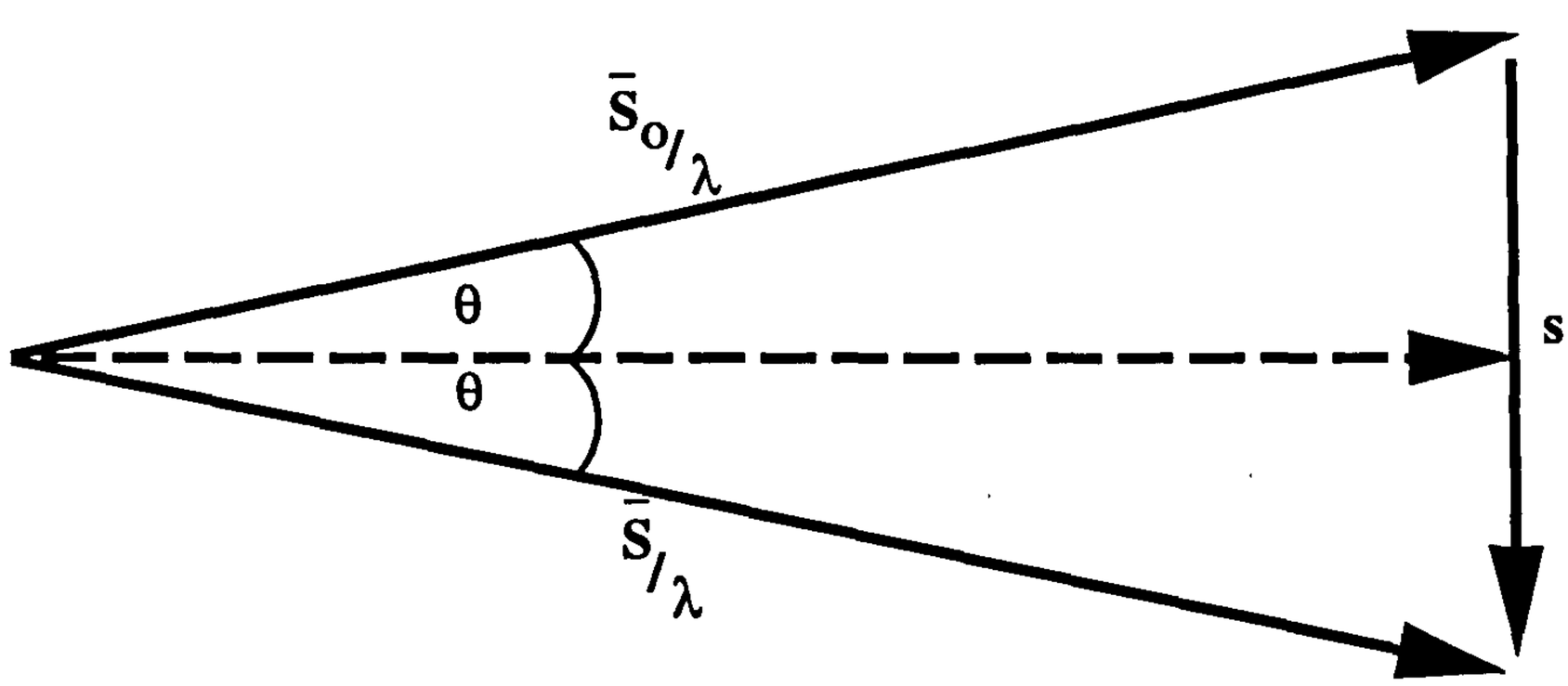


Figure 2.5 The relationship between $\bar{S}$, s and $\bar{S}_o$

2.4.3 Scattering from a general number of centres

Equation 2.15 describes the scattering process relative to O_1. However, to develop a means of determining the scattering event for multiple centres, it is useful to make the origin of all phases to be an arbitrary position, O. The relative positions of O_1 and O_2 to O in the two scattering centre model are given by the vectors r_1 and r_2. Thus equation 2.15 becomes:

$$\eta_2(2\theta,L) = \eta(2\theta,L)\,(\exp 2\pi i r_1.s + \exp 2\pi i r_2.s) \qquad 2.16$$

This is easily extended to a system of multiple, identical scattering centres (O_1, O_2, O_3,....O_n):

$$\eta_n(2\theta,L) = \eta(2\theta,L) \sum_{j=1}^{n} (\exp 2\pi i r_j.s) \qquad 2.17$$

The factor $\eta_2(2\theta,L)$ appears outside of the summation since all the scatters are identical. When the scatters are non-equivalent, the individual point factors need to become part of the summation:

$$\eta_n(2\theta,L) = A/L \sum_{j=1}^{n} (f_{2\theta}) \exp 2\pi i r_j.s \qquad 2.18$$

In X-ray crystallography, the scattering centres are electrons. It is convenient to describe their distribution as a density function ρ(x).

The summation of equation 2.18 becomes an integration in 2.19, since the density function is continuous:

$$F_{(s)} = \int_{j=1}^{n} \rho(x)_j \exp(2\pi i s.x)\, dVx \qquad 2.19$$

2.5 Diffraction and the Fourier transform

Sine and cosine waves combined in the Fourier series can be used to construct a continuous function f(x). Further, the Fourier transform of the function f(x) is essentially complex, describing its wave components:

$$F_{(s)} = \int_{1}^{n} f(x)\, \exp(2\pi i s.x)\, dx \qquad 2.20$$

The amplitude of the wave function is the magnitude of F(s), and the phase relation is the argument of the complex number. It can be determined that the Fourier transform of the wave component is the original function (Blundell and Johnson 1976):

$$f_{(x)} = \int_{0}^{\infty} F(s)\, \exp(-2\pi i s.x)\, ds \qquad 2.21$$

This shows that the function f(x) can be obtained by integrating its Fourier components (F(s) with each appropriate phase angle).

F(s) is the Fourier transform of F(x) and comparison of 2.20 and 2.19 shows that the diffraction pattern (F(s)) of an object (ρ(x)) is the object's Fourier transform. Equation 2.21 shows that the Fourier transform of the diffraction pattern with the correct phase components presents the original object.

2.6 Crystallography

2.6.1 Crystals and liquid-crystals

It is commonly understood that there are several states of matter; the solid, liquid, and gaseous phases are the three most common. Although the full phase transition (as is presently understood), is that of; solid and crystalline, to semi- or liquid-crystalline, to liquid, to gaseous, and then plasma.

A material is regarded as crystalline if it is built up of a continuously translating repetition of some basic structural pattern (motif) in one or more dimensions. This structural pattern may constitute one or more atoms, a molecule, or a complex of several molecules, and is known as the 'unit cell'.

X-ray crystallography is concerned chiefly with fully crystalline solids. Material in the crystalline state contains arrays of highly regularised atoms or molecules, which contrasts with the amorphous solid form where the atoms/molecules are arranged in a more randomised, solid-liquid like (glass) state. The highly ordered arrays of atoms/molecules in any crystal are arranged in parallel planes, and it is the diffraction of X-rays from these planes that gives rise to detectable coherent scattering. In macromolecular crystallography, crystals are grown artificially, the repeating basic unit of the crystal being the relative large molecule that has been induced to crystallise. X-ray fibre diffraction on the other hand is mainly concerned with naturally occurring crystalline systems, which are at best, most often of a semi-crystalline nature.

In a liquid-crystal, there is a degree of crystalline order in one or more dimensions, but with sufficiently weak interactions between the molecules to allow liquid-like properties to be displayed in at least one dimension. This has been described by Hukins (1981) as smectic type A organisation (see Figure 2.6). This can be visualised by taking several cylindrically shaped objects of uniform length, and placing them into a narrow box. If the length of the box is only marginally larger than that of the cylinders, then the cylinders are aligned uniformly within the box, with little or no freedom of movement in the axial direction (Z). The cylinders are however free to move and be rearranged in the x and y directions, demonstrated by the fact that the cylinders will roll over one another to fill in the dimensions of the box.

2.6.2 The crystal lattice and the reciprocal lattice

A crystal or the crystalline components of a crystalline solid can be described as containing an array of points, at which the local environment (the unit cell) is the same. This array is termed the 'crystal lattice'.

The reciprocal lattice is directly related to the crystal lattice, and is frequently referred to in crystallography, requiring a brief definition here.

If the axes of the crystal lattice are defined as being; a,b, and c, then the axes of the reciprocal lattice are a^*,b^*, and c^*. The reciprocal lattice axes point, relative to the crystal lattice axes as follows; a^* is perpendicular to b and c; b^* is perpendicular to a and c; and c^* is perpendicular to a and b. The values of a^*,b^*, and c^* are related to a,b, and c depending upon the shape of the unit cell, for instance, for a triclinic unit cell (where none of the angles relating to the sides are necessarily the same):

$$a^* = (bc \sin\alpha)/V \qquad b^* = (ca \sin\beta)/V \qquad c^* = (ab \sin\gamma)/V$$

Where $V = abc(1+2\cos\alpha\cos\beta\cos\gamma - \cos^2\alpha\cos^2\beta\cos^2\gamma)^{1/2}$

$$\cos\alpha^* = (\cos\beta\cos\gamma - \cos\alpha)/(\sin\beta\sin\gamma);$$

$$\cos\beta^* = (\cos\gamma\cos\alpha - \cos\beta)/(\sin\gamma\sin\alpha);$$

$$\cos\gamma^* = (\cos\alpha\cos\beta - \cos\gamma)/(\sin\alpha\sin\beta)$$

From the International Tables for X-ray Crystallography, Vol. I, International Union for Crystallography.

V=the volume of the crystal lattice unit cell; a, b, and c, and a^*,b^*, and c^*, are the lengths of the axes of the crystal and reciprocal unit cells respectively; α, β, and γ, α^*, β^*, and γ^*, the inter-axial angles of the crystal and reciprocal unit cells respectively.

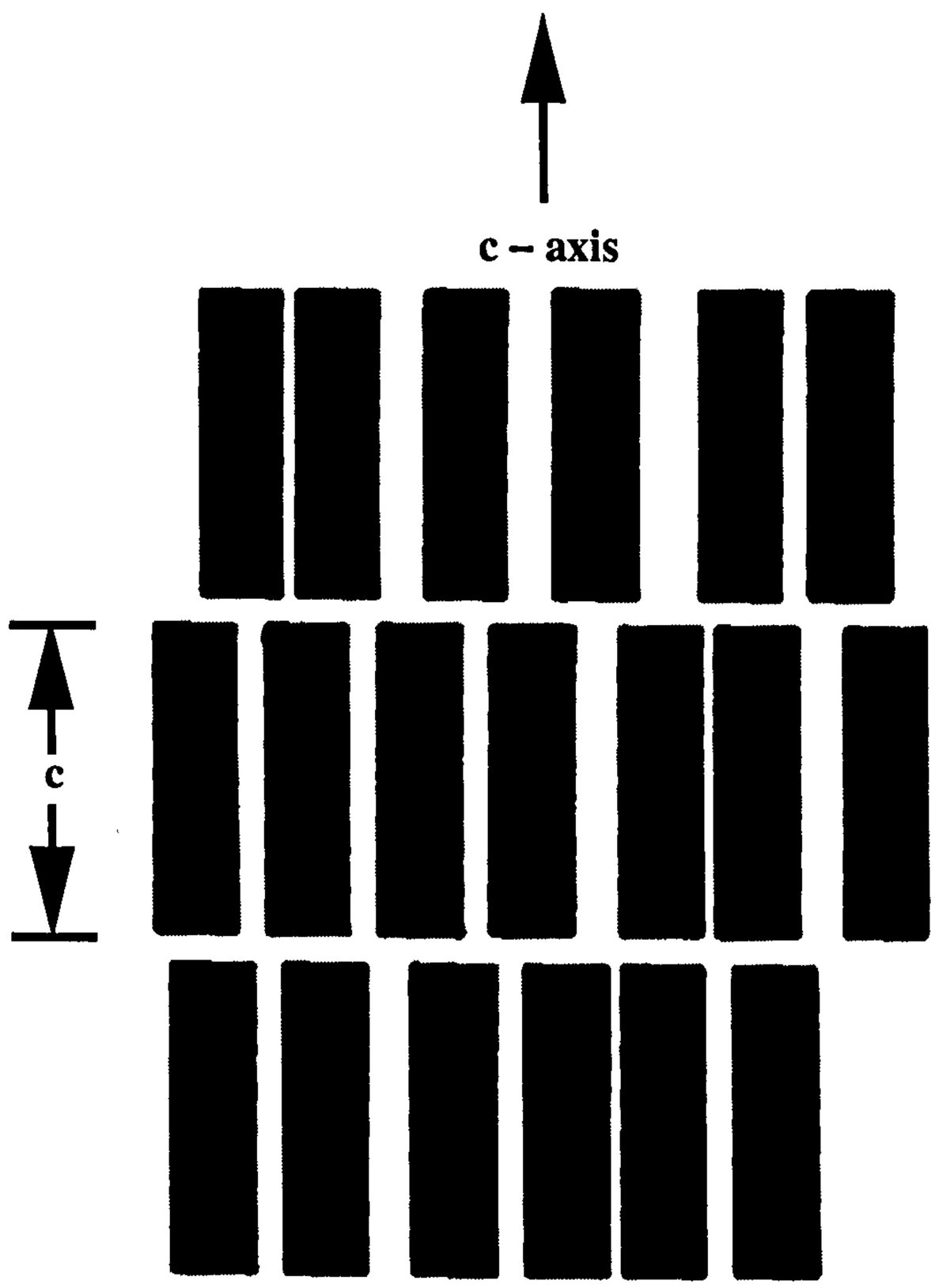


Figure 2.6 Organisation of a smectic type A fibre

The molecules are ordered and confined to layers, each layer having the same thickness, c, so that the structure repeats itself regularly in the c–axis direction. The least ordered smectics are true one–dimensional crystals, others, such as collagen possess limited order in the directions parallel to the c–axis.

2.6.3 The Ewald construction

Ewald (1921), proposed a simple geometric construction that predicts the geometry of diffraction in reciprocal space. A sphere is constructed with its centre at the crystal (C), with a radius of $1/\lambda$ (Figure 2.7). The origin of the reciprocal lattice is the point through which the unreflected X-ray beam passes having first travelled through point C. Point B is a point on the surface of the sphere. X-rays arriving at point B from point C are diffracted X-rays if point B represents a reciprocal lattice point (h,k,l). Therefore, to obtain diffraction of one particular hkl reflection, it is necessary to move or tilt the crystal to bring that reciprocal lattice point into contact with the Ewald sphere.

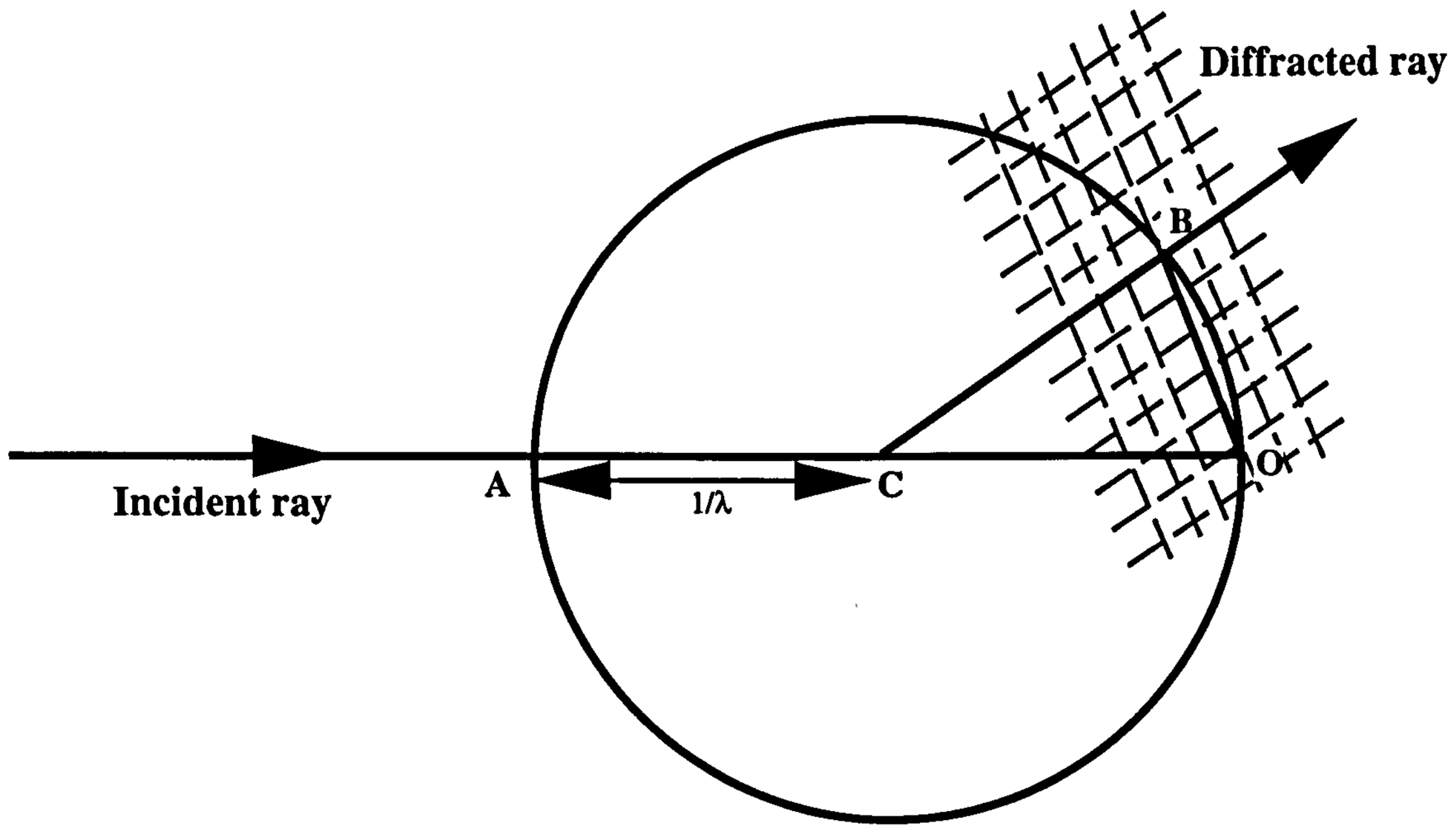


Figure 2.7 The Ewald sphere construction

The sphere has a radius of $1/\lambda$ and centre C. The origin of the reciprocal lattice is marked by O where the unreflected X–ray beam travelling in direction AC meets the sphere. A ray travelling in direction CB is a diffracted ray if the point B represents a reciprocal lattice point (h,k,l).

2.7 Solutions to the phase problem

The structure factor function F(hkl) is composed of both amplitude, *F*(hkl), and phase, α(hkl). However, only the amplitude component can be measured from the recorded diffraction pattern, the phase component information is lost. The phase element is needed for the calculation of the electron density, and because that information is missing, the determination of the hkl phases constitutes the basic problem in crystallography.

There are a variety of means available to the researcher to overcome the phase problem (for review see Woolfson 1970, Blundell and Johnson 1976). The techniques and theory directly applicable to the work presented here are; the Patterson function, model building, and Multiple Isomorphous Replacement/Addition (MIR/A).

2.7.1 Patterson function

The Patterson function is essentially a convolution function of the directionless (scalar) separation between electron dense regions of the diffracting sample. It can be used in the analysis of data where only the intensity maxima (square of the amplitude) are available. It is often used in the initial stages of Multiple isomorphous Replacement (MIR) analysis to locate the labelling sites of heavy atoms.

When described as a series, it has the form;

$$P(r) = 1/v \sum_{h} \sum_{k} \sum_{l} |F_{hkl}| \cos(2\pi hr)$$

Where $r = xa + yb + 2c$, being a point in unit cell space, and $h = ha^* + kb^* + lc^*$, which is a point in reciprocal space.

The significance of the function is that the peaks in a Patterson function refer to the distances between electron dense regions within the structure.

A Patterson map of a molecule is more complex than the structure of the molecule itself. A structure containing N atoms, will produce a Patterson map containing N(N-1)+1 peaks. The complexity of the Patterson map, and its centrosymmetric nature can make interpretation difficult. However, when combined with sequence information and other structural information, it can be a powerful means of determining the position of heavy atoms in isomorphous derivatives.

2.7.2 Model building

Often, during the stages of phase refinement, crystallographers utilise model structures to improve phase data sets. At other times, there is insufficient information available to determine the phases analytically. In these instances, it is common practice to define model structures, Fourier convert the scattering factors (see equations 2.19-2.21 and sections 2.4.3 and 2.5) and compare the amplitudes generated with the observed data set.

Such model building based techniques can be time-consuming and ineloquent, rarely, if ever, producing a single unique solution. Taking the limitations of model building into account, it is still a very useful practice and has been employed to some success in investigations of type I collagen structure (see Hulmes *et al.*, 1977, 1980, Wess *et al.*, 1998a).

2.7.3 Multiple isomorphous replacement/addition

It is often possible to modify the unit cell contents, without altering the crystal lattice dimensions. This usually results in a derivative structure that is structurally similar to the original, differing only by one or more atoms or groups. This is frequently exploited in X-ray crystallography by introducing heavy metal atoms to the sample. The heavy atoms, such as Au, Pd, Hg, etc. bind to the unit cell contents in chemically favourable environments. Since the crystal lattice has remained unaltered, the derivative structure is said to be isomorphous to that of the original or native structure. Strictly speaking, this would be a derivative formed by isomorphous *addition*. Isomorphous *replacement* involves the replacement of atoms with more electron dense atoms of similar size, but as before, the modification being isomorphous, it is the unit cell contents that are altered, not the crystal lattice.

Isomorphous replacement, was first successfully employed by Green *et al.*, (1954) in their determination of the structure of Myoglobin. Since then it has become one of the most common means of structural determination, certainly within the field of macromolecular crystallography. The fundamental strength of the technique is that it provides a theoretically simple means of determining the phase angles. The practical difficulties with the technique lie in creating isomorphous derivatives in the first place, and with the scaling of native and derivative data sets.

In the diffraction pattern of a truly isomorphous derivative the diffraction maxima are found to occur in the same reciprocal spatial positions as those for the native it is the

relative intensities of the diffraction peaks that differ. The amplitudes from the diffraction of the native and derivative are related as shown in the Argand diagram (Figure 2.8). The derivative structure factor (F_{ph}) is the sum of the vectors F_p (the native structure factor) and F_h (the heavy atom label). If the positions of the derivative labels are known within and relative to the unit cell, then it is possible to calculate F_h, and determine the phases of the native and derivative structure factors.

One means of doing this is represented by the Harker construction (Harker 1956). In the Harker construction (Figure 2.9), circles are drawn for the native and derivative, each with a radius corresponding to the observed amplitude for the native ($|F_p|$), and the derivative ($|F_{ph}|$). The origin of the construction is the centre of the native circle, the centre of the derivative circle being determined by the vector F_h. One of the points or point intersected by the circles represents the correct phase. Using two or more different isomorphous derivatives removes the ambiguity of selecting the correct phase.

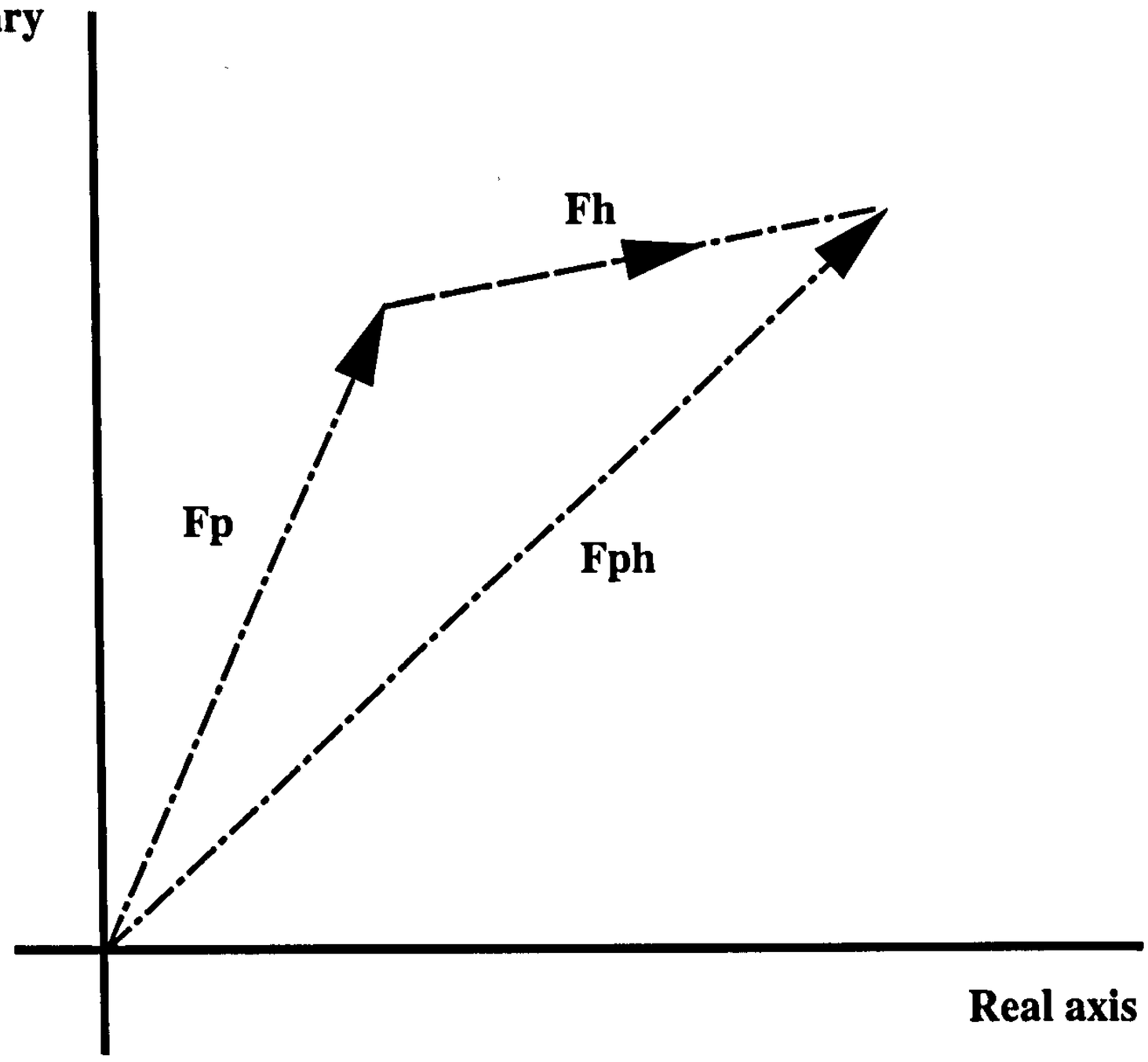


Figure 2.8 Argand diagram

The derivative structure factor (Fph) is the sum of the vectors of the native and heavy atom structure factors (Fp and Fh respectively).

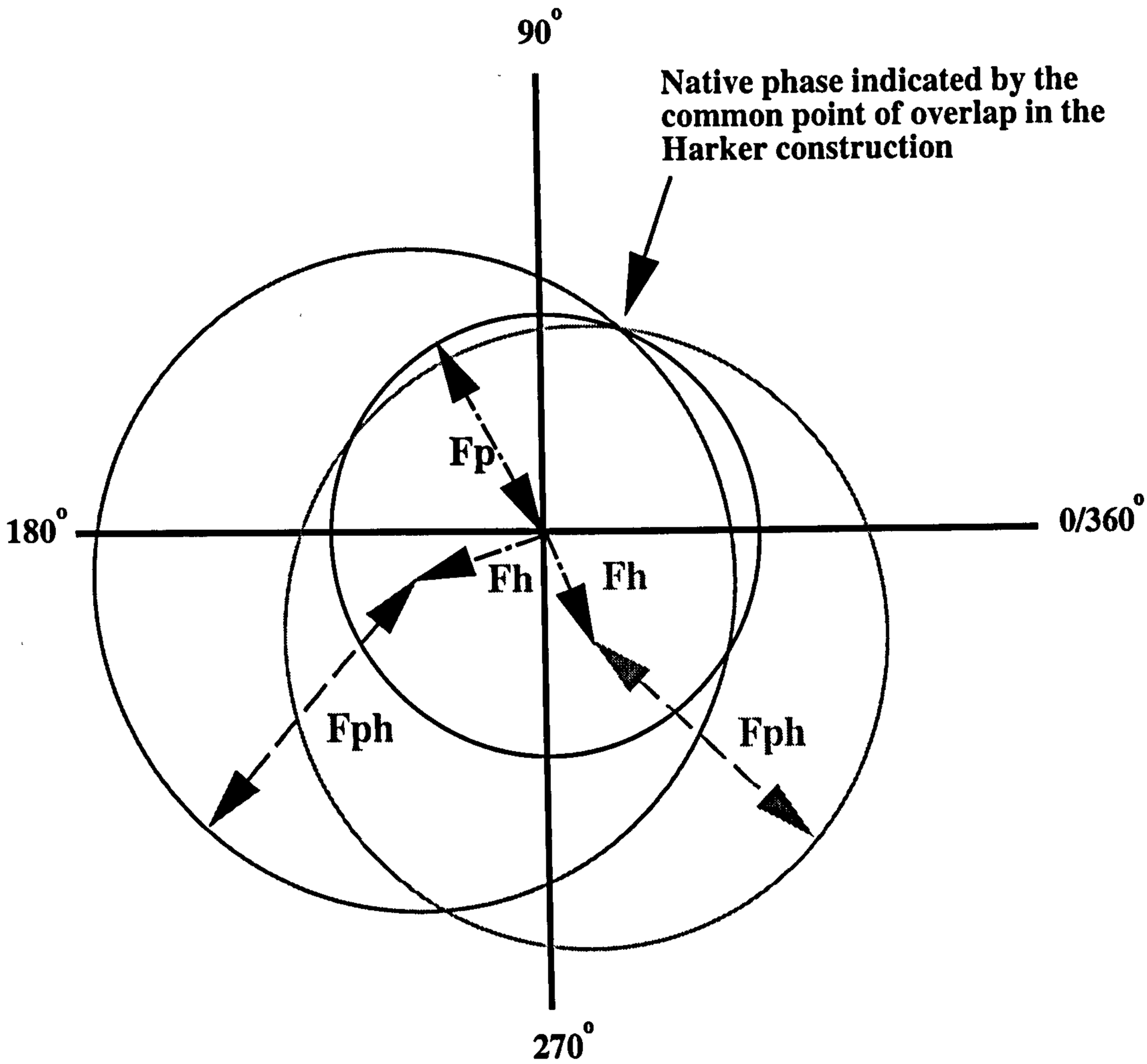


Figure 2.9 The Harker construction

The relationship between Fp, Fph, and Fh (see Argand diagram in Figure 2.8), can be used to calculate the phase component of the native protein structure factors. When the amplitudes of the native protein, two or more isomorphous derivatives and the labelling positions of the heavy atoms are known (so as to be able to calculate the heavy atom structure factors), the Harker construction shows the value of the native phase (where the three circles overlap, this is marked above).

2.8 Fibre diffraction

2.8.1 Order and disorder in natural fibres

The nature of diffraction from fibrous proteins differs significantly to that originating from a macromolecular crystal. Fibrous proteins contain arrays of crystallites and do not show the same high degree of crystallinity in all three dimensions as those used in single crystal crystallography. The degree of ordering within a fibrous protein system is anisotropic, the axial order parallel to the fibre axis is greater than the order found perpendicular to the fibre axis (the lateral plane).

The degree of order in the lateral direction is often short range, and frequently includes a number of distortions that fall broadly (but not necessarily neatly) into four classes (Holmes and Blow 1966):

1. Crystalline fibres. Molecular restrictions may allow efficient packing of molecules. When this does occur, the sample is found to be composed of microcrystalline domains with a common axial direction. The orientation around this direction is however random.

2. Semi-crystalline fibres. Molecular restrictions do not allow the system to form a regular periodic structure in the lateral direction. These systems are only able to form quasi-crystalline lateral lattices, but have regular axial order.

3. Non-crystalline fibres. In the situations where the lateral disorder is so great,

when the only commonality between molecular filaments is that their axial orientation is the same. The lattice effects are weak to the extent of being negligible, so only the molecular transform may be observed.

4. Oriented gels. Some colloidal solutions of long rod-like molecules display an orientation of the rods parallel to each other. The degree of angular variation from their common axis can be less then 1°. Frequently, the rods are separated by large amounts of solvent, hence there is no, or very little three-dimensional orientation.

The long range order of the axial packing structure of tendon and semi-crystalline packing for the fibres within the tendon is characterised by the tendon X-ray fibre diagram (type 2 order according to the system described above). Meridional Bragg reflections (corresponding to the axial packing) extend far into reciprocal space (past 0.48 nm). Whilst the lateral reflections are present, the associated diffuse scatter indicates an inherent degree of disorder. See Fraser *et al.*, (1983).

2.8.2 The cylindrical transform

The organisation of crystallites within a collagen fibre gives rise to a diffraction pattern that is equivalent to the cylindrically averaged projection of the diffraction pattern from a single crystal (Finkenstadt and Millane 1998). Thus the positions of reflections on the collagen fibre diagram can be described by a cylindrical projection of the reciprocal lattice. The reciprocal space polar coordinates R and Z are used to describe the position of Bragg reflections in the fibre diagram, where Z is the fibre axis, and R the radial coordinate. In collagen, the Z axis is not quite parallel with any of the unit cell axes, and hence two further entities are used to describe the tilt of the unit cell to the fibre axis θ and ϕ (discussed in section 5.1.5) , and allow the polar coordinates R and Z to be calculated from the unit cell parameters. Fraser and MacRae (1981), in refining the parameters of the unit cell of collagen molecular packing, defined the polar coordinates R and Z as:

$$R = [(ha^*\sin\theta_{a^*}\cos\phi_{a^*} + kb^*\sin\theta_{a^*}\cos\phi_{b^*} + lc^*\sin\theta_{a^*}\cos\phi_{c^*})^2 + (ha^*\sin\theta_{a^*}\sin\phi_{a^*} + kb^*\sin\theta_{a^*}\sin\phi_{b^*} + lc^*\sin\theta_{a^*}\sin\phi_{c^*})^2]^{1/2}$$

and

$$Z = ha^*\cos\phi_{a^*} + kb^*\cos\phi_{b^*} + lc^*\cos\phi_{c^*}$$

a*, b*, and c* being the reciprocal unit cell axis and h, k, and l the Miller indices.

2.9 Projection theorem

The significance of the distribution of Bragg reflections in the tendon diffraction pattern can be explained in part by the Projection theorem (See Woolfson 1970).

Put simply:

The transform of a three dimensional density, projected onto a two dimensional plane, is the same as the central section of the three dimensional transform, perpendicular to the direction of the projection.

This means that the meridian of the fibre diagram is the transform of the projected electron density on to the axis of the fibre. The meridional reflections can be regarded as belonging to a one dimensional crystallite's diffraction pattern since it is the central section through the three dimensional diffraction pattern. Whilst the equatorial reflections describe the three-dimensional packing structure of collagen molecules in fibrils.

When the equations described at the start of this chapter are expressed in one or three dimensions, they are applicable to the interpretation of the meridional and equatorial reflections respectively, as has been the case in this study.

Chapter 3

Instrumentation, data collection, extraction and correction

3.1 Introduction

This chapter describes the instrumentation used to generate X-rays and record their diffraction by biological samples, specifically collagen fibres.

This encompasses a wide range of topics, further complicated by the fact that five different beamline stations at three different synchrotron light sources were used in experiments described in this thesis. There is a significant degree of diversity in the experimental set-up of each of these light sources and stations, although they are also fundamentally similar. The following sections describe the common features of the light sources and beamline stations. The significant specific details of each synchrotron and beamline station used are tabulated within the following sections.

3.2 Synchrotron light sources

By causing fast moving electrons to hit a metallic target, scientists have been able to generate X-ray radiation for some time (first achieved by Wilhelm Conrad Röntgen whilst working with cathode-ray tubes in 1895). Since then, high-energy physics has produced significant advances in the generation of X-rays. Third generation synchrotron sources today emit X-rays that are a trillion times more brilliant than those produced by X-ray tubes (ESRF source information page, www.esrf.fr). Figure 3.1 shows the progression of technology with time against the brilliance of emitted radiation, brilliance being defined here as the number of photons per second passing through the

smallest spot onto which an X-ray beam can be focused (photons s^{-1} $mrad^{-1}$ mm^{-1} (0.1% BW)$^{-1}$).

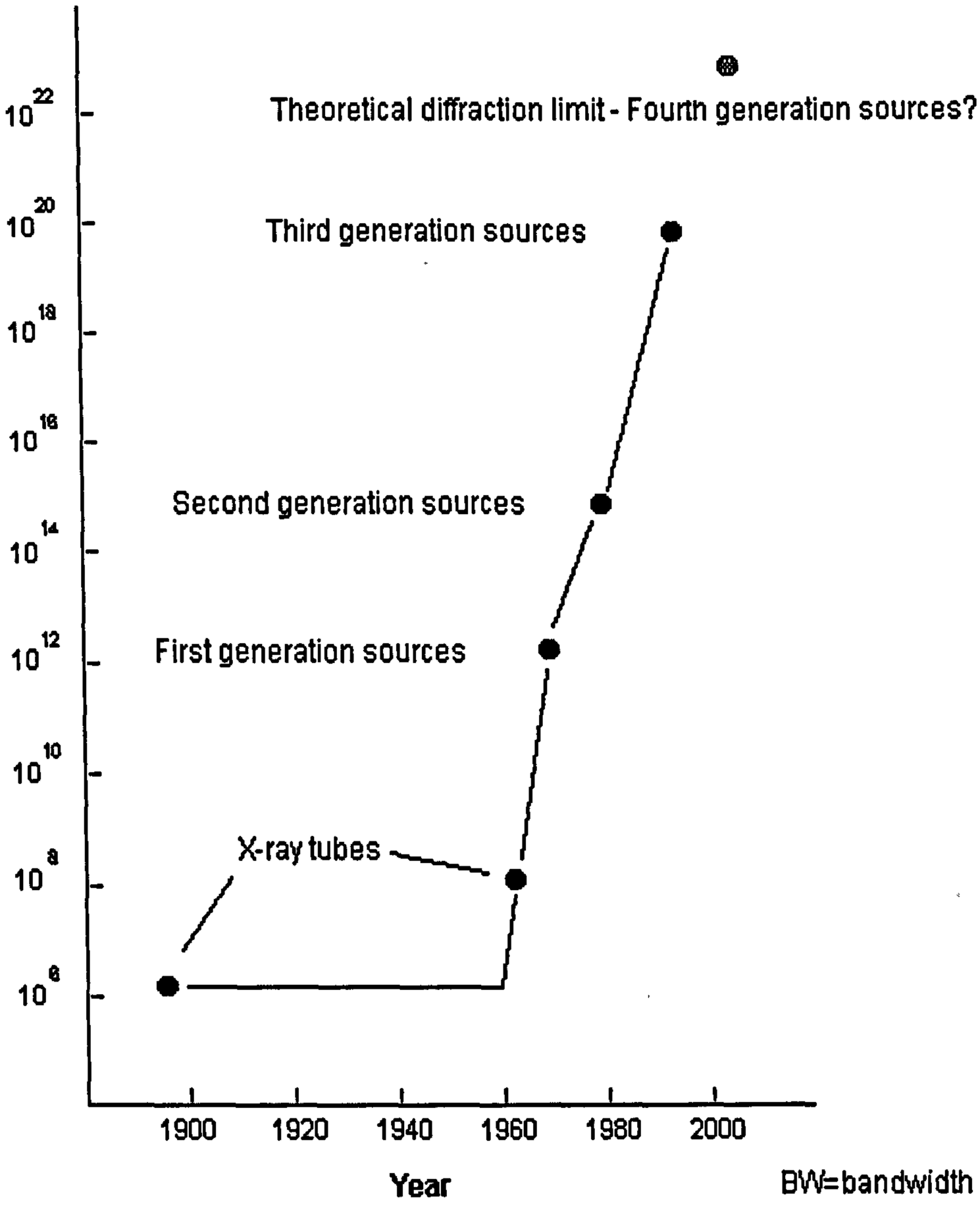


Figure 3.1 The historical development of generating increasingly brilliant X–rays

A synchrotron emits high energy light by accelerating charged particles such as electrons or positrons to near relativistic speeds. A series of electromagnets arranged around a circle alter the direction of the fast moving particles, making them follow a circular path. These electromagnets are termed bending magnets, and as the charged particles are deflected from their original path, they emit intense beams of light (synchrotron radiation). The emitted light partially covers the range of the electromagnetic spectrum, from infra-red to gamma-ray. The operation mode of the synchrotron and specific setup of the stations affect the nature of radiation used in individual experiments.

There are several benefits to the use of radiation emitted from a synchrotron source as opposed to a conventional rotating anode source.

The flux (the number of photons per second passing through a defined area) and brilliance of synchrotron emitted X-rays are many times greater than of those emitted from conventional sources.

Apart from this high flux though, the source size of a synchrotron is also relatively small due to minimal angular divergence, meaning that much more of the emitted radiation reaches the (small) sample. This can be significantly improved upon by the use of focusing mirrors, monochromators, zone plates and refractive lenses (or a combination of these) enabling the diameter of the X-ray beam to be focused to the sub-micron scale, without attenuating the beam and incurring a major loss of flux.

The circulating particles in the storage ring travel in 'bunches' (gapped pulses) with a precise time interval between the bunches. The emitted radiation is consequentially generated in pulses and can be utilised in time-resolved experiments.

Finally, the radiation emitted from the synchrotron source is both highly polarised (100% in the plane of the ring), and polychromatic (UV to gamma) enabling a wide range of experiments to be performed using the same source at the same time.

3.2.1 General synchrotron operation

The second and third generation synchrotron sources used in this project were; the Synchrotron Radiation Source (SRS) CLRC Daresbury laboratory, Daresbury UK; the European Synchrotron Radiation Facility (ESRF), Grenoble, France; and the Advanced Photon Source (APS) Argonne laboratory, Illinois, USA.

Each of these machines (SRS, ESRF, APS) is essentially a complex of three particle accelerators; a pre-injector - linear accelerator (linac), a booster synchrotron, and a storage ring (Figure 3.2). Particles are fired from the linear accelerator into the booster synchrotron where they are accelerated to near relativistic speed before being injected into a larger, storage ring. The particles will then circulate for several hours in a high vacuum at a relatively constant energy (several GeV).

The lattice of bending magnets around the storage ring keeps the particles moving in a (nearly) circular orbit; since however they move in straight lines changing course only

at the bending magnets, it would be more accurate to say that they traverse around a polygon. At each of the bending magnets, the diverted particles emit radiation that can be channelled down beamlines that sit at a tangent to the storage ring, to eventually be delivered to the various experimental stations. This arrangement can be seen in the simplified schematic of the SRS shown in Figure 3.2, which shows the three accelerators as well as the beamlines of the storage ring.

For more specific details of the physical characteristics and operational parameters of the three synchrotrons, see Table 3.1.

3.2.2 Synchrotron radiation and increasing experimental demand

As increasingly ambitious experiments are being designed and attempted, greater demands are made upon the facilities available at synchrotron radiation sources. These demands are both organisational and experimental; the use of synchrotron facilities is competitively sought by scientists world-wide and the resources are limited, increasing the need for faster turnover times for experiments. Similarly, a rising number of experiments require the delivery of maximum intensity of radiation at the sample, due to weakly diffracting materials and/or for those crystals with large unit cell parameters. Increasing the flux delivery to the sample has become a major priority in synchrotron science, greater flux means faster experimental turnover for most scientists, and makes some experiments possible that would otherwise not be attempted. A range of means are employed to achieve this goal, one of them being the use of insertion devices.

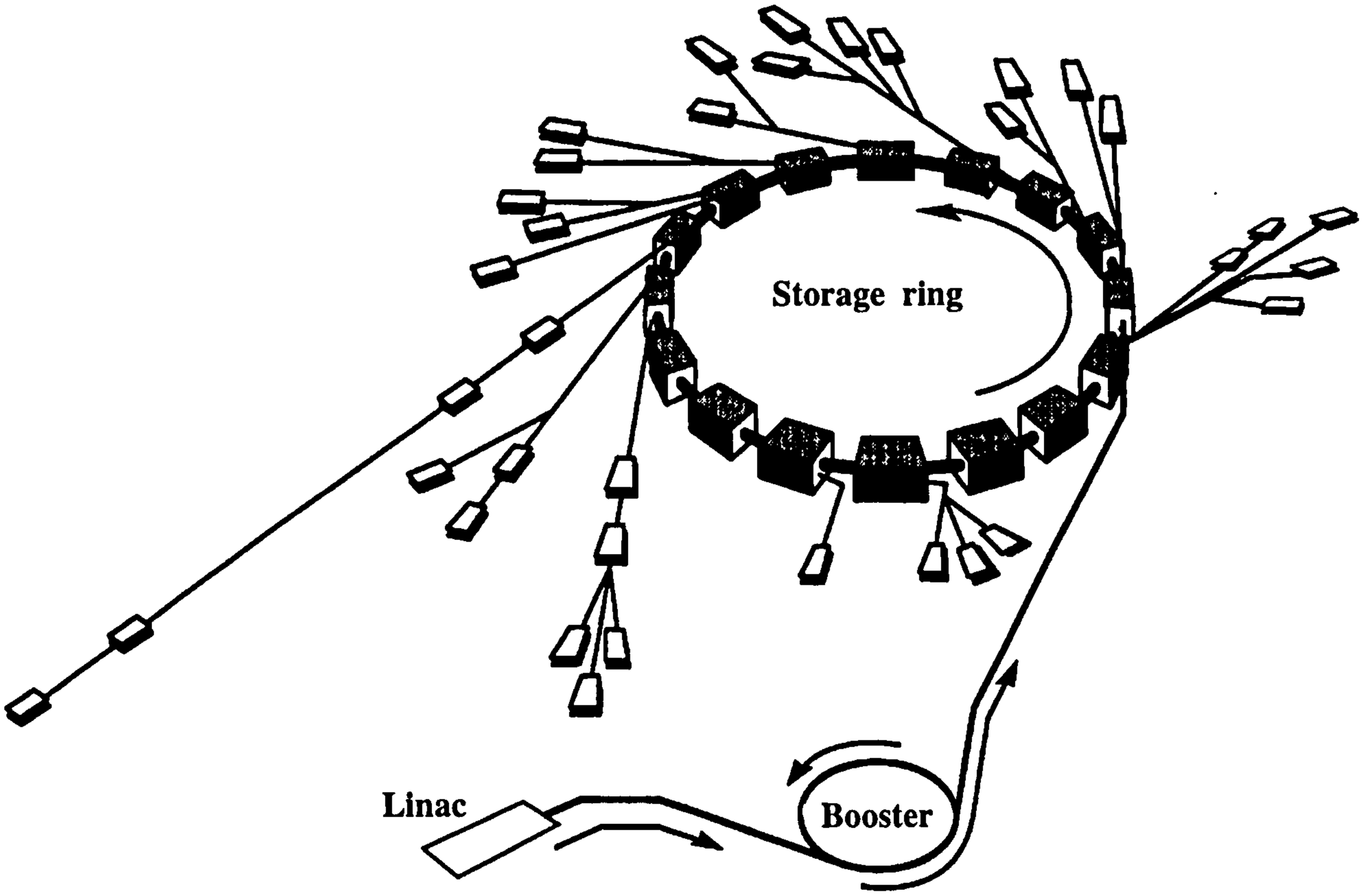


Figure 3.2 A simplified schematic layout of the SRS, Daresbury

Electrons generated from the linac (electron gun) further acclerated in the booster synchrotron before being injected into the storage ring where they travel in 'bunches' (arrows indicate the direction of travel). Electrons are directed down beamlines that are located around the storage ring, where specialised stations on each beamline utilise the radation generated. SRS schematic courtesy of T. Wess.

