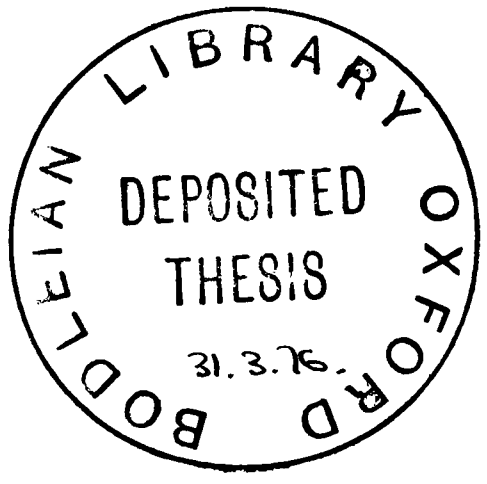
STRUCTURAL STUDIES

ON

FIBROUS BIOLOGICAL MATERIAL



Wolfson College,

Oxford

D.J.S. Hulmes

November 1975

ABSTRACT

This work is concerned with the structure, self-assembly and sequence of collagen. The complete amino acid sequence of the α1 chain, as found in the collagen triple-helical molecular structure, is used in the following three ways. First, the electron microscopic banding patterns, of the many ordered forms of intermolecular collagen assembly, are simulated on the basis of the known distribution of charged residues in the sequence (Chapters 2 and 3). In this way, various models, for the axially projected structures of these forms, are tested. The result is a set of experimentally determined intermolecular staggers or overlaps. Secondly, the distribution of residues along the sequence is analysed to find the origins of collagen self-assembly (Chapters 2 and 3). The specificity for each structure is due to the patterns of interacting residues along the sequence. Different residues, hence different distributions, seem to be important for each structure, depending on the physico-chemical conditions of the mainly in vitro precipitations. Thirdly, the sequence is used to predict the meridional X-ray and neutron diffraction patterns from rat tail tendon (Chapter 4). There is substantial agreement with the observed patterns. This creates the possibility of interpreting the changes in meridional intensities, in terms of the sequence, from collagen fibrils in a variety of tissues and in many metabolic states.

ACKNOWLEDGEMENTS

As far as is possible, the following pages describe my contribution to a group research project concerning molecular packing in collagen. The work would not have been possible without the stimulus and encouragement of the following people: Dr. A. Miller, my supervisor, to whom I am also grateful for reading the manuscript, Dr. J. Woodhead-Galloway, Dr. B. Brodsky-Doyle, Dr. D.A.D. Parry, Dr. K.A. Piez, Dr. C.J. Rattew and Dr. D.W.L. Hukins. Most of the content of Chapters 2 and 3 has been published, with some of the above people as co-authors, and a list of publications is included at the end of this thesis.

I am grateful to Prof. D.C. Phillips and the Medical Research Council for the opportunity to carry out this work in the Laboratory of Molecular Biophysics, University of Oxford. Other people in the laboratory, to whom I am particularly grateful for their help, are Mr. S. White, Miss C. Spencer, Dr. K.R. Leonard, Mr. G.S. French, Dr. J.M. Thornton, Dr. C.D. Rodger, Mrs B.M. Luke, Mrs C.D. van Horn, Mr. A.J. Allen and Mr. C. Blackmore.

Outside the laboratory, I am grateful for discussions with Dr. J.A. Chapman, Mrs P.M. Armitage and Dr. R.D.B. Fraser. Professors B.R. Olsen and K. Kuhn kindly provided electron micrographs for Chapter 3. The neutron diffraction work, described in Chapter 4, was possible only with the help of Dr. A. Miller and Dr. B. Brodsky-Doyle from Oxford and Dr. G.T. Jenkin, Dr. J.W. White, Dr. K. Ibel, Dr. J. Haas, Dr. D. Worcester, Dr. R. Ghosh and colleagues from the ILL, Grenoble.

I am grateful to Wolfson College, Oxford, for a Graduate Award during the latter two years of this work.

This manuscript was skilfully and efficiently typed by Mrs H.M. Holloway.

Lastly, I am indebted to my wife, Pat, for her continual encouragement, patience and help in the preparation of this thesis.

D.J.S.H.

CONTENTS

		Page
ABSTRACT		II
ACKNOWLEDGEMENTS		III
CHAPTER 1	SCOPE OF THE WORK	
	1. Introduction	1
	2. Connective Tissue	1
	3. The Structure of Collagen Assemblies	7
	i) The molecule	7
	ii) The fibril	13
	4. Self-Assembly	21
	5. Sequence	26
CHAPTER 2	THE INTERMOLECULAR D STAGGER IN COLLAGEN FIBRILS	
	1. Introduction	31
	2. Analysis of the Sequence for the Origins of the D stagger	32
	i) Interactions	32
	ii) Method	33
	iii) Results	37
	iv) Discussion	46
	3. Electron Microscopy and the Amino Acid Sequence	49
	i) Introduction	49
	ii) Methods and Results	51
	iii) Discussion	54
	4. Patterns of Residues within the α1 Sequence	56
	i) Introduction	56
	ii) Hydrophobic residues	58
	iii) Charged residues	66
	iv) Other residues	70
	v) Discussion	73
	5. Summary	78
CHAPTER 3	COLLAGEN POLYMORPHISM: ITS ORIGINS IN THE AMINO ACID SEQUENCE	
	1. Introduction	80
	2. The Axially Projected Molecular Arrangements	82
	i) Fibrous long spacing collagen (FLS)	82
	ii) D-periodic symmetric collagen (DPS)	95
	iii) The obliquely striated and 110 Å periodic forms	100
	3. Origins of the Polymorphism in the Sequence	103
	i) DPS	104
	ii) FLS	110
	iii) SLS	115

		Page
	4. Discussion	116
	i) Bridged interactions	116
	ii) Collagen proteoglycan interaction *in vivo*	118
	iii) Collagen morphology in the presence of other connective tissue components	122
	iv) Some general observations on specificity	123
CHAPTER 4	THE CORRELATION OF X-RAY AND NEUTRON DIFFRACTION PATTERNS OF NATIVE COLLAGEN WITH THE AMINO ACID SEQUENCE	
	1. Introduction	126
	2. Historical Background	128
	3. Electron Density Distributions based on the Amino Acid Sequence	137
	i) Parameters used to describe model distributions	137
	ii) The effect of varying the model parameters on the predicted meridional intensities	140
	iii) Difference Fouriers	153
	iv) Refinement	165
	v) Discussion	168
	4. Neutron Diffraction from Collagen	177
	i) Introduction	177
	ii) Data collection	179
	iii) Results	182
	iv) Calculated neutron structure factor amplitudes	183
	v) Discussion	189
	5. Conclusion	197
APPENDIX 1		200
APPENDIX 2		205
REFERENCES		210
LIST OF PUBLICATIONS		220

CHAPTER 1

SCOPE OF THE WORK

1. Introduction

This thesis is devoted to the study of one particular fibrous biological material: collagen, the main protein constituent of animal connective tissues. More precisely the work is concerned with the inter-relationship between the following three aspects of this material:-

a) Structure i.e. the nature of the three-dimensional packing of collagen molecules within the various assemblies observed, both *in vivo* and *in vitro*.

b) Self-assembly i.e. the way in which the genetic information involved in the biosynthesis of collagen is used to specify the formation of the macromolecular assemblies.

c) Sequence of amino acid residues within the polypeptide chains of each collagen molecule, this being the direct manifestation of the genetic information.

These three aspects will be described more fully in the remainder of this Chapter, but first it is necessary to put the collagen problem in perspective in terms of its role in connective tissue.

2. Connective Tissue

Most multicellular living organisms require some sort of supporting framework to maintain their structural integrity. There are three types of system which fulfil this role in nature:-

a) cellulosic - peculiar to plants

b) collagenous - peculiar to animals

c) chitinous - present in animals and fungi.

Though these systems are each referred to by one type of macromolecule, it would be an over-simplification to consider them to be in such an homogeneous state. In all cases, the connective tissue is an extracellular association of a main component, e.g. protein (collagen) or polysaccharide (cellulose, chitin) with other types of molecule, e.g. glycosaminoglycans, glycoproteins, proteoglycans, minerals and water. The main component is more readily detected in electron micrographs as it is always present in the form of very long (~1 μ) filaments of varying diameters (~ 100Å) embedded in a seemingly amorphous matrix.

Despite their different chemical compositions the supporting structures of plants and animals can show a striking similarity. Both algal cell walls and the basement membranes of amphibians, for example, are composed of layers of lamellae each containing parallel fibrils which are stacked with a 90° rotation between sheets (see Mathews, 1975; Preston et al., 1948; Weiss, 1956). This alternating ply arrangement is distributed widely throughout the metazoa (Mathews, 1975; Picken, 1960; Fitton-Jackson, 1968). The following are a few examples:-

a) Coelenterates - in the mesoglea of sea anemones the neighbouring lamellae are rotated by 45°.

b) Echinoderms - orthogonal ply.

c) Nematodes, nemerteans, annelids - as members of these phyla have a cylindrical form, the alternating ply in the basement

membrane of the epidermis leads to the formation of two sets of helices of opposite hand with their axes parallel to the main axis of the animal (Harkness, 1961).

d) Vertebrates - the transparency of the cornea is due to an orthogonal ply composed of lamellae with parallel, uniformly thick collagen fibrils (Maurice, 1969).

Of course, by no means all of the collagenous tissues display such a highly ordered arrangement of fibrils. This level of organisation represents the result of adaptation of each tissue to its particular function, e.g. transparency of the cornea. A tissue such as hyaline cartilage has an apparently random arrangement of collagen fibrils, thus enabling it to resist the forces of shear and compression. When the biological function is resistance to tensile forces in a certain direction, as in tendon, the fibrils are aligned parallel to that direction. Weiss (1959) has shown that the application of tension to a regenerating wound will stimulate newly formed fibrils to become aligned with the lines of stress. Mechanical factors can therefore influence morphology directly (Rodbard, 1970).

It has been suggested (Muir, 1964; Jackson and Bentley, 1968) that the variation in properties of the collagenous connective tissues is a function of their different chemical compositions. The collagen content, for example, of a few vertebrate tissues is listed in Table 1.1 (Harkness, 1961; Mathews, 1975). The remainder of these tissues is made up of water, minerals (in particular, hydroxyapatite in bone (Glimcher and Krane, 1968)) and the so-called ground substance. The latter is a mixture of polysaccharide (glycosaminoglycans) and non-collagenous protein which is usually covalently linked together

TABLE 1.1 Collagen content of various collagenous tissues (from Harkness (1961) and Mathews (1975))

Tissue	Percentage of collagen on a fresh weight basis
Tendon	25-35
Bone, cartilage	10-20
Cornea	15
Blood vessel walls	5-12
Vitreous humour	0.01-0.05

to form either proteoglycans (when polysaccharide predominates) or glycoproteins (when protein predominates). Glycosaminoglycans are linked polymers of repeating disaccharides with the general formula $(A\text{-}B\text{-})_n$ where A is a hexosamine with its amino group either acetylated or sulphated, and B is a neutral sugar or hexuronic acid. The composition and distribution of the main types of glycosaminoglycans amongst vertebrate connective tissues is shown in Table 1.2 (Mathews, 1975). The diversity in glycosaminoglycans is not limited to the above list, for two reasons. First, the polymers may be of variable length and, secondly, the repeating disaccharide may not be identical throughout the whole length of the chain. Dermatan sulphate, for example, contains D-glucuronic acid to some extent and its molecular weight can vary between 15×10^3 and 40×10^3.

As mentioned earlier, the glycosaminoglycans are usually found *in vivo* covalently linked to protein. A model for proteoglycan structure in vertebrate cartilage has been proposed by Mathews and Lozaityte (1958) and Mathews (1971). It consists of a linear polypeptide of 2000-3000 amino acid residues (length ~4000Å) with paired sets of side-chains of chondroitin 4 (or 6) - sulphate containing 10-100 disaccharide units. The glycoproteins consist of shorter, branched polysaccharide chains of hexosamines and neutral sugars (~10 residues) covalently linked to protein (Mathews, 1975). The importance of proteoglycans and glycoproteins in 'cementing' collagen fibrils together has been stressed by Jackson and Bentley (1968) and by Mathews (1965). These components may also have a role in bringing about the orderly arrangement of fibrils in the cornea (Weiss, 1956). Furthermore they may help regulate fibril size in different tissues. The range of fibril widths in some human

TABLE 1.2 The glycosaminoglycans of vertebrate connective tissues (from Mathews, 1975)

Type	Tissue	Repeating disaccharide
1. hyaluronate	Vitreous humour, synovial fluid, umbilical chord, skin, aorta	D-glucuronic acid D-glucosamine
2. chondroitin-4-sulphate	cartilage, bone, cornea, notochord, skin	D-glucuronic acid D-galactosamine
3. chondroitin-6-sulphate	cartilage, umbilical chord, skin, aorta	D-glucuronic acid D-galactosamine
4. dermatan	skin, tendon, aorta	L-iduronic acid D-galactosamine
5. corneal keratan sulphate	cornea	D-galactose D-glucosamine
6. skeletal keratan sulphate	cartilage, nucleus pulposus, annulus	D-galactose D-glucosamine
7. heparan sulphate	aorta, beef lung, liver amyloid	D-glucuronic acid D-glucosamine
8. heparin	lung, liver, skin	D-glucuronic acid D-glucosamine

adult tissues is illustrated in Table 1.3 (Harkness, 1961).

Changing collagen content and fibril size in development and wound healing is accompanied by changes in the relative amounts of different glycosaminoglycans. In regenerating rabbit tendon, for example, Dorner (1968) has made the observations shown in Table 1.4. The growth of fibrils in embryonic chick tendon was studied by Fitton-Jackson (1956). Both the average fibril width and the proportion of extracellular matrix occupied by fibrils increases. Mathews (1975) has discussed the changes in glycosaminoglycan composition of vertebrate cartilage during development and ageing. Finally, the rise in transparency of chick cornea during development is correlated with the occurrence of highly sulphated keratin sulphate and uniformly wide fibrils, whereas fibrils of varying diameters are correlated with the presence of dermatan sulphate in corneal scars (Mathews, 1975; Anseth and Fransson, 1970).

The study of the molecular organisation in collagen fibrils is therefore only one aspect of the problem of connective tissue. However, the fact that the collagen molecule has been chemically well characterised and that it assembles with a high degree of three-dimensional order makes such a study of great interest and most amenable to the use of modern methods of structural investigation.

3. The Structure of Collagen Assemblies

i) The Molecule

The collagen molecule is now known to exist in solution as a rod 3000Å long and 15Å wide, with molecular weight 300,000 (von Hippel, 1967). The molecule is composed of three polypeptide chains each of molecular weight 100,000. These chains are not

TABLE 1.3 The variation of collagen fibril sizes in different tissues (from Harkness, 1961)

Tissue	Fibril width range (Å)
Tendon	300-1300
Skin	600-1100
Cornea	200- 400
Vitreous humour	100- 300

TABLE 1.4 Glycosaminoglycan composition of regenerating rabbit tendon (from Dorner, 1968)

Type	Percentage of total glycosaminoglycans in regenerating tendon since first excision		
	4 days	4 months	mature tendon
hyaluronate	56	20	16
chondroitin-4-sulphate	11	9	7
chondroitin-6-sulphate	8	8	9
dermatan sulphate	12	54	60

necessarily identical in amino acid sequence (see Section 5) but they do have characteristic compositions for collagen of almost exactly one third glycine and a high percentage (~20%) of the imino acids proline and hydroxyproline.

With the exception of the distinctive amino acid composition, the above facts about collagen were not available to the early interpreters of its high angle X-ray diffraction pattern. It was seen that the pattern bore little resemblance to either the α or β forms of the so-called kmef class of fibrous proteins, so Astbury gave collagen a separate classification. Furthermore he also suggested that the unique collagen structure might be related to a possible periodic amino acid sequence of the form $(P\text{-}G\text{-}R)_n$ where P is proline or hydroxyproline, G is glycine and R is any other amino acid (Astbury, 1933; Hannig and Nordwig, 1967; Fraser and McRae, 1973).

Between 1938 and the early 1950's various models were proposed to account for the high angle pattern, but none of them were entirely satisfactory (see Ramachandran, 1967; Fraser and McRae, 1973). It was not until 1955 that the main features of what is now accepted as the correct kind of model were first outlined (Ramachandran and Kartha, 1955; Rich and Crick, 1955; Cowan et al., 1955a). These ideas came as a result of three major advances in the preceding few years. First there was the development of the stereochemical principles of Pauling and Corey (1951) in which they stressed the importance of hydrogen bonding and the co-planarity of the peptide bond. Secondly, the theory of X-ray diffraction by helical structures was elucidated by Cochran et al (1952) and by Stokes (1955). Thirdly there was the synthesis and structural solution of simple polypeptides containing typical collagen amino acids, i.e.

polyglycine and polyproline. These polypeptides each crystallise in two forms, designated I and II. Polyglycine II (Rich and Crick, 1955) and polyproline II (Cowan and McGavin, 1955) are both single helical structures with three residues per turn and unit height of approximately 3.1Å. The former can adopt either a left-handed or right-handed twist but the latter is restricted to the left-handed form.