STORAGE RING PARAMETERS	SRS	ESRF	APS
Accelerated particles (e^-/e^+)	e^-	e^-	e^-/e^+
operational energy	2.0 GeV	6.0 GeV	7.0 GeV
Current (in multi-bunch mode)	200 mA	200 mA	100 mA
Lattice type (bending magnet distribution)	FODO	Chasman-Green	Chasman-Green
Bending magnets	16	64	80
Bend radius	5.6 m	24.95 m	39 m
Bend Field	1.2 T	0.8 T	0.7 T
Circumference	96 m	844 m	1104 m

Table 3.1 General operational parameters of the SRS, ESRF and APS synchrotrons

3.2.3 Insertion devices

Devices inserted into the storage ring straight sections can be used to increase the flux density of photons delivered to the experimental stations. Devices such as undulators and wigglers are used to this end.

Both undulators and wigglers are composed of two sets of smaller magnets of alternating polarity. As the particle beam passes through the magnetic field of the insertion device, the path of the particles is altered. The effect is that of a cumulative increase in the intensity of the emitted radiation (dependent upon the number of times the particle path is changed, and the scale of the path changes). The nature of the effect on the particle path gives rise to the naming of the insertion devices; a large path deviation being called a wiggle, and a relatively small path change being termed an undulation.

Undulators produce a considerable increase in the brilliance of the emitted radiation. However, the small deviations to the path of the charged particles caused by the undulator results in interference and photons are concentrated only at certain wavelengths, although this can be tuned by altering the field strengths of the electromagnets or by changing the size of the gap between them (affecting the extent of particle undulation). (See Figure 3.3).

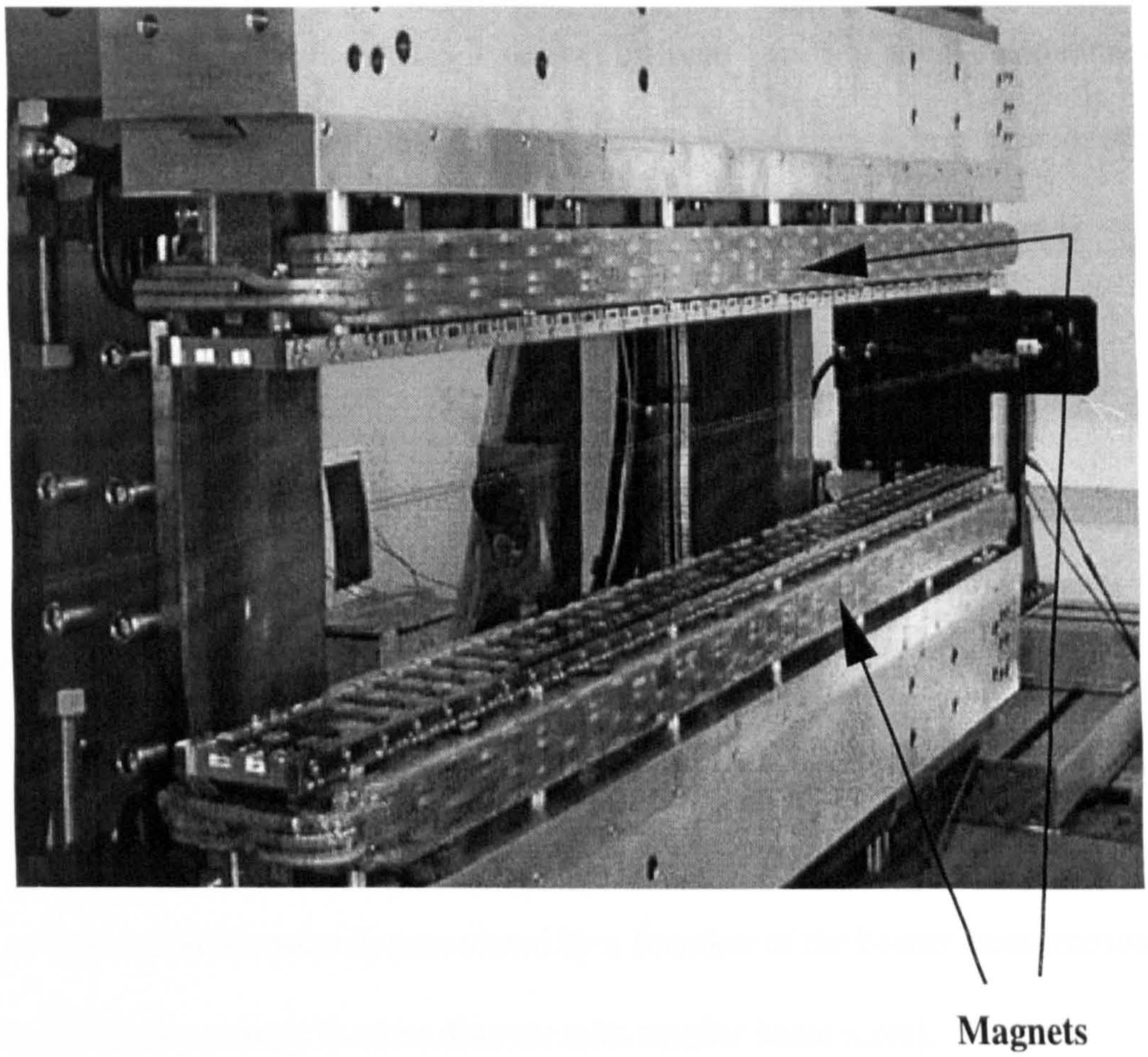


Figure 3.3 Electromagnet/permanent magnet helical undulator under development at the ESRF

Insertion devices have become common features of synchrotron radation stations. The undulator beamlines of the stations used in this project use devices of this type. The magnets can been seen in this image, the distance between the plates is reduced to a few millimeters in operation.

3.3 Brilliance, a necessity for collagen research

The ability of synchrotron sources to deliver brilliant beams of small cross-sectional area is ideally suited to X-ray fibre diffraction studies of collagen. The desire for the delivery of a brilliant X-ray beam of small cross-sectional area for this project is twofold;

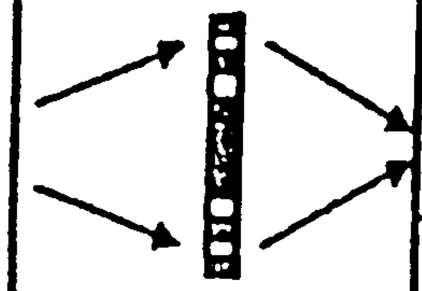
1) Some features of the collagen diffraction pattern are weak, and high flux/brilliance is needed to record diffraction within a suitable time frame (minutes and seconds rather than days or weeks).

2) Large or long molecules such as collagen give rise to diffraction maxima that are closely spaced, beams with a smaller cross-sectional area give rise to better resolution (every reciprocal lattice point is convoluted by a function of the beams cross-sectional area, the diffraction pattern is more discrete with smaller beam sizes).

Reducing the cross-sectional area of the beam without reducing flux is one of the main aims of a beamline designer. There are a variety of means available, some of which are tabulated in Table 3.2.

Table 3.2 Focusing optics for hard X-rays

The range of microfocusing devices that have been used to pass the micrometer barrier. This table has been kindly provided by Irina Snigireva and Anatoly Snigirev and reproduced here with their permission.

	Bent mirror or crystal			Capillaries	Waveguides	Fresnel optics	Bragg-Fresnel Optics		Refractive optics
				Kreger 1948	Feng et al 1993	Baez 1952	Aristov et al, 1986 [2]		Snigirev et al, 1996 [9]
	mirror	multilayer	crystal						
	Kirkpatrick Baez, 1948	Underwood Barbee, 1986	Johann, Johansson, 1931-1933				ML	crystal	
accept	~1 mm	~1 mm	1 mm	0.1 mm	~10 μm	~1 mm	~1 mm		~1 mm
gain	10^4	10^4	10^4	10^2	<10	10^4	10^4		10^2 -10^3
E	<20keV	<80keV	<100keV	<20keV	<20keV	<20keV	<100keV		<100keV
ΔE/E	w. b.	10^{-2}	10^{-4} -10^{-6}	10^{-3}	10^{-3}	10^{-3}	10^{-2}	10^{-4} -10^{-6}	10^{-2} - 10^{-3}
flux(ph/s)	10^8 (9keV)	10^{10} (12keV)	10^9 (90keV)	10^9 (8keV)	10^9 (10keV)	10^9 (9keV)	10^9 (10keV)	10^9 (10keV)	10^9 (20keV)
resol.	**0.7 μm** [24]	**0.7 μm** [25]	**1.2 μm** [26]	**0.05 μm** [27]	**0.15 μm** [28]	**0.15 μm** [7]	**0.9 μm** [29]	**0.5 μm** [4]	**0.3 μm** [12]

3.3.1 Beamline stations

The beamlines and stations located on them, specialise for experiments utilising particular parts of the radiation spectrum emitted from the synchrotron source. In X-ray fibre diffraction experiments on collagen, there is an additional need to provide enough physical space so as to be able to resolve the first meridional order of collagen (hkl = 001, Z=0.0015 nm^{-1}), although the specifics depend upon the exact wavelength of the X-rays, and the spatial resolution capabilities of the detector system used. Some stations that specialise in fibre diffraction also specialise in low or high angle diffraction experiments, although it is more common to combine a large degree of flexibility in the kind of experiments possible, as well as enabling the collection of both high and low resolution data. Some specific details of the stations used for data collection in this project are recorded in Table 3.3

The fixed wavelength SRS stations (2.1 and 7.2) were used in the collection of one dimensional data for the high resolution phase determined structure described in Chapter 4, station 2.1 for low resolution and 7.2 for high resolution. The respective resolution ranges of the stations were sufficiently overlapping so as to allow adequate merging of data sets (2.1, 1-500 nm; 7.2, 0.16-25 nm). The high brilliance beamline ID2 at the ESRF was used in collection of high and low resolution data for the one dimensional structure of collagen, and high spatial resolution data for the three dimensional structure. Primary diffraction experiments towards obtaining data from a single collagen fibril were performed at the Micro-Fluorescence, Imaging and Diffraction (Micro-FID) beamline (ID22 at the ESRF). Whilst the experiments at the

Biological Collaborative Access Team (Bio-CAT) beamline at the APS (ID18) concentrated solely on obtaining high spatial resolution data of the equatorial region of type I collagen (see Chapter 5).

3.3.2 Experimental complications

3.3.2.1 Parasitic scatter

Defects in the optical elements in the beamline give rise to aberrant scatter that contributes to the accumulated background noise in the diffraction pattern.

This can be reduced by the use of horizontal and vertical guard slits on either side of the optical components. These attenuate the parasitic rays whilst allowing the main beam to pass, the loss of flux being minimal, although even limited focusing of the X-ray beam reduces the extent of parasitic scatter.

Parasitic scatter is also caused by gas molecules in the path of the beam, reducing the overall intensity of the beam which is particularly problematic once the X-rays have been diffracted (the intensity of a diffracted X-ray is less then 1% of the incident beam; Blundell and Johnson 1976). The principal problem is the scattering of the main beam by air molecules, creating a 'halo' of diffuse low angle scatter. This loss of flux and increase in diffuse scatter can be reduced by housing the camera array in a vacuum, or by replacing air in the beam path with an electron sparse gas such as helium.

3.3.2.2 Increasing exposure time/minimisation of sample degradation

Obtaining high quality data for both the high angle and equatorial reflections of the collagen fibre diagram required exposure times that caused significant damage to the region of the sample exposed to the X-ray beam. The ionising effect of X-rays generates free radicals in the sample that attack and alter the molecular organisation of the collagen fibres. On the more intense undulator beamlines (ESRF and APS), a beam-sized hole would be quickly burned through the tendon in seconds, the principal concern here, of course, being heat damage.

Both of these problems were resolved by spreading the exposure of the sample to the X-rays along the length of the tendon, limiting the damage caused to any one part of the sample and increasing the exposure time. Remotely controlled motorised sample stages were available at the ESRF and APS stations, but not at the high-resolution beamline (7.2) at the SRS. It was therefore necessary to construct equipment to do this needed to be portable (to be used elsewhere if necessary) and compatible with a range of situations including the sample camera at station 7.2 (shown in Figure 3.4). The schematics of the control board circuitry are shown in Figure 3.5. The stepper motors were controlled from a portable PC using the software 'Motors' that was specifically written to drive a goniometer head, moving the sample cell in the Z plane and tilting the sample to the Z plane. The translation range was limited by the goniometer (20 mm in Z), but this allowed exposure times to be increased by 6-10 times.

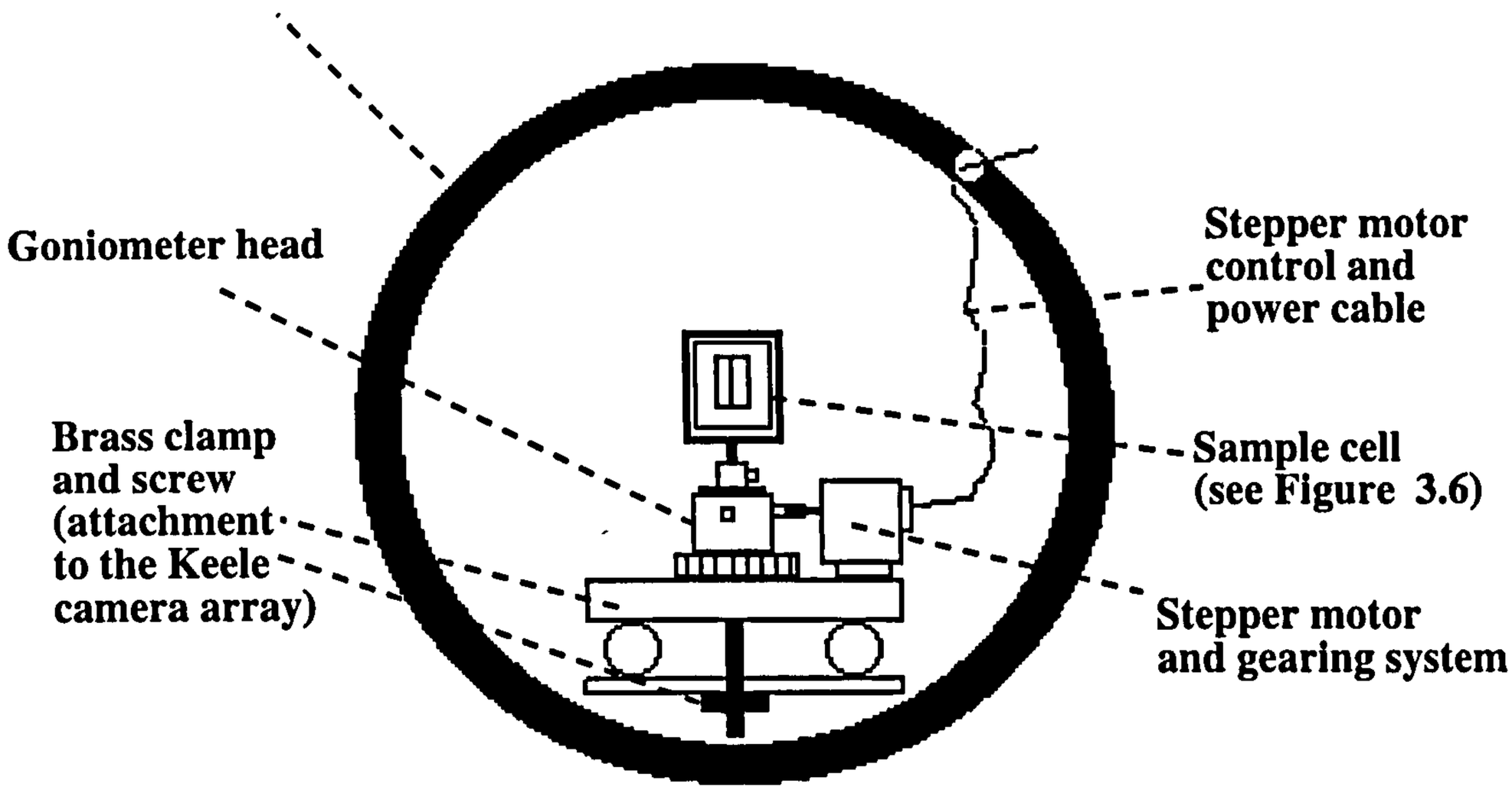
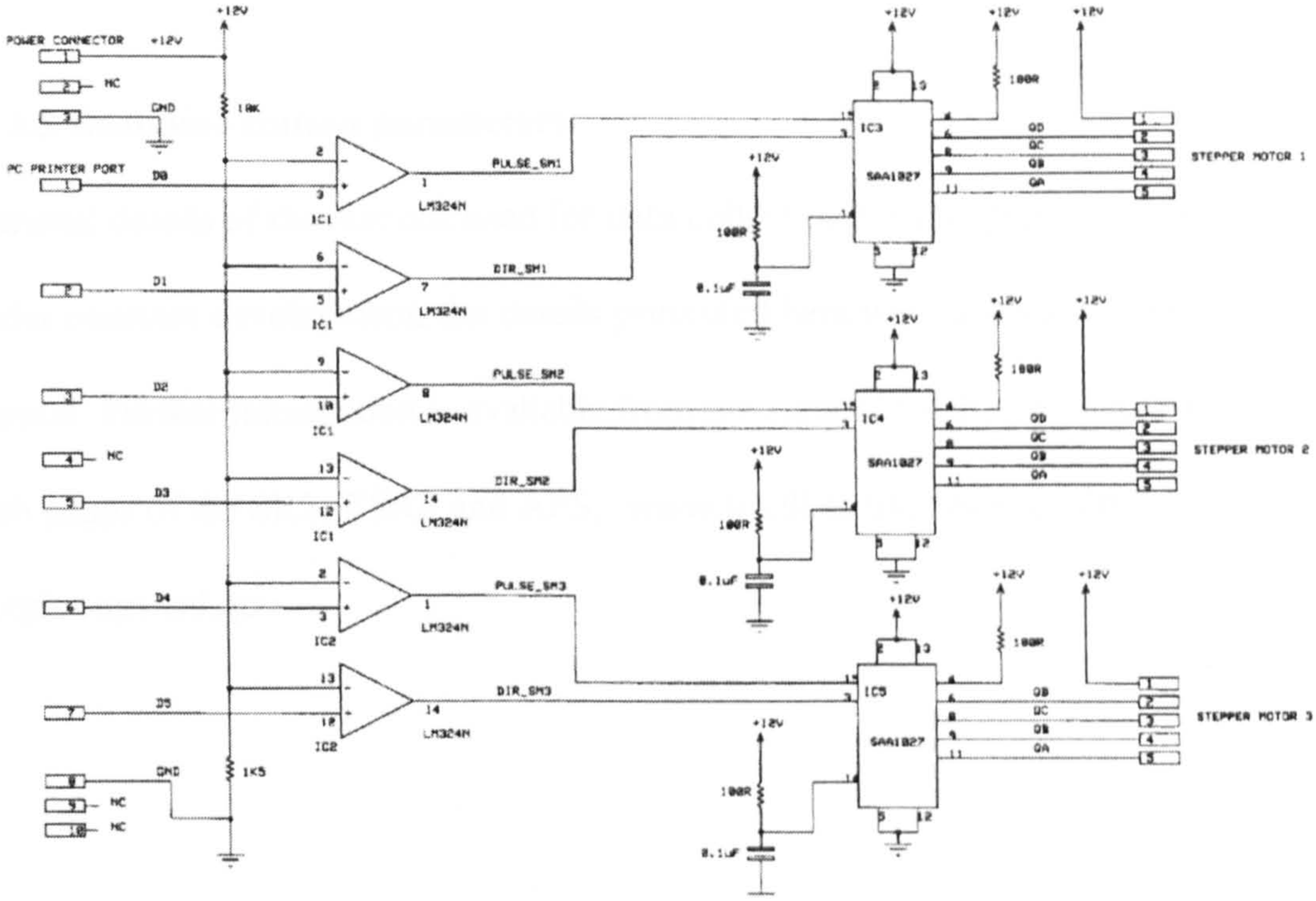


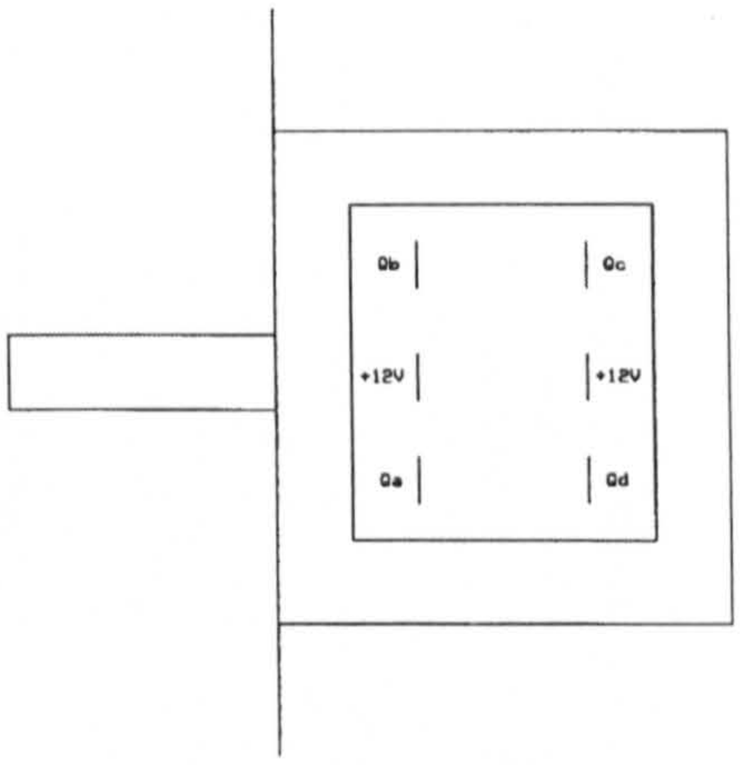
Figure 3.4 The sample rig used for data collection on SRS beamline 7.2

Head on view (down the beam pipe) schematic drawing of the remote sample positioning apparatus constructed for high angle diffraction experiments on beamline 7.2 (SRS). This setup enabled the collection of data sets for the same sample at various tilt angles, as well as extending the life–span of a sample by moving it in the vertical direction during data collection without needing to disturb the camera setup itself. The two stepper motors (one for tilt angle one for vertical movement) were controlled by a laptop PC running the program "Motors".



a

Stepper Motor Connection



Printer Port to Interface

25 Way 'D' Plug	Colour	Signal	15 Way 'D' Plug
Pin 2	Red	D0	Pin 1
Pin 3	Brown	D1	Pin 2
Pin 4	Black	D2	Pin 3
Pin 5	White	D3	Pin 4
Pin 6	Grey	D4	Pin 5
Pin 7	Purple	D5	Pin 6
Pin 18	Green	Gnd.	Pin 9

Interface to Stepper Motor

9 Way 'D' Plug	Signal	Colour
Pin 1	+12V	Yellow
Pin 2	Qd	Black
Pin 3	Qc	Brown
Pin 4	Qb	Red
Pin 5	Qa	Orange

b

Figure 3.5 Schematics of the stepper motor controller and motor interface

The circuit design and construction of the remote sample positioning controller used for data collection at beamline SRS 7.2, was performed by Frank Kelly of the Dept. of Computer Sciences, University of Stirling. Creation of the software and physical interfaces (motors, gearing, calibration etc.) was undertaken during the course of this project.

a) Stepper motor controller. Input connections constituted the 12v power supply, and parallel port interface to the PC. Output constituted the pulsed current to turn the stepper motor clockwise or anticlockwise. The three motors drove horizontal alignment, vertical alignment and sample tilt (see main text).

b) Stepper motor and PC parallel port interfaces, showing signal type transports.

Table 3.3 Beamline station parameters

Operational details of the stations used for data collection for this project. These stations are under constant development, the details presented here were accurate at the time of submission. Further information is available from the station's web sites, accessible via the web pages of the SRS, ESRF and APS; www.srs.dl.ac.uk, www.esrf.fr, www.epics.aps.anl.gov.

BEAMLINE STATION PARAMETERS	2.1 (SRS)	7.2 (SRS)	ID2 (ESRF)	ID22 (ESRF)	ID18 (APS)
Operational energy / wavelength	1.54 Å	1.448 Å or 1.28 Å	8-17 KeV (0.73 - 1.55 Å)	4-70 KeV	3.5-35Kev
Smallest vertical beam size (focused)	0.75 mm	collimated (-)	0.2 mm	0.002 mm	0.02 mm
Smallest horizontal beam size (focused)	5.00 mm	collimated (-)	0.6 mm	0.020 mm	0.01 mm
Focusing system	Asymmetric bent triangular Ge(111) 11.6° monochromator bent for horizontal focusing. Uncoated fused quartz mirror for vertical focusing.	Asymmetric bent triangular Ge(111) 10.5° monochromator for horizontal focusing (the standard, λ = 1.488 Å). Optional 8.5° monochromator for operating at λ = 1.28 Å. Bent platinum coated quartz mirror for vertical focusing.	Monolithic Si(111) double crystal monochromator for vertical focasing. Tororidal Rh coated quartz mirror for horizontal focusing.	Flat Si mirror with Pt and Pd coated (and none coated) strips, for horizontal focusing. Vertical double flat crystal 3° -30° monochromator. For microfocusing; Bragg-Frensel lenses Fresnel zone plates Compound refractive lenses.	Two Rosenbaum-Rock high-flux double crystal 6.5° - 38° monochromators; First Si(111), second Si(400) Rosenbaum-Rock elliptically bent vertical focusing mirror, Pt and Pd coated (and non coated) strips.
Detector systems	Gas-wire realtime detector.	35cm MAR research image plate detector.	Fuji Bas 2500 image plate scanner. 1024x1024 CCD (Princeton) detector .	Si(Li) detector . Si drift diode detector . PIN diodes, ionisation chambers. High resolution CCD cameras. Medium resolution CCD camera. Gas filled (position sensitive) detector.	Fuji Bas 2500 image plate scanner. 1024x1024 CCD (Princeton) detector . Fast-time-slicing scintillator array. Lytle fluorescence detector. Multilayer fluorescence analyser.
Resolution range	Low-angle	High-angle	Small-high angle	Small-high angle	Small-high angle
Camera strategy (if any)	Vacuum path	He filled path	Vacuum path	Air path (very intense beam)	Vacuum path

3.4 Sample environment

It was crucial to ensure that the sample was maintained in a moist environment; dehydration gradually changes the molecular packing arrangement of the collagen fibrils, and eventually leads to axial unit cell changes also. A sample cell was used that was water tight, and contained a small reservoir of buffered salt solution (0.15 M Na Cl) to keep the sample moist. The cell (shown in figure 3.6) also contains a number of pulleys for mounting and allowing a small amount of tension to be applied to the tendon, to ensure the sample is properly positioned in the beam, and to remove the 200 μm crimp (Rowe 1985), the extension needed being approximately 4 %. This limited stretching improves the alignment of the collagen fibres, and consequently improves the diffraction pattern.

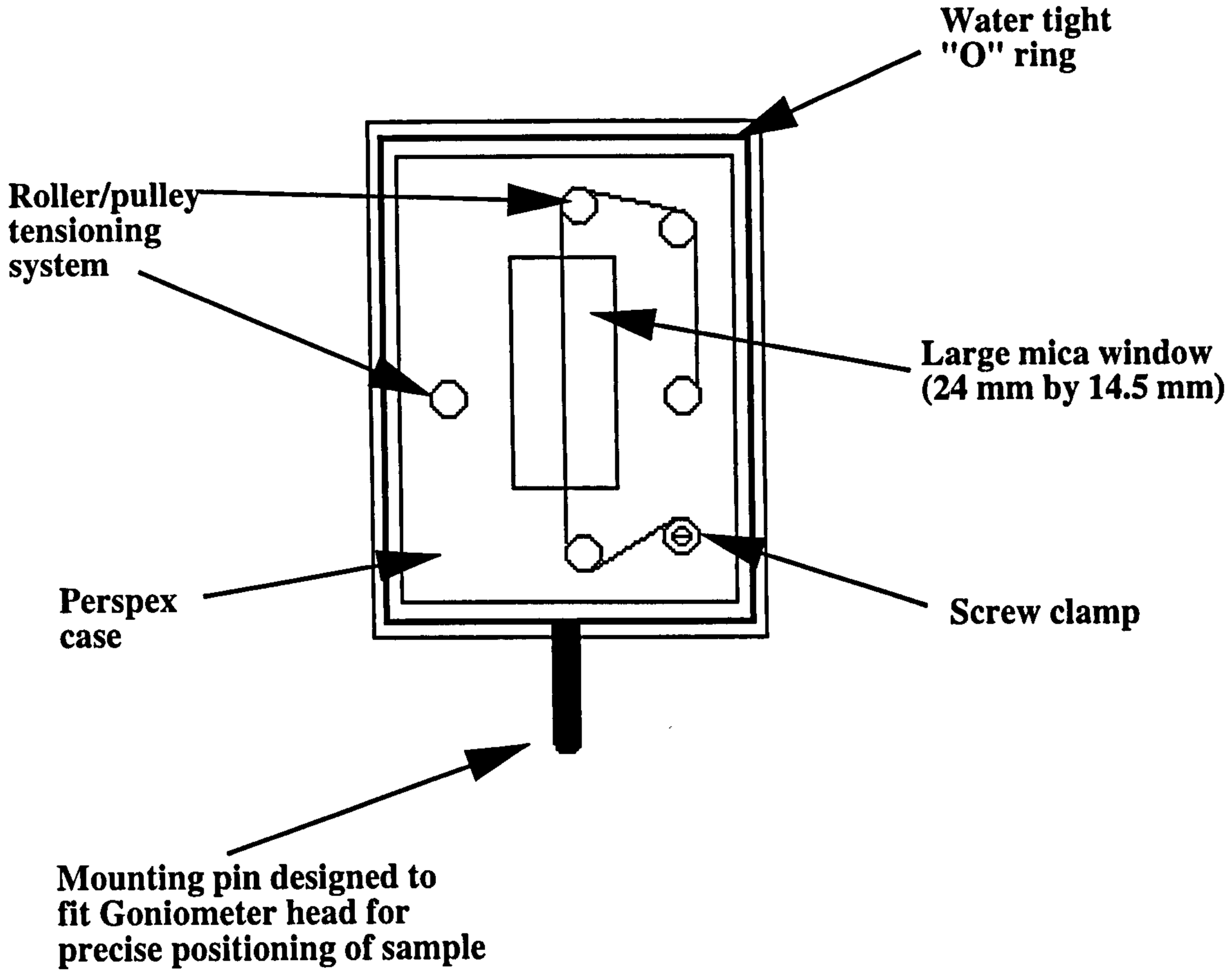


Figure 3.6 The sample cell

The sample cell was designed so as to be able to mount the sample under slight tension and maintain a moist environment.

3.5 Data collection

Data were collected electronically on a range of detectors available at the respective stations, these being; gas-wire area detector, Mar image plate, image plate scanned on Fuji Bas 2500 scanner, Princeton high definition CCD camera.

A data file was automatically generated with each of these detector systems, where the diffraction data is recorded as a two dimensional image, and intensity is represented visually by colour depth. Spatial detector shape corrections were carried out either automatically by the station software in the generation of the diffraction image file, or manually using the software suite FIT2D (also for linearisation of Fuji data files). Specific corrections such as the Lorentz correction, sample to detector distance, absorption etc. are described later in this Chapter.

3.6 Data extraction

The collagen fibre diagram contains a series of intense meridional peaks roughly parallel to a weaker series of equatorial peaks on row-lines superimposed upon a diffuse background. The meridional series of intensities contains detailed information about the axial packing structure of collagen molecules, which in rat tail tendon possess considerable long range order (over 140 orders observed Figure 4.1, and Orgel *et al.*, 2000). At this resolution, structural details down to the two amino acid residue level are observable (more accurately, details of the *projected* structure of 4 or 5 collagen chains (triplexes) are observable). This is certainly sufficient resolution to provide compelling evidence for the telopeptide conformations.

The equatorial pattern contains data relating to the average packing structure of collagen molecules. The discrete Bragg peaks on row-lines parallel to the meridian indicate crystalline order whilst the underlying diffuse scatter shows that there is also a degree of liquid-like disorder in this packing arrangement (Hulmes *et al.*, 1995).

Contribution to the background is made by scattering caused by the sample cell windows and the optical elements of the beamline. It is necessary to remove this background so that the intensity of the Bragg peaks can be more accurately measured. The nature of the X-ray diffraction pattern and the process of investigation of the one and three dimensional structures of type I collagen in this project has meant that the background subtraction and intensity determination have been carried out in two separate approaches; one for the meridional series and one for the equatorial pattern.

3.6.1 The meridional series

3.6.1.1 Background subtraction and peak integration

One-dimensional integration's of the meridional series were prepared from both low-angle and high-angle X-ray fibre diagrams using the software suite 'FIT2D'. Although excellent data treatments exist within the FIT2D suite for background subtraction (discussed in the next section) the process of building a model 2D function to fit the whole meridional series was slow, requiring that the pattern be fitted piecemeal. Hence this was only performed for the most steeply sloping part of the low angle pattern (corresponding to orders 1-25). The background estimation for the remaining high-angle pattern was made from the 1D summation, this having the advantage of dealing with a simpler continuous, one dimensional polynomial rather than a series of two dimensional ones requiring the remerging of data sets. This also proved to be more time efficient.

Two software suites were used to estimate the background and integrate the peaks: 'Fit' (recently integrated into a GUI to become 'XFit'), by Richard Denny of CCP13 (non-crystalline diffraction software accessible via www.srs.dl.ac.uk) and 'Peakfit', a suite written specifically for this project.

In Fit, the background would be fitted according to a polynomial of degree 4 or less whilst the peaks were fitted to the Gaussian type: $y = he^{-4\ln 2(x/w)^2}$, where x and y are Cartesian coordinates, h is the height and w is the full width at half maximum.

Various parameters such as background, peak position, peak height etc. could be set and/or tied to other parameters (e.g. peak position of several consecutive peaks), or left free for the algorithm to determine. Normally it would be appropriate to set and tie the peak positions to ensure an accurate determination of background and peak area.

'Peakfit' was written along similar principles to 'Fit'. The major difference being that unlike Fit, the number of peaks that could be fitted at one time was theoretically unlimited, whilst Fit was limited to only eight peaks at a time. Fit was written with time resolved data in mind, where most data sets are comprised of a small number of peaks per frame but with potentially large numbers of frames. This is a limitation as far as data extraction is concerned in determining the underlying background for a frame containing over a hundred peaks, as was the case in this project. Hence Peakfit was written to fulfil the role of Fit, but allowing the determination of a continuous background to a polynomial function whilst fitting tens of peaks at once.

In the use of both programs, problems of spatial disorder and Lorentz effects, which alter the Gaussian character of the peak shape, were overcome in the same way. Because these effects make the determination of peak shape somewhat complicated, it was judged that measuring the peak height as the intensity function rather than the peak area (integral) would suffice (since they are directly proportional to one another, Fraser *et al.*, 1976).

3.6.2 Equatorial data extraction

3.6.2.1 Background subtraction

The suite of programs FIT2D (FIT2D manual is accessible through www.esrf.fr - Hammersley 1998) was used to remove the diffuse scatter from the low-angle diffraction pattern background. This background was assumed to be smooth and regions containing background were presumed to be identifiable from regions that did not. Interpolation was performed within a user-determined area of the pattern, with pixels containing Bragg peak data masked (excluded from the calculation).

Surface polynomial functions (two-dimensional Chebyshev polynomials) were used to fit a background model to the data (as Wess *et al.*, 1998a). Generally, the area fitted would be larger in the Y direction than the X, hence the X polynomial would be in the order of 3-5 whilst the Y would be approximately 5-10. In either case, the polynomial order would always be much smaller than the number of pixels within the interpolation region.

3.6.2.2 Selection of appropriate background model

A suitable background model is observed when the residue between the model and observed data is low (Hammersley 1998). An important additional check is a visual inspection of the model background and the background subtracted interpolation region. The lowest order polynomial that fitted these criteria well was selected.

3.6.2.3 Intensity determination

The integration of intensity of the Bragg reflections in the low-angle equatorial diffraction pattern has not been possible until recently for several reasons:

The quality of diffraction data has been insufficient, with either too few film grains or pixels for each reflection to be accurately determined or blurred spatial resolution due to large beam sizes. Poor signal to noise ratio has also made it difficult to determine the relative intensity of Bragg reflections.

Even when the insufficiencies of these problems are overcome, the nature of the diffraction pattern with the partial overlapping of several row-lines has made the prospect of measuring the Bragg peak intensities seem unrealistic. Despite this, partial determinations have been previously attempted (Fraser *et al.*, 1987, Wess *et al.*, 1998a), but the vast majority of the diffraction pattern has yet to be quantified.

Wess *et al.*, (1998) developed a model for fitting the group 1 row-lines (where group n (n is an integer 1-4) refers to the groups of row-lines in the collagen fibre diagram, see Figure 4.3) using a two dimensional Gaussian function for each peak, the coordinate positions of each peak being calculated from the unit cell parameters of Wess *et al.*, (1995). It is this strategy that has been adopted and developed here, using a simulated diffraction pattern generated from the unit cell parameters of Wess *et al.*, (1995) to fit the background subtracted observed diffraction pattern, the minimisation between the two being driven by a simulated annealing method.

3.6.3 Simulated annealing

The method of simulated annealing is essentially a means of computational minimisation. A number of elements, N are fitted into a space or array of spaces that are discrete in number and size. The method has been applied successfully to a wide range of purposes, including the arrangement of complex circuitry on minute silicon substrates as well as to biological problems mainly dealt with in structural elucidation. It has proven to be a powerful minimisation method and at the same time, its implementation is quite simple (Press *et al.,* 1989).

The method of simulated annealing is based upon the analogy with thermodynamics, that is, the way that liquids cool and crystallise or metals cool and anneal. At relatively high temperatures the molecules of a liquid are free to move relative to one another, when the liquid is cooled slowly, this thermal mobility is lost. The molecules are often able to arrange themselves according to an energy minimisation scheme with ordered arrays up to billions of times the length of the size of the constituent atoms in all directions. It is the slow cooling that allows the energy minimisation to take place, a sudden 'quenching' of the system could be more accurately described as a snapshot view of a higher-energy state. The slow cooling allows the redistribution of molecules/atoms as they lose mobility (Press *et al.,* 1989).

3.6.3.1 Minimisation algorithms

The algorithm 'Search' was written along these principles, applying the annealing analogy to fitting a model diffraction pattern to the observed diffraction pattern, (the algorithms are summarised in the schematic flow diagram of Figure 3.7), each peak of intensity being treated as an atomic position in the annealing analogy. The coordinate position of the intensities in the diffraction pattern were fixed according to the unit cell parameters of Wess *et al.*, (1995).

Three kinds of intensity change (analogous to atom redistribution in the annealing analogy) were allowed for each peak of the model pattern:

1) A random intensity change between the minimum and maximum pixel values of the observed pattern, of one randomly determined peak.

2) One random intensity change for a selected sequential series of peaks of random length.

3) One random global intensity change for all the peaks being fitted.

The model system has a temperature, just as in the thermodynamic analogy. At high temperature, greater intensity changes where allowed as well as more wide spread change to the number of peaks affected. As the temperature drops, the chances of the options 2 or 3 (above) occurring diminish to reflect the fact that in the thermodynamic

analogy widespread changes affecting arrays of atoms would be less likely to occur as the liquid cools.

The temperature of the model system (model diffraction pattern) dropped each time the residue between the model and observed data showed that there was a better global fit (determined by RMS/R-factor, defined below). To avoid falling into local pockets of minima, the temperature change would not be immediate. The algorithms would assess whether or not the minima was shallow, by the number of alternative intensity configurations that gave lower residues than that of the current configuration. When the number of configuration possibilities producing lower residues than the previous cycles lowest value reached 10 or more, the lowest residue and its associated intensity configuration was selected (one cycle = 30 x 100 configurations). After 100 configurations, if any lower residues had been found, then the intensity configuration that gives the lowest residue was adopted and the algorithm continued to search for a better fit. This ensured that the algorithm generally followed the steepest slope of minimisation without taking a laborious amount of time to make decisions when the minimisation is straight-forward, and should be rapid.

If after 1 cycle there were no lower energy alternatives (improved simulated data fits to the observed), then the temperature of the system was set to the maximum, allowing the full range of intensity configuration changes. This allowed the algorithm to jump out of shallow or dead end minima pockets and find a better fit to the observed data (corresponds to finding a lower energy state in the annealing analogy).

The temperature of the simulated annealing system reduced as the fit between the observed and simulated pattern improved, the velocity of the temperature reduction decelerating inverse-exponentially. That is, steepest from high to low temperature so as to reflect the thermodynamic analogy.

The greatest changes in atom redistribution would occur in the top ranges of the temperature range, so in the algorithm, the greatest changes in intensity occur at the highest 'temperature'. This process continued until no further changes occurred in the relative intensity scale of the Bragg reflections in the simulated diffraction pattern.

Model fits of the diffraction pattern produced in this way possessed residues and R-factors equal to or less than that of 4.21 and 0.07 respectively.

The residue corresponds to the root mean square (RMS) observed, verses simulated mean pixel value of the highest part of the each Gaussian peak and surrounding area (note that by definition of a Gaussian/normal distribution, approximately 80% of the intensity is contained within 30% of the area surrounding the highest part of the peak). The maximum and minimum pixel values of the observed and simulated diffraction patterns were scaled to be real numbers between 0 and 256. Therefore an RMS of less than 10, represents less than 3.9% error, the high quality of the simulated diffraction patterns is indicated by observation (Figure 3.8) and low estimate of error (equal to or less then 1.64%). The R-factor was calculated in a similar way, differing only from the RMS in that the form of the R-factor is:

$$R=\Sigma_i \mid F_o(i) - F_c(i) \mid \quad / \quad \Sigma_i \; F_o(i)$$

Where F_o and F_c are the observed and calculated intensities respectively.

The estimate of error was restricted to the areas surrounding each peak rather than the entire diffraction pattern as a means of optimising the speed of the algorithm, since the majority of pixels within the background subtracted observed and simulated patterns were equal to zero (those areas of the pattern where no Bragg scatter is recorded).

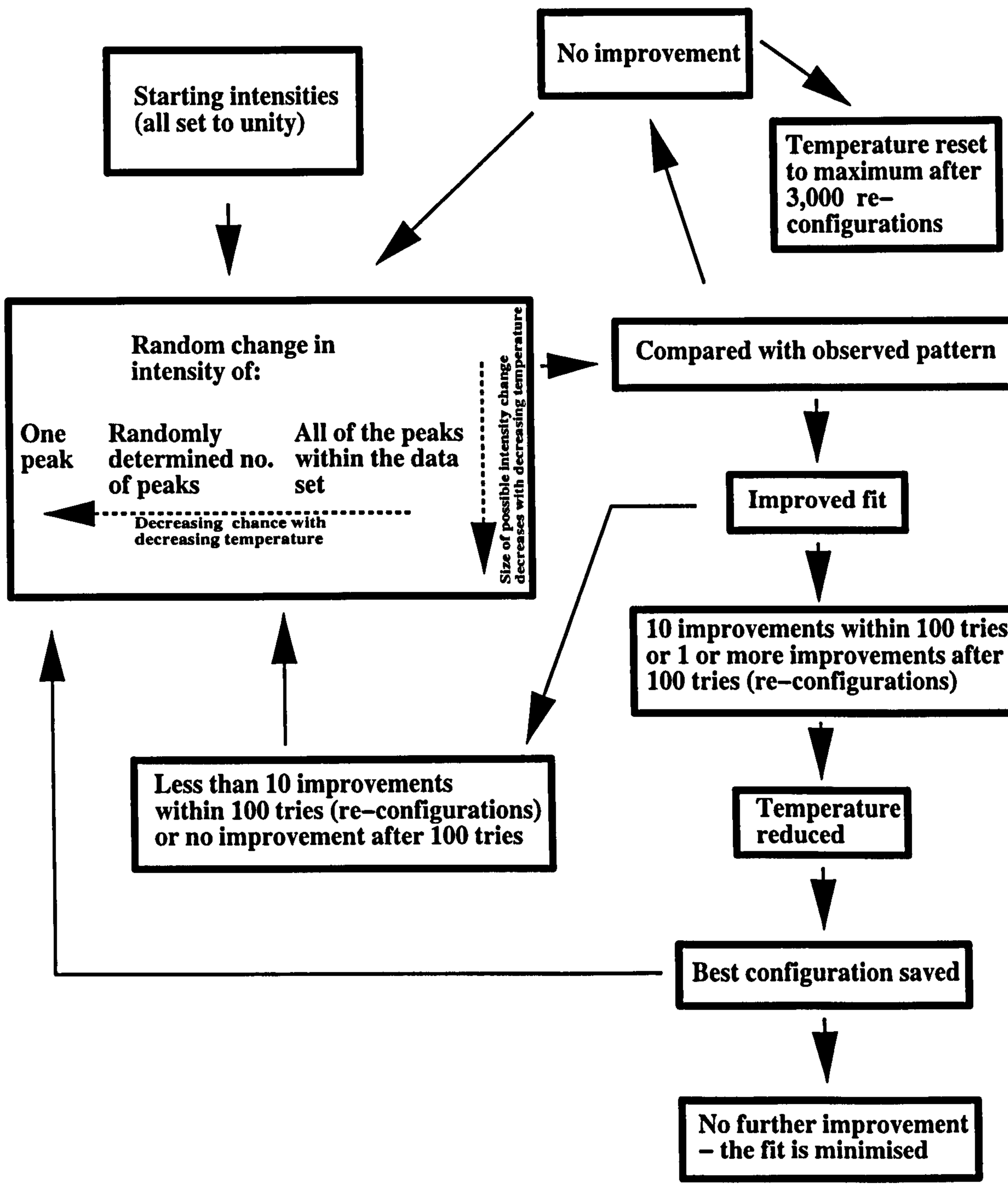


Figure 3.7 Flow diagram describing the operation of the Program "Search"

Figure 3.8 Determination of the equatorial diffraction pattern intensities

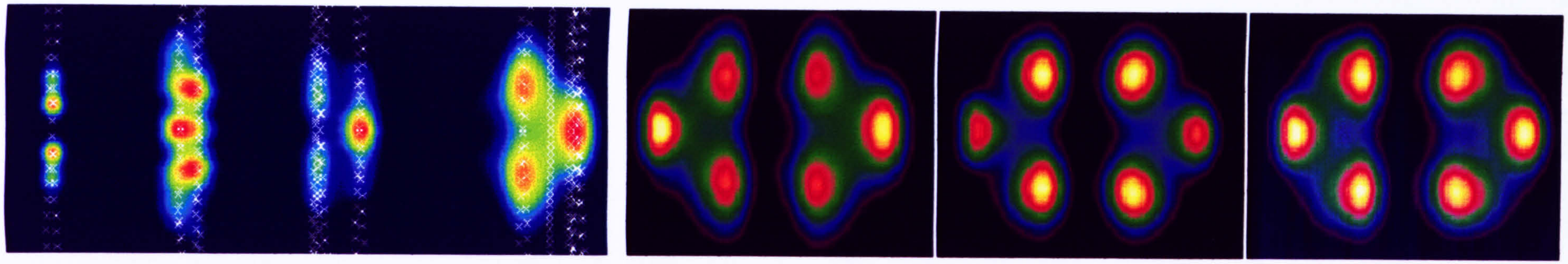
a) Background subtracted native profile with indexed Bragg peak positions (Wess *et al.*, 1995), produced by the program 'Search'.

b) Simulated diffraction patterns restricted to group 4, the most difficult part of the pattern to fit due to large number of overlapping peaks (shown in a).