The above three lines of approach, together with the then almost certain knowledge that the collagen sequence contained glycine as every third residue, led Ramachandran and Kartha, Rich and Crick and Cowan, North and McGavin to suggest triple helical models for the structure of the molecule. All the models contained three polypeptide chains of the form $(gly\text{-}X\text{-}Y)_n$ each in an approximate polyglycine II - polyproline II conformation (i.e. left-handed 3_1 helices), joined together by inter-chain hydrogen bonds and then twisted so that the chains follow a right-handed helical path along the resulting coiled coil. Equivalent tripeptides in the structure were related by a unit height of ~2.9Å and an anti-clockwise rotation of ~110°. Furthermore, the arrangement was such that all the glycine residues (which are unique among the amino acids in having no side chains attached to the α-carbon atom) could fit neatly in the core of the molecule. The only differences between the various models was in the number and pattern of inter-chain hydrogen bonds per tripeptide. In the two models of Rich and Crick (which were similar to the models of Cowan, North and McGavin) there was only one hydrogen bond per tripeptide (Rich and Crick, 1961), whereas Ramachandran and Kartha suggested two per tripeptide, except in those regions of the molecule where proline (or hydroxyproline) occurs in position 2(i.e. gly-X-Y, X

is position 2, Y is position 3) in all three chains, in which case only one hydrogen bond is formed (Ramachandran, 1967). In order to make a choice in favour of any of these models a number of lines of investigation have been carried out. They include:-

a) synthesis and structural study of collagen-like poly-tripeptides and polyhexapeptides (see Traub and Piez, 1971; Fraser and McRae, 1973).

b) various spectroscopic studies (see Fraser and McRae, 1973)

c) energy calculations (see Von Hippel, 1967)

d) hydrogen isotope exchange experiments (see Von Hippel, 1967; Traub and Piez, 1971; Fraser and McRae, 1973)

e) comparison of calculated intensities with the observed pattern (see Fraser and McRae, 1973).

These studies have limited the range of possible models down to basically two kinds, the Rich-Crick type II and the latest Ramachandran-Kartha type. The calculated transforms of these two models for layer lines 0 to 10 are so similar that it is not possible to say that either gives a better fit to the X-ray data. A lot of the synthetic poly-tripeptides and polyhexapeptides conform well to the Rich-Crick II model. The hydrogen isotope exchange experiments have given conflicting results though the most recent work has come out in favour of a two-bonded model (Yee et al., 1974). Fortunately these uncertainties about the precise details of the molecular conformation do not preclude a useful study of the molecular packing in collagen, to which attention is now turned.

ii) The fibril

The fact that collagen molecules are assembled in nature in a regular way was first demonstrated in the late 1930's and early 1940's with the results of low angle X-ray diffraction (Clark et al., 1935; Wyckoff et al., 1935; Clark and Schaad, 1936; Corey and Wyckoff, 1936; Bear, 1942; Kratky and Sekora, 1943) and electron microscopy (Hall et al., 1942; Schmitt et al., 1942; Wolpers, 1943). A low angle X-ray diffraction pattern from a whole rat tail tendon and an electron micrograph of an isolated collagen fibril are shown in plates 1, 1(a) and (b) respectively. The periodicity (D) along the fibre axis varies with moisture content (see Fraser and McRae, 1973). In rat tail tendon, for example, D varies from 671Å when wet to 636Å when dried in vacuo at room temperature (Ericson and Tomlin, 1959). To observe the D periodicity clearly with electron microscopy it is necessary to treat specimens in one of three ways; heavy metal shadowing, negative staining and positive staining. Shadowing is used to show the external shape of the specimens, in collagen this method reveals a corrugated appearance along the axis of the fibril (Gross and Schmitt, 1948). Negative staining was first used on collagen in 1963 (Tromans et al.; Olsen); this technique reveals the variations in stain permeability throughout the specimen. Positive staining shows the distribution of those amino acids with charged side chains. In the collagen fibril this distribution is seen as up to twelve bands transverse to the main axis which are repeated in each D period (Gross and Schmitt, 1948; Nutting and Borasky, 1948; Hodge and Schmitt, 1960; Doyle et al., 1974a). X-ray diffraction has the advantage over electron microscopy in that no such chemical treatment

a

b

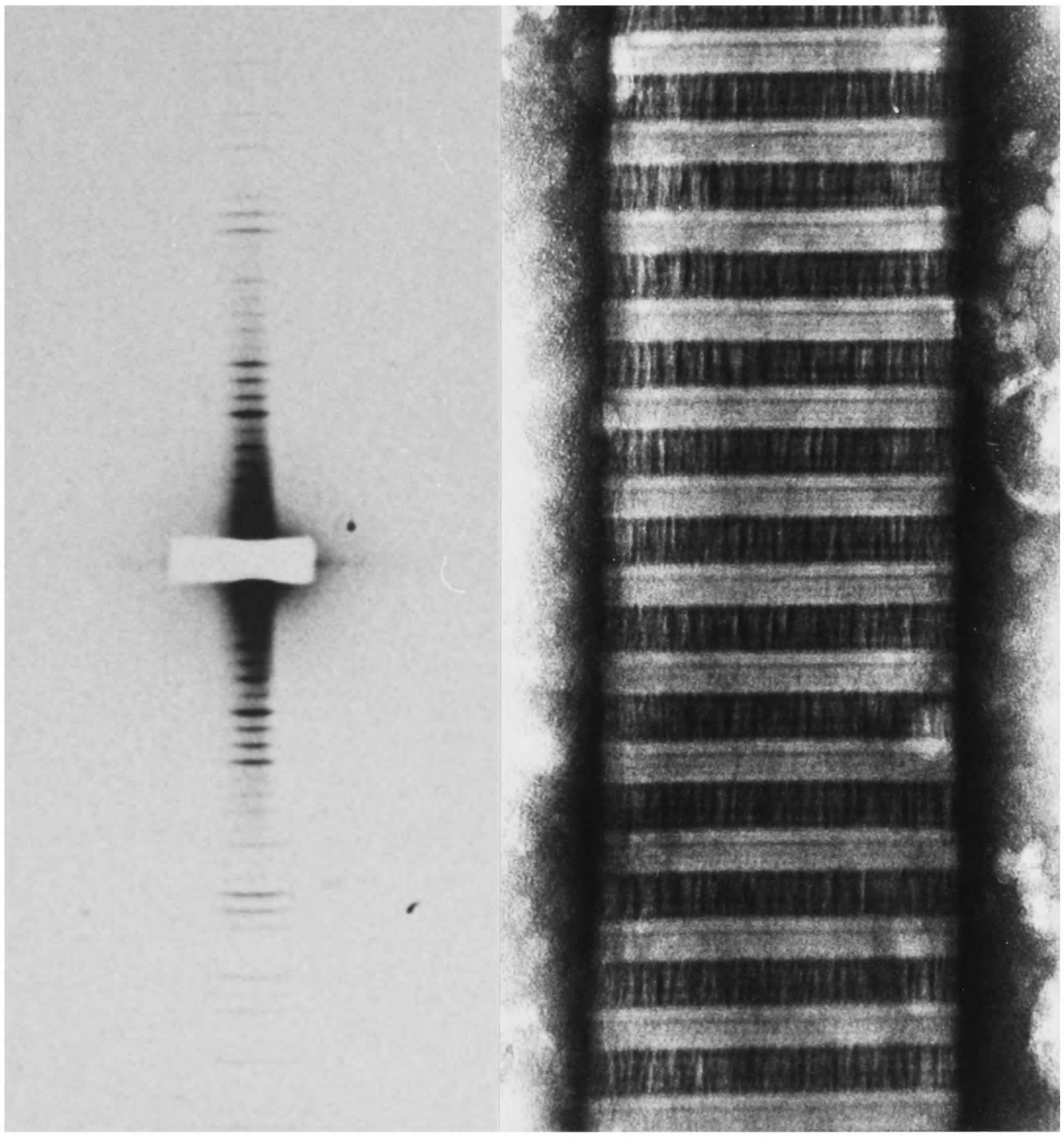


PLATE 1.1

a) The low angle, meridional X-ray reflections from moist rat tail tendon. The specimen was held under tension and suspended over 0.15M NaCl (courtesy of Mr. S. White).

b) An electron micrograph of part of a collagen fibril. The specimen was obtained, by dispersion, from rat tail tendon. It was then negatively stained with 2% phosphotungstic acid (courtesy of Dr. J.Woodhead-Galloway).

is required to observe the characteristic low angle pattern, tendons can therefore be studied in a more natural state. It has the disadvantage, however, that there is not a direct relation between the intensities of the meridional reflections and the electron density within a D period. The interpretation requires additional phase information (see Chapter 4).

At the time of the above X-ray and e.m. studies little was known about the gross physical properties of the collagen molecule; hence without this knowledge any models for molecular packing in collagen fibrils could not proceed very far. A most significant advance was made when it was discovered that collagen molecules in solution could be made to aggregate in a variety of ways, depending on the chemical conditions. Besides the native-type fibrils (Schmitt et al., 1942), the so-called fibrous long spacing (FLS) (Highberger et al., 1950, 1951) and segment long spacing (SLS) forms (Schmitt et al. 1953) could be reconstituted. FLS collagen, when first observed was seen to have a periodicity of approximately 4D, whereas SLS collagen had a length of approximately 4D. These findings led Gross et al. (1954) to put forward the idea of the so-called tropocollagen molecule as the basic unit of collagen. This unit, with a length of ~2600Å would then aggregate in different ways to form native-type FLS or SLS collagen. In SLS the tropocollagen molecules were thought to assemble in parallel register forming segments of fixed length and variable width. In FLS they were thought to assemble again side by side but in an anti-parallel manner with end to end contacts between molecules of adjacent periods. The molecular arrangement in native-type fibrils was first proposed by Schmitt et al. (1955). Since the native periodicity seemed to be one quarter of the molecular length,

these authors suggested that, within the fibril, the molecules were staggered with respect to their neighbours by one quarter of their length. Soon after this work, it became clear from physical chemical studies, that the molecule was indeed of length 2800Å (Boedtker and Doty, 1956). In 1958 isolated molecules of about this length were seen directly in the electron microscope (Hall and Doty, 1958).

The so-called 'quarter stagger' theory received considerable support in 1960 when Hodge and Schmitt showed that the native, type positive staining banding pattern could be synthesised from the SLS pattern by repeated superimposition of the latter upon itself after shifts of one quarter of its length. Kuhn and his co-workers gave a similar demonstration, by addition of microdensitometer traces of SLS for comparison with the native-type trace (Kuhn et al., 1960; Kuhn and Zimmer, 1961). Hodge and Schmitt (1960) were also able to demonstrate SLS crystallites growing at specific points on the surface of native-type fibrils. The specificity was such that most of the SLS positively stained bands were seen to be continuous with the native banding pattern. In particular, the origins of the so-called 'd' bands of the native pattern (see Hodge and Schmitt, 1960; also Chapter 2) could be traced to the four particular bands of the SLS pattern. These bands thus provide an internal calibration for measurements of the SLS crystallite. This information, together with the results of the newly introduced negative staining technique, led Hodge and Petruska (1962, 1963) to produce a more accurate estimate of the length of the collagen molecule in terms of the D periodicity. They found the length to be $4.4 \pm 0.02D$. The stagger in the native-type fibril was therefore not exactly one quarter of the molecular length, and it was necessary to modify the original quarter stagger model. The Hodge-Petruska modified

scheme is shown in Figure 1.1. The consequence of this non-integral molecular length was the appearance of regions of high and low density along the collagen fibrils, the so-called overlap and gap regions respectively. This prediction was seen to be consistent with the results of negative staining, where two bands of high and low stain permeability were observed in each D period of native-type fibrils (Tromans et al., 1963; Olsen, 1963). It was also consistent with the interpretations of the intensities of the first few low angle X-ray meridional reflections (see Chapter 4).

The Hodge-Petruska modified quarter stagger theory is only a two-dimensional representation for molecular packing in the collagen fibril (Hodge et al., 1965). Indeed, insofar as the most important feature of the scheme is that molecules are related axially by multiple staggers of D, the Hodge-Petruska theory is only a one-dimensional model. There are many ways of extending the model into three dimensions. It is not necessary that all neighbouring molecules should be related by a one D stagger. Indeed, Smith (1965) has shown that, for a hexagonal array of collagen molecules in a fibril of diameter 500Å, only 75% of all the nearest neighbour contacts are 'quarter-staggered'. The only restriction imposed on any model by the D periodicity is that the number of molecules contained in every section of height D along the fibril should be constant. A random arrangement, in which nearest neighbours have an equal probability of being staggered by $\pm$nD, where n = o, 1, 2, 3 or 4, would fulfil this requirement. Ross and Benditt (1961) have discussed this type of molecular arrangement and its relation to the minimum width of a collagen fibril displaying a D periodicity. Cox et al. (1967) have proposed

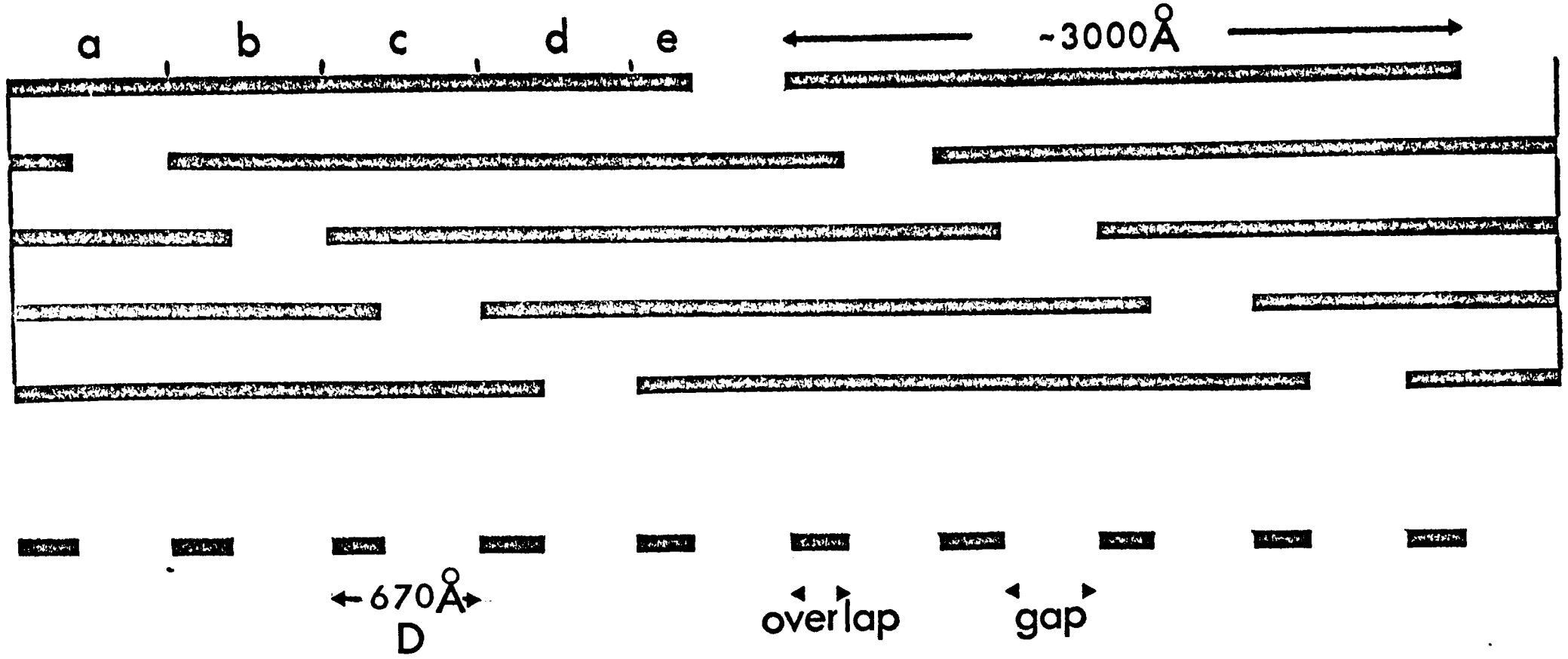


Fig. 1.1 The Hodge-Petruska scheme for molecular packing in collagen fibrils. Ten D repeats are shown. The molecule, of length ~3000Å, can be considered as being composed of five non-identical segments a,b,c,d, and e.

this 'random accretion' as a general model for fibril assembly. Furthermore, Doyle et al. (1974b) have discussed the total number of ways the five molecular strands of the two-dimensional Hodge-Petruska scheme can be mutually staggered whilst preserving the overall D periodicity. There are twelve possible combinations, only two of which have the same stagger between all nearest neighbours.