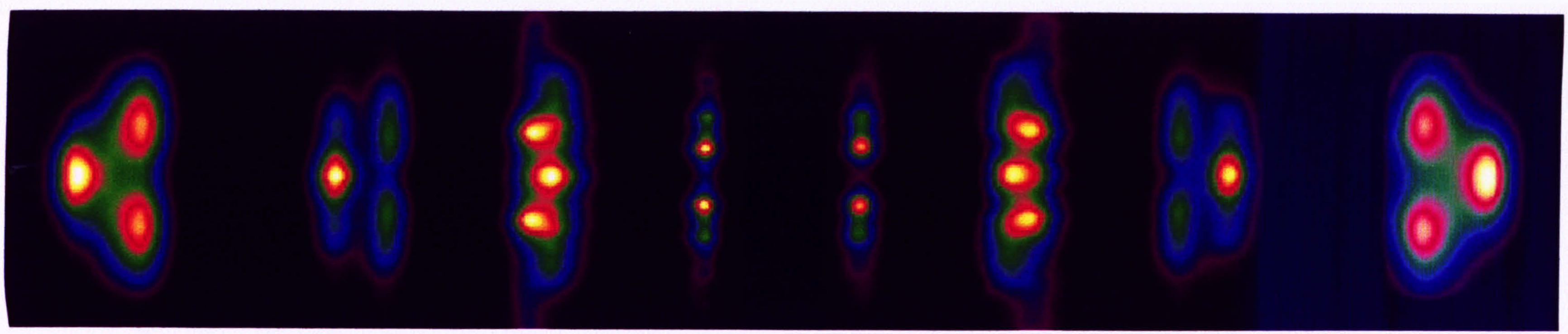
The left side of each image is the simulated pattern, the right side the background subtracted observed pattern, these are as follows;

i) Native, ii) iodide derivative, iii) gold derivative.

c) Native simulated pattern (left) background subtracted observed pattern (right).



a b) i ii iii



c

3.7 Data correction

3.7.1 The meridional series

Systematic corrections were applied to the intensity values obtained from the 1-D profile. The Lorentz correction is the most significant in this case since the range of meridional reflections investigated, show significant decay in intensity as the Ewald sphere (of radius $^1/\lambda$) curves away from the discrete toroids ('doughnut' shaped discs) of constructive interference originating from the diffracting fibre (Vainshtein 1966). The Lorentz factor may be regarded as the term that corrects for the fact that not all the diffracted rays from the meridional reflections have had the same opportunity of being recorded (equally applicable for reflections on the R axis as for the meridional series on the Z axis).

Aside from the decay in recorded intensity due to the curving of the Ewald sphere, an account must be taken of the fact that not one of the unit cell axes coincides with the fibre axis. This means that the unit cell is tilted with respect to the fibre axis (in addition to the angle between the fibre axis and the normal to the X-ray beam). Hence the radius of the meridional toroids or flattened discs in reciprocal space increases in proportion to order number (resolution in Z space), increasing the area over which the intensity is spread, the spread function approximating to a Gaussian. With increasing Z, the Ewald sphere sampling of the Gaussian changes due to the meridional toroid/disc radius increasing and the effective radius of Ewald sphere sampling of the reflection decreasing.

This effect was simulated numerically within the range of meridional orders n=1-150 in the program 'Tilt'. The intensity of each of the simulated reflections was set to unity, which was distributed as a Gaussian function over the area of an annulus of radius r.

The Gaussian was of the form: 2.71828**-((nmax/2. – n) **(2./hw**2)) (where ** represents 'to the power of'; n is an array element (e.g. nmax=100); hw is the halfwidth, equal to 30 in this case). The radius of each torriod (approximated to an annulus) was equal to the expression: $r = (n/d) \tan \theta$ (where n = the order of intensity (1-150), d = the diffraction grating spacing (67 nm), and θ = the sample angle of tilt relative to the line perpendicular to the plane of the incident X-ray beam + 5^0 (the tilt of the unit cell relative to the fibre axis (Fraser *et al.*, 1983), the sample tilt being between 0.0-6.5^0). The radius of the Ewald sphere being $1/\lambda$ (λ -the wavelength of the incident radiation).

The program 'Tilt' calculated the point of intercept of each annulus with the Ewald sphere and the relative intensity sampled for each order in reciprocal space. The resulting intensity profile of the orders 1-150 was plotted after the Lorentz factors described above were calculated (see Figure 3.9). This profile shows the underlying diminishment of meridional intensity with decreasing Z, and the inverse function of this curve provides a correction factor for each meridional order of intensity.

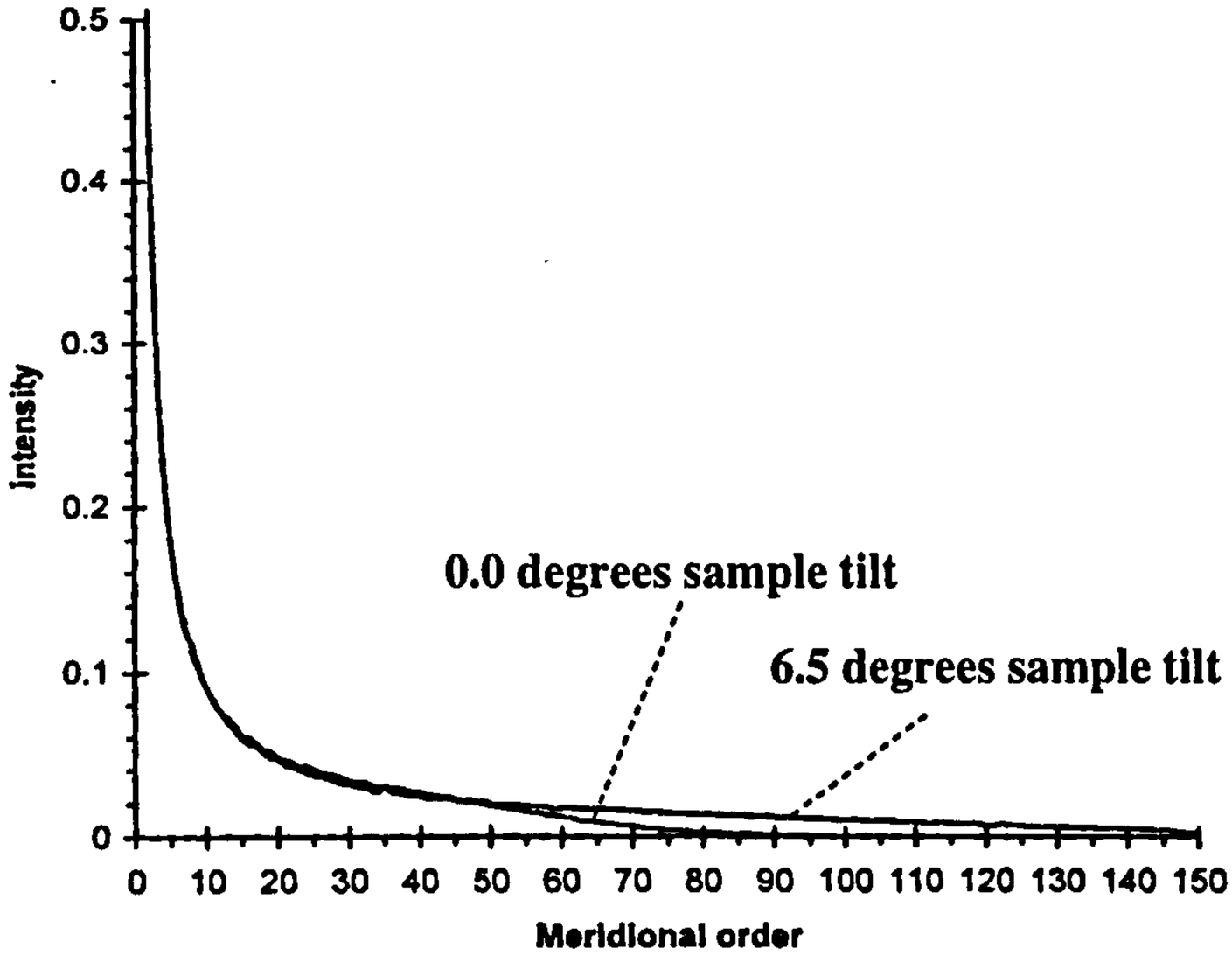


Figure 3.9 The Lorentz fall off of intensity in relation to increasing meridional order

The intensity content of each meridional order annulus was set to 1, this being distributed over the area according to a Gaussian function (see main text). Each annulus was sampled by the Ewald sphere construct, and the total value of the annulus sampled was plotted against the order number of the meridional order annulus in question (1–150 maximum). This was repeated for a sample tilt of 0.0 and 6.5 degrees, and is plotted above. The Lorentz correction is the inverse of these curves (depending on sample tilt).

3.7.2 Equatorial reflections

The low angle equatorial data determined from the native and derivative fibre diagrams were corrected for the cylindrical transform spread of intensity. As in Wess *et al.*, (1998), each intensity was multiplied by its position in R-space. Since the equatorial intensities determined from the fibre diagram are of relatively low order, it was not judged necessary to correct for Lorentz effect as described above for the meridional series. The reduction of intensity due to the curving away of the Ewald sphere is relatively small over this area.

Chapter 4

The one-dimensional structure of type I collagen

4.1 Introduction

The mechanical strength of many connective tissues is conferred by the presence of collagen fibres, the major constituent of these fibres being type I collagen fibrils consisting of (triple helical) collagen molecules organised in a specific axial manner. The type I collagen triplex is a heteropolymer consisting of two α1 chains and an α2 chain of over 1000 residues in length.

Fibrils form by a self-assembly process that requires the short non-helical end regions of the molecules (telopeptides) to facilitate correct molecular registration and crosslink formation. This in turn ensures the development of structural strength and integrity (Eyre *et al.,* 1984, Helseth and Veis 1981). Type I collagen contains specific lysine and hydroxylysine residues which are critical in forming intra and intermolecular crosslinks crucial for the normal configuration and stability of the 67 nm axial repeat of collagen fibrils in the extracellular matrix. The major crosslinkage sites are believed to occur between the non-helical termini regions (telopeptides) and helical segments of adjacent collagen molecules.

The conformation of the telopeptides regions has been sought in order to determine the key role of telopeptide in fibrillogenesis and maintenance of the structural organisation within a single collagen fibril. This requires a high resolution axial study of the 67 nm axial unit cell of fibrillar collagen in order to define the non-helical telopeptide conformation *in situ*.

4.1.1 Telopeptide structure

The Gly-X-Y repeating sequence of amino-acids characteristic of the collagen triple helix does not continue into the N and C termini (telopeptides) of the molecule. Since the two α1chains are longer than the α2 chains, the α1chains extending out past the α2 chains (from the tripeptide regions of the molecule) for several residues (7 at the N and 19 at the C termini). Clearly the telopeptides are incapable of forming a collagen triple helix, and therefore do not necessarily conform to the same axial translation as that of the rest of the molecule.

Residue spacing within the telopeptide regions that is significantly shorter than that of the main chain collagen helix has been reported; Hulmes *et al.*, (1980) suggested a value of 0.282 nm inside the main chain, 0.241 nm in the N-terminal telopeptide and 0.2 nm in the C-terminal telopeptide. However, in theory, the telopeptides could be extended, but in axial projection seem contracted due to the angle of azimuth tilt relative to the main chain. Alternatively, the axial contraction could be due to one or both of the telopeptides adopting a folded conformation.

Previous attempts to determine the conformation of the telopeptides have produced relatively ambiguous results, usually due to their model-based premises. Even studies such as that of Bradshaw *et al.*, (1989) where the phase solutions were not model-based (unlike Hulmes *et al.*, 1980), were unable to clarify the specific conformation of the telopeptides due to limited resolution.

A number of studies based upon the primary structure of type I collagen; Hulmes *et al.,* 1977,1980), Jones and Miller (1987,1991), Vitagliano *et al.,* (1995), Wess *et al.,* (1995, 1998), collectively acknowledge that the telopeptides are capable (in theory) of adopting a range of contracted, extended, or folded conformations or mixed combinations thereof. Most investigators are of the opinion that the telopeptides are axially condensed. Whilst the specific conformation of the telopeptides *in situ* has remained unclear, researchers investigating the conformation of isolated and synthetic peptide fragments have had some success, although the relevance of these studies to the native structure is uncertain.

Otter *et al.,* (1988) working with a single synthetic α1 chain C-telopeptide and NMR spectroscopy were able to demonstrate that the isolated chain was axially condensed, with a possible disposition to folding. Scott (1986), proposed a folded C-telopeptide conformation to account for the spectroscopic (disappearance of tyrosine chromophores on heating and at high pH) data obtained in the analysis of a single α1 chain C-telopeptide isolated from calf skin.

The most likely means available for determining the specific conformations of the telopeptides *in situ* may well be that of a study based upon the data contained in the meridional diffraction pattern. However, to avoid subjectivity, the phase component of the structure factors should be calculated rather than being based upon a model, and the study would need to be of high resolution to make the determination of telopeptide structure more stringent.

4.1.2 A high resolution study

The possibility of obtaining a unique and high-resolution electron density map from the X-ray diffraction data makes overcoming the problems of calculating the phase angles worthwhile. It has been previously demonstrated that it is possible to partially phase the meridional diffraction pattern of type I collagen in an unambiguous way through isomorphous addition (Bradshaw *et al.,* 1989), as has been the case here. In this X-ray fibre diffraction study, the tissue has been maintained in the hydrated fibrillar state, whilst detailed structural information was obtained using highly collimated synchrotron radiation. Over 140 meridional orders corresponding to an axial diffraction resolution of approximately 0.48 nm, have been observed in the collagen fibre diagram (Figure 4.1).

Data presented here show a significant increase in the resolution and also the definition of the axial structure of type I collagen compared with previous studies. This was obtained through a non-model dependent means of calculating the phase component of the axial structure factors. 124 native and isomorphous derivative orders of diffraction were used to generate a one dimensional electron density profile for the native protein and heavy atom derivatives to a resolution of 0.54 nm (2.5 times greater in real space, than any previous study). From this it has been possible to deduce the axial alignment of the collagen chains, the relative ratio of the gap overlap period, and to identify structural features of the telopeptides.

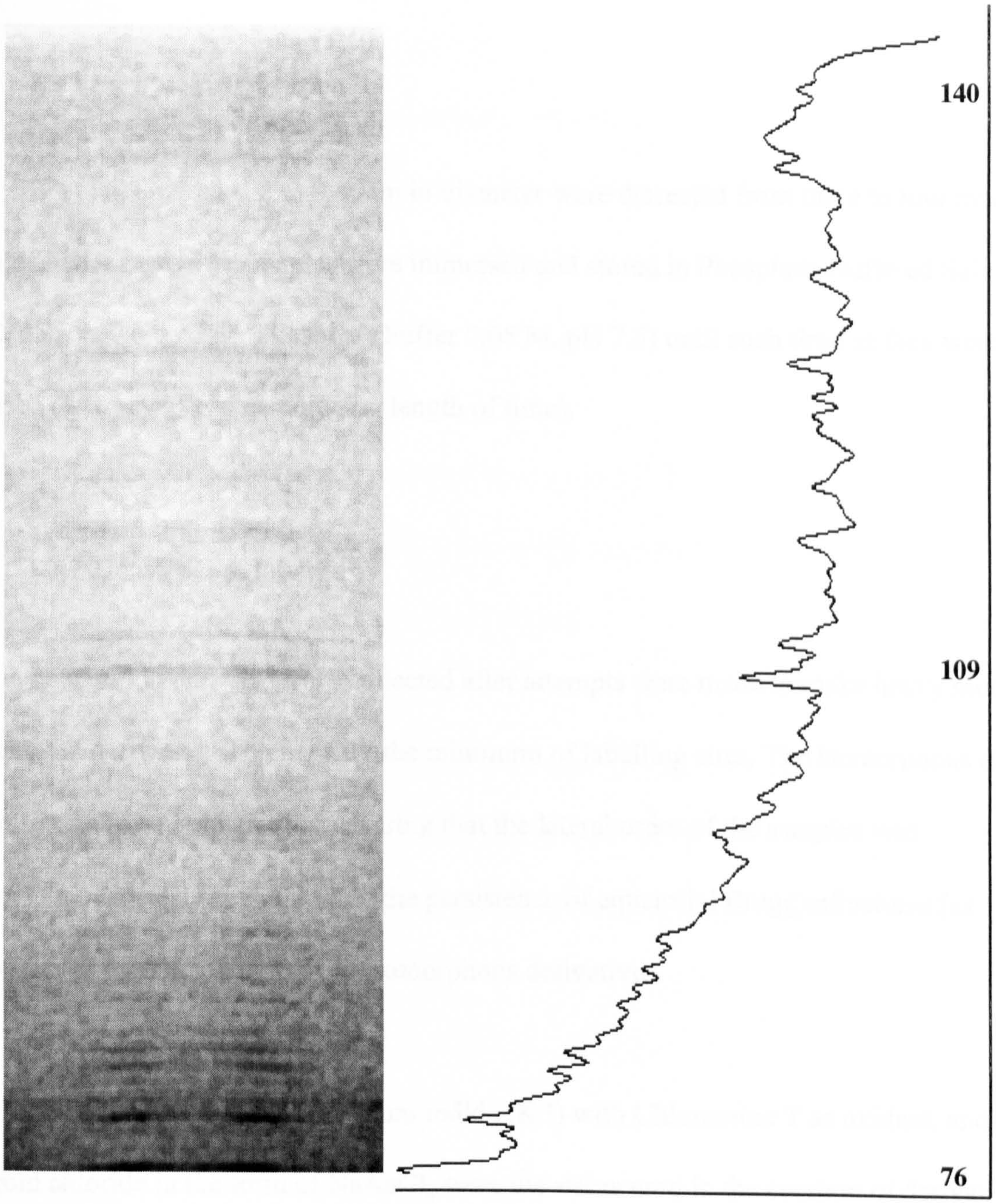


Figure 4.1 High angle diffraction from UV treated iodide derivative rat tail tendon

The high–angle portion of the meridional diffraction pattern is shown here (left), selected orders are labelled in the 1D profile (right), the background has not been removed. This diffraction pattern was recorded at SRS station 7.2.

4.2 Experimental

4.2.1 Sample preparation

Rat tail tendons 0.2 to 0.4 mm in diameter were dissected from three to four month old Wistar rats. The samples were immersed and stored in Phosphate Buffered Saline (PBS, Na Cl 0.15 M, Phosphate buffer 0.05 M, pH 7.5) until such time as they were used (4°C or -40°C depending on length of time).

4.2.2 Heavy atom labelling

Diffraction data sets were collected after attempts were made to make heavy metal labelled derivative proteins with the minimum of labelling sites. The isomorphous data sets were recorded carefully ensuring that the lateral order of the samples was maintained, easily monitored by the persistence of equatorial Bragg reflections (as observed in the native) in truly isomorphous derivatives.

Iodine in the form of potassium iodide (K I) with Chloramine T as oxidant, and gold chloride in the form of $NaAuCl_4$ were the stains used in the creation of derivatives for the successful phase solution. Several other heavy atom labels were attempted but failed to produce suitable isomorphous derivatives (see Appendix 2).

4.2.2.1 Creation of isomorphous derivatives

Potassium iodide or gold chloride was added to PBS to a final concentration of 1 mg/ml in 25 ml of solution. Rat tail tendons were stained within the heavy atom buffered solution for 1hr and then washed two or three times in excess buffer for 10-20 minutes. UV treated tendons were exposed to UV light ($\lambda = 254$ nm) for 1-2 hours and then immediately stained (see above).

Low angle X-ray diffraction data sets were collected for UV treated tendons that had *not* been iodinated, and the meridional intensities compared to those of the native protein. No significant differences were observed.

The sample was mounted above PBS in a sealed cell with 10 μm thick mica windows. The sample cell was also designed to apply a slight tension (approximately 4% extension) to the sample in order to remove the macroscopic crimp observed in such samples.

4.3 Data collection

X-ray diffraction patterns of rat tail tendon were obtained at the Daresbury Laboratory (CLRC) Synchrotron (SRS), UK, on beamlines 7.2 (in fibre diffraction mode, λ 1.448 Å) and 2.1 (small angle fibre diffraction, λ 1.54 Å), and at the European Synchrotron Radiation Facility Grenoble, France, ID2 (high brilliance beamline, λ 0.7-1.0 Å). The sample to detector distance was 0.425 m and 1.1 m for high resolution data (orders 6-146) at the SRS and ESRF respectively. Data sets were recorded on a variety of detector systems. At the SRS 2.1 a gas wire detector, at SRS 7.2 a Marresearch scanner, and at ESRF ID2, a CCD detector and phosphor image plate scanned on a molecular dynamics flatbed scanner. Data sets were carefully screened to ensure repeatability across this broad range of detector systems and samples.

4.4 Data analysis

Data sets were merged to give up to 141 meridional intensities in each native protein or derivative set. Of these, 124 reflections common to all data sets were used in the phase determination. Phases were estimated for each axial reflection and used to produce electron density profiles that were calculated for the native and three derivative structures, by Fourier inversion of the structure factors (Bradshaw *et al.*, 1989). The difference Fourier of the native and derivative fibril structure produced a refined model of the labelling positions of the respective heavy atoms. The refined labelling positions were used to generate new heavy atom structure functions and the process was repeated, until the labelling positions stabilised over a number of cycles (approximately 25).

4.4.1 Phase calculations

The two iodine derivatives described above and gold chloride were used in the determination of the native and derivative phases. The size of the first order from the native protein was estimated from an initial model structure that was Fourier converted (based on the projection of the axial structure proposed by Wess *et al.*, (1998). The remaining observed native amplitudes were scaled by the same scale factor. The derivative intensities were also converted to amplitudes and then put onto the same scale as the native and heavy atom amplitude components according to the relationship:

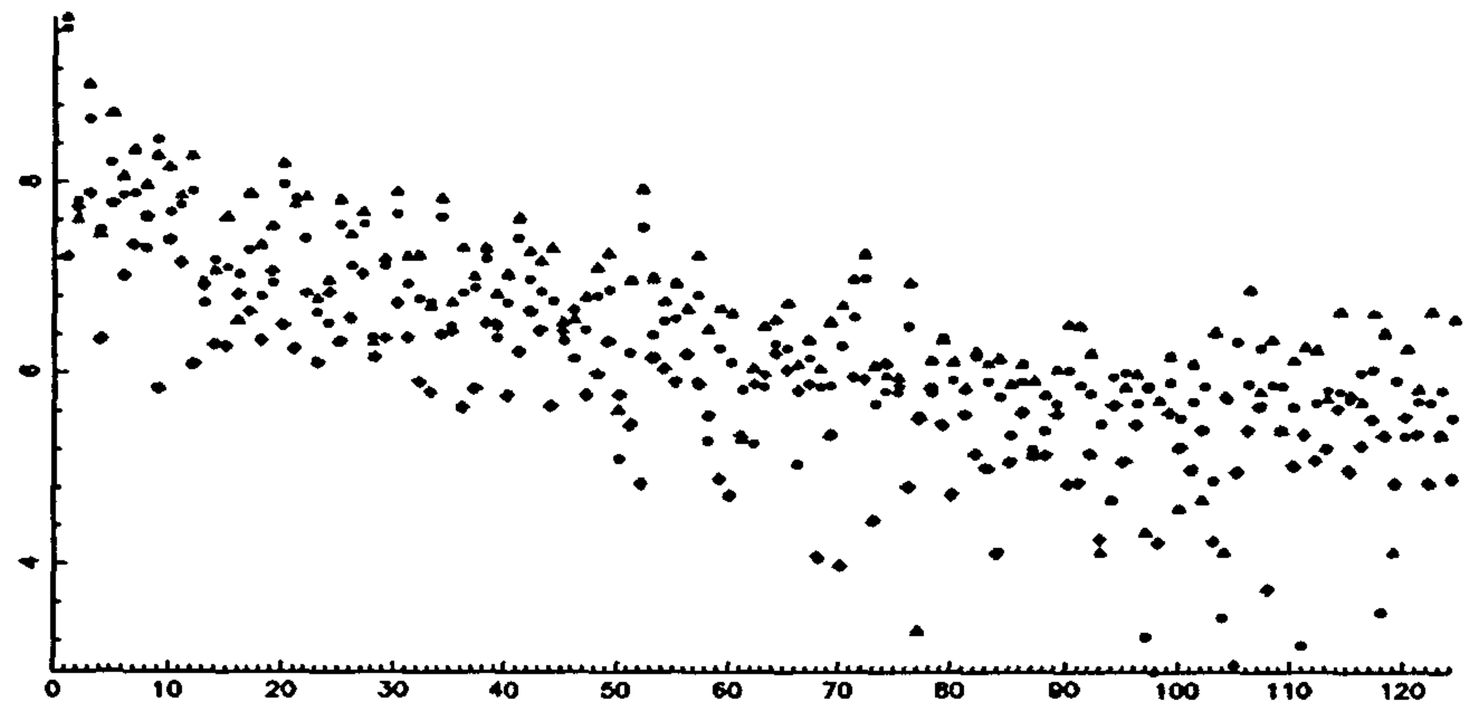
$$K^2_{PH} \cdot \Sigma F^2_{PH} = \Sigma F^2_{P} + \Sigma F^2_{H}$$

Where K_{PH} is the derivative scaling factor (Blundell and Johnson 1976).

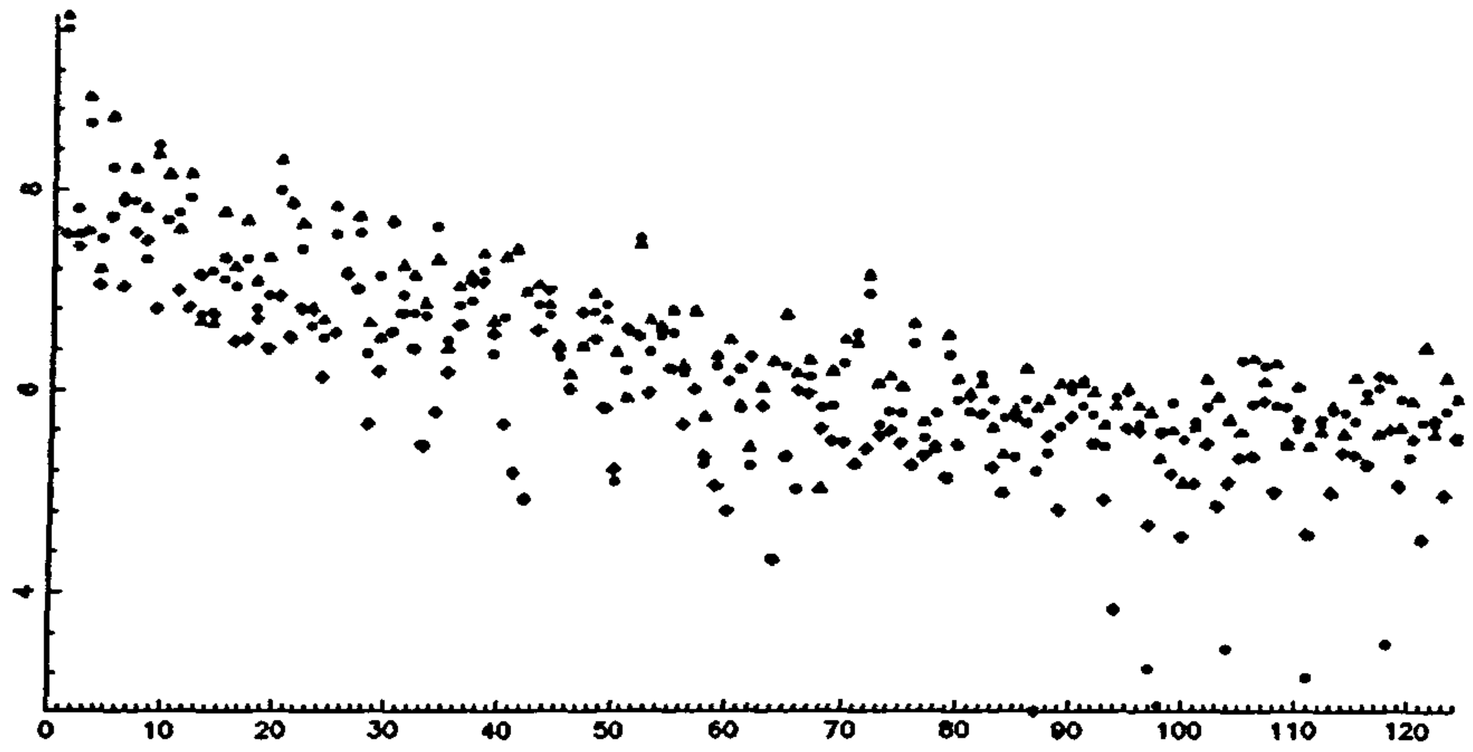
This factor holds true over a large range of amplitudes where the distribution of positive and negative phases is roughly equal, but is not necessarily true at low angles. The scale factor K_{PH} was adjusted in some cycles of the refinement to improve the phase agreement between the derivatives (see Figure 4.2 for graphs of scaled intensities).

Phases were calculated for the native and three isomorphous derivatives on the basis of obtaining one unambiguous phase for each diffraction order. Where the phases were in poor agreement, greater weight was given to two derivatives in better agreement over the remaining one. Electron density profiles were calculated for the native and three derivative structures by Fourier inversion of the structure factors.

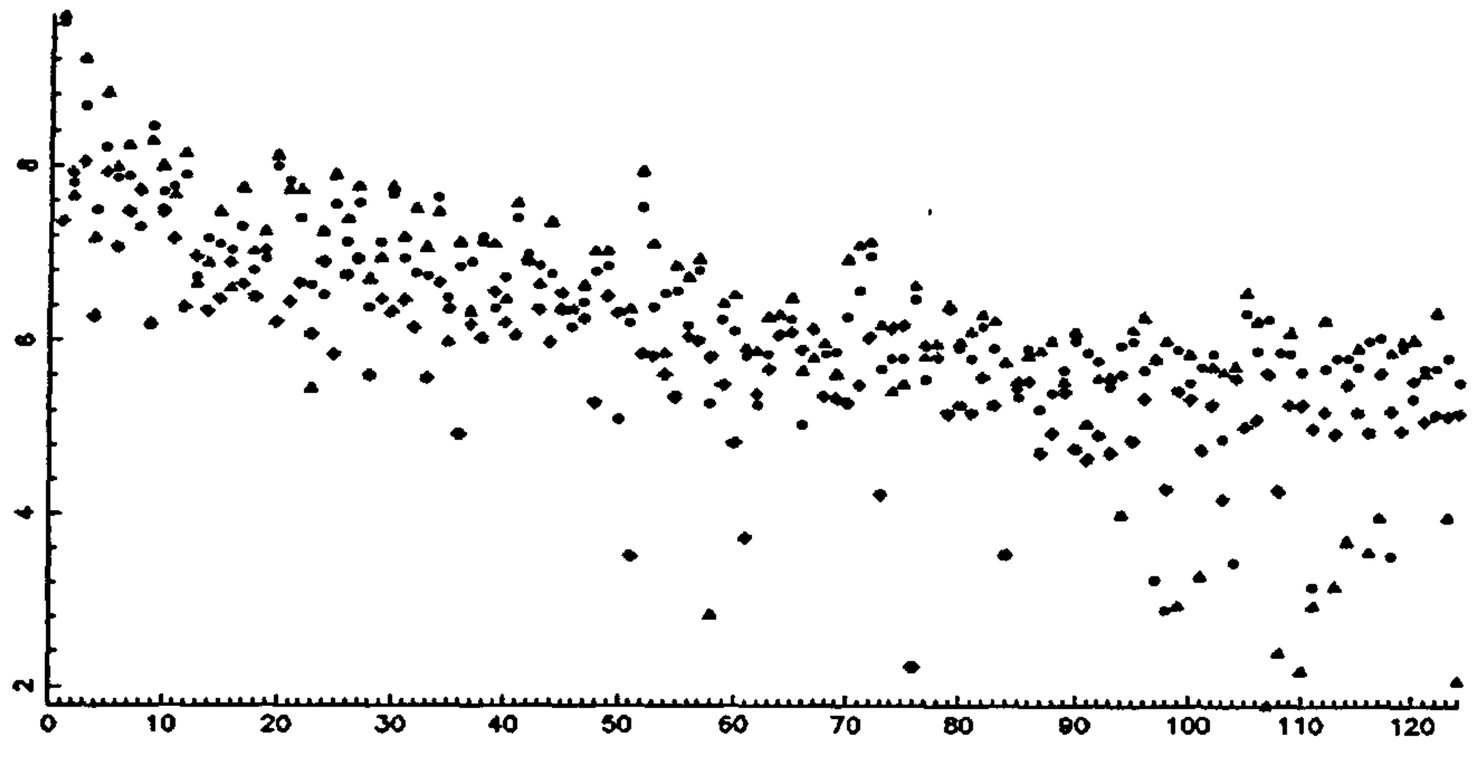
The difference Fourier of the native and derivative fibril structure produced a better-refined model of the labelling positions of the respective heavy atoms. The refined labelling positions were used to generate new heavy atom structure functions, and the process was repeated until the labelling positions stabilised over a number of cycles (approximately 25).



a



b



c

intensities

- native
- derivative
- heavy atom

Figure 4.2 Semi–log plot of intensity vs. order number (1–124)

Log of derivative intensity is plotted with the native and calculated heavy atom structure factors in each of the graphs above.

a) Unmodified iodide derivative; b) gold chloride derivative; c) UV irradiated (pre–stain treatment) iodide derivative.

4.5 Results

4.5.1 X-ray diffraction pattern of type I collagen fibrils in tendon

The X-ray diffraction pattern of rat tail tendon contains a strong meridional series of reflections corresponding to axial molecular organisation; this can be seen in Figure 4.3. This reveals the extent of discrete meridional reflections that can be observed for both native and heavy atom derivative samples. Low and high-angle diffraction patterns were recorded for the native and three derivative proteins (Figures 4.4 to 4.7). The meridional Bragg reflections were integrated to generate a one-dimensional meridional profile from which the individual peak intensities could be fitted. A portion of the meridional intensity profile corresponding to orders 79-124 is shown in Figure 4.8

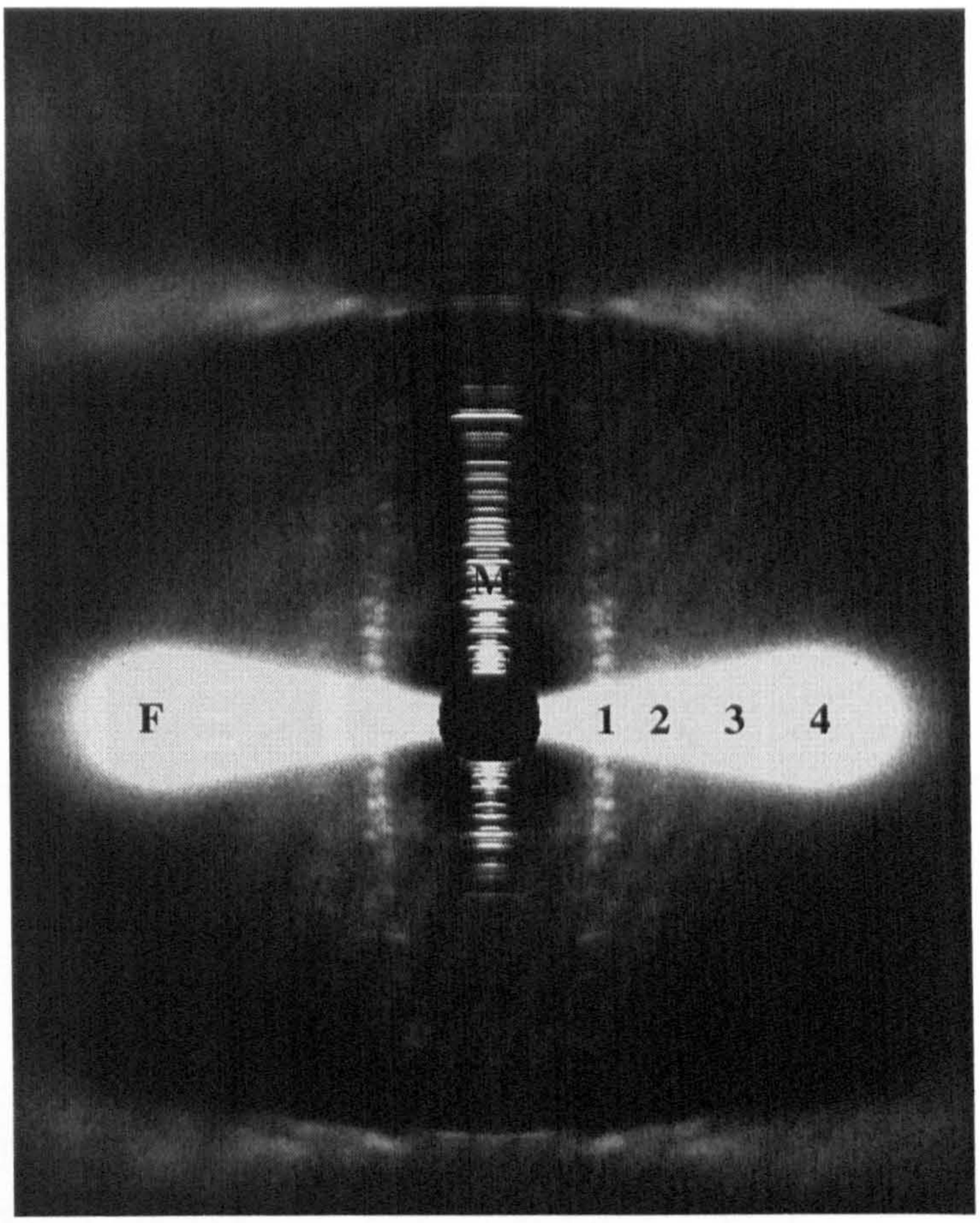


Figure 4.3 Collagen fibre diagram

X–ray diffraction pattern of collagen fibrils in tendon. The X–ray diffraction pattern of rat tail tendon contains a series of intense meridional reflections corresponding to axial molecular organisation and discrete though much less intense equatorial Bragg peaks on row–lines roughly parallel to the meridian that are due to the lateral packing arrangement of the collagen molecules. These equatorial reflections are superimposed upon a diffuse fan of background scatter that lies perpendicular to the meridan. These features are marked as follows:

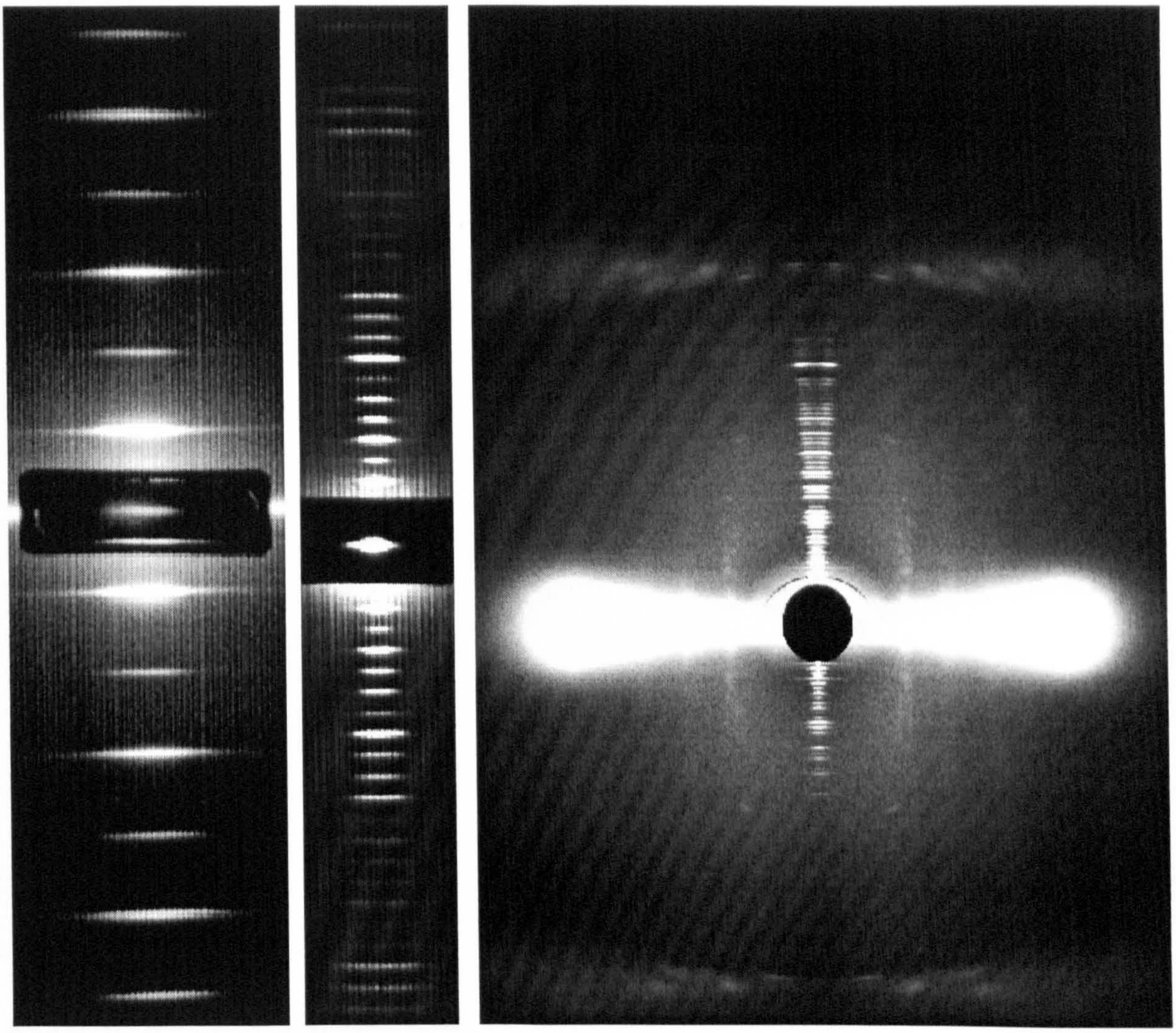
Meridional series = M

Equatorial row–lines, four groups of two or more overlapping rowlines that occur at approximately:

1=3.9 nm, 2=2.7 nm, 3=1.8 nm, 4=1.3 nm

Equatorial fan of diffuse scatter = F

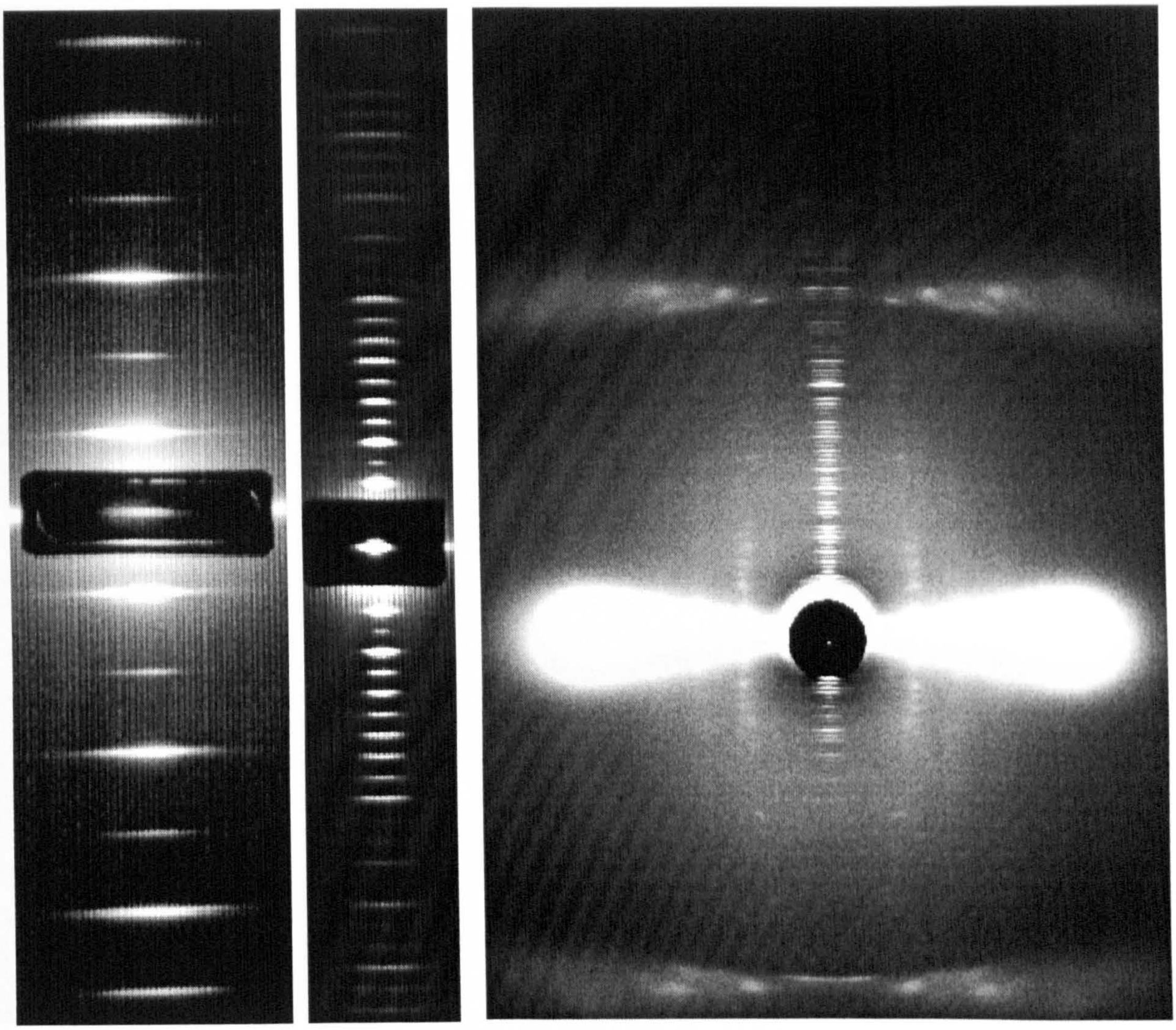
The m=0, n=–1 layer line of the molecular helix is marked with an arrow head.



a b c

Figure 4.4 Low, medium, and high–angle diffraction patterns of native rat tail tendon

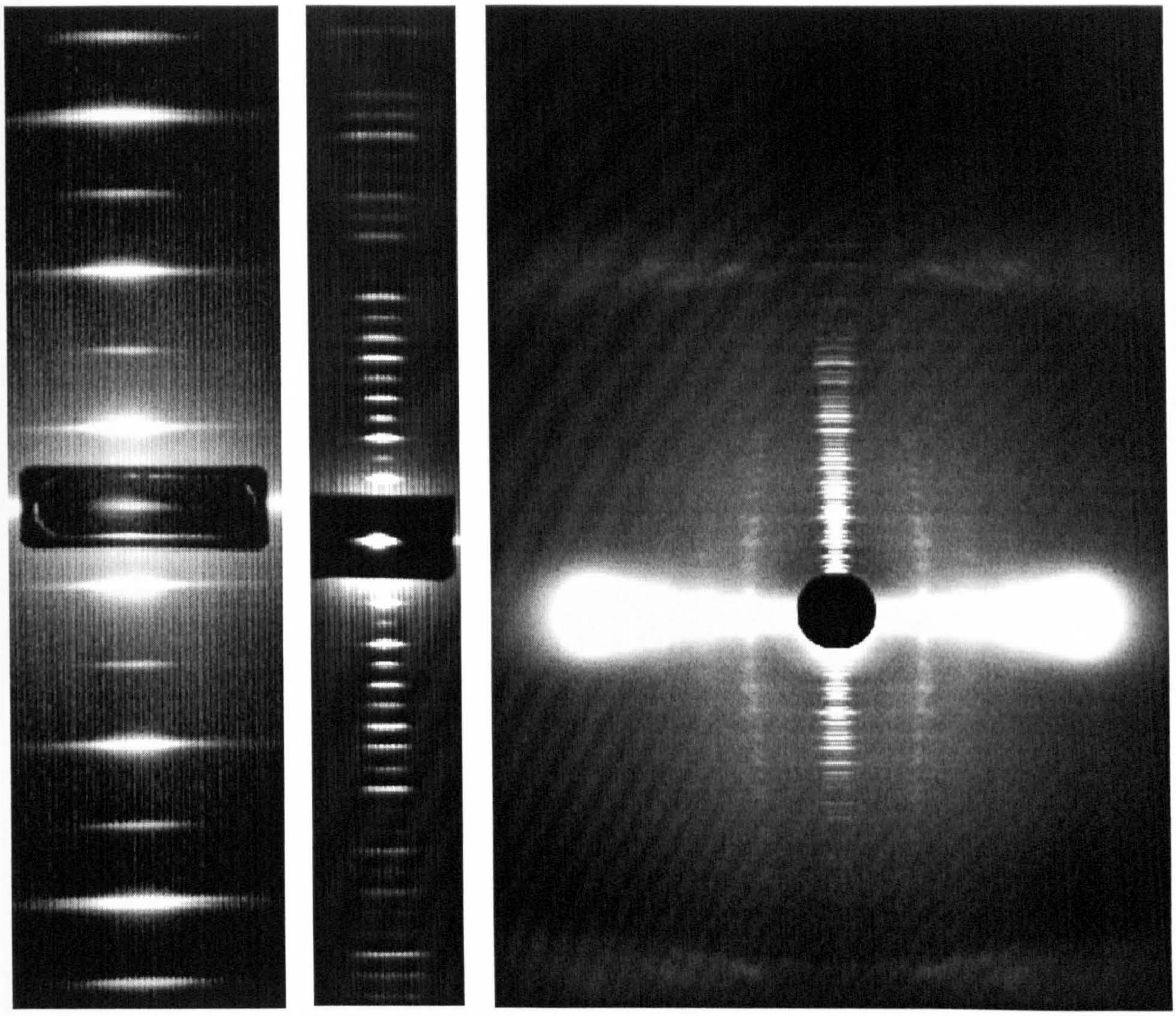
Low (a) and medium (b) angle diffraction patterns collected at SRS 2.1, whilst the high–angle diffraction pattern was recorded at SRS 7.2 (c).



a b c

Figure 4.5 Low, medium, and high–angle diffraction patterns of iodine stained rat tail tendon

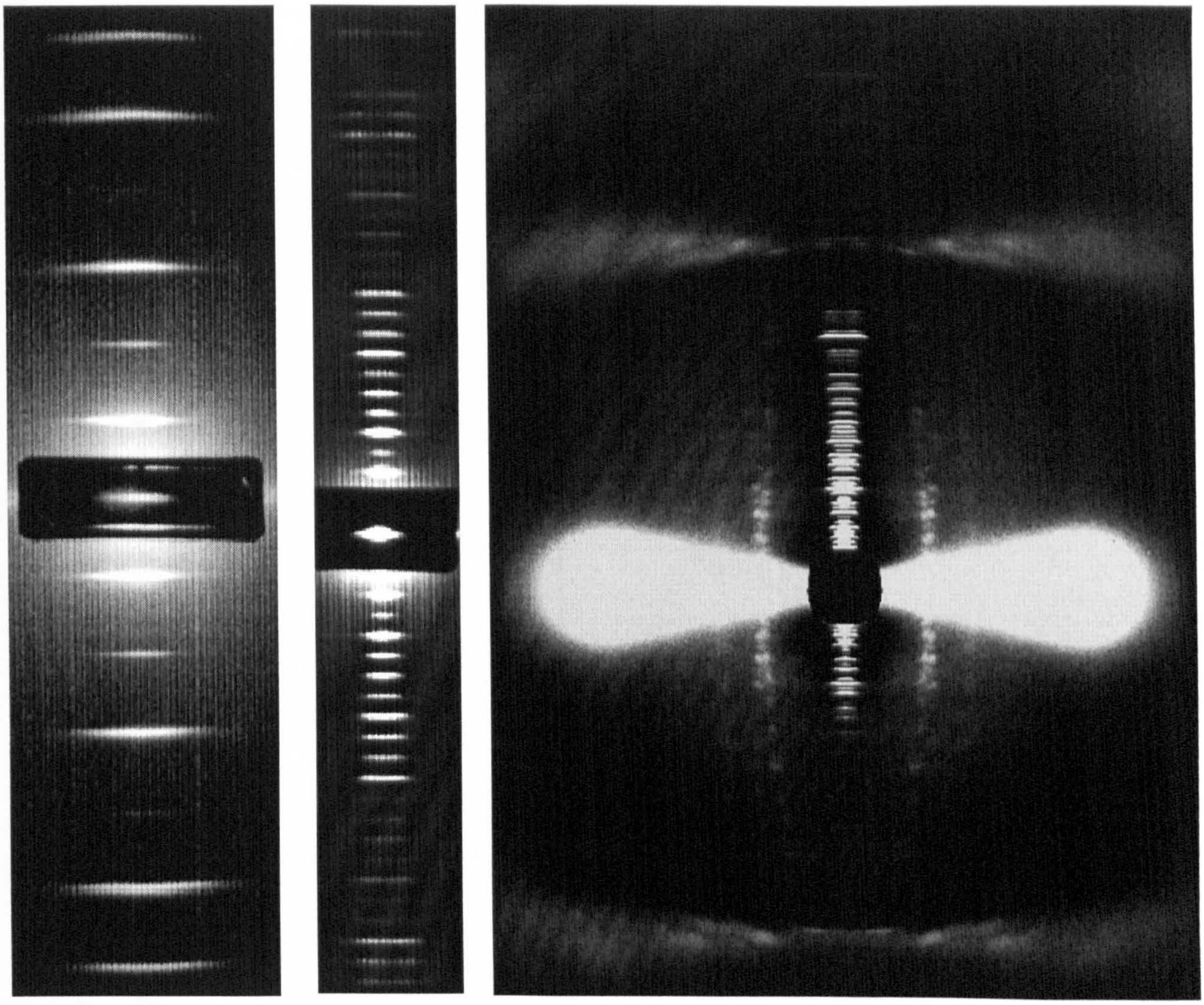
Low (a) and medium (b) angle diffraction patterns collected at SRS 2.1, whilst the high–angle diffraction pattern was recorded at SRS 7.2 (c).



a b c

Figure 4.6 Low, medium, and high–angle diffraction patterns of gold chloride stained rat tail tendon

Low (a) and medium (b) angle diffraction patterns collected at SRS 2.1, whilst the high–angle diffraction pattern was recorded at SRS 7.2 (c).



a b c

Figure 4.7 Low, medium, and high–angle diffraction patterns of UV irradiated, iodine stained rat tail tendon

Low (a) and medium (b) angle diffraction patterns collected at SRS 2.1, whilst the high–angle diffraction pattern was recorded at SRS 7.2 (c).