It has been known since 1954 that the collagen molecules in rat tail tendon fibrils are arranged with some lateral order, since this tendon gives a set of equatorial and near equatorial reflections in its low-angle X-ray diffraction pattern (North et al., 1954). Originally these reflections were indexed in terms of a monoclinic cell ($a = 62\text{Å}$, $c = 76\text{Å}$, $\beta = 125^{\circ}$), but, following North's observation (1955) that the reflections could be indexed as orders of a single dimension, the data was shown to be consistent with the positions of the local maxima arising from diffraction by a limited cylindrical (Ramachandran and Sasisekharan, 1956), spiral (Sasisekharen and Ramachandran, 1957) or hexagonal lattice (Burge, 1963) containing approximately 150 molecules with overall diameter 200Å. Since then the origin of some of the equatorial reflections has been questioned (Miller and Wray, 1971), with the result that some of them have been rejected and others added. This new set of data has been indexed in terms of a tetragonal unit cell of side 77Å, with, tetragonal pseudo unit cells of side 54Å or 38.5Å (Fraser et al., 1974; Miller and Parry, 1973). These cells are aligned with their long axis of 4 x 670Å parallel to the fibre axis. Alternatively Woodhead-Galloway et al. (1975) have chosen a tetragonal cell of side 55Å with a long axis of 5n x 670Å, again aligned with the fibre axis. Finally, Nemetschek and Hosemann (1973) have

chosen an orthorhombic unit cell of sides 38.4Å x 34.9Å with its long axis tilted with respect to the fibre axis by 2.5°. The similarity of these three indexing schemes hides fundamental differences in structural models (see later).

Most of the early models for lateral packing assumed the collagen molecule to be the basic structural unit, which was then packed in an extended lattice of some sort (Ramachandran and Sasisekharan, 1956; Sasisekharan and Ramachandran, 1957; Burge, 1963) though the same principle was used in McFarlane's (1971) model - a triply twinned crystal of hexagonally close packed molecules. The idea that the basic structural unit was a bundle of collagen molecules was proposed by North et al. (1954). They suggested an elementary bundle of seven molecules in cross-section. Burge (1965) then proposed a two strand coiled coil of molecules, this idea has recently been developed by Woodhead-Galloway et al. (1975). Veis et al. (1967) suggested an elementary microfibril of four molecular strands which was itself 5D periodic and hence was packed in the fibril with random nD staggers between neighbours (Veis et al. 1970). This model has recently been modified so that the microfibril is exactly D periodic. This is achieved by considering the microfibril to consist of four strands, where each strand is a continuous length of collagen molecules (a so-called protofibril) which are cross-linked with a 0.4D overlap (Veis and Yuan, 1975; see also Kuhn et al., 1966). This latest version of the four stranded microfibril is topologically equivalent to the five stranded microfibril of Smith (1968) in which the strands are chains of molecules separated by 0.6D gaps. Finally Hosemann et al. (1974) have suggested an eight stranded octafibril, where each strand is a 0.4D overlapped protofibril.

All the recent models for molecular packing in collagen fibrils possess features which explain certain aspects of the low angle X-ray pattern. Electron microscopy has not provided convincing evidence in favour of one particular model, with the possible exception of the five stranded microfibril (Doyle et al., 1974b).

There is uncertainty, then, in both the exact details of the molecular structure of collagen and also in the nature of the lateral packing within collagen fibrils. Fortunately, however, the molecular arrangement in projection on to the fibril axis seems to be well understood in terms of the generalised Hodge-Petruska theory. One of the main concerns in this thesis is the investigation of the precise details of this theory, and how it can be used in conjunction with the amino acid sequence to understand both the experimental observations (i.e. electron microscopy and X-ray, neutron or optical diffraction) and also the intermolecular self-assembly of collagen.

4. Self-assembly

Biological self-assembly is the spontaneous association of macromolecules to form biological structures. It is spontaneous in the sense that, given the appropriate environmental conditions (of pH, temperature, concentration, ionic strength, pressure, etc.), the macromolecular components of a biological structure will associate without the need for external control factors, i.e. macromolecules which direct the course of building the structure but are absent in the final product. The particular nature of the macromolecular components, e.g. the amino acid sequence unique to each protein, is ultimately specified by the nucleotide sequences in the genetic material

DNA (or RNA).

The acquisition of the three-dimensional structure of biological objects is due to the physico-chemical interaction between the basic units of macromolecules (e.g. amino acid residues, monosaccharide residues). A complete understanding of biological self-assembly will only be achieved when the process of information transfer from the gene to the products of gene expression and the subsequent specificity of the three-dimensional interactions are understood at the molecular level.

That biological self-assembly does in fact occur in nature can only be inferred from studies in which purified macromolecular components are seen to self-assemble into recognisable biological structures in vitro. The criteria to judge the identity of these re-assembled structures with their naturally occurring counterparts must be quite stringent. For renaturation of the specific three-dimensional structure of an enzyme molecule from a random polypeptide chain, for example, the criteria are:

a) complete recovery of enzymatic activity.
b) formation of the correct disulphide covalent bonds (if any).
c) re-appearance of characteristic spectroscopic properties.
d) recovery of characteristic immunological properties.
e) re-appearance of characteristic X-ray diffraction pattern when crystallised.

Some or all of these criteria have been satisfied upon renaturation of several enzyme molecules, including bovine pancreatic ribonuclease (Sela et al., 1957; White, 1961; Bello et al., 1961; Anfinsen et al., 1961), staphylococcal nuclease (Jardetsky et al., 1971; Anfinsen, 1972) and hen egg white lysozyme (Epstein, 1970). The specificity of the renaturation process is well illustrated by the formation of the correct

disulphide bonds; in ribonuclease, for example, only four of the 105 possible bonds are formed.

For in vitro self-assembly of larger macromolecular aggregates, such as virus particles, bacterial flagellae, or collagen fibrils, other accuracy criteria may be used, notably electron microscopy. If tobacco mosaic virus (TMV) particles, for example, are dissociated into their component

D.PHIL.THESIS 1975 D.J.S.HULMES

ERRATA

page 23 — for "only four of the 105 possible bonds are formed" read "only one of the 105 possible bonding schemes is formed".

page 215 — reference omitted:
Kellenberger, E. (1966). In 'Principles of Biomolecular Organisation, CIBA Foundation Symposium' (ed. G.E.W. Wolstenholme & M. O'Connor), pp.192-228, Churchill, London.

filaments can be observed (Davson, 1964), though the filaments thus produced do not have the characteristic length of the in vivo filaments. Furthermore, bipolar segments of myosin molecules can be made to self-assemble in vitro in a similar way to their arrangement in muscle (Huxley, 1963). So far, of course, a myofibril has not been shown to self-assemble in vitro. Nevertheless it is possible to produce a complex of actin, tropomyosin, troponin and heavy meromyosin subfragment 1 which shows all the regulatory behaviour of intact muscle, except

that the difference in ATPase activity in the presence and absence of calcium is not as marked as in vivo (see Offer, 1974).

Clearly muscle is a relatively complex macromolecular system compared to, for example, collagen fibrils or TMV. It seems that self-assembly as defined in the first paragraph of this section only occurs for relatively simple structures. Indeed it has been suggested that some external enzymic control is used in nature even during the folding of some disulphide-linked proteins (Venetianer and Straub, 1963; Goldberger et al., 1963). For more complex systems it appears that some mechanisms must exist whereby multi-component structures are built with a definite order of morphogenesis. This has been clearly described by Wood and Edgar (1967) in studies on the T4 bacteriophage. This phage is composed of a head and a tail, the latter consisting of a contractile sheath, an hexagonal base plate and six tail fibres. Wood and Edgar showed that when, for example, there is a mutation in one of the genes responsible for head formation then the tail, plate and tail fibres will not assemble in vivo, they will only do so when the head has been successfully completed. Factors which control the order of morphogenesis have been discussed by Kellenberger (1966).