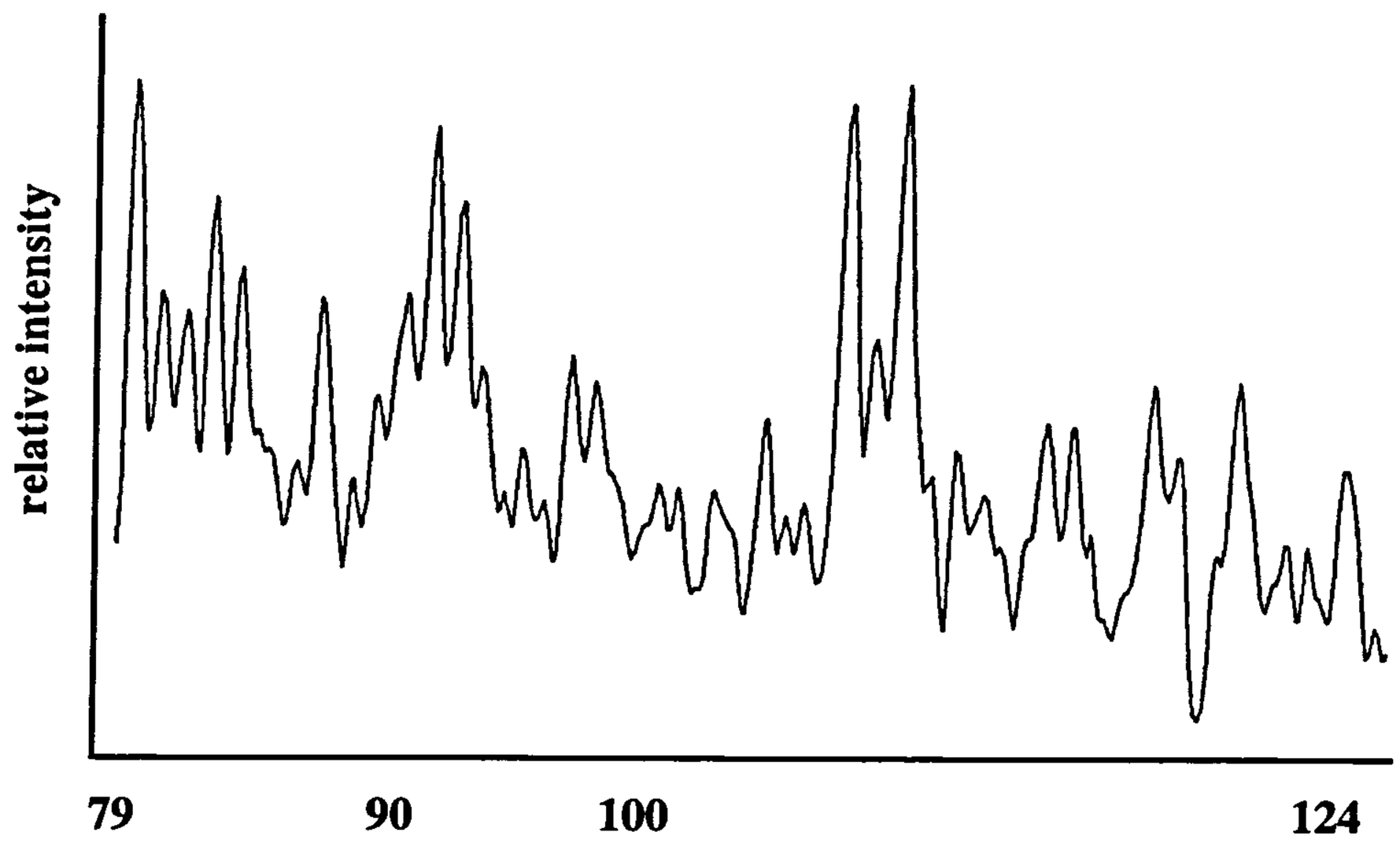


Figure 4.8 Meridional trace of high–angle native intensities

Orders 79–124 are shown here after background subtraction (see Chapter 3).

4.5.2 Interpretation of the heavy atom binding sites

The heavy atom distributions for the derivatives used are shown in the difference maps (Figure 4.9). Methionine, histidine and tyrosine were the principal residues labelled. The gold chloride derivative difference Fourier shows that it exhibits preferential labelling at selected sites (histidine and several methionine residues) in the triple helical region. The relative axial locations of the tyrosine residues are shown by the iodination peaks in the difference Fourier. For these profiles, the best phase agreement between the derivatives was obtained with mono- rather then di-iodide substitution of tyrosine residues. Ultraviolet irradiation was also used to modify the labelling pattern of tyrosine. The attachment of iodine at a selected number of histidine residue sites within the overlap region is also apparent. Both the C and N termini $\alpha 1$ and $\alpha 2$ telopeptides contain tyrosine. However no methionine or histidine residues are found within the N-terminal telopeptide. The location of the iodinated tyrosine peaks in the difference Fourier is of particular relevance to the determination of the C-telopeptide structure, since the tyrosine residues are located at both ends of the telopeptide sequence. The UV pre-treated iodinated difference Fourier contains significant differences in some of the labelling sites indicating the susceptibility of some tyrosine residues to photo-damage.

In the case of the N-terminal telopeptide, the two $\alpha 1$ chains contain two tyrosine residues located at positions 4 and 6. At the N-terminal telopeptide, the position of the tyrosine residues is more difficult to measure due to overlap with the axially projected

position of histidine residues in the helical region of one of the other collagen segments in the D period (where D = 67nm).

The labelling positions of the heavy atoms provide a means of determining the most likely conformation of the C-telopeptide due to the location of the tyrosine residues by iodination within the telopeptide sequence (Figure 4.10). The α1 chains (25 residues in the C-telopeptide) contain one proximal tyrosine residue at position 4 in the telopeptide, and two distal at the end of the telopeptide region (positions 24 and 25), in addition to one histidine residue at position 19. The much shorter α2 C terminal telopeptide chain (9 residues) has one tyrosine at position 4.

The difference Fourier map for the iodine derivative shows two sets of peaks in the C-terminal telopeptide region. The distance between the large peak and the first of the series of 4 peaks in this region (Figure 4.10) is approximately 0.027D (1.8 nm). The two groups of iodine-labelled tyrosine residues are separated by 20 amino acids in the C-terminal telopeptide sequence. The telopeptide sequences do not contain the Gly-Pro-Pro tripeptide motif required for collagen triplex formation. If the C-telopeptide adopted an axially contracted conformation the average axial translation per residue would have to be approximately 0.09 nm. In a hairpin folded conformation, the major iodination peak in Figure 4.10 corresponds to the proximal and distal tyrosine residues (positions 4, 24 and 25) that have been brought into close proximity by a sharp change in the polypeptide direction, possibly between residues 13-14. Caution is needed in assigning exact estimates of distances on the basis of the histidine labelling peaks. The C-terminal

histidine residues at position 19 of the α1 chains are in close axial projection position to three histidine residues within the helical region of adjacent D periodic segments.

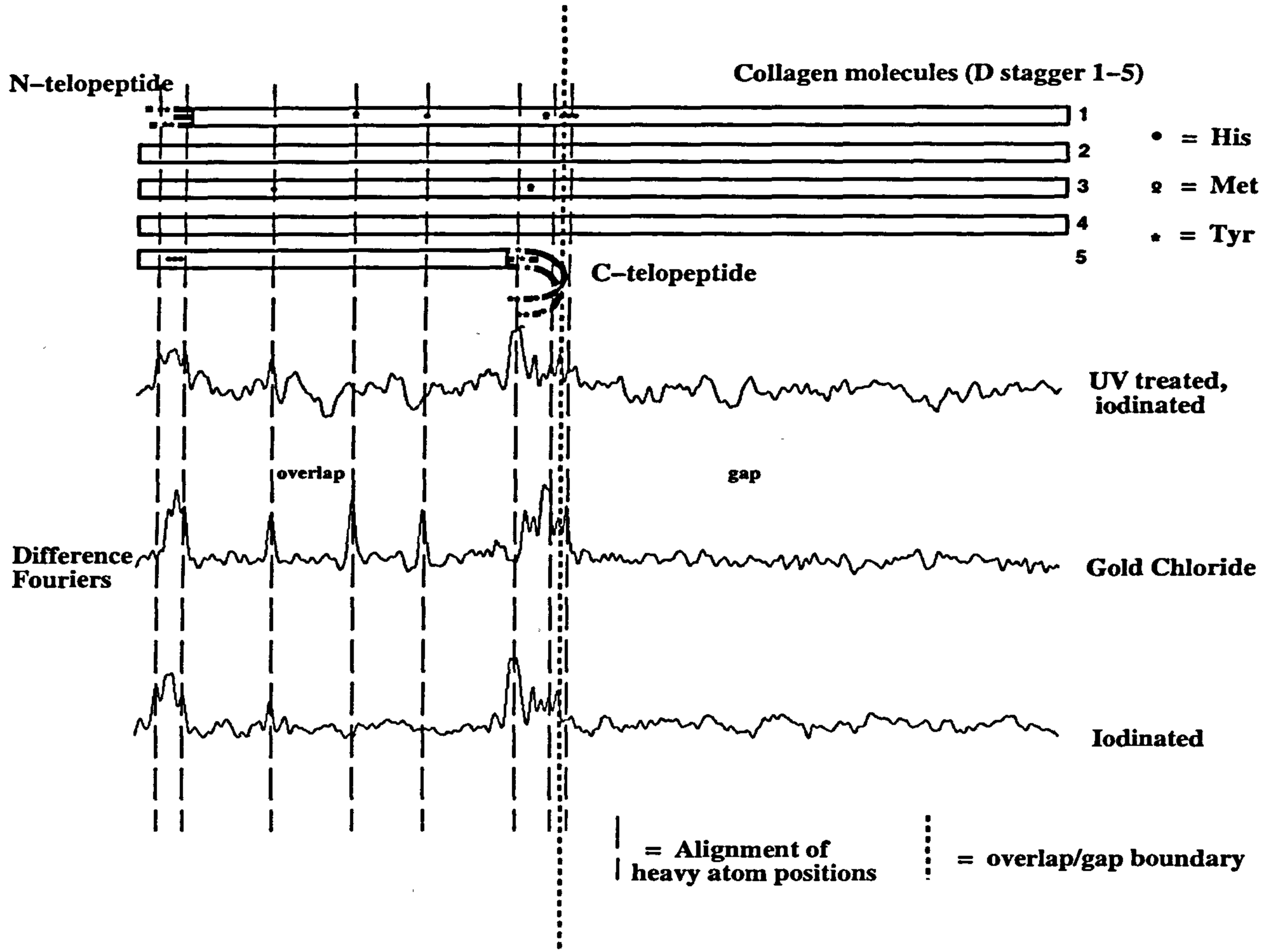


Figure 4.9 Electron density difference maps of one D–period

The axial D periodic function is shown schematically (top), with three electron density difference maps shown below. The putative residues involved in binding heavy atoms are indicated. Vertical lines have been drawn for alignment purposes between selected residues and peaks.

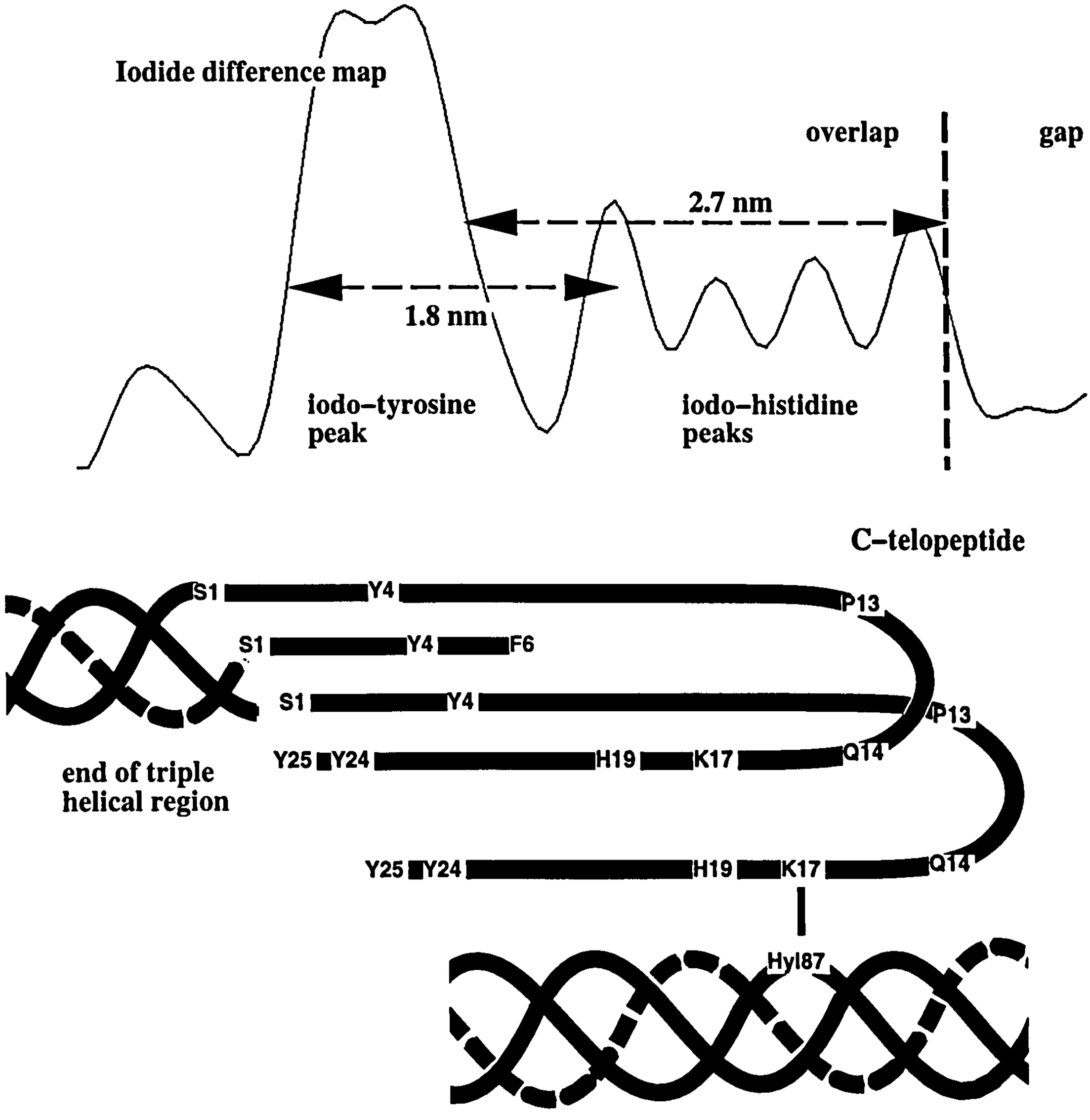


Figure 4.10 Conformation of the C–telopeptide restricted by heavy–atom positions

An insert of the iodide difference Fourier map (above) shows the folded C–telopeptide conformation fits the difference density data well. The tight turn of the peptide at Pro 13/Gln 14 would bring the Tyr 24 residues of the $\alpha 1$ chains into rough axial alignment with the Tyr 4 residues of the same chains. This also brings Lys 17 into a favourable position to form the lysine–hydroxylysine crosslink at Hyl 87 (shown) and help stabilise the microfibrillar structure.

4.5.3 The significance of the native density profile

The native electron density profile is shown in Figure 4.11. It can be used to estimate parameters such as the (axial length) ratio between the gap and overlap as well as giving some information about the telopeptide structure. The length of the overlap (with telopeptides) was measured to be 0.46 D (the gap region therefore being 0.54 D), the limits being judged by the maximum observed extent of the telopeptides. Since this study has a maximum resolution of 0.54 nm, this suggests an estimated error of +/- 0.008 D. This compares with previous gap/overlap estimations; 0.52:0.48 (Bradshaw *et al.*, 1989), 0.52:0.48 (Hulmes *et al.*, 1980), 0.6:0.4 (Hodge and Petruska 1963), 0.54:0.46 (Ericson and Tomlin 1959), 0.53:0.47 (Tomlin and Worthington 1956).

Interpretation of the native axially projected electron density profile can also be aided by an attempt to build a model that matches the electron density profile. The parameters that were used to test a fit were:

1) The number of amino acids in each D periodic segment,

2) The C-terminal telopeptide conformation and

3) N-terminal telopeptide conformation.

The maximum agreement between the electron density profiles determined in this study, and that of the real space models was reached when each D periodic segment contained 234.2 amino acids; this value is exactly the same as found by Meek *et al.*, (1979). The amino acid sequence (Chapman and Hulmes 1984), and electron density (corrected for water occupancy) were treated in an analogous manner to Hulmes *et al.*,

(1977). The C-terminal telopeptide was built in a folded arrangement with a tight turn at residues proline 13 and glutamine 14 and the N-terminal telopeptides contracted to 85% of the nominal axial amino acid translation value for the helical regions.

As can be seen from Figure 4.11, the solution for the native electron density profile contains many common features with the model that has been made solely from amino acid scattering functions. Differences in the density profiles may be due to small local variations in the helical repeat distance. The native electron density profile shows two conspicuous peaks in the region of the C-telopeptide. Unless the C-telopeptide residues are modelled as being in a tight turn, with the tyrosine residues of both ends of the telopeptide in axial alignment, the corresponding two peaks are not observed in the electron density model. It is significant to note that both the solutions of Hulmes *et al.*, (1977), and Bradshaw *et al.*, (1989), also predicted these two peaks in the C-terminus region. For the N-telopeptide, the best fit in the model structure to that of the observed electron density is where the N-telopeptides are contracted to an overall telopeptide length of approximately 85% compared to that of a triple helical region containing the same number of residues.

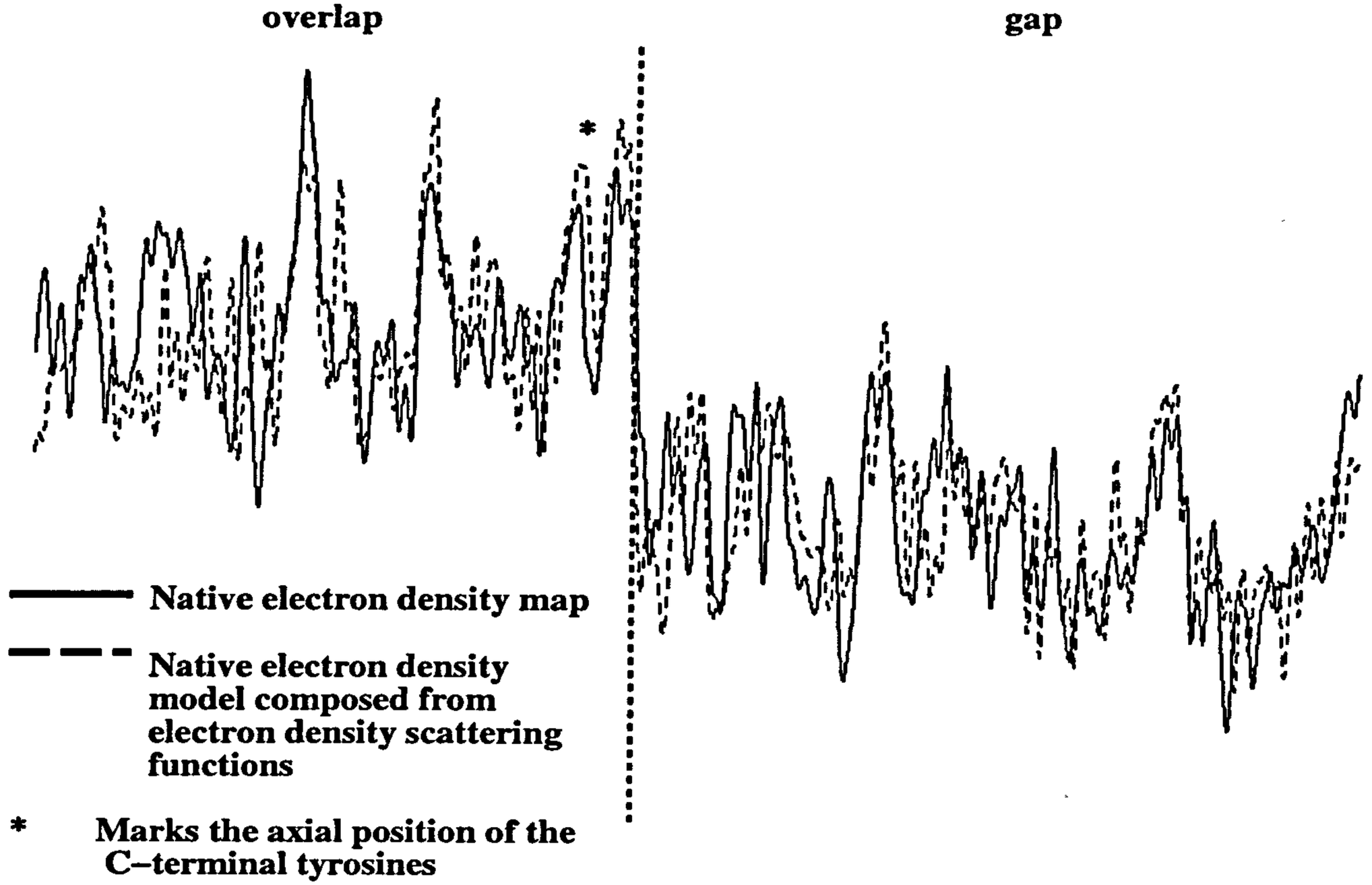


Figure 4.11 The native electron density map and a model based on sequence data

The characteristic gap/overlap step function is seen within the D–repeat axial unit cell. The length of the overlap (with telopeptides) was measured to be 0.46 D (the gap therefore being 0.54 D), the limits being judged by the maximum observed extent of the telopeptides. The model density shows common features with the electron density profile. The difference between the two may be due to local variations in the axial rise per residue along the length of the molecule.

4.5.3.1 Implications for the nature of the molecular helix pitch

The goal of any crystallographic study is the production of a biochemically sound structural interpretation of the observed data. Ultimately, this would be in the form of a electron density map that is used to build a three dimensional molecular model. An analogous procedure has been presented here, where the structure of the collagen in tendon in projection has been determined, and a model of five D-staggered molecules in projection has been fitted to the electron density map. The information that can be directly determined with some confidence from the electron density map and fitted model are discussed in the previous section.

There are however several notable discrepancies between the electron density map, and the model constructed from the amino acid sequence and residue electron densities. An obvious explanation for these differences could be that several phases used to calculate the electron density map are incorrect, leading to the invalid distribution of some of the structure function terms. A more optimistic interpretation that has been confirmed by recent published work (Kramer *et al.,* 1999), is that one of the underlying assumptions of the model is incorrect, that is, that the helix pitch and residue spacing (unit height of the molecule) is constant along the length of the collagen triplex.

The observation that the fit of the model could be significantly improved by altering the relative spacing of residues (increasing or decreasing the unit height) along the collagen chains by up to, but no more than 5% in certain regions has already been commented upon (Orgel - CCP13 Fibre diffraction meeting, May 1999). This change

in residue spacing could correspond to the pitch and therefore unit height difference between a 7/2 and 10/3 collagen triple helix (the differences in unit height between these conformations being approximately 5%).

Kramer *et al.*, (1999) observed in their crystal structure of the collagen-like peptide that regions containing imino rich sequences adopted a 7/2 conformation, whilst the peptide sequence lacking imino acid residues adopted a 10/3 like conformation. The 7/2 helix conformation had previously been observed in model peptide studies (Okuyama *et al.*, 1981, 1999), although the implication that the 7/2 or 10/3 conformations may be due to the concentration of imino acids within the helical region had not been noted previously.

In Figure 4.12, a number of small alterations have been made to the relative spacing of residues in small, defined regions. The increased or diminished projected electron density is seen to significantly improve the fit of the model to the electron density map. It is noteworthy that the regions where the residue spacing was adjusted due to poor local model to map fit, roughly correspond to regions in the sequence where one or more of the five collagen chains is relatively imino acid poor (see Appendix 1).

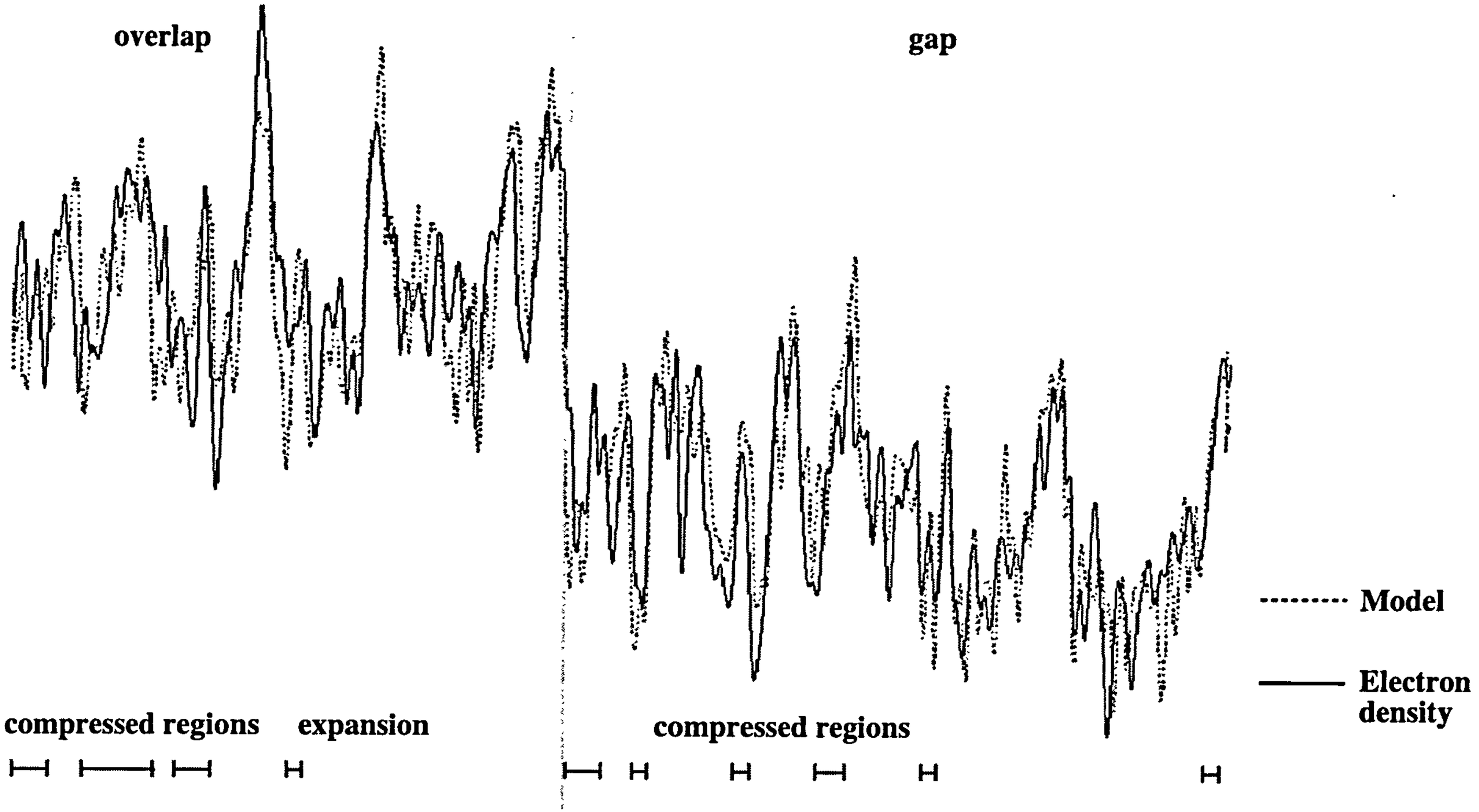


Figure 4.12 The native electron density map and modified model based on sequence data and local variations in residue spacing

4.6 Discussion

The amount of lateral space available in the gap region for peptide folding is limited to the molecular diameter of the 'missing' molecular segment that constitutes the gap region. This places large constraints on the possible molecular conformations of the telopeptides. Previous studies of telopeptide conformation involved model based approaches (Helseth and Veis 1981) or investigations of single isolated synthetic or extracted telopeptide chains (Otter *et al.*, 1988). Analysis of a single synthetic α1 chain C-telopeptide by NMR spectroscopy demonstrated that the isolated chain was axially condensed, with a possible disposition to folding. A folded C-telopeptide conformation was proposed to account for the spectroscopic data obtained from a single α1 chain C-telopeptide isolated from calf skin. Such studies however pay little or no attention to the possible *in situ* interactions of the telopeptides with each other, or with the six surrounding collagen molecules that define the boundary of the gap region. Some modelling studies have examined the molecular packing within a hexagonal lattice (Jones and Miller 1987, Jones and Miller 1991). However the protein prediction algorithms used did not consider interactions with the amino acid side chains that are present *in situ* from the surrounding helices that may interact with the telopeptide region.

X-ray diffraction studies have shown that localised amino-acid sequence and conformation changes within the telopeptide regions significantly affected the intensity distribution of the low-angle meridional orders. Hulmes *et al.*, (1977), reported axial translation values for amino acid residues of 0.282 nm for the collagen triplex, 0.241

nm in the N-terminal telopeptide and 0.2 nm in the C-terminal telopeptide, assuming that the telopeptides were contracted rather then folded. Bradshaw *et al.*, (1989) demonstrated the axial contraction of the telopeptide regions, but the study was not of sufficiently high resolution to determine whether the axial contraction of N and C telopeptides is due to the distances between residues being reduced, or folding of the peptides.

The electron density profiles shown here are interpretable in terms of the reagents used and the available sequence data. The importance of the results is that the conformation of the telopeptide regions has been established *in situ*. Whilst the exact conformation of the N-terminal telopeptide remains unclear, strong evidence has been presented for a tight turn located at proline 13 and glutamine 14 in the α1 C terminal telopeptide and the location of the important intermolecular crosslinks have been determined. The main chain and telopeptide amino acid residues where implicated in the crosslinking scheme by their axial alignment and by their identification as the cross-linking sites in isolated crosslinked telopeptide and main chain fragments (Nakamura 1987), these being; N-terminal Lys 9 and Hyl 927, and C-terminal Lys 17 and Hyl 87 (see section 1.7.3.1).

Publication

The recent publication of Orgel *et al.*, (2000) is the result of the work undertaken and reported within this chapter.

Chapter 5

The three-dimensional molecular packing structure of fibrillar type I collagen

5.1 Introduction

The nature of the molecular packing arrangement of collagen molecules within fibrils has been a matter of speculation and of investigation for some time. Data that appear to be conflicting or even contradictory have been used to explain the molecular packing within a collagen fibril. The major problem has been in accommodating structural evidence that indicates collagen molecular packing possessing combined crystalline and liquid-like properties. This is compounded by biochemical evidence that points toward specific molecular interactions being dictated by crosslinking patterns. The local molecular interactions that occur within regions of a collagen fibril also have to be reconciled with an overall fibril structure that may require specific molecular interactions and organisation. These different aspects of collagen molecular organisation within a fibril are described in the following sections.

5.1.1 Evidence and implications for crystallinity

The low-angle diffraction pattern from rat tail tendon contains a series of very intense meridional reflections (Bear 1944) and discrete though much less intense equatorial Bragg peaks on row-lines parallel to the meridian (North 1954) see Figure 4.3. These equatorial reflections are superimposed upon a diffuse fan of background scatter that lies perpendicular to the meridian.

These features of the fibre diagram demonstrate that the degree of lateral order (packing arrangement of the collagen molecules) in rat tail tendon and other type I

collagen containing tissues (Jesior *et al.*, 1980) must be greater than just that arising from nearest neighbour contacts (Hukins and Woodhead-Galloway 1977). It is the molecular packing that forms the diffraction grating that gives rise to the coherent scatter seen in the equatorial orientation.

Several other fibrillar tissues have been shown to possess crystallinity similar to that of rat tail tendon, both for type I collagen (Jesior *et al.*, 1980), and type II (Eikenberry *et al.*, 1984).

5.1.2 Evidence and implications for liquid crystal arrangements

The presence of the diffuse equatorial fan in the diffraction pattern of these crystallite-containing tissues indicates a degree of liquid-like short-range order in the lateral packing arrangement (Woodhead-Galloway and Machin 1976). This diffuse equatorial scatter is also present in the fibre diagram of other fibrillar tissues where the coherent scatter is absent (such as in bone or skin). In both instances (in tissues that show equatorial crystallinity and those that do not), the position of the maximum intensity of the diffuse scatter is proportional to the intermolecular interference function (Woodhead-Galloway and Machin 1976).

Additional evidence of the liquid-like qualities of the packing arrangement within fibrillar tissues, has been obtained from NMR studies. The NMR data show that there is considerable azimuthal mobility of the molecules within fibrils (Jelinski *et al.*, 1980, Torchia 1982). These studies show that the collagen molecules in fibrils lacking

covalent crosslinking (lathrytic fibrils) twist around their long axis at rates comparable to free molecules in solution, whilst the degree of mobility is decreased in native crosslink containing fibrils (Sarkar *et al.*, 1983).

Implications of liquid-crystallinity are found in the distribution of amino and imino acids between the overlap and gap regions. Fraser *et al.*, (1983) observed that the frequency of Gly-Pro-Pro and Gly-Pro-Hyp triplets was approximately 14% of all triplets in the overlap region, and 8% in the gap region. The reduced packing density of molecules in the gap (4 compared with 5 in the overlap), led Fraser *et al.*, to attribute the continuous scatter in the vicinity of the m=0 n=0 and m=0, n=-1 layer lines of the helix to the gap region containing mobile segments.

Some fibrillar tissues may lack long-range order in their packing arrangements, but they do possess long-range order in the axial (fibre) direction, likening them to that of a smectic A liquid-crystal (Hukins and Woodhead-Galloway 1977). However, the most obvious distinction between those fibrillar tissues that display lateral crystallinity and those that do not, is the collagen typical composition. The former group (those showing lateral crystallinity) are primarily homotypic such as in rat tail tendon (Miller 1976), and the latter group (those not showing lateral crystallinity) are composed of a heterotypic mix of fibrillar collagen types (Kadler 1995). However, the absence of X-ray diffraction evidence for a lateral super-lattice does not necessarily rule out the possibility of some regularity for collagen packing within fibrils, such as the helicoidal arrangement in packing along the fibre axes, revealed by freeze fracture electron microscopy (Ruggeri *et al.*, 1979, Raspanti *et al.*, 1989, Katsura *et al.*, 1991).

5.1.3 Structural information implicated by previous model based studies

Several key model studies have attempted to elucidate the three-dimensional packing structure of collagen and provided partial explanations that are consistent with the observed data. Apart from the inherent ambiguity of model studies, that they cannot be depended upon to necessarily produce one 'correct' solution, few of these studies have attempted to explain the duality of the packing structure - crystalline arrays of molecules organised within a liquid-like system, although they have been significant in defining the general character of the crystalline packing system and the unit cell. This section will now examine the results of these key studies.

5.1.4 The compatibility between order and disorder in collagen fibrils

Indirect evidence for the presence of disorder in a collagen fibril was given in Fraser *et al.*, (1987) where the contribution to diffraction from the gap region was judged to be diminished due to the lower packing density. Additionally, as speculated upon by Fraser *et al.*, (1983), the contribution to diffuse X-ray scattering is implicit in the arrangement of molecules within the gap region.

Taking this into account, Hulmes *et al.*, (1995) attempted to explain the molecular packing in collagen fibrils by conducting a model based study that incorporated both the long-range (crystalline) and short-range (liquid-like) order of the three dimensional molecular packing.

Comparing two-dimensional Fourier calculations of various packing models based on concentric ring and spiral (with radial nearest neighbour contacts) systems with the observed equatorial data, they found that concentric packing models based on the work of Silver *et al.*, (1992) and Galloway (1985) produced poor correlation with the observed data, whilst an energy minimised cylindrical structure with radial molecular packing produced good agreement. This was particularly relevant since in the observations of Hulmes *et al.*, (1981,1985) transverse sections of collagen fibrils prepared for electron microscopy showed the largest lattice spacing had a preference for radial orientation.

Implicitly important to the approach of Hulmes *et al.*, (1995) was the use of models that were composites of crystalline ordered and liquid-like disordered regions. The model fit was further improved when disorder (liquid-like order) was introduced into the overlap region in addition to that of the gap region. To quantify their approach, Hulmes *et al.*, (1995) identified four distinct regions within the unit cell (shown in Figure 5.1):

1. Disordered partial gap region.
2. Disordered partial overlap region.
3. Relatively ordered N-terminal gap/overlap interface region.
4. Relatively ordered C-terminal gap/overlap interface region.

The optimal fit was found when 60% of the overlap and 80% of the gap regions were organised according to liquid-like packing, the significance being that the gap

region whilst 'missing' a molecular segment is likely to be somewhat more disordered than the overlap region as a result of relatively low packing density. The greater concentration of imino acids within the overlap region, combined with a higher packing density (5 molecular segments rather then 4) would be expected to contribute to greater stability/order within the overlap region of the fibril.

In order to advance the understanding of collagen molecular packing, the potential compatibility between crystallinity and static disorder must be more fully understood. The framework of molecular packing as a discrete unit cell and the molecular topology within the unit cell must be discussed before a full appreciation of all components concerned in molecular packing can be obtained.

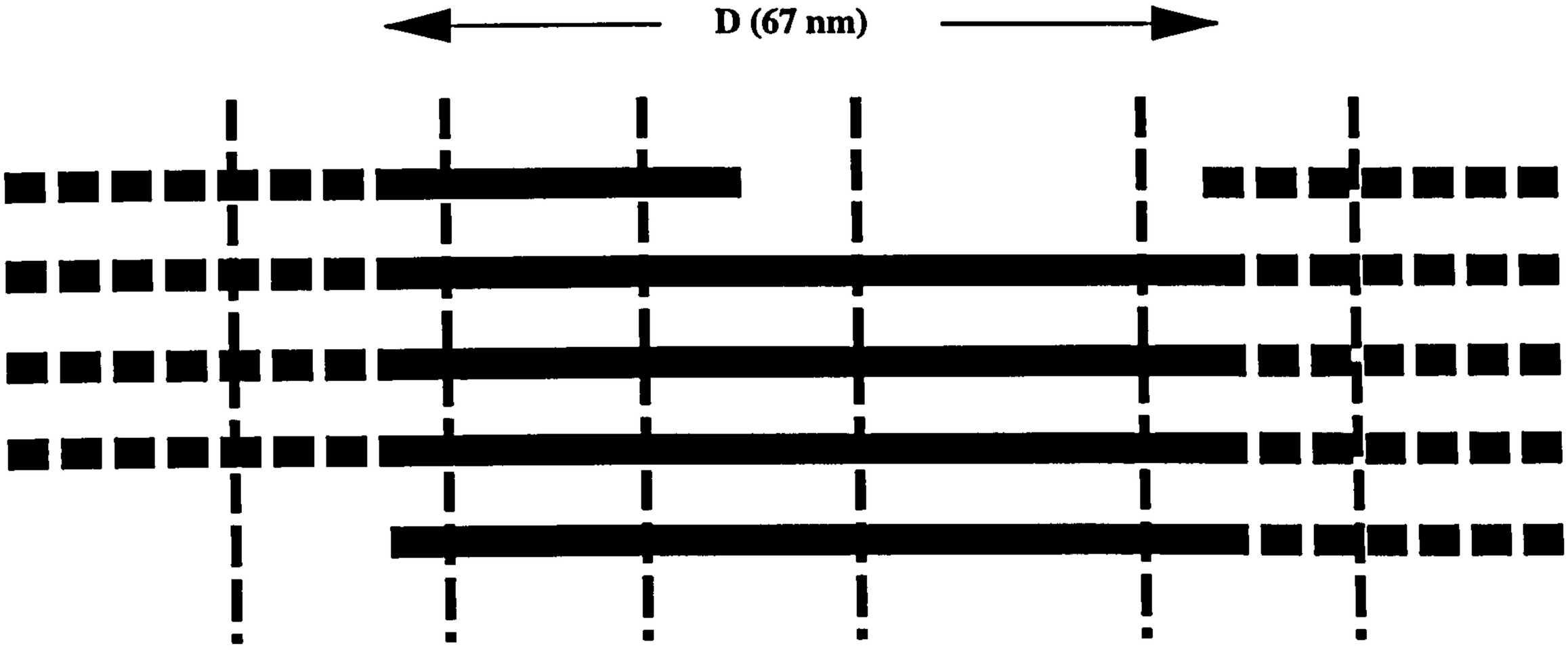


Figure 5.1 Order and disorder within a D–repeat

A single D–repeat in the collagen fibril is shown here with four distinct regions (marked). The gap/overlap interfaces are relatively ordered, high–contrast regions thought to contribute most to the sharp (Bragg) maxima in the low angle equatorial X–ray diffraction pattern. The solid black lines represent molecular segments 1–5, whilst the dotted lines are the continuation of collagen molecules into neighbouring D–repeats (since a collagen molecules is approximatly 4.4 D long). Adapted from Hulmes *et al.*, (1995).

5.1.5 The unit cell

The spacing of the discrete equatorial Bragg reflections in the fibre diagram provides information regarding the packing arrangement of the molecules and hence the conformation of the unit cell (type, dimensions, symmetry), whilst the intensity of each peak is determined by the unit cell contents. The coordinates of the Bragg peaks in the fibre diagram of type I collagen have been indexed in terms of a three-dimensional triclinic unit cell (Fraser *et al.*, 1983, Wess *et al.*, 1995), a consequence of quasi-hexagonal packing of collagen molecules (Hulmes and Miller 1979).

This development came from the observation that the intense 'triplet' resulting from the overlapping of Bragg peaks of several row-lines in the region of R= 0.7 to 0.8 nm^{-1} could be interpreted as reflections resulting from three principal lateral spacings (see Figure 5.2), indicating that the molecules are arranged on a quasi-hexagonal lateral lattice (Hulmes and Miller 1979). With the enhancement of the X-ray diffraction pattern using PTA stain, the positions of further Bragg peaks were determined. A least squares refinement to the unit cell parameters fitting at least 40 Bragg peaks allowed the unit cell to be defined as triclinic (Fraser and MacRae 1981, Fraser *et al.*, 1983). Fraser *et al.*, (1983) were also able to determine the relative orientation of the unit cell to the fibre axis using the polar coordinate angles θ, and ϕ (see Figure 5.3). The parameters were:

$a = 3.997$ nm, $b = 2.695$ nm, $c = 67.79$ nm, $\alpha = 89.24°$, $\beta = 94.59°$, $\gamma = 105.58°$

$\theta_a = 92.79°$, $\theta_b = 88.99°$, $\theta_c = 1.82°$, $\phi_a = -15.60°$, $\phi_b = 90.0°$, $\phi_c = 155.72°$

Fraser *et al.*, (1983) commented on the systematic deviations between observed and ideal lattice spacings (mainly the 3.8 and 2.6 nm equatorial spacings according to Hulmes *et al.*, 1995), leading to their suggestion that the deviations might be due to curvature of the unit cell.

A significantly different unit cell was proposed by Kajava (1991). In this model the axial stagger of 67 nm (Hodge and Petruska 1962) is maintained, by arranging non-staggered bunches of collagen molecules head to tail, which penetrate into each other by 30 nm to form microfibrils. The microfibrils in this arrangement possess an axial periodicity of 4*67 nm, but have a significantly different lateral structure and hence unit cell parameters to that of the triclinic system, with each Kajava unit cell containing four Kajava microfibrils.

Wess *et al.*, (1995), using superior quality data obtained using synchrotron radiation sources, improved spatial resolution (via shorter specimen to film distance) and PTA stain to enhance the diffraction pattern and new image analysis algorithms, tested the validity of the Kajava and triclinic models. The cylindrical polar RZ coordinates (see section 2.8.2) generated from the parameters of the two unit cells were plotted on the background- subtracted fibre diagram obtained from PTA stained rat tail tendon. The RZ set of coordinates identified and indexed every observable reflection and row-lines of overlapping reflections, whilst the Kajava unit cell produced a much poorer fit to the observable data. Several extra row-lines predicted by the Kajava unit cell were not identifiable in the diffraction pattern and with the c-axis of the unit cell being

approximately four times greater than in the triclinic model, there was a greater number of predicted peak positions for l (index of the Miller indices) than actually observed in the diffraction pattern.

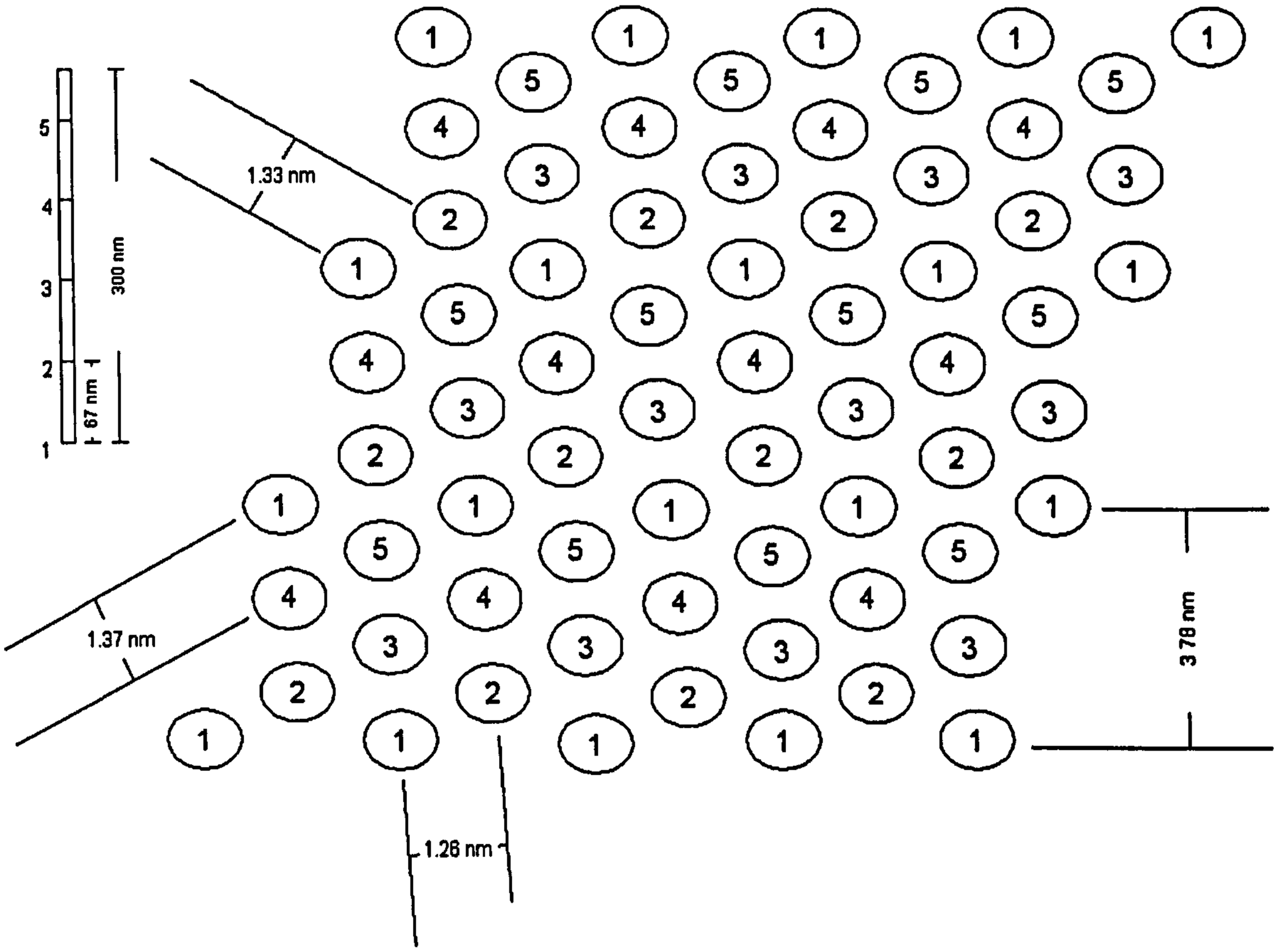
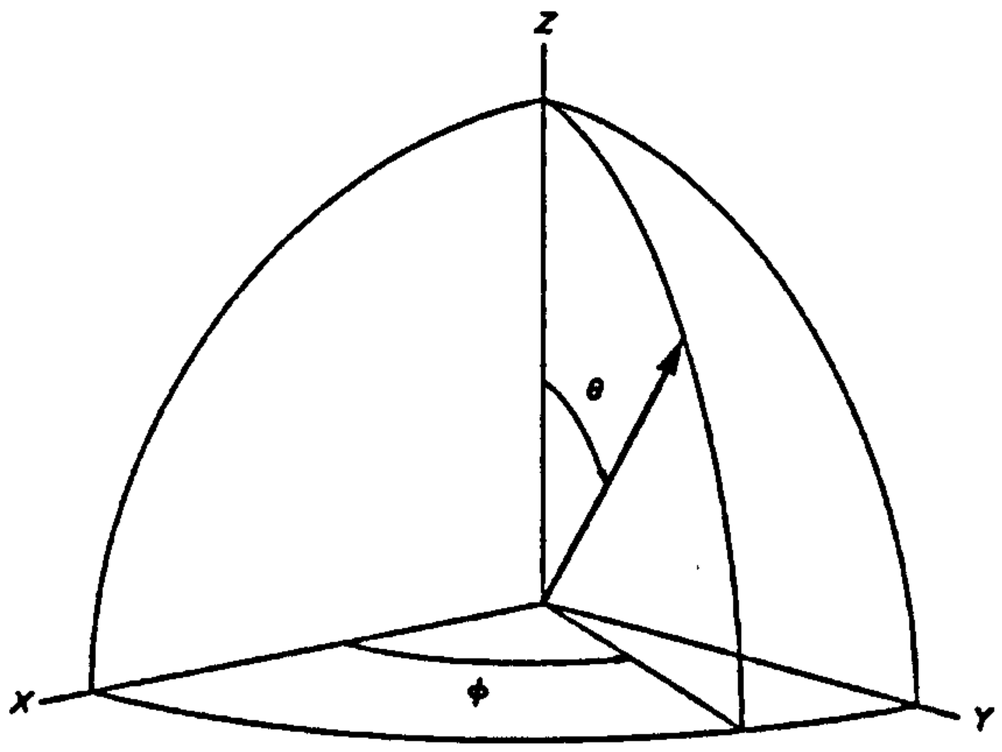
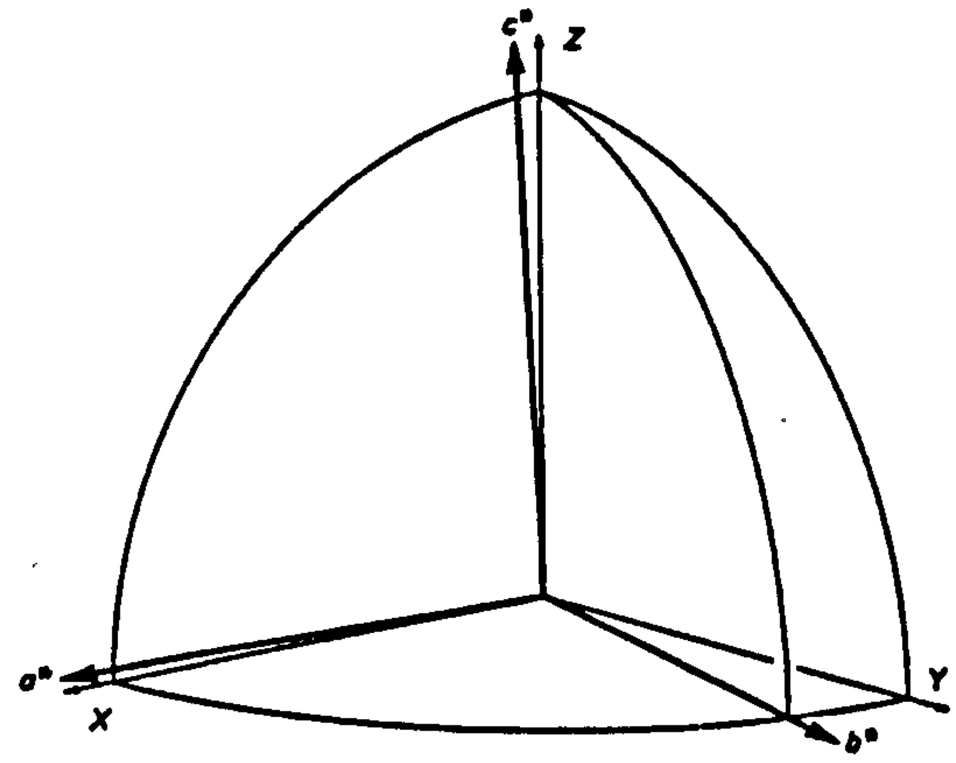


Figure 5.2 Quasi–hexagonal packing scheme

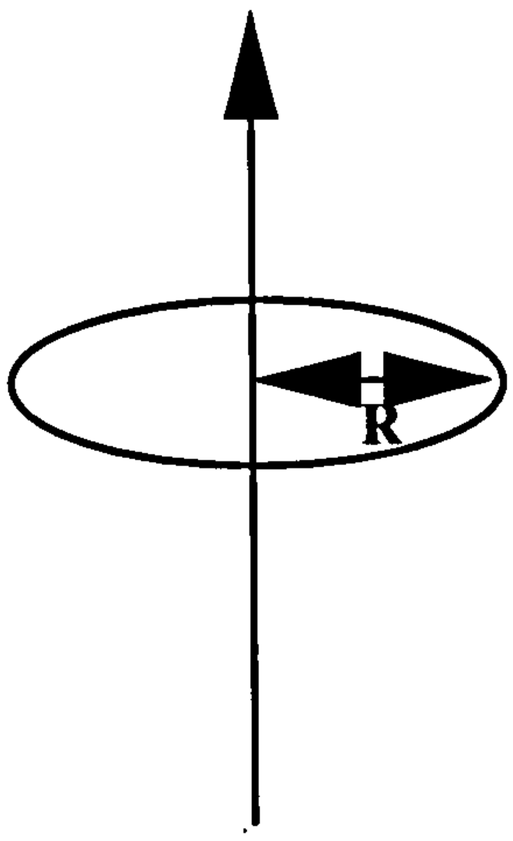
The three principal lattice spacings that correspond to the 'triplet' of intensity seen at approximately 1.3 nm in the fibre diagram are marked here. The molecular segments as defined by the D–period are labelled, but their assignment in the quasi–hexagonal scheme here is purely arbitrary. Adapted from Hulmes and Miller (1979).



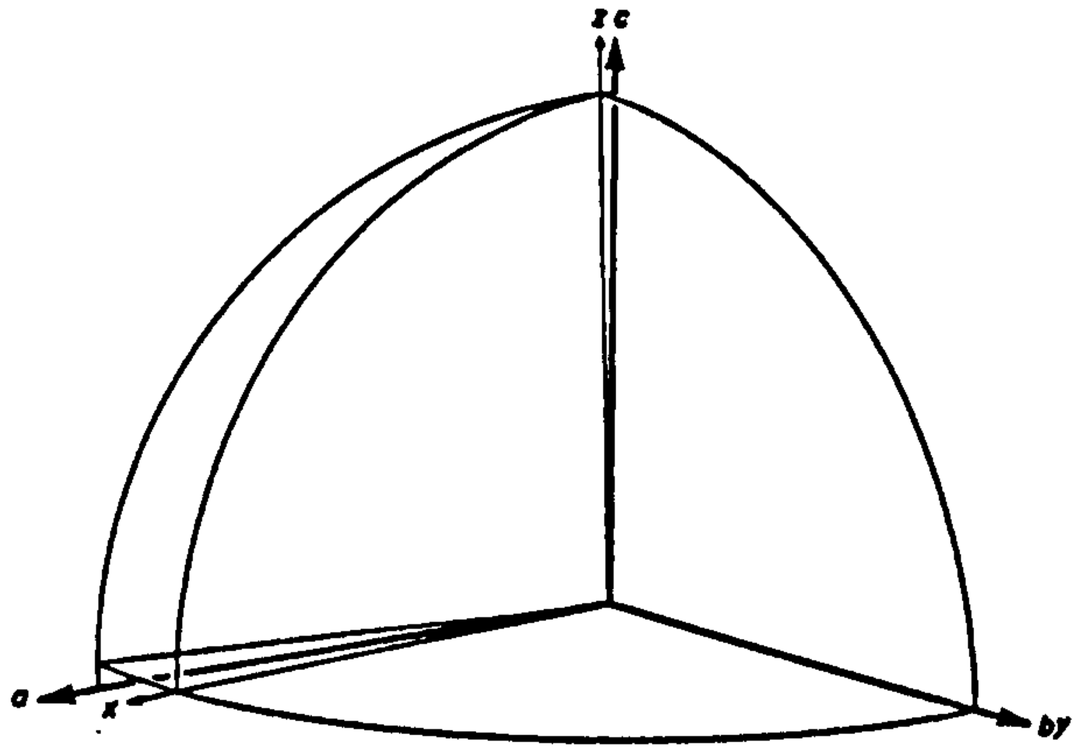
a



b



d



c

Figure 5.3 Spherical polar coordinate system

a) The spherical polar coordinates used by Fraser and MacRae (1981) to describe the orientation of the unit cell axes with respect to the fibre axis.

Orientation with respect to the fibre axis of:

b) the reciprocal cell axes and

c) the real cell axes in PTA–treated rat–tail tendon

d) polar coordinates R and Z

Adapted from Fraser and MacRae (1981).

5.1.6 Molecular tilts

The Bragg peak sampling at the layer-line (m=0, n=-1) indicates that the five molecular segments of the overlap region run parallel to a line from the uvw coordinates (0,0,0) to (0,2,1) (Fraser *et al.,* 1983, 1987), meaning that the collagen molecules are tilted (within the overlap) relative to the c-axis of the unit cell. Due to the increased mobility of the gap region, the direction of the molecular tilt (within the gap) has not been similarly assigned.

Observation of an interference maximum in the vicinity of R=0.77 nm^{-1} within the continuous scatter in the vicinity of the m=0, n=0 layer line has been taken to mean that the molecules within the gap region remain in 'clumps', that is that their path and orientation remain similar. Fraser *et al.,* (1983), demonstrated that the relative tilt of the gap region molecular segments was unlikely to be the same as the overlap, since it would not be possible to maintain the molecular segments positioning within successive unit cells as the length of the collagen molecules are traversed (4.4D). They concluded that the gap region molecular segments were crimped, and observed that this would be possible since the imino acid content of the gap region is relatively low compared to that of the overlap.

5.1.7 Molecular topology

5.1.7.1 Microfibril models

Evidence for a filamentous intermediate structure between that of the collagen molecules and the fibril has come from electron microscopy (negative staining; Bear 1952, Franc (1993), freeze fracture; Ruggeri *et al.*, 1979), and microfibrillar, X-ray data based models have confirmed and refined this idea (Miller and Wray 1971, Fraser *et al.*, 1987, Wess *et al.*, 1998a), thus leading to the wide acceptance of the concept of a D-periodic microfibril.

Several attempts have been made to reconcile the three-dimensional packing structure with that of the D-staggered one dimensional structure. Two particular models have received much attention; the hollow cylindrical 1-D staggered, 5 stranded microfibril (the 'Smith' microfibril; Smith 1968), and the compressed (non-standard packing coordinate) D-periodic microfibril (Piez and Trus 1981, Fraser *et al.*, 1983), a development of the (standard packing coordinate) quasi-hexagonal packing scheme of Hulmes and Miller (1979). See Figure 5.4

The observation that regular pentagonal ('Smith' microfibril; Smith 1968) packing does not fit with the observed quasi-hexagonal packing (Hulmes and Miller 1979) and does not rest well with the concept of molecular *packing* (the emphasis being in efficiently fitting a number of molecules into a finite space), led to the proposal of a compressed (pentagonal) D-periodic microfibril packing arrangement that was consistent with the unit cell of Fraser and MacRae (Trus and Piez 1980, Piez and Trus

1981, Fraser *et al.*, 1983). Here, the idealised packing coordinates of a regularly spaced quasi-hexagonal packing scheme were modified to a non-standard arrangement (see Figure 5.5) that still packs onto a quasi-hexagonal lattice. This arrangement also presented a structure that is energetically more favorable than the hollow cylindrical pentagonal construct of Smith by more than 31.9 Kcal/mol (Lee *et al.*, 1996).

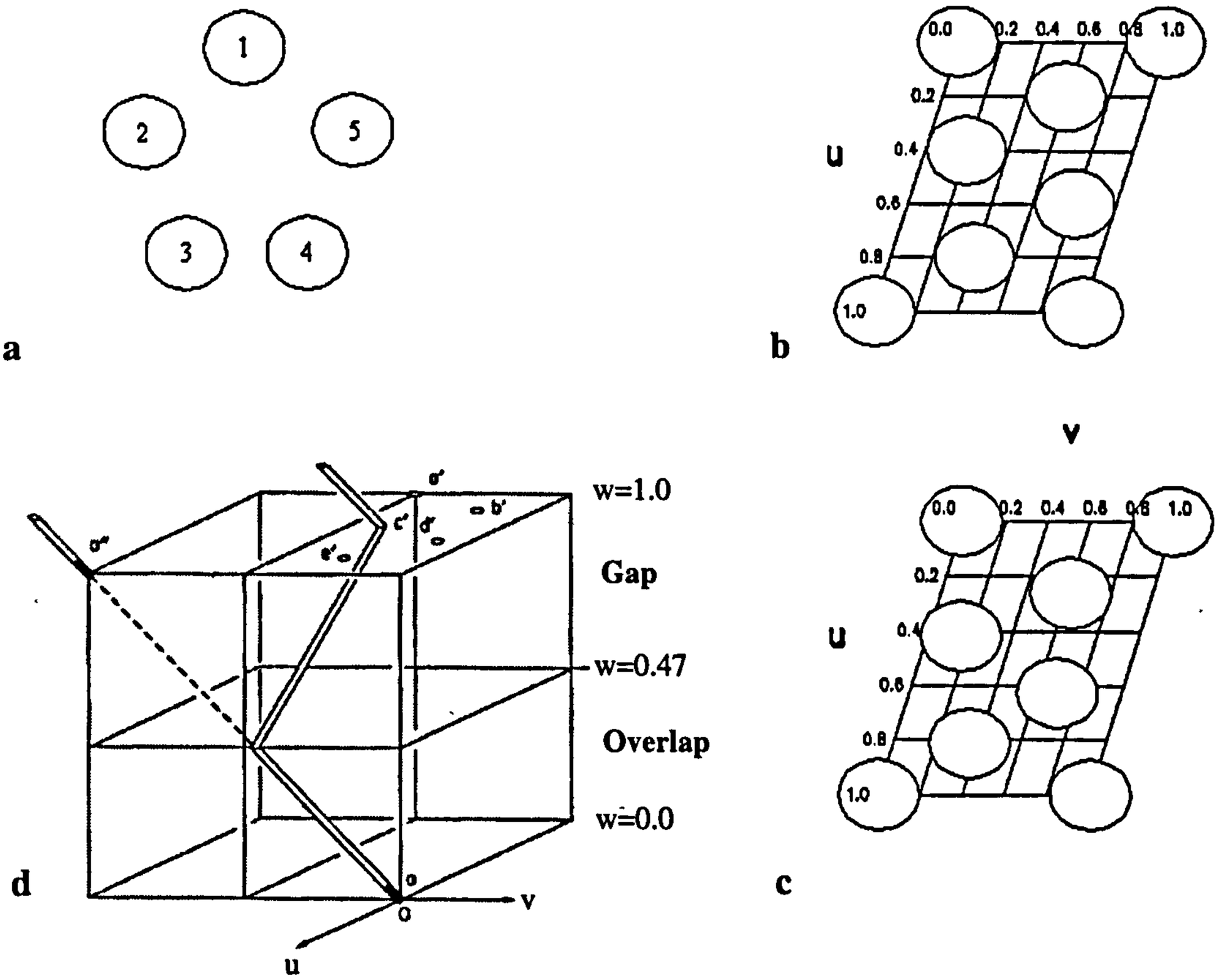


Figure 5.4 Packing models, and the molecular tilt of segments within the overlap region

a) D–periodic 'Smith' microfibril

b) Molecules arranged on a quasi–hexagonal lattice (standard coordinates)

c) Compressed quasi–hexagonal packing (non–standard coordinates)

d) Molecular segment tilt within the overlap, and required progression of a molecule through a regular series of the cell positions a–e (1–5). This progression is consistent with the sheet model S3A (up sense) and S3B (down sense) Fraser *et al.*, (1983). The tilt of the molecules in the overlap means that it is not possible to accommodate straight molecules in the fibrillar packing, if the sheet packing model is valid, the molecules must be bent in the vicinity of the overlap/gap interface (Fraser *et al.*, 1987). Adapted from Fraser *et al.*, (1987).

The 'Smith' theoretical microfibril arrangement contrasted with packing models that fit the observed X–ray data. The standard coordinates for the molecular positions of the segments that pack onto a quasi–hexagonal lattice come from Hulmes and Miller (1979). Whilst the non–standard coordinates are from Fraser *et al.*, (1983).

The non-standard coordinates for the compressed microfibril model were optimised by Fraser *et al.*, (1983) using a least squares refinement for the intensity data observed on the (2,0,l) and (0,2,l) row lines. Fraser *et al.*, observed that the sampling of row-lines at the m=0 n=0 and m=0 n=-1 layer lines indicate a non-uniform molecular spacing distribution. They further observed that Bragg sampling on the (2,0,l) and (0,2,l) row-lines must correspond to an electron density contrast due to these spacings within the unit cell, and that the molecular segments are bunched together in the overlap region.

Following the path of a single collagen molecule, it would be expected to traverse five unit cells along the direction of the lattice coordinate w (parallel to unit cell axis c), and a number of unit cell positions in the direction of the lattice axis v and u (parallel to unit cell axis a and b).

5.1.7.2 Microfibril based vs. sheet based models

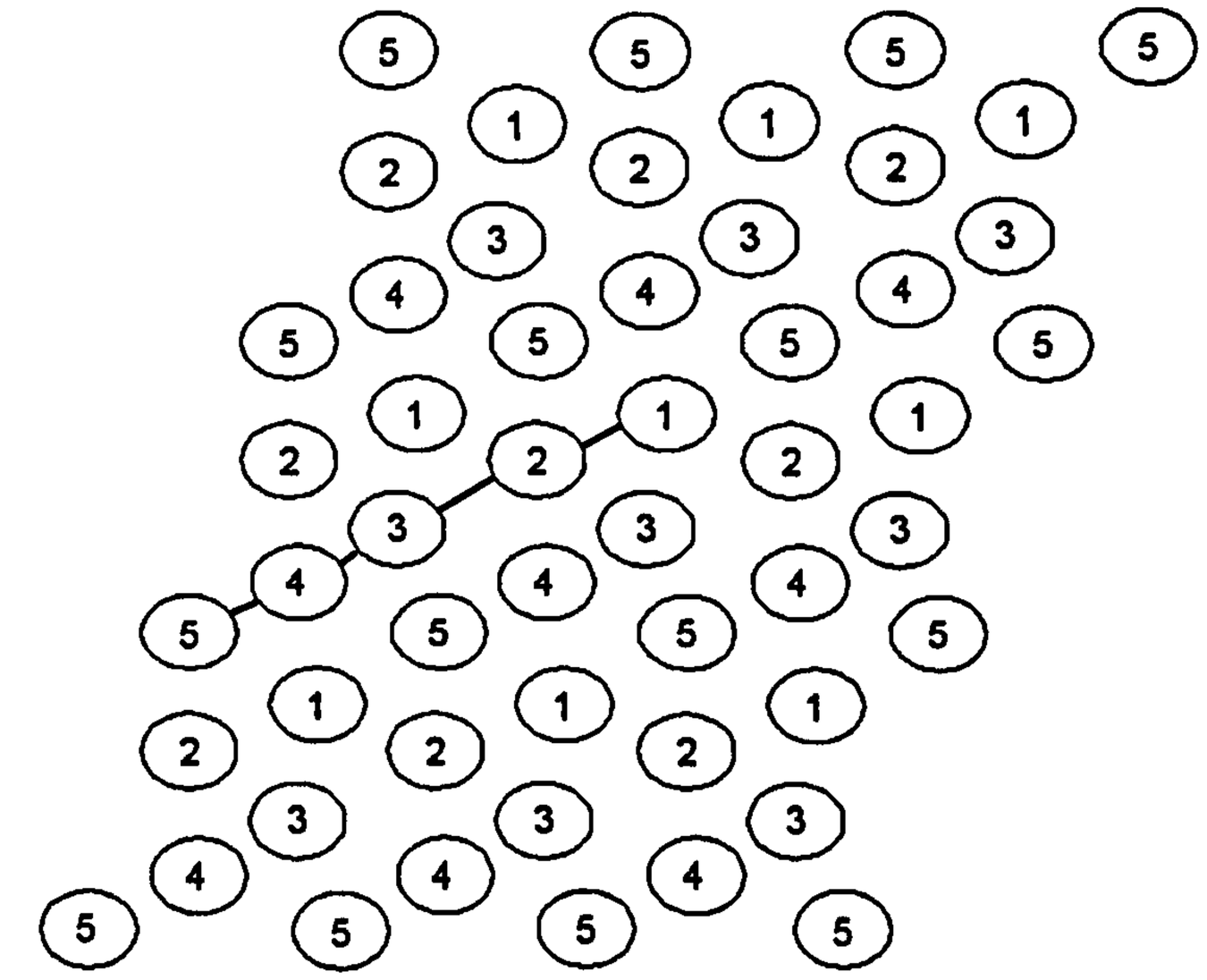
Since Fraser *et al.*, (1983) predict a linear translation of all five molecules in the overlap region running parallel to a line from (0,0,0) to (0,2,1), then the major rearrangement of the molecular segments must occur within the gap region.

The nature of this arrangement is described by one of the two following principal types that in turn describe the molecular hierarchy of the fibril:

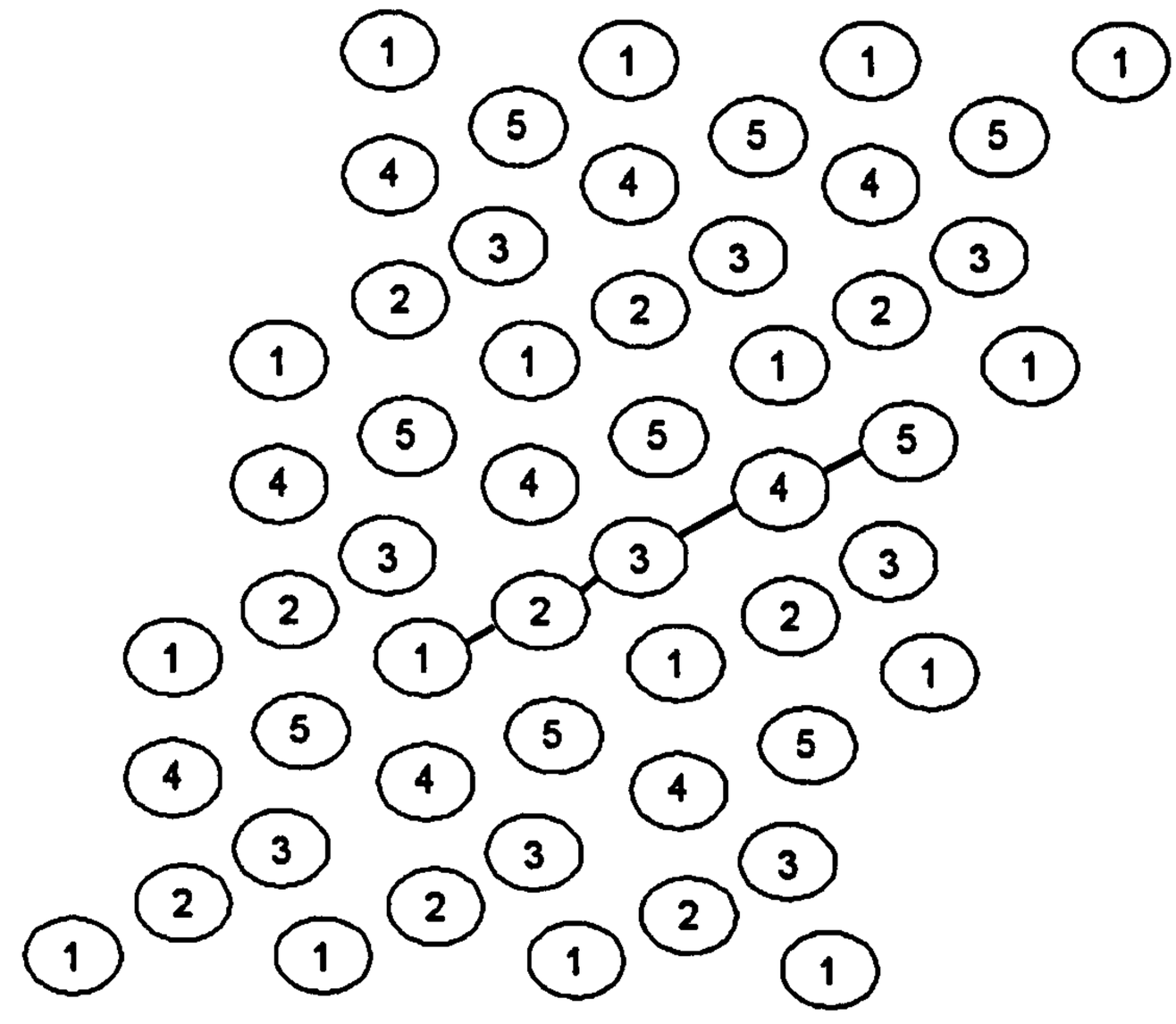
1. Structures based on the collagen molecules forming sheets where the molecular translations are linear (Hulmes and Miller 1979). See Figure 5.5

and

2. Microfibrillar structures where the molecular translations are cyclic (Smith 1968, Piez and Trus 1981, Fraser *et al.*, 1983, Fraser *et al.*, 1987, Wess *et al.*, 1998a). (Discussed above). See Figure 5.6



a



b

Figure 5.5 Segment assignments consistent with sheet models

Assignment of molecular segments to specific cell positions based on the least–squares refinement of Fraser *et al.*, (1983). Numbers between 1 and 5 correspond to molecular segments, which are nD staggered (n = a whole number, in this instance n=1) and packed on a compressed quasi–hexagonal lattice. These structures are sheet based, the topmost (a) being the S2A and (b) being the S2B models described by Fraser *et al.*, (1983).

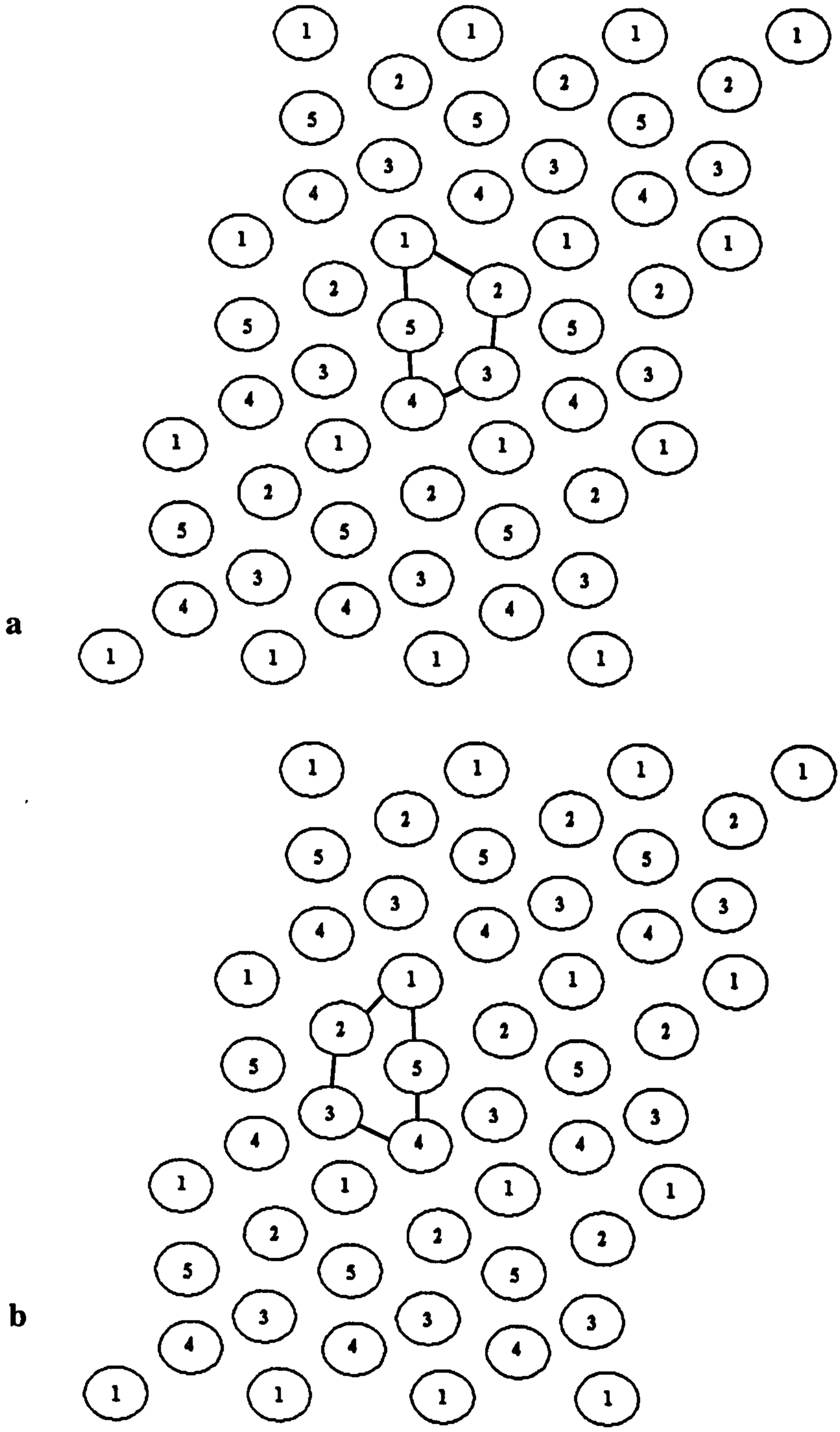


Figure 5.6 Segment assignment consistent with the microfibril model

Assignment of molecular segments to specific cell positions based on the least–squares refinement of Fraser *et al.*, (1983). Numbers between 1 and 5 correspond to molecular segments, which are nD staggered (n = a whole number, in this instance n=1) and packed on a compressed quasi–hexagonal lattice. These structures are microfibril based, (a) being the CM1(rh,1D) and (b) being the CM1 (lh,1D) models (Fraser *et al.*, 1987), where rh and lh represent right–handed and left–handed structures.

The sheet model is the simplest in terms of the rearrangement, where the gap region segments run parallel. Fraser *et al.*, (1983) observed that in order for the Hulmes-Miller sheet model to be correct, the molecules must be regularly crimped and follow a different direction to the segments in the overlap. This would allow a single collagen molecule to progress through the segment assignments (1,2,3,4,5) and assume the correct positions within each successive unit cell. Amino acid sequence evidence supports the possibility of more 'flexible' segments within the gap (Fraser *et al.*,1987), but the sheet model predicts that the gap molecular segments still follow the same path together, as they did in the overlap ('clumped' together). The contrast between this accumulated electron density with the water-filled space of the missing molecular segment 5 would produce only a limited number of low angle terms based on these two vectors. However, additional low-angle terms can be provided for in sheet models (so as to better to fit the observed data), by:

a) periodicites in the axial distribution of particular types of residue,

b) localised accumulated electron density maxima due to the telopeptides.

(Fraser *et al.*, 1987).

The compressed microfibril model does fulfil some of the observed data rather better than the sheet model. The increased mobility of the gap region demonstrated by Hulmes *et al.*, (1995) and indicated by the amino acid sequence rests well with gap region molecular translation as described by the microfibril model. The small differences in the molecular vectors between each segment in the gap region (whilst still following the same general direction) allows for more small angle diffraction terms to

be generated (Wess *et al.*, 1998a). Modelling studies (Fraser *et al.*, 1983, 1987, Wess *et al.*, 1998a) show a preference for models based on the compressed microfibril model.

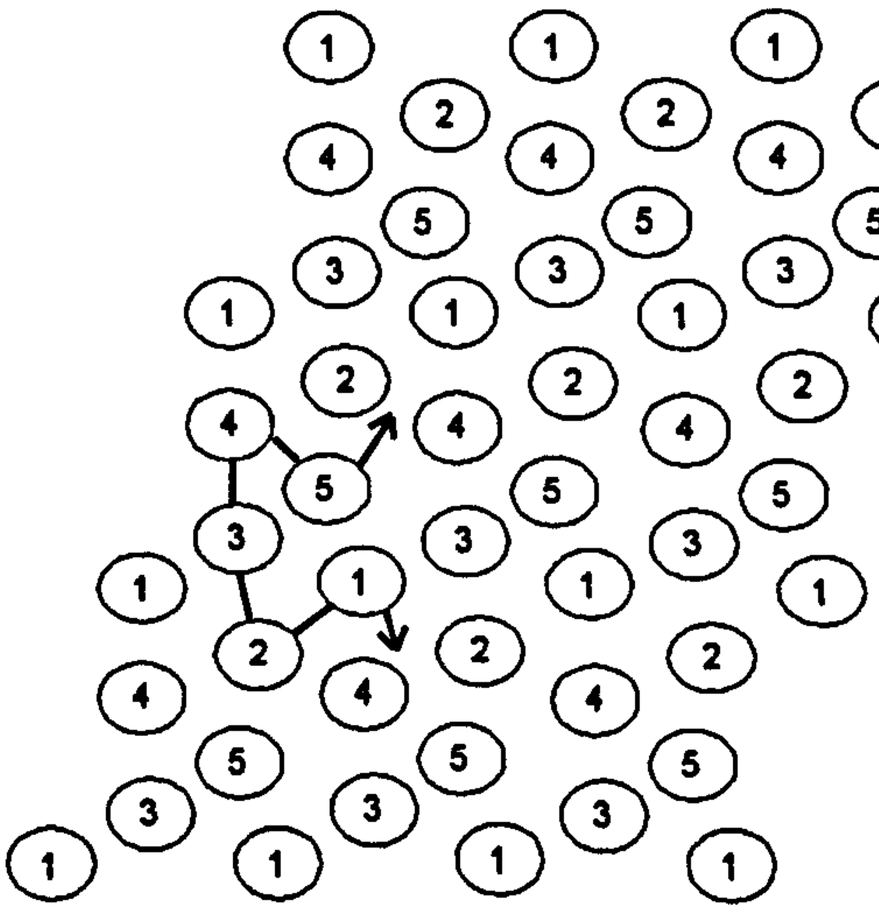
Wess *et al.*, (1998a) proposed a model for a compressed microfibril that showed excellent agreement with the observed Bragg scatter. They fixed the translation of molecules within the overlap region to that of parallel to a line (0,0,0) to (0,2,1), and the compressed microfibril non-standard coordinates of the segment positions in the global search of packing arrangements, the major variables being:

1) The conformation of the telopeptides,

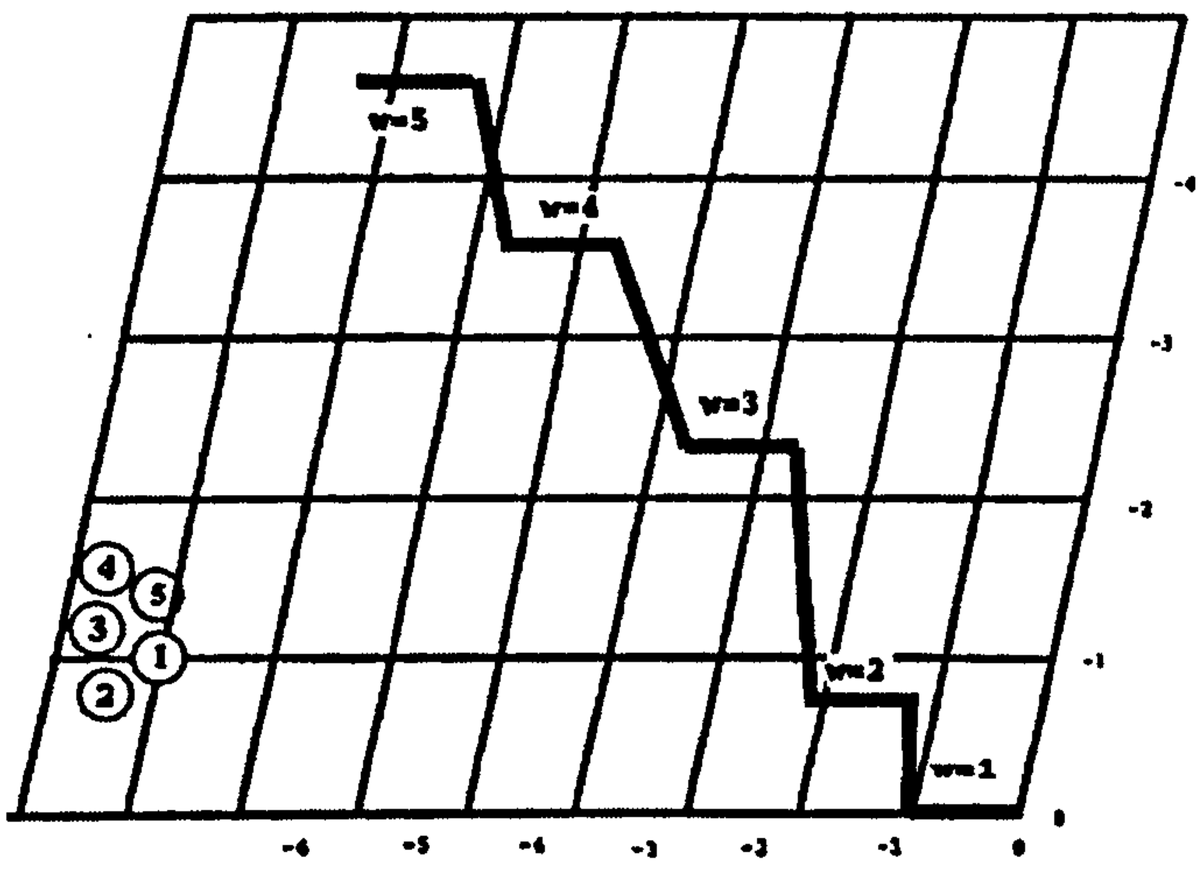
2) The assignment of molecular segment to the coordinate positions determined by Fraser *et al.*, (1983).

Fourier terms were calculated from each model and compared to the background subtracted observed low-angle diffraction pattern. The row-lines of groups 1 and 2 were fitted, whilst the remaining row-lines where left free. The resulting fit was in good agreement with both the fitted and free portions of the diffraction pattern. The subsequent assignment of molecular segments to the non-standard fractional coordinate positions allowed the course of a single collagen molecule to be followed over several unit cells (see Figure 5.7) and was in good agreement with the biochemical data. Segments 1 and 5 were in close proximity allowing for the intermolecular crosslinking between the two segments seen in the isolation of proteolytically cleaved segments of type I collagen (Nakamura 1987, Erye 1987). The assignment of the cyclic set in this way suggests that the formation of crosslinks between collagen molecules is not

contained within the same group of 5 molecules (i.e. there is inter rather than intra-microfibril crosslinking). This may explain why it has not yet been possible to isolate single microfibrils.



a



b

Figure 5.7 Segment assignments and topology of Wess *et al.*, (1998) microfibril model

Assignment of molecular segments to specific cell positions based on the global search of possible conformations of Wess *et al.*, (1998). Numbers between 1–5 correspond to molecular segments, D–staggered and arranged on a compressed quasi–hexagonal lattice (a). The structure is compressed microfibril based, and the path of a single molecule is shown in b. The lateral direction in which N– (within segment 1) and C– (within segment 5) terminal telopeptides point are shown as arrows in a, their lateral direction indicating that they form inter–microfibrillar crosslinks rather than crosslinks within the same pentamer.
Part b is adapted from Wess *et al.*, (1998).

5.1.8 Remaining questions

The nature of the Bragg scattering has been interpreted and used to calculate the packing arrangement of fibrillar type I collagen. This does not however provide a suitable explanation as to how the liquid-like packing of collagen fibrils relates to that of the crystalline packing. Additionally, the manner in which the packing structure relates to the molecular topology of a single collagen molecule has yet to be described. An understanding of this is necessary to appreciate the manner in which fibrils may be constructed from microfibrils or sheets of molecules and the location of the intermolecular crosslinks. Both of these issues will now be discussed here.

Although electron microscopy techniques are capable of visualising the packing arrangement of collagen molecules within prepared samples, this very preparation may produce artefacts that distort the lattice arrangement (e.g. drying and shearing of the sample) although Hulmes *et al.*, (1981, 1985) have demonstrated that it is possible to maintain a degree of lateral crystallinity in samples whilst studied by freeze fracture electron microscopy. In both studies (Hulmes *et al.*, 1981, 1985), evidence of crystalline structure was given by the observation of a periodic repeat of ~3.8 nm preferentially oriented in a radial direction (in transverse sections of collagen fibrils). In contrast to the drying and fixing effects of most electron microscopy methods, X-ray diffraction provides a means of studying the packing arrangement of the tissue in the hydrated state - the equatorial diffraction peaks providing detailed information of the packing structure.

X-ray diffraction has been used in this study to obtain data relating to the three dimensional packing arrangement of collagen molecules in the hydrated fibrillar state. These data have been used to calculate a three dimensional electron density map that visualises this packing structure (particularly well at the interfaces of the overlap/gap regions), and consequently sheds light on the structural relationship between individual collagen molecules and the (micro)fibril, revealing the location of the intermolecular crosslinks. The significance of these results is discussed, especially in the context of the question of the validity of the compressed microfibril model.

Although it has been claimed that filamentous structures have been observed within fibrils via electron microscopy (Bear 1952, Ruggeri *et al.*, 1979, Franc 1993), microfibrils have not yet been isolated, and this has shed doubt as to the existence of an intermediate structure between that of the fibril and collagen molecule. The possibility that the intermolecular covalent crosslinks may be between collagen molecules belonging to separate microfibrillar units may reconcile the existence of microfibrils with the difficulties encountered with separating individual microfibrils from fibrils - since they may in fact be strongly bound together.

5.2 Methods

Sample preparation, the production of isomorphous derivatives, and collection of X-ray diffraction data was performed as described in Chapters 3 and 4.

Trial experiments were performed at ESRF station ID2 and SRS station 7.2, before optimal data sets were collected at APS station ID18 (λ 1.033 Å, sample to detector distance 1.029 m). Diffraction patterns were recorded using Fuji BAS V image plates and scanned using a Fuji BAS2500 reader.