In the morphogenesis of collagen fibrils in vivo it has recently been suggested that one of the controls is achieved by enzymatic cleavage of the molecular precursor, procollagen (see Bornstein, 1974; Tanzer et al. 1974). The procollagen molecule, with non-triple helical extensions at possibly both the N- and C-termina, is secreted into the extracellular matrix. In this state it does not form collagen fibrils with the characteristic mechanical properties of healthy tissue. This can only occur when the procollagen extension has been

enzymatically removed by procollagen peptidase. Absence of this enzyme results in the disease known as dermatosporaxis (Lapière et al., 1971). Procollagen is an example of a so-called zymogen, other examples are proinsulin, trypsinogen and fibrinogen.

The forces which are involved in biological structures have been discussed by many authors (see, for example, Kauzmann, 1959). Essentially, the thermodynamic principle is that the formation of a biological structure is accompanied by a decrease in the free energy of the system. The free energy change (ΔF) is defined by

$$\Delta F = \Delta H - T\Delta S$$

where ΔH is the change in enthalpy
T is the absolute temperature
ΔS is the change in entropy.

The enthalpy term is associated with electromagnetic interactions (i.e. hydrogen bonds, interactions between ionised groups and van der Waals (or dispersion) interactions), and thermal energy. The entropy term is associated with the amount of disorder in the system. Self-assembly is possible when the free energy tends to at least a local minimum.

The conditions required for self-assembly of biological structures are very critical. TMV protein, for example, can aggregate in one of many different forms depending on the pH and temperature of the solution (Lauffer, 1975). Collagen molecules too can aggregate in at least thirteen different ways depending on pH, temperature, ionic strength and the presence of other macromolecules (Doyle et al., 1975 ; also Chapter 3). The particular conditions that are required can to some extent be related to the important interactions involved in the self-assembly. Polymerisation of globular actin (G-actin) monomers

to actin filaments, for example, is produced by increasing the ionic strength of a G-actin solution (Davson, 1964). This is consistent with the extra ions binding to charged sites on the G-actin molecules, hence suggesting that salt linkages are not involved in the polymerisation. Furthermore the enthalpy and entropy changes accompanying both actin polymerisation and TMV protein aggregation have been measured. In each case the entropy increases, even though self-assembly has resulted in the formation of ordered structures (Kushner, 1969). This observation is explained by the change in structure of water molecules from an ice-like state around exposed apolar side-chains to a disordered state when these side-chains become buried in the macromolecular aggregates. This tendency of apolar side-chains to aggregate in an aqueous solvent is usually termed a hydrophobic interaction. This interaction is most important in the folding of globular proteins, where hydrophobic groups are found in the core of the structure and hydrophilic (charged) groups are found on the surface (see Dickerson and Geis, 1969).

An attempt is made in subsequent chapters of this thesis to show how the particular amino acid sequence of collagen specifies its intermolecular interactions through the interaction properties of certain types of amino acid residues.

5. Sequence

As mentioned in Section 2, the amino acid composition of collagen is very characteristic. Approximately one third of the residues are glycine and there is a high percentage (~20%) of the imino acids proline and hydroxyproline. Hydroxylation of some prolyl and some lysyl residues either during or after protein synthesis is another characteristic of collagen.

Furthermore some residues are notably absent, in particular tryptophan, and the content of tyrosine, methionine or cysteine is very low. For a comprehensive summary of the amino acid composition of collagen from vertebrates and invertebrates see Mathews (1975).

There are three factors giving rise to inter-tissue variability that occur at the post-ribosomal level. They are:

a) degree of hydroxylation of lysine and proline (see Traub and Piez, 1971; Mathews, 1975)

b) nature and amount of intra- and inter-molecular cross-linking, which usually involves lysine, hydroxylysine or histidine (see Mathews, 1975)

c) degree of glycosylation of hydroxylysine, usually by glucose and/or galactose (Spiro, 1972; Mathews, 1975).

At the pre-ribosomal level, variation within each species is also a result of different genes coding for at least five types of α-chain. Three α-chains are incorporated in each collagen molecule. In vertebrates skin, bone and tendon have the molecular composition $\alpha 1(I)_2, \alpha 2$, cartilage is $\alpha 1(II)_3$, embryonic skin and immature skin and aorta are $\alpha 1(III)_3$, and basement membranes are $\alpha 1(IV)_3$ (for references see Mathews, 1975).

Each α-chain contains approximately 1000 amino acid residues, thus sequencing is a formidable task. To date only the α1 chain of rat/calf skin has been completely sequenced, though the sequences of peptide fragments from many vertebrate collagens are known. The complete α1 sequence was compiled partly from rat skin (residues 1-418; Kang et al., 1967; Bornstein (1967); Butler, 1970; Butler and Ponds, 1971) and partly from calf skin

(residues 419-1052; Fietzek et al., 1972a,b,c; Wendt et al., 1972; Rauterberg et al., 1972) and published by Hulmes et al., (1973), see also Figure 1.2.

Most of this α1 sequence can be described by the repeating tripeptide gly-X-Y. 16 residues at the N terminus and 25 residues at the C terminus do not conform to this description. As glycine is so necessary for the formation of a triple helical structure (see Section 3) it seems likely that these terminal peptides do not have this conformation. Within the central helical region some amino acid residues have a tendency to occupy either position X or Y (see Salem and Traub, 1975). Hydroxyproline, for example, is found almost exclusively in the Y position, whereas proline is almost always in the X position. Glutamic acid favours the X position, whereas arginine favours the Y position. Alanine is evenly divided between both positions.

It has been suggested that the non-helical terminal peptides (or telo-peptides) play an important role in cross-linking, antigenicity, and control of fibril formation (Schmitt et al., 1964; Rubin et al., 1965; Drake et al., 1966; Leibovich and Weiss, 1970). Cross-linking has since been demonstrated in these regions (see Tanzer, 1973; Gallop et al., Rauterberg et al., 1972). The characteristic species variability of the N-terminal telopeptide is consistent with its role as a species-specific determinant (Highberger et al., 1971). Finally it is likely that the non-helical peptides are at least indirectly associated with fibril formation, in so far as they are remnants of the biological precursor to collagen, procollagen (Bornstein, 1974; Tanzer et al., 1974).

The remaining chapters of this thesis are concerned with the relation of the sequence to the self-assembly and structure