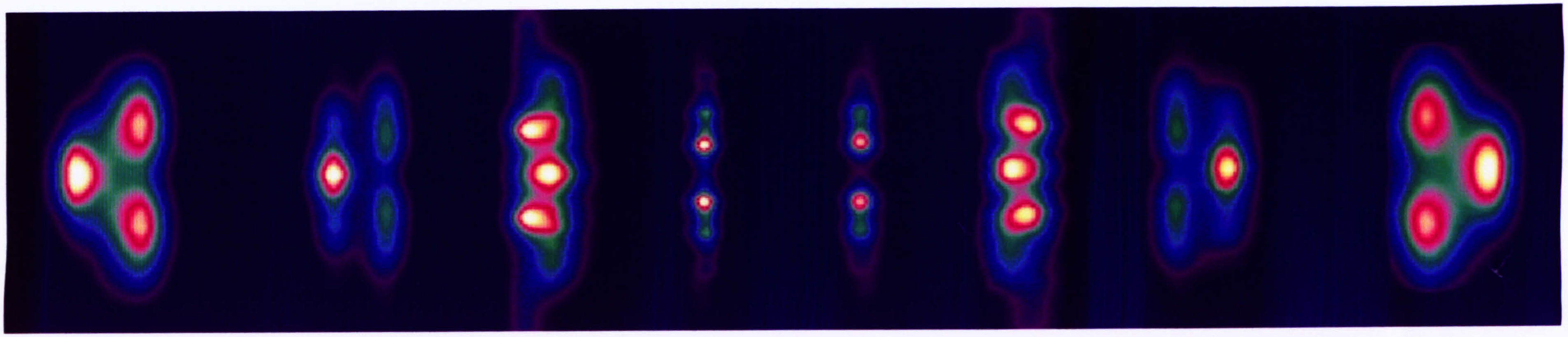
5.3 Results

The intensity sets obtained from the fitting procedure described in Chapter 3 are shown here in the form of a simulated diffraction pattern (Figure 5.8). The residuals between the simulated pattern and the background subtracted observed patterns are low (RMS equal to or less than 4.21, an error of less than 3.9%), and with visual inspection it can be clearly seen that the general fit is good.

The intensities were corrected for the intensity spread of the cylindrical transform (Bragg peaks at higher resolution being spread over a larger area than those of lower resolution) as described in section 3.7.2, and used to calculate the difference Patterson maps shown in Figure 5.9.

Putative information about the relative distances between the 1st and 5th molecular segments was obtained from the iodine and gold chloride difference Patterson maps

with reference to the known axial distribution of the stain vectors (Chapter 4 and Orgel *et al.*, 2000), and the amino acid sequence. This information was then employed to solve the phase problem for the three dimensional unit cell using the Xtalview crystallographic software suite (McRee 1993, 1999), to produce a visualisation of the three dimensional molecular packing arrangement of collagen molecules



Simulated pattern **Background subtracted observed pattern**

Figure 5.8 Simulated and observed lateral reflections of the native fibre diagram

Lateral reflections from the native type I collagen diffraction pattern compared with the simulated diffraction pattern used in determining the intensities of the observed pattern (see Chapter 3). The simulated and observed patterns are marked and shown above (the meridian is not shown). The intensities of 286 off-meridional reflections were determined for each data set (native and two derivatives), hkl indices of (-3,-2,0) to (3,2,12).

Data collection for the observed pattern is discussed in section 5.2, data extraction and treatments are discussed in section 3.6.2.

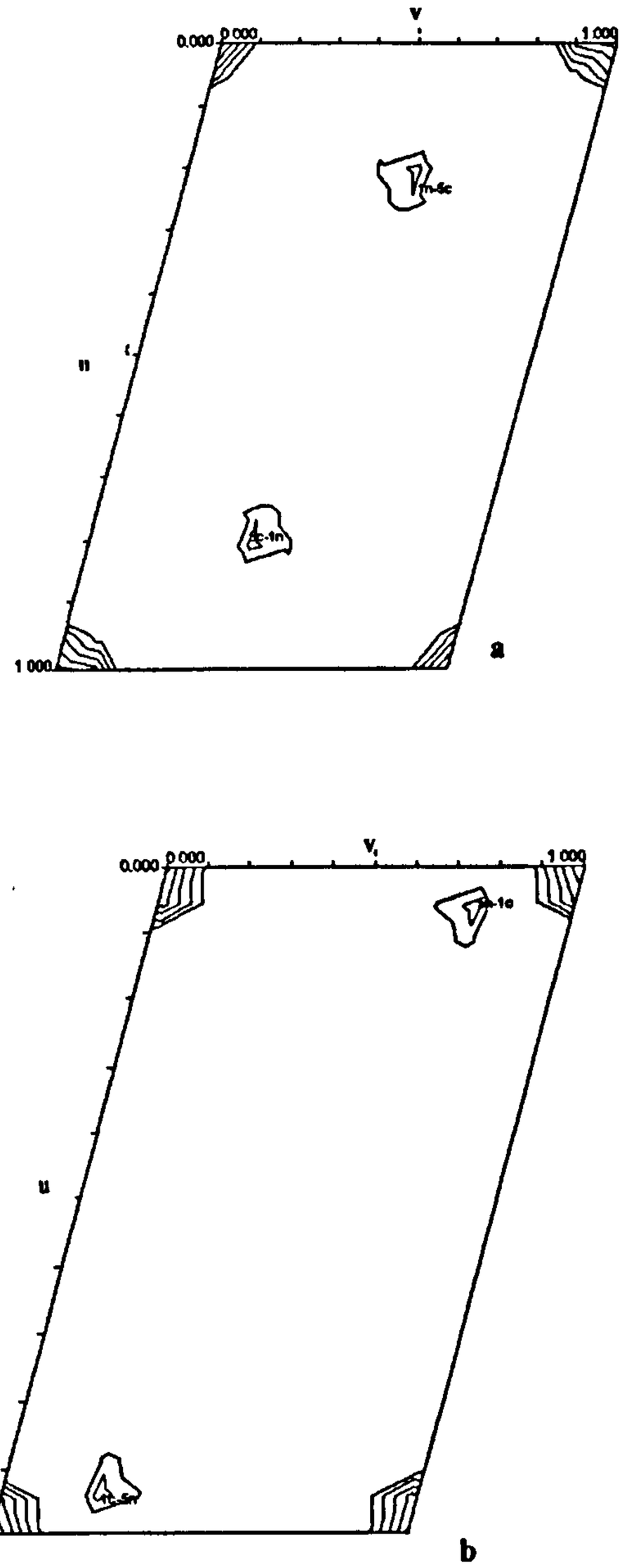


Figure 5.9 Iodide and gold chloride derivative difference Patterson maps

Difference Patterson maps corresponding to w (correspondent with the long axis of the unit cell)=0.33-0.41 (the distance between the telopeptide regions/main axial stain distribution (see Figure 4.9 and Appendix 1). Only the top 50% of peak height is shown for the iodide (a) and the gold chloride (b) derivative difference Patterson maps, so as to clearly show those peaks that correspond to heavy atom vectors rather than protein-protein or protein-heavy atom vectors. The position of the autocorrelation function derived from values in Table 5.1 are shown to coincide with the peak positions.

5.4 Discussion

5.4.1 Difference Patterson maps

The distribution of stain vectors within the collagen unit cell has been discussed for the axial unit cell (Chapter 4). The principal labelling sites for both of the heavy atom labels (iodide and gold chloride) are in the axial location of the telopeptides, although there are, in addition, a series of three 'minor' sites along the axis of the overlap region (see Figure 4.9). In the three dimensional unit cell, the labelling pattern is less straight forward, although knowledge of the amino acid sequence shows that the residues binding with heavy atom label are principally located on segments 1 and 5, the relative location of these segments within the unit cell is unknown. Only one previous study has made a credible attempt to assign coordinates to all of the molecular segments (Wess et al., 1998a), but due to the model based nature of the study, the assignment positions cannot be determined without some ambiguity.

The difference Patterson maps shown in Figure 5.9 reveal the relative distances between the stain attachment sites. Principal vectors in both the iodide and gold chloride difference Patterson maps that relate to the relative spacing between segments 1 and 5 at opposite ends of the overlap region appear in Patterson space between the fractional coordinates $w = 0.33 - 0.41$ (w corresponds to the long axis of the unit cell).

Figure 5.9a shows that a significant vector for the iodide derivative occurs between the two segments (1 and 5) where the relative spacing (the fractional distance) between

them must be approximately u=0.2, v=0.4. Whilst the equivalent vector between the gold chloride derivative segments (5 and 1) is shown in Figure 5.9b, where the relative spacing (fractional distance) is approximately u=0.05, v=0.25.

The information provided by the two difference Patterson maps is extremely useful when combined with what is known of the axial heavy atom distribution, and the amino acid sequence, which provides information regarding the axial distribution of stain binding amino acid residues, and their axial location within different molecular segments.

For the iodide derivative, the principal labelling sites are the tyrosine residues contained only within the N and C-telopeptide sequences (although minor histidine labelling has been shown to occur (Orgel *et al.*, 2000, Chapter 4)). The gold chloride derivative possesses two major binding sites that are also located at either end of the overlap region; segment 5 at the N-terminal end, and segment 1 at the C-terminal end (the exact opposite of the iodide labelling scheme, although there are stain binding histidine residues within the folded C-telopeptide sequence, see Figure 4.10).

The difference Patterson maps shown were used in the process of locating the two molecular segments containing the telopeptides, necessary so that heavy atom structure factors could be calculated in the process of determining the native phases. Heavy atom attachment sites (shown Table 5.1) were calculated and tested using an autocorrelation function (XtalView software suite) for comparison with the difference Patterson maps (Figure 5.9). Those arrangements that agreed well with the difference Patterson

functions were used to calculate the heavy atom structure functions and calculation of the native phases.

5.4.2 Calculation of the phase component of the structure factors

Using the crystallographic suite of programs XtalView, the phases for 286 native protein amplitudes (hkl from -2,-3,0, to 2,3,12) were calculated, and combined with the 124 meridional phases (00l) calculated previously (Chapter 4 and Orgel *et al.*, 2000), to produce the electron density map shown in Figure 5.10. The phases for the native were combined with the amplitudes of each derivative to produce difference Fourier maps that are also shown in Figure 5.11. Clear agreement between the difference Fourier maps and the difference Patterson maps is seen as confirmation of the correct calculation of the phases. The segments are identified where possible, the information being tabulated in Table 5.1. The electron density is rendered in three dimensions in Figure 5.12 where two successive unit cells are displayed, showing the Hodge-Petruska axial packing scheme.

5.4.3 Difference Fourier calculations

The starting assumptions of heavy atom distribution contained in the heavy atom structure factors can affect the distribution of density in the native structure factors in an ambiguous way - essentially weighting the density distribution making it possible to produce results that are merely a reflection of the starting assumptions. This is clearly dangerous if not treated with due care. This potential problem is well understood within the field of crystallography and is often approached through the process of 'cross-phasing', that is to produce difference Fourier maps for each derivative, using phases that were calculated without the contribution of that particular derivative or

assumptions of its heavy atom distribution (Dickerson *et al.*, 1967, Blundell and Johnson 1976). A heavy atom distribution that is significantly different between each derivative and in good agreement with the difference Patterson, is evidence of a properly scaled and calculated difference Fourier (McRee 1993). The technique of multiple isomorphous replacement (or addition) requires more than two derivatives to perform rigorous 'cross-phasing'. The limitation of this study to two isomorphous derivatives is unfortunate and due in part to the limited time available for use of facilities capable of producing data of sufficient quality to tackle this already difficult problem. However, evidence of the validity of the difference Fourier maps is shown here (Figure 5.11) since the pattern of stain distribution is clearly different between the two distinct derivative difference Fourier maps which were calculated using the same phase set. This, combined with the good agreement with the difference Patterson maps fulfils the expectation of a properly performed phase calculation, and validity of the difference Fourier maps.

Since the attachment of iodide stain is known to occur in both molecular segments 1 and 5, at the telopeptide regions (Bradshaw *et al.*, 1989, Orgel *et al.*, 2000), it might at first seem impossible to distinguish one segment from the other with the limited resolution of this study. However, the *major* sites of heavy atom labelling for iodide are at the tyrosine residues contained *only* within the telopeptides themselves. Whilst the general distribution of gold chloride stain is similar to that of iodide (but different in terms of which residues are chemically labelled), it is significantly different in that no major attachment site is expected to occur for molecular segment 1 at a location axially parallel to the N-telopeptide.

The combined information of the difference Fourier maps with the prior knowledge of the axial stain distribution and specific residue labelling sites makes it possible to assign fractional unit cell coordinates for the positions of molecular segments 1 and 5, and a tentative suggestion for the position of segment 3 (Figure 5.11 and Table 5.1).

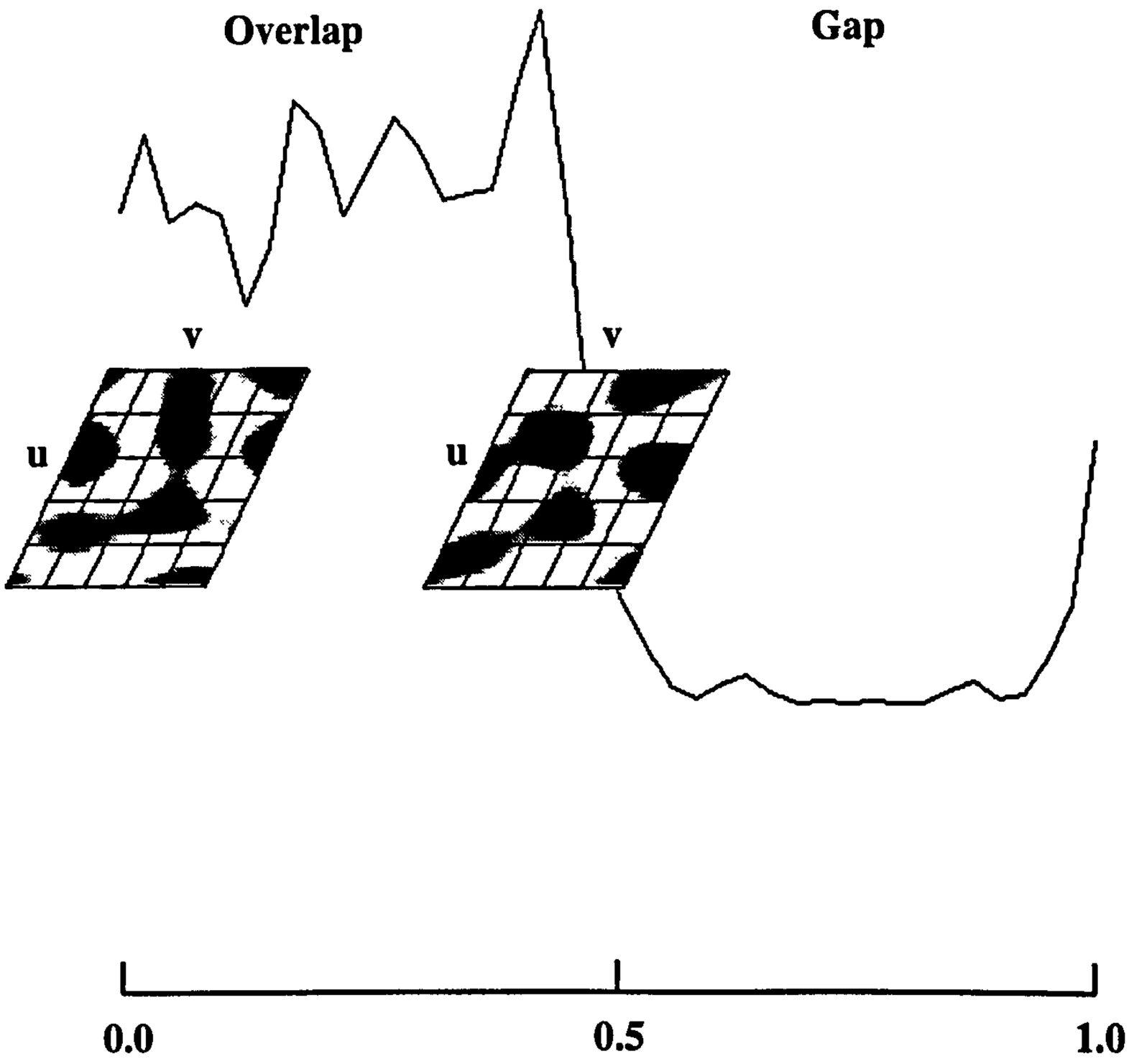
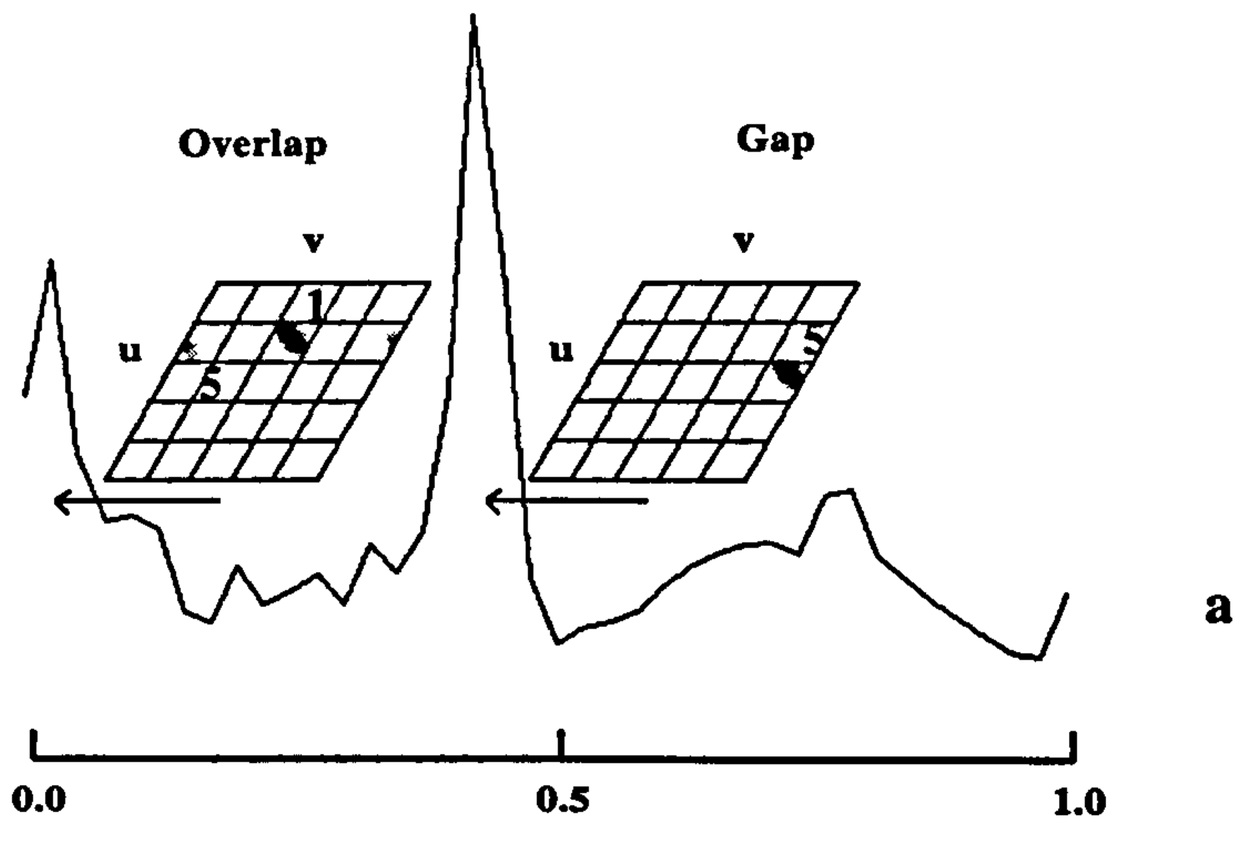


Figure 5.10 Native electron density map shown as a 1D profile with 2D lateral inserts

The distribution of density along the long axis of the unit cell is shown as a one dimensional summation when the electron density has been calculated with the contribution of the meridional structure functions. The two maxima are located in the plane of the telopeptide regions, which correspond to highly ordered areas within the unit cell, possibly due to higher packing density. The packing scheme is clearly discernible at these points and is shown in the trapeziums located at the peak maxima which show the fractional coordinate position, w (equivalent to the c–axis) of the lateral (u by v, fractional coordinates) 2D slab, which are approximately 0.05 w thick.



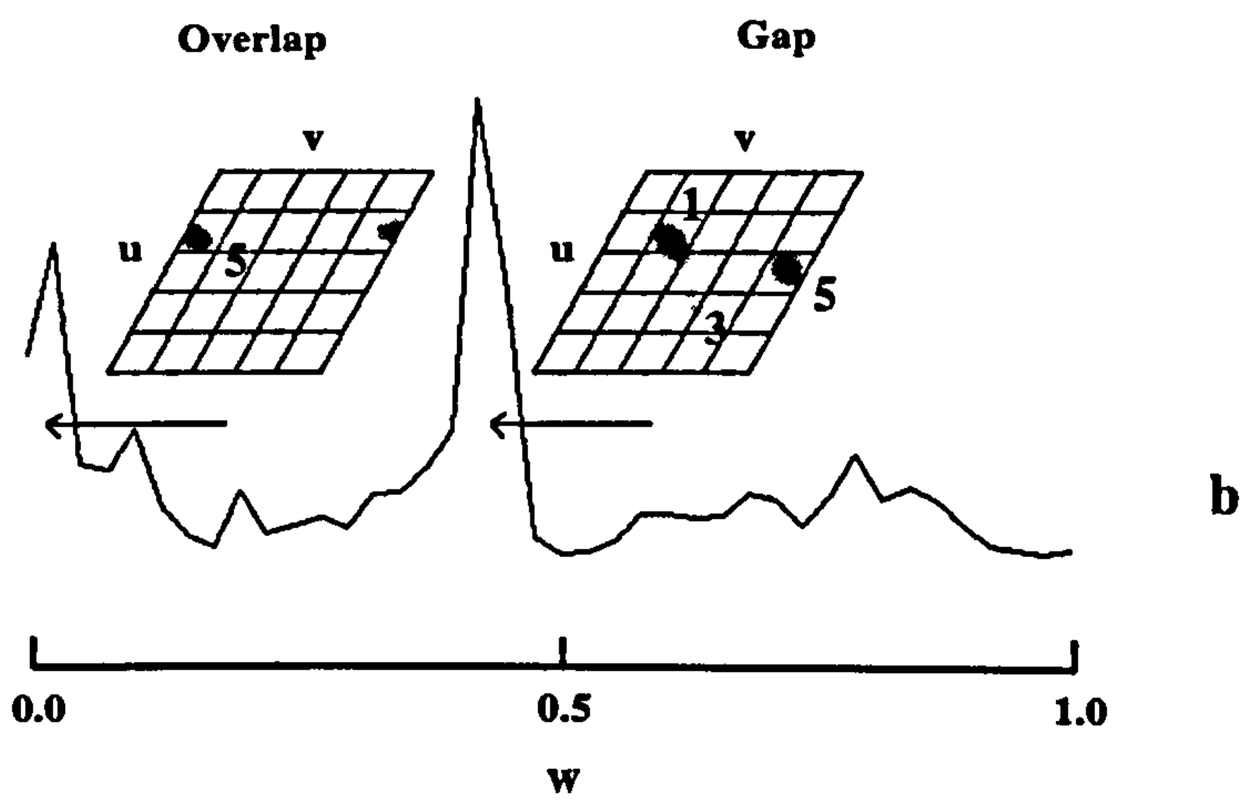


Figure 5.11 Derivative difference maps shown as 1D profiles with 2D lateral inserts

The derivative difference Fourier maps are shown here for the iodide and gold chloride isomorphous derivatives, plotted similarly to that of the native electron density map (Figure 5.10), except the slab sections are shown further to the right of their c–axis position (so as not to obscure the 1D profile, arrows mark true positions).

The iodide (a) and the gold chloride (b) difference maps reveal the locations of molecular segments 1 and 5 (see Figure 5.10 and Table 5.1), whilst the gold chloride derivative difference map shows the possible location of segment 3.

The position of segment 1 is shown by the darker spot in a) at the N–terminal end of the D–period whilst the dark spot at the C–terminal end shows the location of segment 5.

In b), segment 5 is identified at the N–terminal end, whilst the darkest spot shows the location of segment 1, the faintest the possible position of segment 3, and the intermediate intensity spot, the location of segment 5 (see main text for explanation).

SEGMENT	N-TERMINAL LEVEL	(FRACTIONAL COORDINATE)	C-TERMINAL LEVEL	(FRACTIONAL COORDINATES)
	v	u	v	u
5	0.087	0.372	0.929	0.449
1	0.508	0.231	0.310	0.308
a (2)	0.222	0.731	0.175	0.833
b (3)	0.651	0.641	0.556	0.660
c (4)	0.897	0.000	0.635	0.083

a)

SEGMENT	FRASER *ET AL.*,(1983)	(FRACTIONAL COORDINATE)	PIEZ AND TRUS (1981)	(FRACTIONAL COORDINATES)
	v	u	v	u
a	0.000	0.000	0.000	0.000
b	0.619	0.262	0.650	0.300
c	0.162	0.418	0.225	0.450
d	0.697	0.627	0.725	0.650
e	0.356	0.813	0.425	0.850

b)

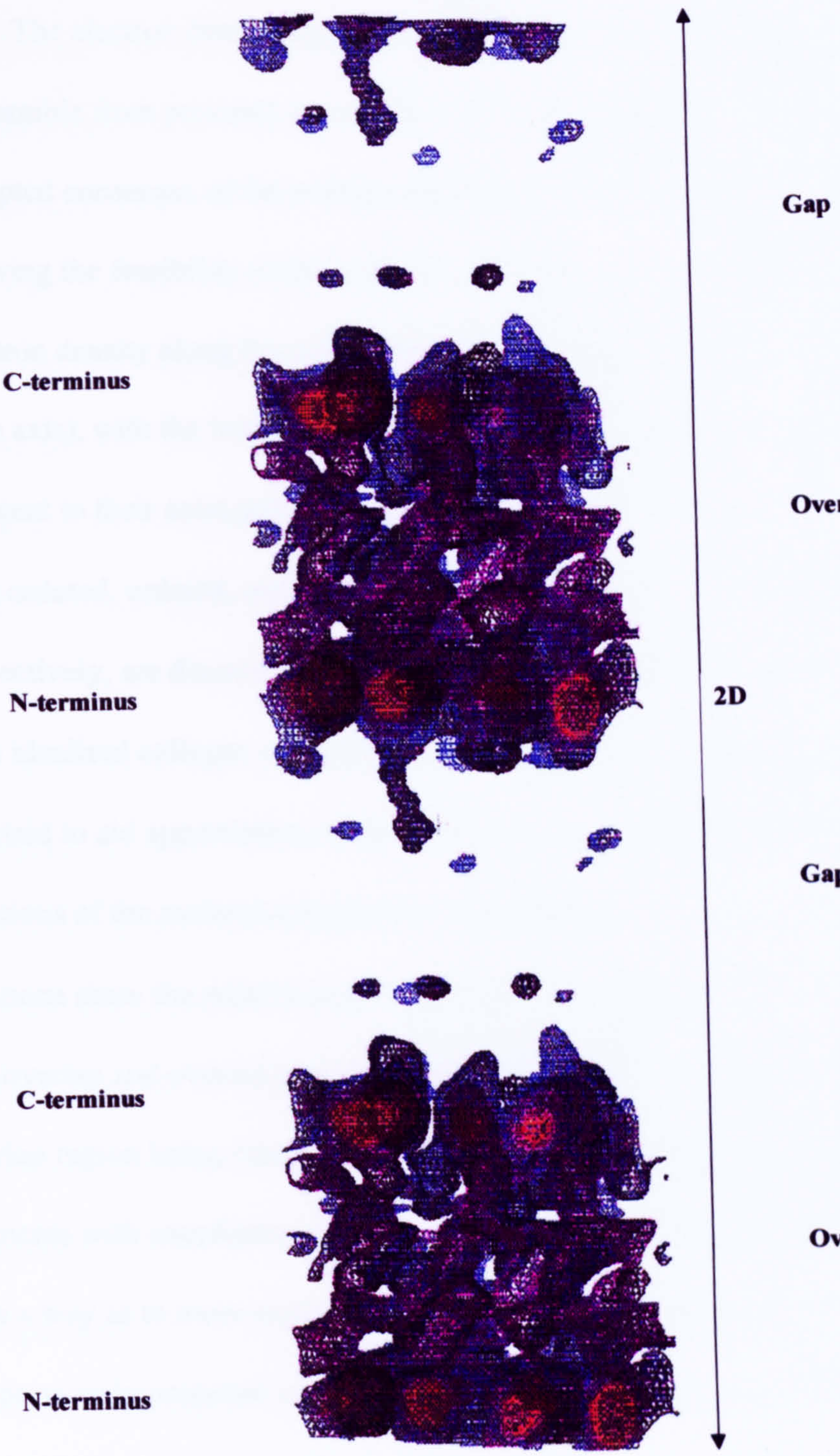
Table 5.1 Fractional coordinate positions of molecular segments

a) The fractional coordinates of the assigned and non-assigned molecular segments are shown here. These coordinates are taken from the w axis corresponding to 0.025 (N-terminal level) and w = 0.4375 (C-terminal level). 1 and 5 correspond to molecular (telopeptide containing) segments 1 and 5. Whilst a, b, and c are the positions of the remaining segments (2-4) although the correct assignment between the remaining segments and positions are not certain. Although there is evidence to suggest that segment labelled b may be segment 3 (see Figure 5.11 and section 5.4.5), if this is the case, then this tentative assignment and the suggested structure shown in Figure 5.17 would make segments a and c correspond to molecular segments 2 and 4.

b) Segment coordinates determined by pervious workers.

Figure 5.12 View of two unit cells perpendicular to the fibril axis

The molecular packing arrangement is most clearly discernible within the plane of the telopeptides, and also discernible within the rest of the overlap region. The molecular segments follow a common path within the overlap region, and due to a higher packing density than that of the gap region and formation of intermolecular cross–links at the interfaces of the overlap/gap regions (at the telopeptides), the overlap region is well ordered, particularly in the plane of the telopeptides in contrast to the disordered state of the gap. The gap region shows a large degree of lateral disorder of the molecules, as predicted in several model studies (Fraser *et al.*, 1987, Hulmes *et al.*, 1995, Woodhead–Galloway and Machin 1976). Disorder is commonly encountered in macromolecular crystallography in regions of crystal structures subject to thermal motion (such as chain loops). Here the disorder is seen on a larger scale of the triple helices, which is due to an as yet unknown contribution of static and thermal disorder. An analogous mixture of ordered and disordered regions comparable to that found in collagen was also reported in crystals of tropomyosin (Phillips *et al.*, 1980).


Gap
C-terminus
Overlap
N-terminus
2D
Gap
C-terminus
Overlap
N-terminus

5.4.4 The significance of the electron density map

The electron density map shown in Figure 5.10 reveals several features not discernible from previous studies, as well as being in good agreement with the generally accepted consensus of the packing structure in type I collagen (Wess *et al.*, 1998b), showing the feasibility of this solution. Figures 5.10 and 5.12 show the distribution of electron density along the unit cells longest axis (approximately parallel to that of the fibre axis), with the lateral packing arrangement at the locus of the telopeptides shown adjacent to their axial positions within the unit cell in Figure 5.10. In Figure 5.12, the well ordered, ordered, and disordered; telopeptide regions, overlap and gap regions respectively, are discernible. The same lateral arrangements are shown in Figure 5.13 with idealised collagen molecules (circles placed around density maxima centroids) overlaid to aid appreciation of the packing arrangement. These overlaid coordinate positions of the molecular segments in the plane of the N and C-telopeptide axial locations show the relative positions of the molecular segments at the start of the gap/overlap and overlap/gap interfaces (Figure 5.14), the tilt of the molecules in the overlap region being clearly revealed in Figure 5.15. Finally, the idealised molecular segments with coordinates obtained from the electron density map have been plotted in such a way as to more easily compare the packing arrangement of the molecules with the previously proposed models, especially the compressed microfibril model (Fraser *et al.*, 1983, Piez and Trus 1981, 1983). See Figure 5.16 and Table 5.2.

Visual inspection of the electron density map reveals several points of interest:

1. The electron density map reveals that there are three levels of order/disorder within the packing arrangement of the molecules; well ordered overlap/gap interfaces (at the axial level of the telopeptides), relatively ordered overlap region, and the relatively disordered gap region. The packing scheme is clearly discernible within the overlap region, particularly at the axial level of the N-telopeptide and of the C-telopeptide. This is indicated by the distribution of electron density along the long axis of the unit cell in the 1-D profile shown in Figure 5.10, and as seen in Figure 5.12 (consistent with the results of the study of Hulmes *et al.*, 1995).

Evidence of the increased mobility of the four molecular segments in the gap region, as well as limited mobility in the overlap zone can be drawn from the electron density map (Figures 5.10 and 5.12). It is however possible that the overlap region is more highly ordered than the electron density map implies. The most highly ordered region of the collagen packing scheme appears to be that in the plane of the telopeptides. It is possible that the correctness of the phase solution is biased towards the packing structure in the plane of the telopeptides (very highly ordered region), and is less stringent in producing correct phases for the rest of the overlap region, and still less stringent for the gap region.

However, NMR studies show rapid azimuthal motion along the long axis of the collagen molecule (in collagen fibrils) over a range approximately equal to 30° - 40° in azimuth (Jelinski 1980, Torchia 1982). It is reasonable to speculate that this

molecular motion occurs largely within the gap region and to a lesser extent within the overlap region (where molecular movement of any one collagen molecule is limited by the interactions of the six closely-packed surrounding collagen molecules). Molecular movement within the overlap region may be further restricted at least for segments 1 and 5 by the presence of the intermolecular crosslinks, this might also affect the overall resistance to movement of the remaining (non-crosslinked) collagen molecules within the vicinity of the telopeptides. This may be due in part to the higher packing density of telopeptides in cross-section (as implied by the folded conformation of the C-telopeptide). Certainly this would help to explain the dual ordered/disordered nature of the system, and the suitability of a combined crystalline/liquid-like approach to delineating the fine structure of the microfibril.

2. Visual inspection of the distribution of molecular segments at both the N and C-termini shows the packing arrangement to be compressed (non-normalised/non-standard). The arrangement bears similarity to both the standard (Figure 5.4c) and non-standard (Figure 5.4b) compressed microfibril arrangements, in that the packing arrangement is quasi-hexagonal, but is significantly distinct from both. This is due to: a) nearest neighbour contacts of the molecular packing (see Figure 5.13), and b) the difference in packing arrangement between the start and end of the overlap region (see Figure 5.14). Although it could be argued that the compressed arrangement of Fraser *et al.*, (1983) (Figure 5.4b) is the average packing arrangement of the overlap region, the prediction of the position of the segments in Fraser *et al.*, (1983) that are closest to the position of segments a and c (Figure 5.15) remains

distinctly different between the modelling studies of Fraser *et al.*, and that shown through the electron density map presented here. Since segments a and b and 1 and 5 (of the electron density map shown stylistically in Figure 5.15) are now known (from this study) to be similarly arranged and separated from one another, it is possible that the model calculations of Fraser *et al.*, (1983) fell into a local minimum whilst fitting the model structure factors to that of the observed diffraction pattern, their model fit being biased towards the segments a/b and 1/5 vector, and failing to resolve all of the molecular segment positions accurately.

3. The relatively limited movement of segments 1 and 5 relative to one another from the start and the end of the overlap region compared with the range of movement displayed by segment c for instance, is evidence for the stabilising effect of the covalent crosslinks on the lateral packing structure of type I collagen. The slightly closer arrangement of the crosslinking segments (1 and 5) and the C-telopeptide, may be a reflection of the tight-turn conformation of the C-telopeptide, that could be stabilised by self-interactions (tyrosine stacking - see alignment of tyrosine residues in Figure 4.10), and holds the two segments in closer proximity than does the N-terminal telopeptide.

4. The molecular tilt of all five collagen molecules in the gap region was observed to follow a vector parallel to the line (0,0,0) to (0,2,1) by Fraser *et al.*,(1987). Overlaying the sections of the electron density map corresponding to w=0.025 and w=0.4375 (the telopeptide regions), indicates that all five molecules have progressed through the gap region largely parallel to one another (Figure 5.14). The

displacement of the molecules at w=0.4375 from the original positions at w=0.025 is approximately 0.06 v. This would be consistent with the path of the molecules following the line (0,0,0) to (-0.08,2.29,1.00), or (-0.4,11.46,5.00) compared with (0.125,10.062,5.000) (Fraser *et al.*, 1983), a tilt of approximately 5.4°. Since the observation of the molecular tilt of the collagen chains in the overlap made by Fraser et al., (1983) has been confirmed here (shown clearly in Figure 5.16), their observations of the physical impossibility of sheets of straight collagen molecules fitting into the compressed, quasi-hexagonal lattice arrangement over the course of five unit cells is confirmed also.

The apparent 'misalignment' of the segments between the start and finish of the overlap region is also consistent with observations of limited movement available to the collagen molecules within the unit cell as discussed above in point 1. See Figure 5.14 for an overlay of packing arrangement of molecules in the plane of the telopeptide regions.

5. The small shift in relative molecular positions of the segments in the overlap and their (compressed, non-idealised) arrangement implies that a significant degree of rearrangement must occur for each molecular segment within the gap region. Because of this non-uniform arrangement of the collagen molecules, and the increased mobility of the gap region, the molecular rearrangement that occurs within the gap (required to maintain continuity in segment positions across multiples of unit cells) must be more like that of a compressed microfibril than that of sheets of collagen molecules (discussed in the following section).

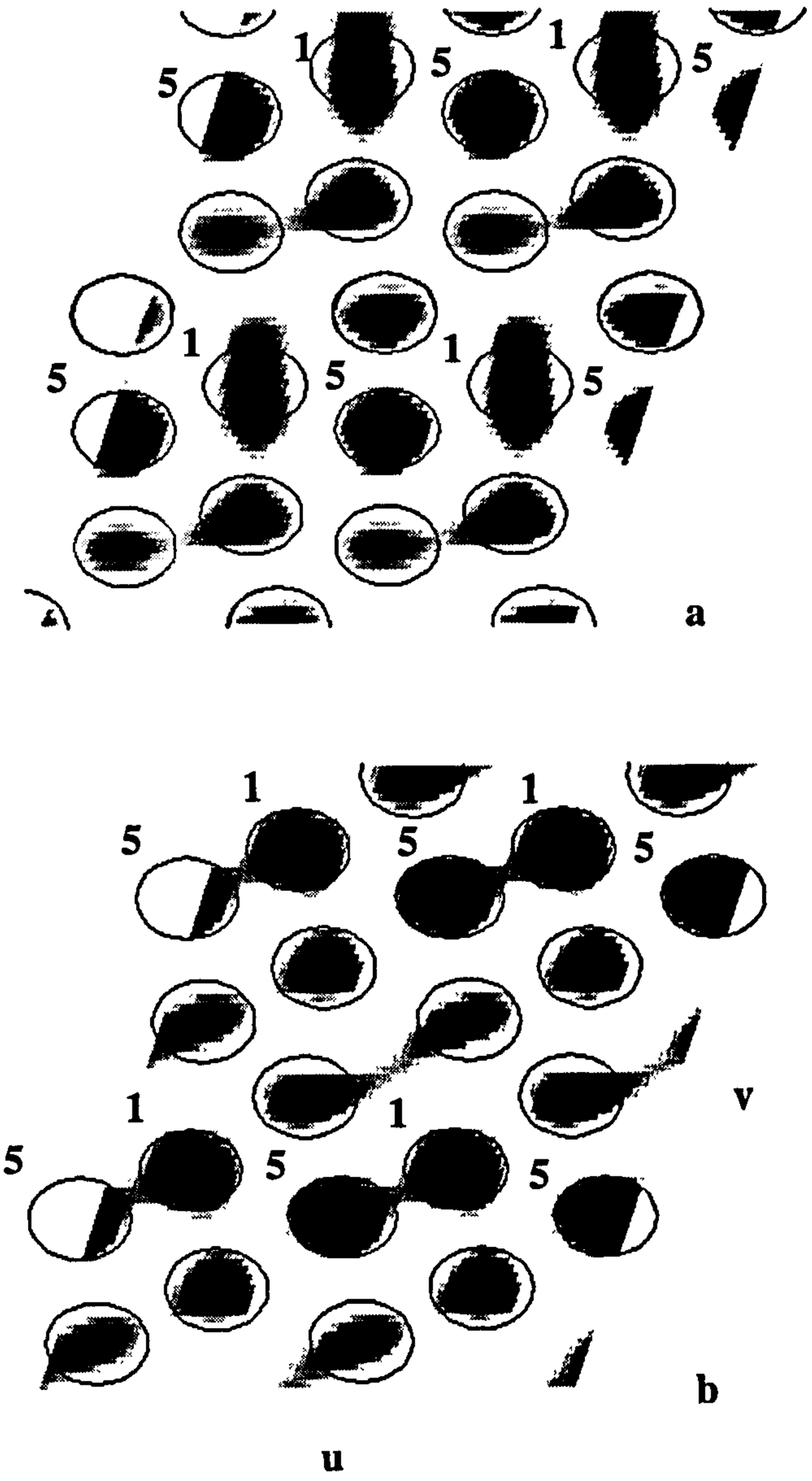


Figure 5.13 Slab sections of the native electron density map at the axial level of the telopeptides

The cross–sections of the electron density map of the unit cell shown in Figure 5.10 are shown here; the N–terminus (a), and the C–terminus (b). The cross–sections are 2x2 unit cells (u,v), and the electron density maxima corresponding to molecular segments have been overlaid with circles (idealised shape of a molecular segment) as a guide to the eye. Molecular segments 1 and 5 have been labelled (label to the left of the segment).

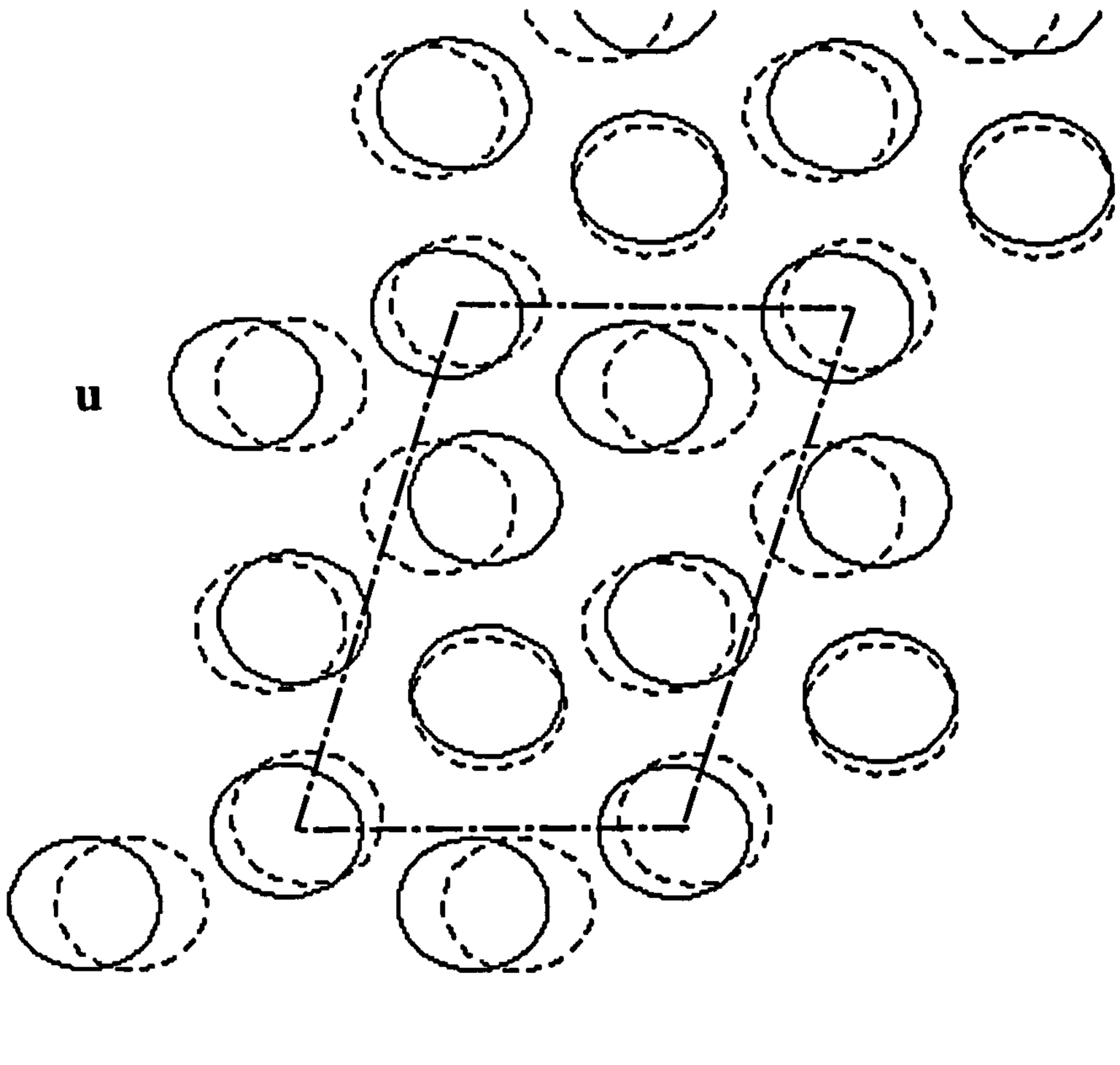


Figure 5.14 Overlay molecular segment positions of the N and C terminal regions

The positions of the molecular segments at the N (solid lines) and C–termini (dashed lines) are shown here relative to one another. The overlay of the molecular segment positions at the interfaces (start and finish) of the overlap region shows that the general progression of the segments within the overlap region are similar (they follow approximately the same vector path). One unit cell is shown (approximately) by the dot–dashed lines.

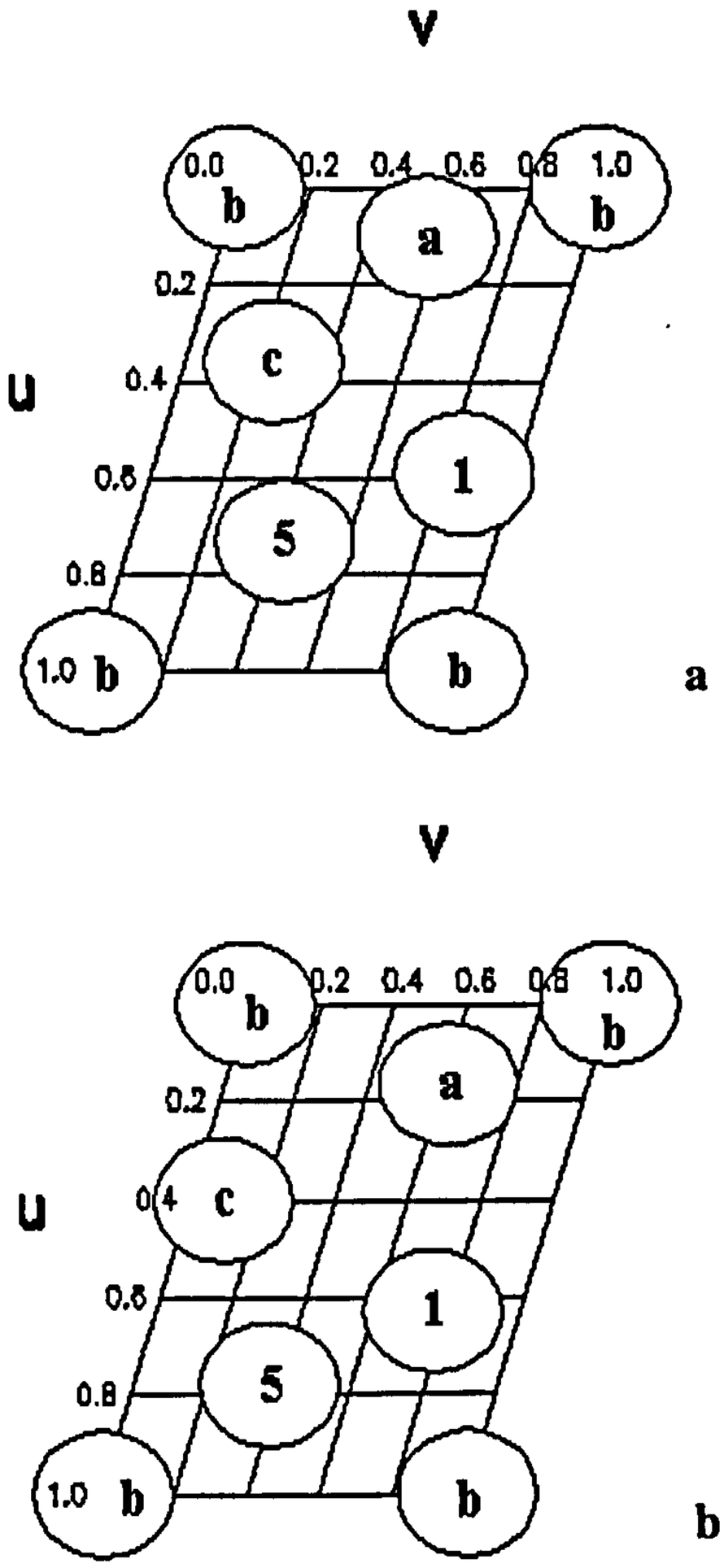


Figure 5.15 Packing arrangement and partial segment assigment

The fractional coordinate positions of the segments have been uniformly shifted to emphasise the similarity between the packing seen in the electron density map, and the compressed packing scheme of Fraser *et al.*, (1983). Molecular segments 1 and 5 are labelled at the N (a) and C–terminal (b) regions of the unit cell; a,b, and c, are the positions of the remaining segments (of which b could possibly be segment 3, see main text).

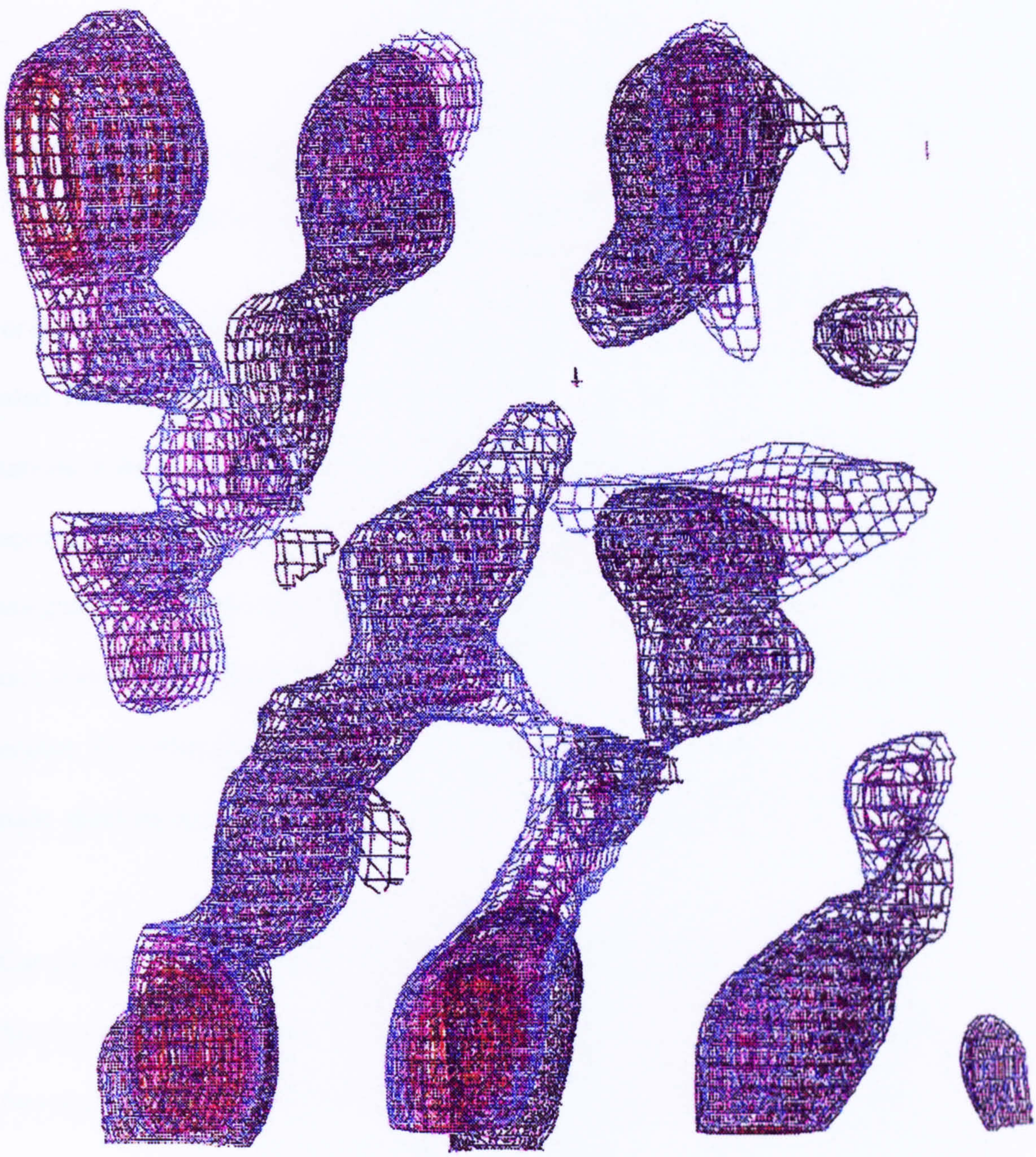


Figure 5.16 Tilt of the molecular segments within the overlap region

The molecular tilt of the collagen molecular segments within the overlap region is observed to follow a vector approximately parallel to the line (0,0,0) to (0,2,1) (u,v,w; Fraser *et al.,* 1987). This corresponds to a tilt of about 5 degrees relative to the c-axis of the unit cell. Only the overlap region of a single unit cell is shown here with the c-axis compressed by 13 times to show the molecular segments as clearly as possible. The variation in the electron density along each segment is due to bands of electron dense amino acids, but may also be due (in part) to disorder/phase error.

5.4.5 The path of a single collagen molecule, topology and a proposition for the packing structure of collagen based on evidence presented here

For the first time, the three dimensional packing structure of collagen has been revealed. Although it is not possible to assign all of the molecular segments due to limitations in the interpretation of the electron density map (impossible to trace the collagen chains through the gap region and hence identify all the molecular segments), the assignment of the nearest neighbours to segments 1 and 5 (in the C-telopeptide region), leaves only a limited number of possible assignments for the remaining three molecules. To further reduce the number of possible assignments some speculation can be made about the specific location of segment 3.

Careful inspection of Figures 4.9 and 5.11b shows that a gold chloride binding methionine residue is located in segment 3 that may correspond to the weak peak below the two strong peaks shown at the C-terminus in Figure 5.11b. This has been given limited attention until now due to the uncertainty in making the assignment of segment 3 to this position. Whereas the location of segments 1 and 5 are confirmed by both the iodide and gold chloride difference Fourier maps, only the latter could be expected to show the location of segment 3. Because of this, and the possible ambiguity of the difference Fourier map, it would not be wise to go beyond making a speculative assignment for segment 3 at this stage. It may be possible to make a more definite assignment to the remaining three segments at some point in the future by producing multiple derivatives that have major labelling sites on one or more of these molecular

segments, or by improving the electron density map (if possible) so that the course of a whole collagen molecule could be traced.

In the meantime, if the location of the weak peak in the gold chloride difference Fourier map corresponds to the segment position labelled b' in Figure 5.15 and Table 5.1, then this would be the location of segment 3. Assuming that this is correct, since only one more segment assignment is needed before the whole structure is known, there are only two possible segment arrangements remaining. Of these two possible segment assignments, only one of them presents possible molecular translations that are cyclic (microfibrillar), and neither of them give rise to structures consistent with sheet models (linear translations).

The segment assignments consistent with microfibrillar models present two possible cyclic translations, both of which are 1D staggered (which is equivalent to 4D stagger). The significant difference between them is due to the non-standard quasi-hexagonal packing arrangement revealed by the electron density map. Segments 1 and 5 are closer together within the unit cell (as defined by Figures 5.15 and 5.17) than they are to segments 1 and 5 in neighbouring unit cells. This implies that the covalent crosslinking is responsible for the non-uniform packing distribution of the telopeptide-containing segments, and means that one of the cyclic structures of Figure 5.17 would give rise to intra-microfibrillar crosslinking and the other inter-microfibrillar crosslinking. This is an extremely important observation, since it shows that there is at least as much chance that microfibrils are actually covalently bound to one another as there is the possibility

that they are not. If the former is true, this might explain the difficulty researchers have encountered in trying to biochemically-isolate microfibrils.

Figure 5.17 Possible molecular topologies

The assignment of segments 1, 3 and 5 to unit cell positions as revealed by the electron density map (Figures 5.10 and 5.11) greatly reduces the number of possible conformations of the packing arrangement. Only two remaining segment assignments are possible, only one of which gives rise to cyclic progressions, neither assignment is consistent with linear progressions. This cyclic progression (shown here) is of the 1D/4D type, consistent with biochemical observations on intermolecular crosslinks (Bailey *et al.*, 1980). There are two possible progressions, one left handed (a;i and b) the other right handed (a; ii,iii, and c). The left handed progression presents a slightly compressed pentagonal structure, whilst the right handed progression shows the energetically more favourable compressed microfibril structure (Fraser *et al.*, 1983, Lee *et al.*, 1996, Wess *et al.*, 1998). The close proximity of the crosslinking segments is believed to be the result of the covalent bonding. This suggests that the compressed microfibril of (c) is crosslinked to neighbouring microfibrils rather than the intermolecular crosslinking being confined within the microfibril. This would explain why microfibrils have not as yet been isolated.

a) i, unit cell of collagen molecular packing showing left–handed connectivity.

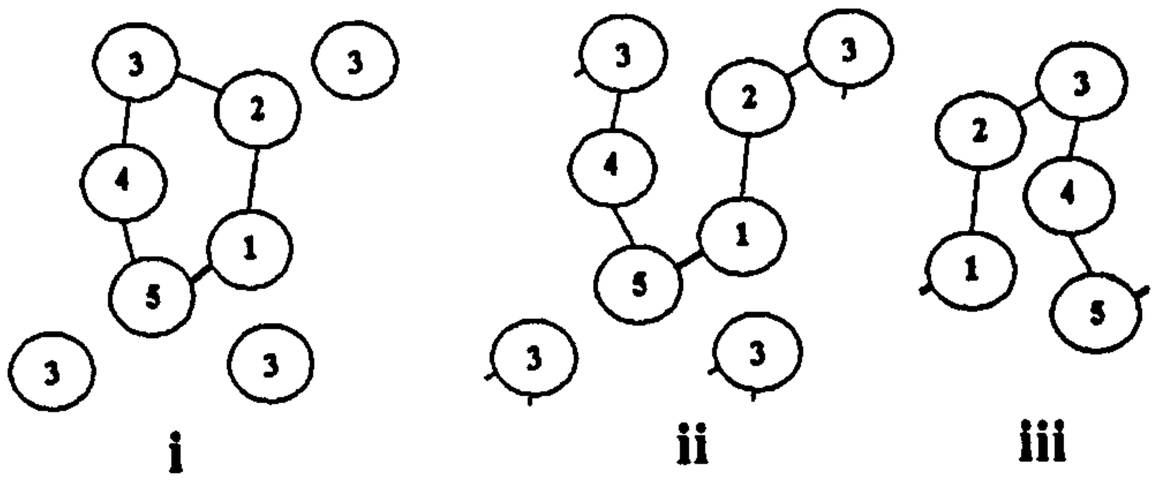
ii, as above except right handed connectivity.

iii, the right–handed 'microfibril'.

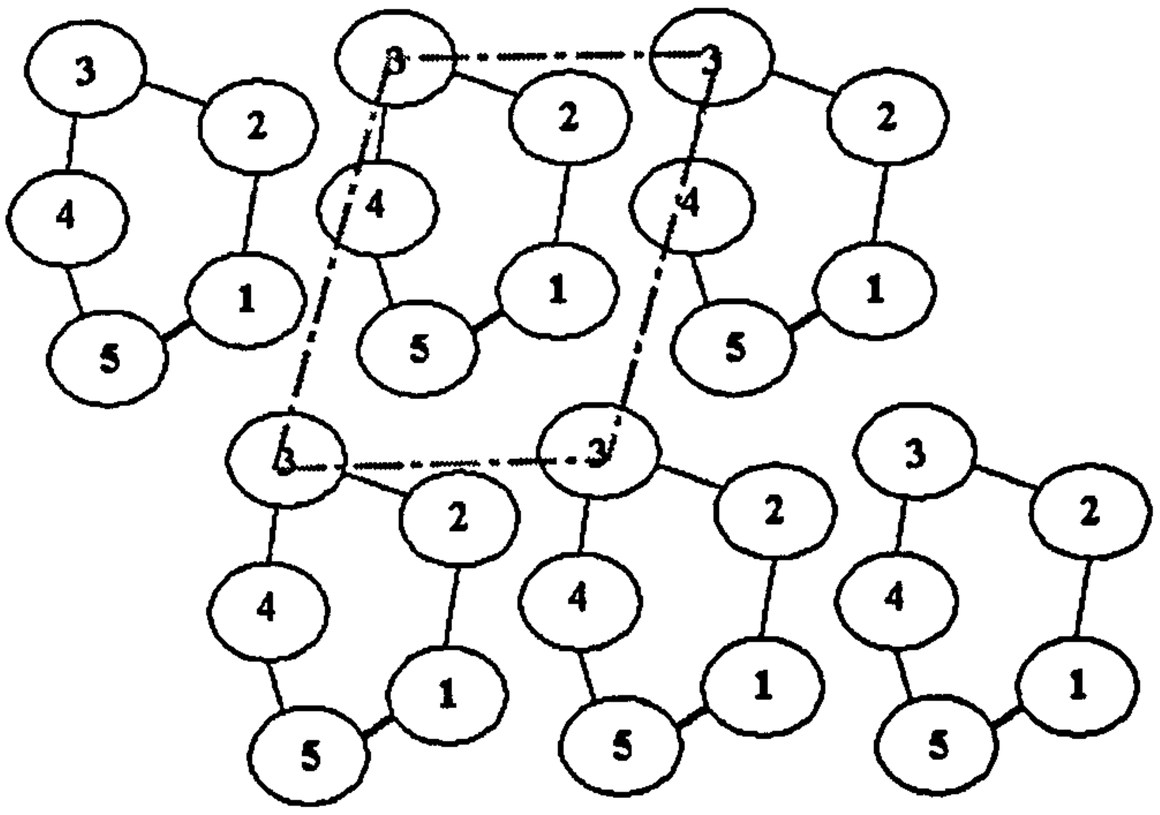
b) Quasi–hexagonal packing arrangement with left–handed topology/connectivity.

c) As above except right–handed topology/connectivity.

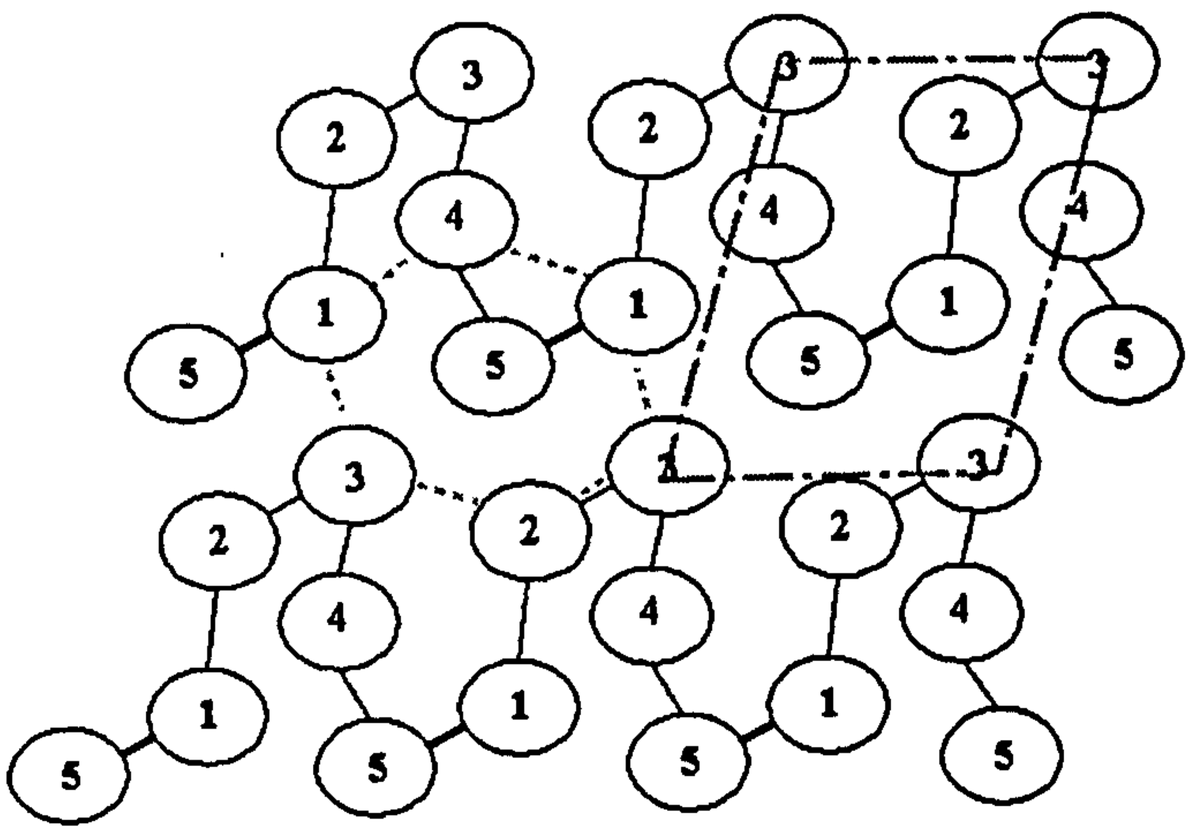
Crosslinks are assumed to be formed between segments 1 and 5 at their nearest approach.



a



b



c

..... Shows the quasi–hexagonal packing of collagen molecules. Each collagen chain is surrounded by six neighbours.

——— Shows crosslinks

——— Shows progression of a single collagen molecule through the molecular segment positions within each unit cell.

—·—·—·—·— Shows the unit cell

5.5 Conclusion

The hydrated fibrillar conformation of the three-dimensional packing arrangement in type I collagen has been defined as belonging to that of a compressed (non-standard packing coordinate) 1D-periodic microfibril type. The telopeptide-containing molecular segments have been identified and their coordinate positions determined, revealing that the intermolecular crosslinked molecular segments are packed closer together than the remaining three segments, particularly in the C-terminal region.

The possible interconnectivity of collagen microfibrils has been demonstrated by the observation that the formation of intermolecular crosslinks are at least as likely to occur outwith a single microfibril than as within. This arrangement of inter-microfibrillar crosslinking would in turn add a degree of integrity and strength to the fibril, as does the fact that the packing arrangement of segments 1 and 5 produces a pleated sheet arrangement of crosslinked molecules that must contribute a significant degree of stability and resistance to shearing forces within and outside of the fibril, see Figure 5.17

The methods employed to determine the three dimensional structure have proved extremely successful. It has been shown, for the first time, that it is possible to determine the phase components for structure factors where the amplitude components are found only in highly overlapping Bragg reflections. The production of an interpretable electron density map from the amplitudes obtained from the mixture of

diffuse scatter and highly overlapping Bragg reflections is compelling proof of the validity of the approach used to estimate the intensities.

It may be possible to improve the quality of the electron density map however. Work has already commenced to attempt to obtain improved diffraction patterns from fibrils that are cryofrozen. The freezing of the tendons in this way will allow the sample to be bathed in the X-ray beam for longer to obtain a higher signal to noise ratio, and further reduce the thermal disorder of the molecular system within the sample. It may be possible to reduce the disorder of the fibrillar packing arrangement using a series of freezing/partial thawing steps to anneal the packing system to a lower energy state.

An altogether different approach, still using X-ray fibre diffraction techniques, involves the attempts to obtain diffraction from a *single* fibril. This in effect would be true macromolecular crystallography, and would, if it can be made to work, produce diffraction patterns with improved spatial resolution of the Bragg peaks (reduced range of fibril orientations within the fibre – only one fibril). Initial experiments have shown the necessity of cryofreezing the sample whilst using powerful X-ray beams of small cross-sectional area (sub micron scale). As the number of unit cells sampled by the incident X-ray beams is diminished in samples of this size (micron scale), the exposure time needed to collect data is increased. The cumulative effect is the rapid destruction of delicate samples, and inspiration will be needed to be drawn from the related field of macromolcular crystallography for techniques that preserve the life-span of samples.

Both of these techniques involve treating the collagen sample in a way that falls outside of physiological conditions. It is possible that the freezing of sample in the annealing studies, or the disruption of the sample to obtain single fibrils would fundamentally change the molecular packing arrangement within the fibril. However, solutions devised from these and other means, can now be compared directly to the visualisation of the hydrated native protein presented here. The first structure determined by MIR for a natural fibre, and the first three dimensional visualisation of the 1D periodic lateral packing structure of collagen.

Chapter 6

Conclusions: A new understanding of collagen structure

6.1 Conclusion

The results of a detailed study of the molecular structure of fibrillar type I collagen have been presented here. The collection of data by use of X-ray fibre diffraction techniques and the maintenance of the samples in the hydrated state have revealed important information regarding the *in situ* structure of type I collagen fibrils within rat tail tendon. By calculating the phase component of the structure factors it has been possible to deduce the structure of collagen in an unambiguous way, producing a single interpretable solution for each of the two lines of investigation. These were:

1. A high angle axial diffraction study.

   High angle axial diffraction data were recorded to a resolution of 0.48 nm, and was used in conjunction with isomorphous derivative data to produce phase sets and an axial electron density profile of 0.54 nm resolution. The electron density map and the difference Fourier maps of the derivatives identify the axial position of the intermolecular attachment sites at high resolution, and the conformation of the telopeptides. The folded conformation of the C-telopeptide, and the contracted nature of the N-telopeptide bring the lysine residues present in both telopeptides into close axial alignment with the hydroxylysine residues in the helical region biochemically identified as being those involved in forming the intermolecular cross-links.

2. An X-ray diffraction study of the lateral packing arrangement of collagen.

   Superior quality equatorial data were obtained for native and derivative tendons. Modern algorithms were used for the removal of the diffuse background and

estimations of the highly overlapping Bragg reflections. From this, phase sets were calculated via MIA, and used to produce a partial visualisation of the overlap region in the vicinity of each telopeptide of anisotropic resolution. Producing the first ever visualisation of the lateral packing arrangement of type I collagen *in situ*. From this, it has been possible to deduce the coordinates of the cross-linking molecular segments and determine that these are nearest neighbours in the molecular packing arrangement. The quasi-hexagonal packing arrangement of type I collagen would seem to be a result of the formation of the intermolecular crosslinks at the telopeptides. This has direct significance on the molecular architecture of tissues in individuals suffering from diseases that affect the formation and stability of these cross-links, since a disruption to the packing arrangement of molecules in the formation of the tissue would imply a limitation to the thickness (or stability) of fibrils that could be grown.

On reflection of the significance of both these studies on the molecular structure of collagen, it is apparent that the role of the telopeptides as the sites of intermolecular cross-linking has been further shown to be all important. The unexpected finding of the folded conformation of the C-telopeptide is of particular significance. The crosslinking lysine residues in the C-telopeptide are present towards the end of the peptide sequence. Without the folded conformation, it might be imagined that the C-telopeptide would be relatively mobile. The folding of the telopeptide brings the tyrosine residues at both ends of the peptide into close axial alignment, and it is possible that the structure is stabilised by electrostatic interactions between them (tyrosine stacking).

The techniques used have made it possible to gain detailed information of type I

collagen *in situ*. It is hoped that these techniques could be further employed to push the resolution boundary of such *in situ* studies to that of approaching atomic resolution. In the meantime, it is apparent that these same techniques can be applied directly to other tissues. Lampray notochord (type II collagen) displays crystallinity similar to that shown by rat tail tendon, and various other type I collagen tissues. Further, the use of heavy atom stain to create isomorphous derivatives to a relatively high resolution can be applied to other tissue systems that show long range crystallinity in the axial plane, without the need to resort to model based approaches.

The study of the structure of type I collagen continues, and it is hoped that the details learnt here will be of use to researchers as they continue to try and elucidate its structure and the structures of similar systems.

Appendix 1 Amino acid sequence used as basis of model for 1D structure

The amino acid sequence has been organised into D-staggered segments for convenience of reference, the regions of compression discussed in section 4.5.3.1 and shown in Figure 4.11 are indicated with straight lines. Pro and Hyp have been represented with P** and H** respectively so as to make imino rich regions more immediately obvious. The protein sequence is based upon that of Chapman and Hulmes (1984), with modification to reflect available sequence data for rat, or when this was lacking, from available mouse amino acid sequence data (taken from the Swiss-Prot Expasy web site, www.expasy.ch). Where information was lacking regarding the post-transitional modification of amino acids for the rat or mouse genes, the corresponding amino acid from the Chapman and Hulmes sequence was used.

Sequence type	Accession no.	Sequence length	Notes
Rat α1(1) (fragment)	P02454	1-672	Amino acids 647-671 are actually from the C-telopeptide *not* from the main chain as the database record inaccurately implies
Mouse α1(1) (precursor)	P11087	1-1440	*
Rat α2(1) (fragment)	P02466	1-184	*
Mouse α2(1) (precursor)	Q01149	1-1373	*

i	D 1			D 2			D 3			D 4			D 5			I
1	Glu			Lys	Ala	Gly	Arg	Glu	Gly	Arg	P**	Gly	Thr	Asp	Gly	1
2	Met			Gly	Hyl	Ala	Gly	Lys	Glu	Gly	Arg	P**	Gly	Lys	Glu	2
3	Ser		Glu	Ala	Gly	Lys	Glu	Gly	Arg	Ala	Gly	Arg	Glu	Gly	Thr	3
4	Tyr		Met	Asn	Ala	Gly	Gln	Glu	Gly	Ala	Asp	Gly	Glx	Asp	Gly	4
5	Gly		Ser	Gly	Ala	Ala	Gly	Gln	Glu	Gly	Ala	Ala	Gly	Hyl	Glu	5
6	Tyr		Tyr	Ala	Gly	Asn	P**	Gly	Gln	P**	Gly	Ala	Asp	Gly	Glx	6
7	Asp		Gly	H**	Leu	Gly	Ala	P**	Gly	H**	P**	Gly	Arg	Ile	Gly	7
8	Glu		Tyr	Gly	H**	Ala	Gly	Ala	P**	Gly	Thr	P**	Gly	Hyl	Asp	8
9	Lys	Glu	Asp	Ile	Gly	H**	Ser	Gly	Ala	Ala	Gly	H**	Ile	Gly	Arg	9
10	Ser	Phe	Glu	Ala	Val	Gly	H**	P**	Gly	Thr	Met	Gly	Lys	Asp	Gly	10
11	Thr	Asp	Lys	Gly	Ala	Ile	Gly	H**	Ser	Gly	Thr	Ala	Gly	Arg	Ile	11
12	Gly	Ala	Ser	Ala	Gly	Ala	Phe	Gly	H**	Phe	Gly	Thr	His	Gly	Lys	12
13	Ile	Lys	Thr	H**	Ala	Gly	Gln	Phe	Gly	H**	Phe	Gly	Arg	His	Gly	13
14	Ser	Gly	Gly	Gly	H**	Ala	Gly	Gln	Phe	Gly	H**	Phe	Gly	Asn	His	14
15	Val	Gly	Ile	Phe	Gly	H**	Leu	Gly	Gln	Ala	Gly	H**	Phe	Gly	Arg	15
16	P**	Gly	Ser	H**	Leu	Gly	H**	Leu	Gly	Ala	Ala	Gly	Ser	Leu	Gly	16
17	Gly	P**	Val	Gly	H**	Phe	Gly	H**	Leu	Gly	Ala	Ala	Gly	Gln	Phe	17
18	P**	Gly	P**	Ala	Gly	H**	P**	Gly	H**	Arg	Gly	Ala	Leu	Gly	Ser	18
19	Met	P**	Gly	Arg	P**	Gly	Ala	P**	Gly	Val	Arg	Gly	Glx	Leu	Gly	19
20	Gly	Met	P**	Gly	Arg	Ala	Gly	Ala	P**	Gly	Thr	Arg	Gly	H**	Leu	20
21	P**	Gly	Met	P**	Gly	Arg	P**	Gly	Ala	P**	Gly	Val	P**	Gly	Glx	21
22	Ser	Leu	Gly	Ser	Ile	Gly	H**	P**	Gly	H**	P**	Gly	H**	Leu	Gly	22
23	Gly	Met	P**	Gly	H**	P**	Gly	Ala	P**	Gly	H**	P**	Gly	Ala	P**	23
24	P**	Gly	Ser	P**	Gly	Ser	Glu	Gly	H**	P**	Gly	H**	Ser	Gly	H**	24
25	Ala	P**	Gly	Glu	P**	Gly	Ala	Glu	Gly	Ser	P**	Gly	H**	His	Gly	25
26	Gly	Arg	P**	Gly	Val	P**	Gly	Ala	Glu	Gly	Ser	P**	Gly	His	Ser	26
27	Leu	Gly	Ala	P**	Gly	Glu	Lys	Gly	Ala	Asn	Gly	Ser	Ser	Gly	H**	27
28	H**	P**	Gly	Ser	Ala	Gly	H**	Lys	Gly	Ala	Leu	Gly	P**	Asp	Gly	28
29	Gly	H**	Leu	Gly	Ala	P**	Gly	H**	Lys	Gly	Thr	Asn	Gly	Gln	Ser	29
30	P**	Gly	H**	Ala	Gly	Ser	Glu	Gly	H**	P**	Gly	Ala	Glu	Gly	P**	30
31	H**	Ala	Gly	H**	Ala	Gly	Gln	Glu	Gly	H**	P**	Gly	Glx	Ala	Gly	31
32	Gly	Ser	P**	Gly	Thr	Ala	Gly	Arg	Glu	Gly	H**	P**	Gly	H**	Glu	32
33	Ala	Gly	H**	P**	Gly	H**	Val	Gly	Gln	P**	Gly	H**	P**	Gly	Glx	33
34	H**	Ala	Gly	Lys	Ala	Gly	H**	Leu	Gly	H**	P**	Gly	Ser	Ala	Gly	34
35	Gly	H**	Ala	Gly	Arg	P**	Gly	His	Val	Gly	H**	P**	Gly	Val	P**	35
36	P**	Gly	H**	Asn	Gly	Lys	Asp	Gly	H**	P**	Gly	H**	Ala	Gly	Ser	36
37	Gln	P**	Gly	Ser	Leu	Gly	Leu	Glu	Gly	Ala	Ala	Gly	Ser	P**	Gly	37
38	Gly	Gln	P**	Gly	Val	Asn	Gly	Phe	Asp	Gly	Ala	P**	Gly	Ala	Ala	38
39	Phe	Gly	Gln	Glu	Gly	Ser	Ala	Gly	Leu	Lys	Gly	Ala	P**	Gly	Ser	39
40	Gln	Phe	Gly	H**	Glu	Gly	H**	Leu	Gly	Glu	Lys	Gly	Ala	P**	Gly	40
41	Gly	Gln	Phe	Gly	H**	Glu	Gly	H**	Ala	Gly	Glu	Lys	Gly	Arg	P**	41
42	P**	Gly	Gln	Ala	Gly	H**	P**	Gly	H**	Gly	Gly	Glu	P**	Gly	Ala	42
43	H**	P**	Gly	H**	P**	Gly	Ser	P**	Gly	Lys	Ile	Gly	Arg	P**	Gly	43
44	Gly	H**	P**	Gly	Ala	Ala	Gly	Ala	P**	Gly	Arg	Gly	Gly	Ala	P**	44
45	Glu	Gly	H**	Asn	Gly	H**	Ala	Gly	Ser	P**	Gly	Lys	Ser	Gly	Arg	45
46	H**	Glu	Gly	Lys	Ser	Gly	Arg	Ala	Gly	Arg	P**	Gly	Ala	P**	Gly	46
47	Gly	H**	Glu	Gly	Hyl	Asn	Gly	Arg	Ala	Gly	Arg	P**	Gly	Ser	Ser	47
48	Glu	Gly	H**	Asp	Gly	Lys	Glu	Gly	Arg	Glu	Gly	Arg	Ser	Gly	Ala	48
49	H**	Glu	Gly	Thr	Glu	Gly	Arg	Glu	Gly	Thr	Asp	Gly	H**	P**	Gly	49
50	Gly	H**	Glu	Gly	Ser	Asp	Gly	Arg	Glu	Gly	Gln	Glu	Gly	Ala	Ser	50
51	Ala	Gly	H**	Ala	Gly	Thr	Phe	Gly	Arg	P**	Gly	Thr	Lys	Gly	H**	51
52	Ser	Gln	Gly	Lys	Asn	Gly	H**	P**	Gly	Ala	P**	Gly	Asp	Lys	Gly	52
53	Gly	Thr	Ala	Gly	Lys	Ala	Gly	H**	Phe	Gly	Val	P**	Gly	Ile	Lys	53
54	P**	Gly	Ser	Glu	Gly	Lys	Glu	Gly	H**	Arg	Gly	Ala	Leu	Gly	Asp	54
55	Met	P**	Gly	H**	Glu	Gly	Arg	Glu	Gly	H**	Arg	Gly	Asn	Arg	Gly	55
56	Gly	Ala	P**	Gly	H**	Glu	Gly	Arg	Glu	Gly	Ser	Arg	Gly	Ile	Leu	56
57	P**	Gly	Met	P**	Gly	H**	Val	Gly	Arg	Glu	Gly	H**	Leu	Gly	Asn	57
58	Arg	Ala	Gly	Ala	Ala	Gly	Glu	Ala	Gly	Val	Glu	Gly	H**	Ala	Gly	58
59	Gly	Arg	P**	Gly	Val	P**	Gly	Ala	Val	Gly	Thr	Glu	Gly	Val	Leu	59
60	P**	Gly	Arg	Val	Gly	Ala	P**	Gly	Glu	P**	Gly	Val	P**	Gly	H**	60
61	H**	P**	Gly	Gln	Gln	Gly	H**	P**	Gly	H**	Ala	Gly	Ile	P**	Gly	61
62	Gly	H**	P**	Gly	H**	Val	Gly	Thr	P**	Gly	Thr	P**	Gly	Ala	P**	62
63	P**	Gly	H**	P**	Gly	Gln	P**	Gly	H**	P**	Gly	H**	H**	Gly	Ile	63
64	H**	P**	Gly	H**	P**	Gly	Ala	P**	Gly	H**	P**	Gly	H**	Ala	Gly	64

65	Gly	H**	P**	Gly	H**	P**	Gly	Ile	P**	Gly	Thr	P**	Gly	Ala	H**						65
66	Lys	Gly	H**	P**	Gly	H**	P**	Gly	Ala	P**	Gly	H**	P**	Gly	H**						66
67	Asn	Lys	Gly	Ala	P**	Gly	Arg	Ser	Gly	Ala	Phe	Gly	Arg	Ile	Gly						67
68	Gly	Ala	Lys	Gly	Ser	P**	Gly	Arg	P**	Gly	Val	P**	Gly	Arg	P**						68
69	Asp	Gly	Asn	Glu	Gly	Ala	Asp	Gly	Arg	Glu	Gly	Ala	Arg	Gly	Arg						69
70	Asp	Glu	Gly	Glu	Glu	Gly	Asp	P**	Gly	Lys	Glu	Gly	Thr	Ser	Gly						70
71	Gly	Asp	Asp	Gly	Glu	Glu	Gly	Ser	Asp	Gly	Lys	Glu	Gly	Gln	Arg						71
72	Glu	Gly	Asp	Lys	Gly	Glu	Ala	Gly	Asp	Ser	Gly	Lys	Asp	Gly	Thr						72
73	Ala	His	Gly	Arg	Lys	Gly	H**	P**	Gly	H**	P**	Gly	Ala	Ser	Gly						73
74	Gly	H**	Glu	Gly	Arg	Lys	Gly	H**	Ala	Gly	Thr	Ser	Gly	Gln	Asp						74
75	Lys	Gly	Ala	Ala	Gly	Arg	Asn	Gly	H**	Ala	Gly	H**	P**	Gly	Ala						75
76	P**	Lys	Gly	Arg	Ser	Gly	Asp	P**	Gly	Asp	Glu	Gly	Ala	P**	Gly						76
77	Gly	P**	Lys	Gly	Thr	Ala	Gly	Asp	Asn	Gly	H**	Ala	Gly	Ala	P**						77
78	Arg	Gly	P**	Glu	Gly	Arg	Ala	Gly	Asp	P**	Gly	Asp	P**	Gly	Ala						78
79	H**	Arg	Gly	H**	Glu	Gly	Lys	Asn	Gly	Ala	Ser	Gly	H**	P**	Gly						79
80	Gly	H**	Arg	Gly	Ile	Glu	Gly	Lys	Ala	Gly	Ala	P**	Gly	H**	P**						80
81	Glu	Gly	H**	P**	Gly	H**	Asp	Gly	Lys	Ser	Gly	Ala	P**	Gly	H**						81
82	Arg	Glu	Gly	Ser	P**	Gly	Thr	Glu	Gly	H**	P**	Gly	H**	P**	Gly						82
83	Gly	Arg	Glu	Gly	Ala	P**	Gly	Ala	Asp	Gly	H**	Ser	Gly	H**	P**						83
84	P**	Gly	Arg	Leu	Gly	Ser	Ala	Gly	Thr	Thr	Gly	H**	P**	Gly	H**						84
85	H**	Val	Gly	H**	P**	Gly	H**	Ala	Gly	P**	P**	Gly	H**	P**	Gly						85
86	Gly	P**	P**	Gly	H**	Leu	Gly	Val	Ala	Gly	H**	Thr	Gly	H**	P**						86
87	P**	Gly	H**	P**	Gly	H**	Ala	Gly	H**	P**	Gly	P**	P**	Gly	H**						87
88	Gln	P**	Gly	H**	P**	Gly	H**	P**	Gly	Gln	P**	Gly	H**	P**	Gly						88
89	Gly	Gln	P**	Gly	H**	P**	Gly	Ala	Ala	Gly	Gln	P**	Gly	H**	P**						89
90	Ala	Gly	Gln	Glu	Gly	H**	Ser	Gly	H**	Ile	Gly	Gln	P**	Gly	H**						90
91	Arg	Ala	Gly	Arg	Leu	Gly	Gln	Ala	Gly	Ala	Leu	Gly	P**	P**	Gly						91
92	Gly	Arg	Ala	Gly	Arg	Glu	Gly	H**	Ser	Gly	Leu	Ile	Ser	P**	P**						92
93	Leu	Gly	Arg	Gly	Gly	Arg	Ala	Gly	Gln	Gln	Gly	Ala	Gly	Ser	P**						93
94	H**	Phe	Gly	H**	Asn	Gly	H**	P**	Gly	Arg	Ala	Gly	Gly	Gly	Ser						94
95	Gly	H**	Leu	Gly	H**	Gly	Gly	Ala	Ala	Gly	H**	Gln	Tyr	Gly	Gly						95
96	Thr	Gly	H**	Ser	Gly	H**	Leu	Gly	H**	Val	Gly	Arg	Asp	Tyr	Gly						96
97	Ala	Thr	Gly	Arg	Ser	Gly	Gln	P**	Gly	Val	Phe	Gly	Leu	Glu	Tyr						97
98	Gly	H**	Thr	Gly	Arg	Ser	Gly	H**	Leu	Gly	Leu	Val	Ser	Phe	Asp						98
99	Leu	Gly	Ala	Phe	Gly	Arg	Met	Gly	Gln	Leu	Gly	Val	Phe		Leu						99
100	H**	Leu	Gly	H**	Leu	Gly	H**	Ile	Gly	H**	Leu	Gly	Leu		Ser						100
101	Gly	H**	Leu	Gly	H**	Phe	Gly	H**	Met	Gly	H**	Leu	P**		Phe						101
102	Met	Gly	H**	Ala	Gly	H**	Glu	Gly	H**	Gln	Gly	H**	Gln		Leu						102
103	Hyl	Phe	Gly	Asp	Ala	Gly	Arg	Glu	Gly	Arg	Ser	Gly	P**		P**						103
104	Gly	Hyl	Met	Gly	Asp	Ala	Gly	Arg	Glu	Gly	Arg	Gln	P**		Gln						104
105	His	Gly	Hyl	Val	Gly	Asp	Ala	Gly	Arg	Glu	Gly	Arg	Gln		P**						105
106	Arg	Ile	Gly	Ala	Val	Gly	Ala	Val	Gly	Arg	Glu	Gly	Gln		P**						106
107	Gly	Arg	His	Gly	Ala	Val	Gly	Ala	Ala	Gly	Arg	Glu	Gln		Gln						107
108	Phe	Gly	Arg	P**	Gly	Ala	Leu	Gly	Ala	Phe	Gly	Arg	Lys		Gln						108
109	Ser	His	Gly	Lys	Val	Gly	H**	Val	Gly	H**	Leu	Gly	Ala		Gln						109
110	Gly	Asn	Phe	Gly	Met	P**	Gly	H**	Leu	Gly	H**	Phe	His		Lys						110
111	Leu	Gly	Ser	P**	Gly	Lys	P**	Gly	H**	Leu	Gly	H**	Asp		Ala						111
112	Asp	Leu	Gly	Ala	P**	Gly	Lys	Gly	Gly	H**	Val	Gly	Lys		His						112
113	Gly	Asp	Leu	Gly	Ala	P**	Gly	Lys	P**	Gly	Ala	Leu	Gly		Asp						113
114	Ala	Gly	Asp	Glu	Gly	Ala	Asp	Gly	Lys	P**	Gly	H**	Arg		Lys						114
115	Lys	Leu	Gly	Arg	Ser	Gly	Arg	Glu	Gly	Ser	Ser	Gly	Tyr		Gly						115
116	Gly	Thr	Ala	Gly	Arg	Glu	Gly	Lys	Asp	Gly	Val	P**	Tyr		Arg						116
117	Asn	Gly	Lys	Ser	Gly	Arg	Asp	Gly	Arg	Glu	Gly	Ser			Tyr						117
118	Thr	Gln	Gly	H**	Thr	Gly	Ala	Ala	Gly	H**	Glu	Gly			Tyr						118
119	Gly	H**	Asn	Gly	Ser	Ser	Gly	H**	Asp	Gly	H**	Glu									119
120	P**	Gly	Thr	P**	Gly	H**	P**	Gly	Ala	Lys	Gly	H**									120
121	Ala	Ala	Gly	Ala	P**	Gly	Lys	Leu	Gly	Gln	P**	Gly									121
122	Gly	H**	P**	Gly	Ala	P**	Gly	Arg	P**	Gly	Leu	Lys									122
123	P**	Gly	Ala	P**	Gly	Ala	Ala	Gly	Lys	P**	Gly	Gln									123
124	Lys	Val	Gly	Lys	Val	Gly	Asp	Asp	Gly	Ser	Ile	Gly									124
125	Gly	Hyl	P**	Gly	Arg	P**	Gly	Thr	Ala	Gly	Ala	P**									125
126	Glu	Gly	Lys	Ser	Gly	Lys	Ser	Gly	Asp	Ala	Gly	Ser									126
127	H**	Glu	Gly	H**	P**	Gly	P**	Ala	Gly	Ser	P**	Gly									127
128	Gly	H**	Glu	Gly	Asn	Ser	Gly	Thr	Ser	Gly	H**	Ala									128
129	Ser	Gly	H**	Glu	Gly	H**	Lys	Gly	P**	Glu	Gly	Ser									129
130	H**	Ala	Gly	Ala	Asp	Gly	Asp	Arg	Gly	Arg	Ala	Gly									130

131	Gly	H**	Ser	Gly	Ser	Glu	Gly	Asp	Lys	Gly	Arg	Glu	131
132	Glu	Gly	H**	Arg	Gly	Ala	Ala	Gly	Asp	P**	Gly	Arg	132
133	Asn	Glu	Gly	H**	Arg	Gly	Arg	Ala	Gly	H**	P**	Gly	133
134	Gly	Asn	Glu	Gly	H**	Arg	Gly	Arg	Ala	Gly	H**	P**	134
135	Ala	Gly	Asn	Glu	Gly	H**	Leu	Gly	Arg	P**	Gly	H**	135
136	H**	Thr	Gly	Ala	Glu	Gly	Thr	Leu	Gly	Met	Asn	Gly	136
137	Gly	H**	Ala	Gly	H**	Glu	Gly	H**	Leu	Gly	Val	P**	137
138	Gln	Gly	H**	Leu	Gly	Ala	P**	Gly	Thr	P**	Gly	Met	138
139	Met	Gln	Gly	H**	Leu	Gly	Ile	Ala	Gly	H**	Asn	Gly	139
140	Gly	Hyl	Gln	Gly	Met	Leu	Gly	Ile	P**	Gly	H**	P**	140
141	P**	Gly	Met	Ala	Gly	H**	P**	Gly	Ile	Leu	Gly	H**	141
142	Arg	Ala	Gly	Lys	P**	Gly	H**	Ala	Gly	Ala	Val	Gly	142
143	Gly	Arg	P**	Gly	Arg	Ala	Gly	H**	P**	Gly	Asn	Leu	143
144	Leu	Gly	Arg	Leu	Gly	Lys	P**	Gly	H**	P**	Gly	Ala	144
145	H**	Leu	Gly	Thr	Phe	Gly	Ala	P**	Gly	H**	Ala	Gly	145
146	Gly	H**	Leu	Gly	H**	Leu	Gly	Ala	P**	Gly	H**	P**	146
147	Glu	Gly	H**	Ser	Gly	Thr	Ala	Gly	Ala	Glu	Gly	H**	147
148	Arg	Glu	Gly	H**	Ser	Gly	H**	Gly	Gly	Ser	Glu	Gly	148
149	Gly	Arg	Glu	Gly	H**	Ser	Gly	Ala	Ala	Gly	Ala	Glu	149
150	Arg	Gly	Arg	Ser	Gly	H**	Asp	Gly	H**	Arg	Gly	Ser	150
151	H**	Arg	Gly	H**	Asn	Gly	Lys	Asp	Gly	Glu	Arg	Gly	151
152	Gly	Val	Arg	Gly	Ile	Ser	Gly	Arg	Asp	Gly	Asp	Arg	152
153	P**	Gly	H**	P**	Gly	H**	Glu	Gly	Lys	Ser	Gly	Glu	153
154	H**	Ala	Gly	Asp	P**	Gly	Ala	Glu	Gly	H**	Asn	Gly	154
155	Gly	H**	P**	Gly	Ala	P**	Gly	Gly	Glu	Gly	H**	Ser	155
156	Ser	Gly	H**	Lys	Gly	Asp	P**	Gly	Ala	Ala	Gly	H**	156
157	Ala	P**	Gly	Thr	Lys	Gly	Ser	P**	Gly	Glu	Asn	Gly	157
158	Gly	Ala	Ser	Gly	Glu	Lys	Gly	Ala	P**	Gly	Asp	Ala	158
159	Ala	Gly	Ala	P**	Gly	Thr	P**	Gly	Ser	Ser	Gly	Glu	159
160	Arg	Ala	Gly	H**	P**	Gly	P**	P**	Gly	H**	P**	Gly	160
161	Gly	Arg	Ala	Gly	Val	P**	Gly	Ala	P**	Gly	H**	Ser	161
162	Asp	Gly	Arg	P**	Gly	H**	P**	Gly	P**	Arg	Gly	H**	162
163	Asp	Ser	Gly	Ala	Leu	Gly	Thr	P**	Gly	Asp	Arg	Gly	163
164	Gly	Asp	Asp	Gly	H**	P**	Gly	Ala	P**	Gly	Asp	Arg	164
165	Ala	Gly	Asp	Glu	Gly	Ala	Ala	Gly	Thr	Ala	Gly	Asp	165
166	Val	Ser	Gly	Asp	Ile	Gly	Arg	Ala	Gly	H**	Gln	Gly	166
167	Gly	Val	Ala	Gly	Asp	Glu	Gly	Arg	Ala	Gly	H**	Ala	167
168	Ala	Gly	Val	Arg	Gly	Asp	Ala	Gly	Arg	Ala	Gly	H**	168
169	Ala	P**	Gly	H**	Arg	Gly	H**	Ile	Gly	Lys	His	Gly	169
170	Gly	Val	Ala	Gly	H**	Arg	Gly	Arg	Ala	Gly	Lys	Ala	170
171	P**	Gly	Ala	P**	Gly	H**	Asp	Gly	H**	Asp	Gly	Lys	171
172	H**	P**	Gly	Ala	P**	Gly	Arg	Glu	Gly	Arg	Glu	Gly	172
173	Gly	Ala	P**	Gly	Ile	P**	Gly	Arg	Asp	Gly	Arg	Asp	173
174	P**	Gly	H**	P**	Gly	Ala	Glu	Gly	Arg	Glu	Gly	Arg	174
175	Thr	P**	Gly	H**	P**	Gly	Ala	Glu	Gly	Thr	Thr	Gly	175
176	Gly	Ile	P**	Gly	Ala	P**	Gly	H**	Glu	Gly	H**	Glu	176
177	P**	Gly	Thr	Ala	Gly	H**	P**	Gly	Ala	P**	Gly	Thr	177
178	Thr	Ser	Gly	Arg	P**	Gly	H**	P**	Gly	Ala	Asn	Gly	178
179	Gly	Ala	P**	Gly	Arg	Ala	Gly	Val	P**	Gly	Ala	P**	179
180	P**	Gly	Thr	Gln	Gly	Arg	P**	Gly	H**	P**	Gly	Ala	180
181	H**	P**	Gly	Ala	Glu	Gly	Ala	P**	Gly	H**	P**	Gly	181
182	Gly	H**	P**	Gly	Ala	Gln	Gly	Ser	P**	Gly	H**	P**	182
183	Phe	Gly	H**	Val	Gly	Ala	Phe	Gly	Ala	Ala	Gly	H**	183
184	H**	Phe	Gly	Met	Ala	Gly	Ala	Phe	Gly	H**	Ala	Gly	184
185	Gly	H**	Phe	Gly	Ile	Val	Gly	Ala	Phe	Gly	Val	Ala	185
186	Ala	Gly	H**	Phe	Gly	Met	P**	Gly	Ala	Ala	Gly	H**	186
187	Ala	Ala	Gly	H**	Phe	Gly	H**	P**	Gly	H**	P**	Gly	187
188	Gly	H**	Ala	Gly	H**	Phe	Gly	Ala	P**	Gly	Val	Ala	188
189	Ala	Gly	Ala	P**	Gly	H**	Ala	Gly	H**	Ala	Gly	H**	189
190	Lys	P**	Gly	Lys	P**	Gly	Asp	Ala	Gly	P**	P**	Gly	190
191	Gly	Hyl	Ala	Gly	Lys	P**	Gly	Ala	Ala	Gly	Val	Ala	191
192	Glu	Gly	Lys	Thr	Gly	Lys	Gln	Gly	Asp	P**	Gly	P**	192
193	Ala	Glu	Gly	Ala	P**	Gly	H**	Gln	Gly	Val	Glu	Gly	193
194	Gly	Leu	Glu	Gly	Thr	Thr	Gly	H**	Gln	Gly	P**	P**	194
195	P**	Gly	Ala	Glu	Gly	Ala	Ala	Gly	H**	P**	Gly	Val	195
196	Glu	P**	Gly	H**	Glu	Gly	Lys	Ala	Gly	Ala	Lys	Gly	196