pGlu-Met-Ser-Tyr-Gly-Tyr-Asp-Glu-Lys-Ser-Ala-Gly-Val-Ser-Val-Pro-

```
   1  Gly-Pro-Met-Gly-Pro-Ser-Gly-Pro-Arg-Gly-Leu-Hyp-Gly-Pro-Hyp-Gly-Ala-Hyp-Gly-Pro-Gln-Gly-Phe-Gln-Gly-Pro-Hyp-
  28  Gly-Glu-Hyp-Gly-Glu-Hyp-Gly-Ala-Ser-Gly-Pro-Met-Gly-Pro-Arg-Gly-Pro-Hyp-Gly-Pro-Hyp-Gly-Lys-Asn-Gly-Asp-Asp-
  55  Gly-Glu-Ala-Gly-Lys-Pro-Gly-Arg-Hyp-Gly-Gln-Arg-Gly-Pro-Hyp-Gly-Pro-Gln-Gly-Ala-Arg-Gly-Leu-Hyp-Gly-Thr-Ala-
  82  Gly-Leu-Hyp-Gly-Met-Hyl-Gly-His-Arg-Gly-Phe-Ser-Gly-Leu-Asp-Gly-Ala-Lys-Gly-Asn-Thr-Gly-Pro-Ala-Gly-Pro-Lys-
 109  Gly-Glu-Hyp-Gly-Ser-Hyp-Gly-Glx-Asx-Gly-Ala-Hyp-Gly-Gln-Met-Gly-Pro-Arg-Gly-Leu-Hyp-Gly-Glu-Arg-Gly-Arg-Hyp-
 136  Gly-Pro-Hyp-Gly-Ser-Ala-Gly-Ala-Arg-Gly-Asp-Asp-Gly-Ala-Val-Gly-Ala-Ala-Gly-Pro-Hyp-Gly-Pro-Thr-Gly-Pro-Thr-
 163  Gly-Pro-Hyp-Gly-Phe-Hyp-Gly-Ala-Ala-Gly-Ala-Lys-Gly-Glu-Ala-Gly-Pro-Gln-Gly-Ala-Arg-Gly-Ser-Glu-Gly-Pro-Gln-
 190  Gly-Val-Arg-Gly-Glu-Hyp-Gly-Pro-Hyp-Gly-Pro-Ala-Gly-Ala-Ala-Gly-Pro-Ala-Gly-Asn-Hyp-Gly-Ala-Asp-Gly-Gln-Hyp-
 217  Gly-Ala-Lys-Gly-Ala-Asn-Gly-Ala-Hyp-Gly-Ile-Ala-Gly-Ala-Hyp-Gly-Phe-Hyp-Gly-Ala-Arg-Gly-Pro-Ser-Gly-Pro-Gln-
 244  Gly-Pro-Ser-Gly-Ala-Hyp-Gly-Pro-Lys-Gly-Asn-Ser-Gly-Glu-Hyp-Gly-Ala-Hyp-Gly-Asn-Lys-Gly-Asp-Thr-Gly-Ala-Lys-
 271  Gly-Glu-Hyp-Gly-Pro-Ala-Gly-Val-Gln-Gly-Pro-Hyp-Gly-Pro-Ala-Gly-Glu-Glu-Gly-Lys-Arg-Gly-Ala-Arg-Gly-Glu-Hyp-
 298  Gly-Pro-Ser-Gly-Leu-Hyp-Gly-Pro-Hyp-Gly-Glu-Arg-Gly-Gly-Hyp-Gly-Ser-Arg-Gly-Phe-Hyp-Gly-Ala-Asp-Gly-Val-Ala-
 325  Gly-Pro-Lys-Gly-Pro-Ala-Gly-Glu-Arg-Gly-Ser-Hyp-Gly-Pro-Ala-Gly-Pro-Lys-Gly-Ser-Hyp-Gly-Glu-Ala-Gly-Arg-Hyp-
 352  Gly-Glu-Ala-Gly-Leu-Hyp-Gly-Ala-Lys-Gly-Leu-Thr-Gly-Ser-Hyp-Gly-Ser-Hyp-Gly-Pro-Asp-Gly-Lys-Thr-Gly-Pro-Hyp-
 379  Gly-Pro-Ala-Gly-Gln-Asp-Gly-Arg-Hyp-Gly-Pro-Ala-Gly-Pro-Hyp-Gly-Ala-Arg-Gly-Gln-Ala-Gly-Val-Met-Gly-Phe-Hyp-
 406  Gly-Pro-Lys-Gly-Ala-Ala-Gly-Glu-Hyp-Gly-Lys-Ala-Gly-Glu-Arg-Gly-Val-Hyp-Gly-Pro-Hyp-Gly-Ala-Val-Gly-Pro-Ala-
 433  Gly-Lys-Asp-Gly-Glu-Ala-Gly-Ala-Gln-Gly-Pro-Hyp-Gly-Pro-Ala-Gly-Pro-Ala-Gly-Glu-Arg-Gly-Glu-Gln-Gly-Pro-Ala-
 460  Gly-Ser-Hyp-Gly-Phe-Gln-Gly-Leu-Hyp-Gly-Pro-Ala-Gly-Pro-Hyp-Gly-Glu-Ala-Gly-Lys-Hyp-Gly-Glu-Gln-Gly-Val-Hyp-
 487  Gly-Asp-Leu-Gly-Ala-Hyp-Gly-Pro-Ser-Gly-Ala-Arg-Gly-Glu-Arg-Gly-Phe-Hyp-Gly-Glu-Arg-Gly-Val-Glu-Gly-Pro-Hyp-
 514  Gly-Pro-Ala-Gly-Pro-Arg-Gly-Ala-Asn-Gly-Ala-Hyp-Gly-Asn-Asp-Gly-Ala-Lys-Gly-Asp-Ala-Gly-Ala-Hyp-Gly-Ala-Hyp-
 541  Gly-Ser-Gln-Gly-Ala-Hyp-Gly-Leu-Gln-Gly-Met-Hyp-Gly-Glu-Arg-Gly-Ala-Ala-Gly-Leu-Hyp-Gly-Pro-Lys-Gly-Asp-Arg-
 568  Gly-Asp-Ala-Gly-Pro-Lys-Gly-Ala-Asp-Gly-Ala-Pro-Gly-Lys-Asp-Gly-Val-Arg-Gly-Leu-Thr-Gly-Pro-Ile-Gly-Pro-Hyp-
 595  Gly-Pro-Ala-Gly-Ala-Hyp-Gly-Asp-Lys-Gly-Glu-Ala-Gly-Pro-Ser-Gly-Pro-Ala-Gly-Thr-Arg-Gly-Ala-Hyp-Gly-Asp-Arg-
 622  Gly-Glu-Hyp-Gly-Pro-Hyp-Gly-Pro-Ala-Gly-Phe-Ala-Gly-Pro-Hyp-Gly-Ala-Asp-Gly-Gln-Hyp-Gly-Ala-Lys-Gly-Glu-Hyp-
 649  Gly-Asp-Ala-Gly-Ala-Lys-Gly-Asp-Ala-Gly-Pro-Hyp-Gly-Pro-Ala-Gly-Pro-Ala-Gly-Pro-Hyp-Gly-Pro-Ile-Gly-Asn-Val-
 676  Gly-Ala-Hyp-Gly-Pro-Hyl-Gly-Ala-Arg-Gly-Ser-Ala-Gly-Pro-Hyp-Gly-Ala-Thr-Gly-Phe-Hyp-Gly-Ala-Ala-Gly-Arg-Val-
 703  Gly-Pro-Hyp-Gly-Pro-Ser-Gly-Asn-Ala-Gly-Pro-Hyp-Gly-Pro-Hyp-Gly-Pro-Ala-Gly-Lys-Glu-Gly-Ser-Lys-Gly-Pro-Arg-
 730  Gly-Glu-Thr-Gly-Pro-Ala-Gly-Arg-Hyp-Gly-Glu-Val-Gly-Pro-Hyp-Gly-Pro-Hyp-Gly-Pro-Ala-Gly-Glu-Lys-Gly-Ala-Hyp-
 757  Gly-Ala-Asp-Gly-Pro-Ala-Gly-Ala-Hyp-Gly-Thr-Pro-Gly-Pro-Gln-Gly-Ile-Ala-Gly-Gln-Arg-Gly-Val-Val-Gly-Leu-Hyp-
 784  Gly-Gln-Arg-Gly-Glu-Arg-Gly-Phe-Hyp-Gly-Leu-Hyp-Gly-Pro-Ser-Gly-Glu-Hyp-Gly-Lys-Gln-Gly-Pro-Ser-Gly-Ala-Ser-
 811  Gly-Glu-Arg-Gly-Pro-Hyp-Gly-Pro-Met-Gly-Pro-Hyp-Gly-Leu-Ala-Gly-Pro-Hyp-Gly-Glu-Ser-Gly-Arg-Glu-Gly-Ala-Hyp-
 838  Gly-Ala-Glu-Gly-Ser-Hyp-Gly-Arg-Asp-Gly-Ser-Hyp-Gly-Ala-Lys-Gly-Asp-Arg-Gly-Glu-Thr-Gly-Pro-Ala-Gly-Ala-Hyp-
 865  Gly-Pro-Hyp-Gly-Ala-Hyp-Gly-Ala-Hyp-Gly-Pro-Val-Gly-Pro-Ala-Gly-Lys-Ser-Gly-Asp-Arg-Gly-Glu-Thr-Gly-Pro-Ala-
 892  Gly-Pro-Ile-Gly-Pro-Val-Gly-Pro-Ala-Gly-Ala-Arg-Gly-Pro-Ala-Gly-Pro-Gln-Gly-Pro-Arg-Gly-Asx-Hyl-Gly-Glx-Thr-
 919  Gly-Glx-Glx-Gly-Asx-Arg-Gly-Ile-Hyl-Gly-His-Arg-Gly-Phe-Ser-Gly-Leu-Gln-Gly-Pro-Hyp-Gly-Pro-Hyp-Gly-Ser-Hyp-
 946  Gly-Glu-Gln-Gly-Pro-Ser-Gly-Ala-Ser-Gly-Pro-Ala-Gly-Pro-Arg-Gly-Pro-Hyp-Gly-Ser-Ala-Gly-Ser-Hyp-Gly-Lys-Asp-
 973  Gly-Leu-Asn-Gly-Leu-Hyp-Gly-Pro-Ile-Gly-Hyp-Hyp-Gly-Pro-Arg-Gly-Arg-Thr-Gly-Asp-Ala-Gly-Pro-Ala-Gly-Pro-Hyp-
1000  Gly-Pro-Hyp-Gly-Pro-Hyp-Gly-Pro-Hyp-Gly-Pro-Pro-
```

Fig. 1.2 Ser-Gly-Gly-Tyr-Asp-Leu-Ser-Phe-Leu-Pro-Gln-Pro-Pro-Gln-Gln-Glx-Lys-Ala-His-Asp-Gly-Gly-Arg-Tyr-Tyr

. **The amino acid sequence of the α1 chain from rat (residues 1–402) and calf (residues 403–1011) skin collagen. The N-terminal (from rat) and the C-terminal (from calf) non-helical regions are separated from the triplet region and are not numbered. See text for references.**

of ordered aggregates of collagen. Chapter 2 deals with the origins of the intermolecular D stagger in the sequence, and how a theoretical value for D can be used to simulate an accurate representation of the electron microscopic banding pattern of normal collagen fibrils. Chapter 3 is an extension of this approach to the study of the structure and self-assembly of mainly in vitro polymorphic forms of ordered aggregation. Chapter 4 is concerned with the relation of the sequence to the low angle X-ray and neutron diffraction patterns of collagen.