197	Gly	Val	P**	Gly	H**	Glu	Gly	Hyl	Ala	Gly	Leu	P**	197
198	Ala	Gly	Glu	Lys	Gly	H**	Glu	Gly	Lys	Lys	Gly	Ala	198
199	Arg	Asn	Gly	Ala	Lys	Gly	H**	Glu	Gly	Asn	Asn	Gly	199
200	Gly	H**	Ala	Gly	H**	Lys	Gly	Arg	Glu	Gly	Arg	Lys	200
201	Ser	Gly	Arg	Glu	Gly	Ala	Asp	Gly	H**	Asp	Gly	Asn	201
202	Glu	P**	Gly	Arg	Glu	Gly	Thr	Thr	Gly	Arg	Glu	Gly	202
203	Gly	Ala	Ser	Gly	Lys	Glu	Gly	Lys	Asp	Gly	H**	Asp	203
204	P**	Gly	Glu	Val	Gly	Arg	Val	Gly	Thr	Glu	Gly	Arg	204
205	Gln	P**	Gly	H**	Asn	Gly	Lys	P**	Gly	Thr	P**	Gly	205
206	Gly	Ala	P**	Gly	Val	Val	Gly	Val	Val	Gly	Ala	Glu	206
207	Val	Gly	Gln	P**	Gly	H**	Asp	Gly	Lys	P**	Gly	Thr	207
208	Arg	P**	Gly	H**	Leu	Gly	Ala	Glu	Gly	Ala	Ala	Gly	208
209	Gly	Arg	Val	Gly	Ala	P**	Gly	Gln	Asp	Gly	Val	P**	209
210	Glu	Gly	Arg	Ala	Gly	H**	P**	Gly	Ala	P**	Gly	Ala	210
211	H**	Glu	Gly	Val	Ala	Gly	H**	P**	Gly	Ala	P**	Gly	211
212	Gly	Val	Glu	Gly	Arg	Ala	Gly	Val	P**	Gly	Ala	P**	212
213	P**	Gly	H**	P**	Gly	Val	P**	Gly	H**	P**	Gly	Ala	213
214	H**	Leu	Gly	Ala	Ala	Gly	Ala	P**	Gly	Ile	Ala	Gly	214
215	Gly	H**	P**	Gly	H**	P**	Gly	Gln	P**	Gly	Val	P**	215
216	P**	Gly	H**	Lys	Gly	Ala	P**	Gly	Ala	P**	Gly	Ile	216
217	Ala	Leu	Gly	Asp	P**	Gly	Ala	P**	Gly	Ala	P**	Gly	217
218	Gly	Ser	P**	Gly	Asp	Lys	Gly	Val	P**	Gly	Arg	P**	218
219	Ala	Gly	Ala	Glu	Gly	Asp	P**	Gly	Ala	Ala	Gly	Ala	219
220	Ala	P**	Gly	Ala	Asn	Gly	H**	Ala	Gly	Arg	P**	Gly	220
221	Gly	Val	Ala	Gly	Asn	Glu	Gly	Ala	P**	Gly	Ser	Ala	221
222	P**	Gly	Ala	Ala	Gly	Ala	P**	Gly	H**	P**	Gly	Arg	222
223	Ala	P**	Gly	Gln	Ala	Gly	Ile	P**	Gly	Ala	P**	Gly	223
224	Gly	H**	P**	Gly	Gln	Ala	Gly	H**	P**	Gly	Gln	P**	224
225	Asn	Gly	Ala	Ala	Gly	Gln	Asn	Gly	Ile	P**	Gly	Ala	225
226	H**	Asn	Gly	H**	P**	Gly	Val	P**	Gly	Glx	Ile	Gly	226
227	Gly	Ala	Asn	Gly	H**	Ala	Gly	Val	Asn	Gly	Arg	P**	227
228	Ala	Gly	H**	P**	Gly	H**	Ala	Gly	Val	P**	Gly	Glx	228
229	Asp	P**	Gly	Ala	Leu	Gly	H**	Ala	Gly	Arg	Asp	Gly	229
230	Gly	Asn	Ala	Gly	Gln	P**	Gly	Ala	Ala	Gly	Asp	P**	230
231	Gln	Gly	Asp	P**	Gly	Ala	P**	Gly	H**	Asp	Gly	Arg	231
232	H**	Leu	Gly	Ala	Val	Gly	Hyl	P**	Gly	Lys	Glu	Gly	232
233	Gly	H**	Gln	Gly	Gln	P**	Gly	Ala	P**	Gly	H**	Asp	233
234	Ala	Gly	H**	Glu	Gly	Ala	P**	Gly	Hyl	Glu	Gly	Lys	234

Appendix 2 A summary of the search for isomorphous derivatives

Multiple isomorphous replacement/addition (MIR/A) techniques require the modification of the sample in such as way as to alter the unit cell contents but to leave the parameters of the unit cell unchanged. More than one such derivative is needed, and the greater the number of isomorphous derivative data sets available, the greater the degree of confidence in the calculation of the native phases.

Several chemical modification regimes were used in the attempt to produce isomorphous derivatives, the criteria used to determine the isomorphic nature of a derivative being that the relative intensities of the Bragg peaks of the collagen fibre diagram be altered without changing the positions of the Bragg reflections (evidence of distortions to the crystal lattice).

The chemical species used, the buffer system and the success of the experiment are listed in the table below:

CHEMICAL SPECIES	PRE-TREATMENT (IF ANY)	BUFFER SYSTEM	ISOMORPHOUS OR REASON FOR UNSUITABILITY
Platinum chloride $PtCl_4$	None	PBS (pH 7.5)	Little or limited evidence of labelling (inconsistent results), evidence of small distortions to crystal lattice.
Platinum diammine chloride Cis or trans $Pt(NH_3)Cl_4$	None	PBS (pH 7.5)	See above
Cerium tri-chloride $Ce\ Cl_3$	None	TBS (pH 7.5)	Significant disruption to the crystal lattice (disappearance of low-angle equatorial Bragg reflections).
Gadolinium tri-chloride $Gd\ Cl_3$	None	TBS (pH 7.5)	Isomorphic at low resolution, but evidence of a large number of labelling positions – difficult to determine labelling sites.
Gold Chloride $AuCl_4$	None	PBS (pH 7.5)	Consistent, repeatable isomorphic labelling – no distortion to low or high angle diffraction pattern.
Potassium Iodide K I	None	PBS (pH 7.5)	Consistent, repeatable isomorphic labelling – no distortion to low or high angle diffraction pattern.
Potassium Iodide K I	UV irradiation $\lambda = 254$ nm	PBS (pH 7.5)	Consistent, repeatable isomorphic labelling – no distortion to low or high angle diffraction pattern, different labelling sites to that of unmodified iodide derivative.
Potassium Iodide K I	DEPC	PBS (pH 7.5)	Consistent, repeatable isomorphic labelling – but inconclusive as to whether the derivative is unique (not the same as the non-pretreated iodide derivative).

Because only two heavy atom stain treatments were initially successful, it was attempted to produce further isomorphous derivatives by chemical pre-treatment of a specimen before exposure to one of two working heavy atom stains.

Iodide is believed to bond covalently with tyrosine residues, although iodohistidine can also form under the same reaction conditions (Means and Feeney 1971). It was

speculated that it might be possible to block the formation of iodohistidine through the use of Diethylpyrocarbonate DEPC in one derivative, and prevent the covalent bonding of iodide to some tyrosine residues by photo-treatment. These two additional derivatives would be expected to have fewer stain vectors than the original iodide derivative, and provide further clarification in the phase calculation.

DEPC pre-treatment was performed by diluting DEPC to 5% vol/vol in PBS and immersing a tendon in the solution for 4 minutes with periodic vigorous shaking. The tendon was then immediately treated with iodide stain. Histidine would be expected to be significantly discriminated against labelling due to the ortho- substitution of the imidazole group (Kyle 1995), although it does not seem as if this treatment has been performed for an insoluble protein before.

The pre-stain photo-treatment was produced by exposing tendons to UV light before iodination. Exposure of samples to UV light of wave length λ=254nm (4 Watt lamp, distance of 4 cm) was performed in the attempt to cause free radical attack principally upon tyrosine residues, although histidine and methionine could be expected to be significantly susceptible also (Guptasarma *et al.*, 1992).

Specimens pretreated by UV exposure and DEPC treatment were examined visually and X-ray patterns obtained. There appeared to be no significant differences between the meridional intensities of these specimens and that of native protein data sets.

Iodination of the two pretreated samples was performed and the data carefully

examined. The DEPC treated iodination appeared to show little, if any, difference from that of the non-pretreated iodine derivative, and it was feared that the pretreatment was ineffective. The DEPC derivative was therefore not used in the phase calculation. The UV pretreatment did however show significant differences from that of the native and iodine derivatives, and evidence of constant labelling at high resolution.

With gold chloride, known to bind with cysteine, methionine, and histidine (only the latter two are found in type I collagen), and the two iodine derivatives (one produced through UV pretreatment before staining, one produced through derivative labelling only), three different derivatives with a limited number of chemically predictable stain vectors were produced. Hence phase calculations could be performed with reasonable confidence.

Appendix 3 Publications

"The *in situ* conformation and axial location of the intermolecular crosslinked non-helical telopeptides of type I collagen". J. Orgel, T. Wess, A. Miller. Structure with folding and design. 8, 137-142 (2000).

"X-ray studies on biological fibres". A. Miller, J.P., Orgel, and T.J., Wess. Fibre diffraction review. 8, 27-31 (2000).

"The three dimensional molecular packing of native type I collagen". Joseph P.R.O. Orgel, Andrew Miller, Tom Irving, Robert Fischetti, Andrew P. Hammersley, and Tim J. Wess. Under review, Nature, June 2000.

"Heat induced changes in collagen molecular packing". T. Wess and J. Orgel. Accepted for *Thermachimica acta*, June 2000.

Conferences presentations

CCP13 conference 1998 (poster).

CCP13 conference 1999 (conference speaker).

IUCr congress 1999 (poster).

Bibliography

Ackerman, M.S., Bhate, M., Shenoy, N., Beck, K., Ramshaw, J.A.M., Brodsky, B., (1999). "Sequence dependence of the folding of collagen-like peptides". J. Biol. Chem. 274, 7668-7673.

Astbury, W.T. (1933). "Some problems in the X-ray analysis of the structure of animal hairs and other protein fibers". Trans. Faraday. Soc. 29, 193-211.

Bailey, A.J., Light, N.D., and Atkins, E.D.T., (1980). "Chemical cross-linking restrictions models for the molecular organization of the collagen fiber". Nature, 288, 408-413.

Barnes, M.J., (1988). Collagens of normal and diseased blood vessel wall. In: Collagen, Vol1: Biochemistry (Nimni, M.E., ed), CRC Press, Boca Raton, p275-290.

Bartlett, M.W., Egelstaff, P.A., Holden, T.M., Stinson, R.H., and Sweeny, P.R., (1973). "Structural changes in tendon collagen resulting from muscular dystrophy". Biochim. Biophy. Acta. 328, 213-220.

Bear, R.S., (1942). "Long X-ray diffraction spacings of collagen". J. Am. Chem. Soc. 64, 727. (One page only).

Bear, R.S., (1944). "X-ray diffraction studies on protein fibers. I. The large fiber-axis period of collagen". J. Am. Chem. Soc. 66, 1297-1305.

Bear, R.S., (1952). "The structure of collagen fibrils". Adv. Prot. Chem. 7, 69-160.

Beck, K., Brodsky, B., (1998). "Supercoiled protein motifs: The collagen triple-helix and the α-helical coiled coil". J. Struct. Biol. 122, 17-29.

Bella, J., Brodsky, B., Berman, H.M., (1995). "Hydration structure of a collagen peptide". Structure, 3, 893-906.

Bella, J., Eaton, M., Brodsky, B., Berman, H.M., (1994). "Crystal-structure and molecular-structure of a collagen-like peptide at 1.9-angstrom resolution". Science, 266, 75-81.

Bella, J., Brodsky, B., Berman, H.M., (1996). "Disrupted collagen architecture in the crystal structure of a triple-helical peptide with a Gly->Ala substitution". Conn. Tiss. Res. 35, 455-460.

Blundell, T.L., and Johnson, L.N., (1976). "Protein crystallography". Academic Press, London.

Bornstein, P., and Traub, W., (1979). The chemistry and biology of collagen. In The proteins, third edition, vol. 4 (Neurath, H., Hill, R.L., eds.), Academic Press, London. p411-632.

Bowes, J.H., and Kenton, R.H., (1948). "The amino-acid composition and titration curve of collagen". Biochem. J. 43, 358-365.

Bradshaw, J.P., Miller, A., Wess T.J., (1989). "Phasing the meridional diffraction pattern of type I collagen using isomorphous derivatives". J. Mol. Biol. 205, 685-694.

Bragg, W.L., (1913). "The diffraction of short electromagnetic waves by a crystal". Proc. Camb. Phil. Soc. 17, 43-57.

Brodsky, B., (1999). "Hydrogen bonding in the triple-helix". Proc. Ind. Acad. Sciences. Chem. Sciences. 111, 13-18.

Brodsky, B., and Eikenberry, E.F., (1982). "Characterization of fibrous forms of collagen". In Methods in Enzymology, (Cunningham, L., and Fredriksen, D., eds.), Academic press, New York. Vol. 82. p127-174.

Brodsky, B., and Ramshaw, J.A.M., (1997). "The collagen triple-helix structure". Matrix Biology, 15, 545-554.

Bruckner, P., Eikenberry, E.F., and Prockop, D.J., (1981). "Formation of the triple helix of type I procollagen incelluo, a kinetic model based on cis-trans isomerization of peptide bonds". Eur. J. Biochem. 118, 607-613.

Bruns, R.R., and Gross, J. (1973). "Band pattern of the Segment-Long-Spacing form of collagen. Its use in the Analysis of primary structure". Biochemistry, 12, 808-815.

Bruns, R.R., and Gross, J., (1974). "High resolution analysis of the modified quarter stagger model of the collagen fibril". Biopolymers, 13, 931-941.

Burgeson, R.E., Morries, N.P., Murray, L.W., Duncan, K.G, Keene, D.R., and Sakai, L.Y., (1985). "The structure of type VII collagen". Ann. NY. Acad. Sci. 460,47-57.

Chan, V.C., Ramshaw, J.A.M., Kirkpatrick, A., Beck, K., Brodsky, B., (1997). "Positional preferences of ionizable residues in Gly-X-Y triplets of the collagen triple-helix". J. Biol. Chem. 272, 31441-31446.

Chan, V.C., Ramshaw, J.A.M., Shah, N.K., Yang, W., Brodsky, B., (1996). "The effects of polar/ionizable residues on the triple helix". Matrix Biology, 15, 152-153.

Chapman, J.A., and Hulmes, D.J.S., (1984). "Ultrastructure of the connective tissue matrix". (Ruggeri, A., and Motta, P.M., eds), Martinus Nijhoff, Boston. p1-33.

Chapman, J.A., (1984). Molecular organisation in the collagen fibril. In Connective tissue Matrix. (Hukins, D.W.L., ed), Macmillan, New York, p89-132.

Chapman, J.A., and Hardcastle, R.A., (1974). "Staining pattern of collagen fibrils. II. Comparison with patterns computer-generated from the amino acid sequene". Conn. Tiss. Res. 2, 151-159.

Clark, G.L., and Schaad, J.A., (1936). "X-ray diffraction studies of tendon and intestinal wall collagen". Radiology, 27, 339-356.

Clark, G.L., Parker, E.A., Schaad, J.A., Warren, W.J., (1935). "New measurements of previously unknown large interplaner spacings in natural materials". J. Amer. Chem. Soc. 57, 1509. (one page only).

Cochran, W., Crick, F.H.C., Vand, V. (1952). "The structure of synthetic polypeptides (I): The transforms of atoms on a helix". Acta. Cryst. 5. 581-586.

Cohen, C., and Bear, R.S., (1953)."Helical polypeptide configuration in collagen". J. Am. Chem. Soc. 75, 2783-2784.

Cohn, D.H., Apone, S., Eyre, D.R., Starman, B.J., Andreassen, P., Charbonneau, H., Nicholls, A.C., Pope, F.M., Byers, P.H, (1988). "Substitution of cysteine for glycine within the carboxy-terminal telopeptide of the α1 chain of type I collagen produces mild Osteogenesis Imperfecta". J. Biol. Chem. 263, 14605-14607.

Corey R.B., and Wyckoff, R.W.G., (1936). "Long spacings in macromolecular solids". J. Biol. Chem. 114, 407-414.

Cott, P.G., (1986). "Spectroscopic study of environment-dependent changes in the conformation of the isolated carboxy-terminal telopeptide of type I collagen". Biochemistry, 25, 974-980.

Cowan, P.M., North, A.C.T., and Randall, J.T., (1955). "X-ray diffraction studies of collagen fibres". Symp. Soc. Exp. Biol. 9, 115-126.

Dickerson, R.E., Kopka, M.L., Varnum, J.C., and Weinzierl, J.E. (1967). "Bias, feedback, and reliability in isomorphous phase analysis". Acta. Cryst. 23, 511-522.

Doyle, B.B., Hulmes, D.J.S., Miller, A., Parry, D.A.D., Peiz, K.A., and Woodhead-Galloway, J., (1974a). "Axially projected collagen structure". Proc. Roy. Soc. ser. B. 187, 37-46.

Doyle, B.B., Hukins, D.W.L., Hulmes, D.J.S., Miller, A., Rattew, C.J., Woodhead-Galloway, J., (1974b). "Origins and implications of the D stagger in collagen". Biochem. Biophy. Res. Comun. 60, 858-864.

Doyle, B.B., Hukins, D.W.L., Hulmes, D.J.S., Miller, A., Woodhead-Galloway, J., (1975). "Collagen polymorphism: Its origins in the amino acid sequence". J. Mol. Biol. 91, 79-99.

Eikenberry, E.F., Childs, B., Sheren, S.B., Parry, D.A.D., Craig, A.S., and Brodsky, B., (1984). "Crystalline structure of type II collagen in lamprey notochord sheath". J. Mol. Biol. 176, 261-277.

Engel, J., and Prockop, D., (1991). "The zipper-like folding of collagen triple helices and the effects of mutations that disrupt the zipper". Ann. Rev. Biophys. Biophys Chem. 20, 137-152.

Ericson, L.G., and Tomlin, S.G., (1959). "Further studies of low-angle X-ray diffraction patterns of collagen". Proc. Roy. Soc. Ser. A., 252, 197-216.

Ewald, P.P., (1921). "Das 'reziproke Gitter' in der strukturtheorie". Zeitschrift für Kristallographie. 56, 129-156.

Eyre, D.R., (1987). "Collagen cross-linking amino acids". Methods Enzymo. 144, 115-139.

Eyre, D.R., Paz, M.A., and Gallop, P.M., (1984). "Cross-linking in collagen and elastin". Ann. Rev. Biochem. 53, 717-748.

Fan, P., Li, M.H., Brodsky, B., Baum, J., (1993). "Backbone dynamics of (Pro-Hyp-Gly)10 and a designed collagen-like triple-helical peptide by 15N NMR Relaxation and hydrogen-exchange measurements". Biochemistry, 32, 13299-13309.

Fietzek, P.P., and Rexrodt, F.W., (1975). "The covalent structure of collagen. The amino-acid sequence of α2-CB4 from Calf-Skin collagen". Eur. J. Biochem. 59, 113-118.

Finkenstadt, V.L., and Millane, R.P., (1998). "Fiber diffraction patterns for general unit cells: the cylindrically projected reciprocal lattice". Acta Cryst. A, 54, 240-248.

Franc, S.J., (1993). "Ultrastructure evidence of a distinct axial domain within native rat tail tendon collagen fibrils". J. Submicrosc. Cytol. Pathol. 25, 85-91.

Fraser, R.D.B., and MacRae, T.P., (1981). "Unit cell and molecular connectivity in tendon collagen". Int. J. Biol. Macromol. 3, 193-200.

Fraser, R.D.B., MacRae, T.P., Miller, A., and Rowlands, R.J., (1976). "Digital processing of fibre diffraction patterns". J. Appl. Crystallogr. 9, 81-94.

Fraser, R.D.B., MacRae, T.P., Moller, A., and Suzuki, E., (1983). "Molecular conformation and packing in collagen fibrils". J. Mol. Biol. 167, 497-521.

Fraser, R.D.B., Macrae, T.B. and Miller, A. (1987). "Molecular packing in type I collagen fibrils". J. Mol. Biol. 193, 115-125.

Fraser, R.D.B., Macrae, T.P., and Suzuki, E., (1979). "Chain conformation in the collagen molecule". J. Mol. Biol. 129, 463-481.

Fratzl, P., Fratzl-Zelman, N., and Klaushofer, K., (1993). "Collagen packing and mineralization". Biophys. J. 64, 260-266.

Galloway, J., (1985). Structure of collagen fibrils. In, Biology of invertebrate and lower vertebrate collagens (Bairati, A., and Garrone, R., eds). Plenum Publishing Corp. New York. 73-82.

Gelman, R.A., and Piez, K.A., (1980). "Collagen fibril formation in vitro. A quasielastic light-scattering study of early stages". J. Biol. Chem. 255, 8098-8102.

George, A., Malone, J.P., Veis, A., (1999). "The secondary structure of type I collagen N-telopeptide as demonstrated by Fourier transform IR spectroscopy and molecular modelling". Proc. Ind. Acad. Sci. Chem. Sci. 111, 121-131.

Green, D.W., Ingram, V.M., Perutz, M.F., (1954). "Sign determination by the isomorphous replacement method". Proc. Roy. Soc. A 225, 287-307.

Grynpas, M., (1977). "Three dimensional packing of collagen in bone". Nature. 265, 381-382.

Guptasarma, P., Blasubramanian, D., Matsugo, S., Saito, I., (1992). "Hydroxyl radical mediated damage to proteins, with special reference to the crystallins". Biochemistry. 31, 4296-4303.

Halfter, W., Dong, S.C., Schurer, B., Cole, G.J., (1998). "Collagen XVIII is a basement membrane heparan sulfate proteoglyan". J. Biol. Chem. 273, 25404-25412.

Harker, D., (1956). "The determination of the phases of the structure factors of noncentrosymmetric crystals by the method of double isomorphous replacement". Acta Cryst. 9, 1-9.

Hanson, D.A., and Eyre, D.R., (1996). "Molecular site specificity of pyridinoine and pyrrole cross-links in type I collagen of human bone". J. Biol. Chem. 271, (43), 26508-26516.

Helseth, D.L., and Veis, A., (1981). "Collagen self-assembly *in vitro*. Differentiating specific telopeptide dependent interactions using enzyme modifications and the addition of free amino telopeptide". J. Biol, Chem., 256, 7118-7128.

Henkel, W., and Glanville, R.W., (1982). "Covalent crosslinking between molecules of type I and type III collagen". Eur. J. Biochem. 122, 205-213.

Hodge, A.J., and Petruska, J.A., (1962). In "Electron Microscopy" (Breese, S.S. Jr, ed.), vol.1 Paper QQ-1. Academic Press, New York.

Hodge, A.J., and Petruska, J.A., (1963). In "Aspects of Protein Structure" (Ramachandran, G. N., ed.), Academic Press, New York. p. 289-300.

Hofmann, H., Fietzek, P.P., Kuhn, K., (1980). "Comparative analysis of the sequences of the three collagen chains α1(I), α2 and α1(III); function and genetic aspects". J. Mol. Biol. 141, 293-314.

Holmes, K.C., and Blow, D.M., (1966). "The use of X-ray diffraction in the study of protein and nucleic acid structure". Interscience publishers, London.

Hukins, D.W.L., (1981). "X-ray diffraction; By disordered and ordered systems". Pergamon Press (Oxford).

Hukins, D.W.L., and Woodhead-Galloway, J., (1977). "Collagen fibrils as examples of smectic A biological fibres". Mol. Cryst. Liq. Cryst. 41, 33-39.

Hulmes, D.J., (1992). "The collagen superfamily - diverse structures and assemblies". Essays Biochem. 27, 49-67.

Hulmes, D.J.S., and Miller, A., (1979). "Quasi-hexagonal molecular packing in collagen fibrils". Nature. 282, 878-880.

Hulmes, D.J.S., Miller, A., Parry, A.D., Piez, K.A., and Woodhead-Galloway, J., (1973). "Analysis of the primary structure of collagen for the origins of molecular packing". J. Mol. Biol. 79, 137-148.

Hulmes, D.J.S., Holmes, D.F., and Cummings, C., (1985). "Crystalline regions in collagen fibrils". J. Mol. Biol. 184:473-477.

Hulmes, D.J.S., Miller, A., White, S.W., and Brodsky-Doyle, B., (1977). "Interpretation of the meridional X-ray diffraction pattern from collagen fibres in terms of the known amino acid sequence". J. Mol. Biol. 110, 643-666.

Hulmes, D.J.S., Miller, A., White, S.W., Timmins, P.A. Berthet-Colominas, C., (1980). "Interpretation of the low-angle meridional neutron diffraction patterns from collagen fibres in terms of the amino acid sequence". Int. J. Biol. Macromol. 2, 338-345.

Hulmes, D.J.S., Wess, T.J., Prockop, D.J., and Fratzl, P., (1995). "Radial packing, order and disorder in collagen fibrils". Biophys. J. 68, 1661-1670.

Hulmes, D.J.S., Jesior, J.-C., Miller, A., Berthet-Colominas, C., and Wolff, C., (1981). "Electron microscopy shows periodic structure in collagen fibril cross-sections". Proc. Ntl. Acad. Sci. USA-Biol. Sci. 78, 3567-3571.

Hulmes, D.J.S., Mould, A.P., Kadler, K.E., Chapman, J.A., Prockop, D.J., (1989). Procollagen processing control of type I collagen fibril assembly. In Cytoskeletal and extracellular matrix proteins. Structure, interactions and assembly. (Aebi, U., Engel, J., eds.), Springer, Berlin, p292-301.

International Tables for X-ray Crystallography (Henry, N.F.M., and Lonsdale, K., eds). Volume I. Symmetry Groups. Kynoch Press, 1952.

Jelinski, L.W., Sullivan, C.E., and Torchia, D.A., (1980). "2H NMR study of molecular motion in collagen fibrils". Nature, 284, 531-534.

Jesior, J.C., Miller, A., and Berthet-Colominas, C., (1980). "Crystalline three dimensional packing is a general feature of type I collagen fibrils". FEBS Letters. 13, 238-240.

Jones, E.Y., and Miller, A., (1987). "Models for the N-terminal and C-terminal telopeptide regions of interstitial collagens". Biopolymers. 26, 463-480.

Jones, E.Y., and Miller, A., (1991), "Analysis of structural design-features in collagen". J. Mol. Biol., 218, 209-219.

Juvonen, M., Sandberg, M., Pihlajaniemi, T., (1992). "Patterns of expression of the 6 alternatively spliced exons affecting the structures of the COL1 and NC2 domains of the alpha-1(XIII) collagen chain in human tissues and cell-lines". J. Biol. Chem. 267, 24700-24707.

Kadler, K.E., (1995)."Extracellular matrix.1. Fibril-forming collagens". Protein Profile. 2, 491-619.

Kajava, A.V., (1991). "Molecular packing in type I collagen fibrils". J. Mol. Biol. 218, 815-823.

Kastelic, J., and Baer, E., (1980). Deformation in tendon collagen. In The mechanical properties of biological materials. (Vincent, J.R.V., and Curry, J.I., eds.), Cambridge University Press, Cambridge. p397-435.

Kastelic, J., Galelski, A., and Baer, E. (1978). "The multicomposite structure of tendon". Conn. Tiss. Res. 6. 11-23.

Katsura, N., Osamu, T., and Yokoyama, M., (1991). "Three dimensional structure of type I collagen and mineralization". Connect. Tissue. 22, 92-98.

Kefalides, N.A., (1971). "Isolation of a collagen from basement membranes containing three identical α-chains". Biochem. Biophys. Res. Comm. 45, 226-234.

Kielty, C.M., Hopkinson, I., Grant, M.E., (1993). Collagen: The collagen family: structure, assembly, and organization in the extracellular matrix. In connective tissue and its heritable disorders. Wiley-liss Inc. (London). p103-147.

Kivirikko, K.I., and Myllyla, R., (1989). "The biosynthesis of collagen: intracelluar enzymes". Methods Enzymol, 82, 245-304.

Klein, T.E., Huang, C.C., (1999). "Computation investigations of structural changes resulting from point mutations in a collagen-like peptide". Biopolymers, 49, 167-183.

Kramer, R.Z., Bella, J., Mayville, P., Brodsky, B., Berman, H.M., (1999). "Sequence dependent conformational variations of collagen triple-helical structure". Nature. struc. Biol. 6, 454-457.

Kramer, R.Z., Vitagliano, L., Bella, J., Berisio, R., Mazzarella, L., Brodsky, B., Zagari, A., Berman, H.M., (1998). "X-ray crystallographic determination of a collagen-like peptide with the repeating sequence (Pro-Pro-Gly). J. Mol. Biol. 280, 623-638.

Kuivaniemi, H., Tromp, G., and Prockop, D.J., (1991). "Mutations in collagen genes - Causes of rare and some common diseases in humans". FASEB J. 5, 2052-2060.

Kyle, J., (1995). "Structure in protein chemistry". Garland Pub. inc. New York and London. P364.

Last, J.A., Armstrong, L.G., Reiser, K.M., (1990). "Biosynthesis of collagen crosslinks". Int. J. Biochem. 22, 559-564.

Lamande, S.R., and Bateman, J.F., (1995). "The type-I collagen pro-alpha-1(I) COOH-terminal propeptide N-linked oligosaccharide - functional-analysis by site-directed mutagenesis". J. Biol. Chem. 270,17858-17865.

Le Pape, A., Guitton, J.-D., and Muh, J.-P. (1984). "Distribution of non-enzymatically bound glucose *in vivo* and *in vitro* glycosylated type I collagen molecules". FEBS letters. 170, 23-27.

Lee, J., Scheraga, H.A., Rackovsky, S., (1996). "Computational study of packing a collagen-like molecule: Quasi-hexagonal vs 'Smith' collagen microfibril model". Biopolymers. 40, 595-607.

Li, M.H., Fan, P., Brodsky, B., and Baum, J., (1993). "Two-dimensional NMR assignments and conformation of (Pro-Hyp-Gly)10 and designed collagen triple-helical peptide". Biochemistry. 32, 7377-7387.

Linsenmayer, T.F., Chen, Q., Gibney, E., Gordon, M.K., Marchant, J.K., Mayne, R, and Schmd, T.M. (1991). "Collagen Type-IX and Type-X in the developing chick tibiotarsus – analyses of messenger-RNAs and proteins". Development. 111, 191-196.

Linsenmayer, T.F., Fitch, J.M., and Birk, D.E. (1990). "Heterotypic collagen fibrils and stabilizing collagens - controlling elements in corneal morphogenesis". Annals of the New York Academy of science. 580, 143-160.

Long, C.G., Braswell, E., Zhu, D., Apigo, J., Baum, J., Brodsky, B., (1993). "Characterization of collagen-like peptides containing interruptions in the repeating Gly-X-Y sequence". Biochemistry. 32, 11688-11695.

McBride, D.J., Choe, V., Shapiro, J.R., Brodsky B., (1997). "Altered collagen structure in mouse tail tendon lacking the alpha2(I) chain". J. Mol. Biol. 270, 275-284.

McRee, D.E., (1993). "Practical Protein Crystallography". Academic Press, San Diego CA.

McRee, D.E., (1999). "XtalView/Xfit - A versatile program for manipulating atomic coordinates and electron density". J. Struct. Biol. 125,156-165.

Meek, K.M., Chapman, J.A., and Hardcastle, R.A., (1979). "The staining pattern of collagen fibrils". J. Biol. Chem. 254, 10710-10714.

Means, G.E., and Feeney, R.E., (1971). "Chemical modification of proteins". Holden-Day, San Francisco. p175-182.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E., (1953). "Equation-of-state calculations by fast computing machines". J. Chem. Phys. 21, p1087. (one page only).

Miller, A. (1976). "Biochemistry of collagen" (Ramachandran, G. N., and Reddi, A., eds), Plenum Press, New York. p85-136.

Miller, A., and Parry, D.A., (1973). "Structure and packing of microfibrils in collagen". J. Mol. Biol. 75, 441-447.

Miller, A., and Tochetti, D., (1981). "Calculated X-ray diffraction pattern from a quasi-hexagonal model for the molecular arrangement in collagen". Int. J. Macromol. 3, 9-18.

Miller, A., and Wray, J. S., (1971). "Molecular packing in collagen". Nature. 230, 437-439.

Monboisse, J.C., and Borel, J.P., (1992). Oxidative damage to collagen. In Free Radicals and Ageing. (Emerit, I., and Chance, B., eds), Birkhauser Verlag Basel, Switzerland. p323-327.

Morgan, P.H., Jacobs, H.G., Segrest, J.P., and Cunningham, L.W., (1970). "A comparative study of glycopeptides derived from selected vertebrate collagens". J. Biol. Chem. 245, 5042-5048.

Moro, L., and Smith, B. D., (1977). "Identification of Collagen α1(I) Trimer and normal type I collagen in a polyomaric virus-induced mouse tumour". Arch. Bioch. Bioph. 182, 33-41.

Nagarajan, V., Kamitori, S., Okuyama, K., (1998). "Crystal structure analysis of collagen model peptide (Pro-Pro-Gly)(10)". J. Biochemistry, 124, 1117-1123.

Nagarajan, V., Kamitori, S., Okuyama, K., (1999). "Structure analysis of a collagen-model peptide with a (Pro-Hyp-Gly) sequence repeat". J. Biochemistry, 125, 310-318.

Nakamura, Y. (1987), "Structure of type I collagen dimers", Int. J. Biol. Macromol. 8, 281-290.

Nemetschek, T., Grassman, W., Hoffman V., (1955). "The highly subdivided cross-striations of collagen". Zeitschrift fur. Naturforschung. B. 10, 61-68.

Nimni, M.E., and Harkness, R.D., (1988). Molecular structures and Functions of collagen. In: Collagen, Vol1: Biochemsitry (Nimni, M.E., ed), CRC Press, Boca Raton, p1-78.

Nold, J.G., Kang, A.H., Gross, J., (1970). "Collagen molecules: Distribution of alpha chain". Science, 170, 1096-1098.

North, A.C., Cowan, P.M., and Randall, J.T., (1954). "Structural units in collagen fibrils". Nature. 174, 1142-1143.

Okuyama, K., Arnott S., Takayanagh, M., Kakudo M., (1981). "Crystal and molecular structure of collagen-like polypeptide (Pro-Pro-Gly)10". J. Mol. Biol. 152, 427-443.

Okuyama, K., Nagarajan, V., and Kamitori, S., (1999). "7/2-helical model for collagen - evidence from model peptides". Proc. Ind. Acad. Sci. Chem. Sci. 111, 19-34.

Orgel, J.P., Wess, T.J., and Miller, A., (2000). "The *in situ* conformation and axial location of the intermolecular cross-linked non-helical telopeptides of type I collagen". Structure with folding and design. 8, 137-142.

Otter, A., and Scott, P.G., (1988). "Type-I collagen alpha1 chain C-telopeptide - solution structure determined by 600-MHZ proton NMR-spectroscopy and implications for its role in collagen fibrillogenesis". Biochemistry. 27, 3560-3567.

Paterlin, M.G., Nemethy, G., Scheraga, H.A., (1995). "The energy of formation of internal loops in triple-helical collagen polypeptides". Biopolymers. 35, 607-619.

Phillips, G.N., Fillers, J.P., and Cohen, C., (1980). "Motions of tropomyosin. Crystals as metaphor". Biophys. J. 32, 485-502.

Piez, K.A., (1997). "History of extracellular matrix: A personal view" Matrix Biology, 16, 85-92.

Piez, K. A., and Trus, B. L., (1981). "A new model for packing of type I collagen molecules in the native fibril". Biosci. Rep. 1, 801-810.

Piez, K.A., Eigner, E.A., and Lewis, M.S., (1963). "The chromatographic separation and amino acid composition of the subunits of several collagens". Biochemistry. 2, 58-66.

Piez, K.A., Lewis, M.S., and Martin, G.R., (1961). "Subunits of the collagen molecule". Biochim. Biophys. Acta. 53, 596-599.

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T., (1989). Numerical Recipes; The art of scientific computing. Cambridge University Press, Cambridge. p326-334.

Prockop, D.J., and Fertala, A., (1998). "The collagen fibril: the almost crystalline structure". J. Struc. Biol. 122, 111-118.

Prockop, D.J., Berg, R.A., Kivirikko, K.I., and Utto, J., (1976). "Intracellular steps in the biosynthesis of collagen". (Ramachandron, G.N., Reddi, A.H., eds.), Plenum, New York. p163-273.

Purslow, P.P., Wess, T.J., and Hukins, D.W.L. (1998). "Collagen orientation and molecular spacing during creep and stress-relaxation in soft connective tissues". J. Exp. Biol. 201, 135-142.

Ramachandran, G.N., (1967). "Treatise on Collagen", vol. 1 (Ramachandran, G.N., ed.), Plenum Press, New York. p 45-84.

Ramachandran, G.N., and Kartha, G., (1954). "Structure of collagen". Nature. 174, 269-270.

Ramachandran, G.N., and Sasisekharan, V. (1968) Adv. Protein Chem. 23, 283-437.

Ramshaw, J.A.M., Shah, N.K., Brodsky, B., (1998). "Gly-X-Y tripeptide frequencies in collagen: A context for host-guest triple-helical peptides". J. Struc. Biol. 122, 86-91.

Raspanti, M., Ottani, V., and Ruggeri, A., (1989). "Different architectures of the collagen fibril: Morphoiogical aspects and functional implications". Int J. Biol. Macromol. 11,367-371.

Rich, A., and Crick, F.W.C., (1955). "The structure of collagen". Nature. 176, 915-916.

Rich, T., and Crick, F.H.C., (1961). "The molecular structure of collagen". J. Mol. Biol. 3, 483-506.

Rosenberg, H., Modrak, J.B., Hassing, J.M., Al-Turk, W.A., and Stohs, S.J., (1979). "Glycosylated collagen ". J. Biochem. Biophys. Res. Comm. 91, 498-501.

Rowe, R.W.D. (1985). "The structure of rat tail tendon". Conn. Tissue. Res. 14, 9-20.

Ruggeri, A., Benazzo, F., Reale, E., (1979). "Collagen fibrils with straight and helicoid microfibrils: a freeze fracture and thin section study". J. Ultrastruct. Res. 68, 101-108

Ryhanen, L., Zaragoza, E.J., and Uitto, J., (1983). "Conformational stability of type-1 collagen triple helix - evidence for temporary and local relaxation of the protein conformation using a proteolytic probe". Arch. Bioch. Bioph. 223, 562-571.

Sarker, S.K., Sullivan, C.E., and Torchia, D.A., (1983). "Solid state 13C NMR study of collagen molecular dynamics in hard and soft tissues". J. Biol. Chem. 258, 9762-9767.

Schacke, H., Schumann, H., Hammami-Hausali N., Raghunath M., Bruckner-Tuderman L., (1998). "Two forms of collagen XVII in Keratiocyte". J. Biol. Chem. 273, 25937-25943.

Schmitt, F.O., Gross, J., Highberger, J.H., (1955). "Tropocollagen and the properties of fibrous collagen". Exp. Cell. Res. Suppl. 3, 326-334.

Schmitt, F.O., Hall, C.E., Jakus, M.A., (1942). "Electron microscope investigations of the structure of collagen". J. Cell. Comp. Physiol. 20, 11-33.

Silver, D., Miller, J., Harrision, R., and Prockop, D.J., (1992). "Helical model of nucleation and propagation to account for the growth of type I collagen fibrils from the symmetrical pointed tips: a special example of self assembly of rod like monomers". Proc. Natl. Acad. Sci. USA. 89, 9860-9864.

Smith, J.W., (1968). "Molecular packing in native collagen". Nature. 219, 157-158.

Spiro, R.G., (1969). "Characterization and quantitative determination of the hydroxylysine-linked carbohydrate units of several collagens". J. Biol. Chem. 244, 602-612.

Timpl, R., and Engel, J., (1987). Type VI collagen. In Structure and function of collagen types. (Mayne, R., Burgeson, R.E., eds), Academic press, Orlando. p105-153.

Tomlin, S.G., and Worthington, C.R., (1956). "Low-angle X-ray diffraction patterns of collagen". Proc. Roy. Soc. ser. A. 235, 189-201.

Torchia, D.A., (1982). "Solid-state NMR studies of molecular motion in collagen fibrils". In Methods of Enzymology. Vol. 82. (Cunningham, L., and Fredriksen, D., eds), Academic Press, New York. 174-186

Trus, B.L., and Piez, K.A., (1980). "Compressed microfibril models of the native collagen fibril". Nature. 286, 300-301.

Uzawa, K., Grzesik, W.J., Nishiura, T., Kuznetsov, S.A., Robey, P.G., Brenner, D.A., and Yamauchi, M. (1999). "Differential expression of human lysyl hydroxylase genes, lysine hydroxylation, and cross-linking of type I collagen during osteoblastic differentiation *in vitro*". J. Bone. Min. Res. 14, 1272-1280.

Vainshtein, B.K., (1966). "Diffraction of X-rays by chain molecules". Elsevier.

Vanderrest, M., Dublet, B., Labourdette, L., Ricardblum, S., (1999). "Mechanisms of collagen trimer assembly". Proc. Ind. Aca. Sci. Chem. Sci. 111, 105-113.

Venugopal, M.G., Ramshaw, J.A.M., Braswell, E., Zhu, D., Brodsky, B., (1994). "Electrostatic interactions in collagen-like triple-helical peptides". Biochemistry, 33, 7948-7956.

Vitagliano, L., Nemethy, G., Zagari, A., Scheraga, H.A. (1993). "Stabilization of the triple-helical structure of natural collagen by side-chain interactions". Biochemistry, 32, 7354-7359.

Vitagliano, L., Nemethy, G., Zagari, A., Scheraga, H.A. (1995). "Structure of the type I collagen molecule based on conformational energy computations: The triple-stranded helix and the N-terminal telopeptide". J. Mol. Biol. 247, 69-80.

Vogel, H.,G., (1978). "Influence of maturation and age on mechanical and biochemical parameters of connective tissue of various organs in the rat". Conn. Tiss. Res. 6, 161-166.

Weckmann, A.L., and Cabral, A.R., (1996). "New molecular and clinical aspects of collagens". Revista De investigacion Clinica. 48, 207-221.

Wess, T.J., Miller, A., and Bradshaw, J. P., (1990). "Cross-linkage sites in type I collagen fibrils studied by neutron diffraction". J. Mol. Biol. 213, 1-5.

Wess, T.J., Wess, L., Miller, A., Lindsay, R.M., and Baird, J.D., (1993). "The *in Vivo* glycation of diabetic tendon collagen studied by neutron diffraction". J. Mol. Biol. 230, 1297-1303.

Wess, T.J., Hammersley, A., Wess, L., Miller, A., (1995). J. Mol. Biol. "Type I collagen packing, conformation of the triclinic unit cell". 248, 487-493.

Wess, T.J., Hammersley, A., Wess, L., Miller, A., (1998a). J. Mol. Biol. "Molecular packing of type I collagen in tendon". 275, 255-267.

Wess, T.J., Hammersley, A., Wess, L., Miller, A. (1998b). "A consensus model for molecular packing of type I collagen". J. Struct. Biol. 122, 92-100.

Whittaker, E.J.W. (1955). "The diffraction of X-rays by a cylindrical lattice. II". Acta Crystallogr. 8, 261-265.

Woodhead-Galloway, J., and Machin, P., (1976). "Modern theories of liquids and the diffuse equatorial x-ray scattering from collagen". Acta Cryst. A. 32,368-372.

Woodhead-Galloway, J., and Young, H. (1978). "Probabilistic aspects of the structure of the collagen fibril". Acta. Cryst. A. 34, 12-18.

Woolfson, M.M., (1970). "Introduction to X-ray crystallography". Oxford University Press, London.

Wyckoff, R.W.G., Corey, R.B., Biscoe, J., (1935). "X-ray reflections of long spacing from tendon". Science, 82, 175-176.

Yang, W., Battineni, M.L., Brodsky B., (1997). "Amino acid sequence environment modulates the disruption by osteogenesis imperfecta glycine substitutions in collagen-like peptides". Biochemistry. 36, 6930-6935.

Yang, W., Chan, V.C., Kirkpatrick, A., Ramshaw, J.A.M., Brodsky B., (1997). "Gly-Pro-Arg confers stability similar to Gly-Pro-Hyp in the collagen triple-helix of host-guest peptides". J. Biol. Chem. 272, 28837-28840.

Yonath, A., and Traub, W., (1969). "Polymers of tripeptides as collagen models; Structure analysis of Poly(L-prolyl-glycyl-L-proline)". J. Mol. Biol. 43, 461-477.

Yurchenco, P.D., and Schittny, J.C., (1990). "Molecular architecture of basement-membranes". FASEB Journal. 4, 1577-1590.