CHAPTER 2

THE INTERMOLECULAR D STAGGER IN COLLAGEN FIBRILS

1. Introduction

In the previous chapter it was shown that collagen molecules in solution are capable of self-assembly into recognisable ordered aggregates. Under appropriate conditions of temperature, pH and ionic strength (37°C, pH 7, I ~0.2) fibrils are formed which show all the characteristic electron microscopic features of fibrils extracted from native tissue (Hodge and Schmitt, 1960). When the conditions are altered, and/or other macromolecules are added, the collagen molecules aggregate in up to thirteen ordered forms (see Chapter 3). To some extent certain chemical conditions are specific to the formation of particular structures. As this process of ordered aggregation is spontaneous, the intermolecular interactions involved in lowering the free energy to a minimum must be due to the properties of the chemical groups within collagen molecules, as mediated by the prevailing physicochemical conditions. The chemical character of a single protein molecule is determined by the particular sequence(s) of amino acids and the arrangement of the polypeptide chain(s) in three dimensions. As pointed out in the previous chapter, however, the spontaneous acquisition of a protein molecule's three-dimensional structure from a random chain is, at least in some proteins, entirely specified by its amino acid sequence. The formation of the triple-helical structure of collagen molecules has also been shown to be a spontaneous process (see Traub and Piez, 1971). Thus both the structure of individual protein molecules and their mode of association

are determined by the properties of the amino acid sequence in the prevailing chemical conditions. This is the central assumption for the rest of the analysis in this and the next chapter. The ultimate aim is to understand the specificity of intermolecular self-assembly in terms of amino acid sequence, and therefore how the genetic information, which codes for the sequence, is used to build biological structures.

The Hodge-Petruska scheme (Fig. 1.1) for native-type collagen fibrils involves interactions between molecules which are axially staggered with respect to each other by multiples of 670Å (D).

The next section describes an attempt to determine the specificity for these interactions in the α1 sequence.

2. Analysis of the Sequence for the Origins of the D Stagger

i) Interactions

The types of non-covalent interactions that are involved in the stabilisation of protein structures in aqueous solution have been discussed by, for example, Kauzmann (1959) and Scheraga (1963). The main categories are:-

a) Hydrogen bonds, mainly between carbonyl oxygens and amide hydrogens. There may be other hydrogen bonds involving hydroxyl and carboxylate groups or carboxy and carboxylate groups.

b) Dispersion (London) Forces, usually called van der Waal's interactions.

c) Salt linkages between positively charged amino and guanidino groups and negatively charged carboxyl groups.

d) Hydrophobic bonds due to the large gain in entropy which

occurs when non-polar side chains associate by the exclusion of water.

An understanding of structure stabilisation via either hydrogen bonds or dispersion forces is only possible when the precise details of a structure are known. This is so since dispersion forces are very short range (energy varying as $\frac{1}{r^6}$, where r is the interatomic distance) and hydrogen bonds require a precise linear alignment of donors and receptors (Pauling, 1960). The salt-linkage (electrostatic) interactions, however, involve more isotropic potential distributions from individual charges and the energy of interaction is long range (varying as $\frac{1}{r}$). It is therefore not as necessary to know the precise structural details for studying possible salt-linkages as it is for hydrogen bonds or dispersion forces. This is also true for the hydrophobic interactions, which involve the bulk removal of apolar side-chains from contact with water. As there is still dispute over the exact structure of the collagen molecule, and even more uncertainty in the precise mode of molecular packing in fibrils (see Chapter 1), it was felt that, for the reasons given above, only the possible electrostatic and hydrophobic interactions should be studied with regard to the origins of the D stagger.

ii) Method

In the subsequent analysis the collagen molecule is treated as a linear sequence of evenly spaced amino acid residues. Interactions between two such 'molecules' are examined, with the aid of a computer, at all possible axial staggers. The approach is therefore a one-dimensional analysis. It is justified for the following reasons:-

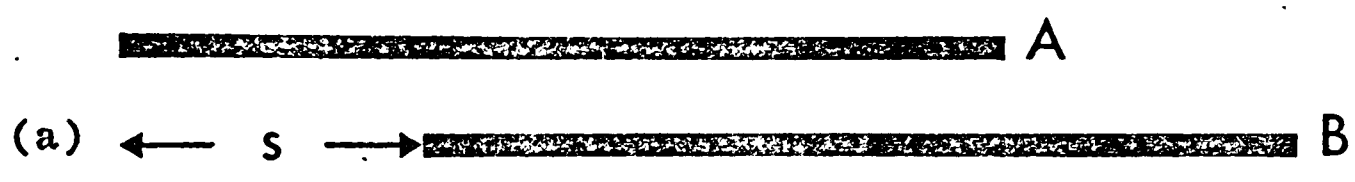
a) The main part of the collagen molecule, the helical region, consists of three regularly supercoiled α chains. The helical symmetry is such that any particular residue on one of the chains is related to the equivalent residue on the other two chains by either an axial shift (h) of 2.9Å and a rotation (τ) of 110° or by h = -2.9Å and τ = -110°. This is so throughout the whole 2900Å length of the helical region. Thus, with the assumption that the unknown α2 sequence can be approximated by the α1 sequence (see discussion), the interaction property of any one residue is repeated twice at approximately equi-angular intervals around the molecule. Furthermore all the non-glycine side chains face outwards from the core of the molecule and are thus available for possible intermolecular interactions. The triple helical structure, therefore, is such that the approximation to a linear, regular sequence of residues is probably reasonably valid, certainly much more so than in those other fibrous protein molecules which are composed of a smaller number of polypeptide chains and have some residue side-chains involved in intramolecular interactions (e.g. tropomyosin, see Stone et al., 1974). Both these features tend to accentuate the azimuthal variation in planes perpendicular to the fibre axis.

b) Some amino acid residue side-chains used in the following calculations are large and somewhat flexible. Thus even though the peptide groups of two interacting residues on adjacent molecules may be far apart, their side-chains may be capable of interaction. The azimuthal variation imposed by the backbone of the collagen triple-helical structure may therefore be considerably modified by

the side-chains. This is also true in the axial direction, where side-chain flexibility would add to the slight variation in the 2.86Å residue translation of the backbone. (A constant 2.86Å translation is only maintained over an average length of 90 to 140Å, as measured from high angle X-ray patterns (Miller et al., 1975; Wray, 1972)). The variability in side chain position is allowed for in the subsequent analysis by including, for each axial stagger between linear sequences, a range of interaction. This idea is explained with reference to Fig. 2.1. Fig. 2.1(a) shows two linear sequences, A and B, staggered by a distance S. For every value of S (i.e. from S = 0 to S = 1011, the number of residues in the helical region, see Fig. 1.2), the total number of intermolecular interactions are counted. The relative position of the residues in adjacent segments of sequences A and B for a particular value of S are shown in Fig. 2.1(b). Clearly for any one residue in sequence A there is one residue in sequence B that is exactly adjacent. A range of interactions is introduced when the residue on sequence A is allowed to search for possible interactions with not only the exactly adjacent residue on sequence B but also a limited number of residues to either side. For the present analysis a range of ±3 residues was used for electrostatic interactions and ±2 residues for hydrophobic interactions. The side-chains of the charged residues are longer than those of the larger apolar residues.

The interactions between 'molecules' were assessed using the following scoring system:-

a) An interaction between a negatively charged side-chain (aspartic or glutamic acid) and a positively charged side-chain (arginine or lysine) scores +1.



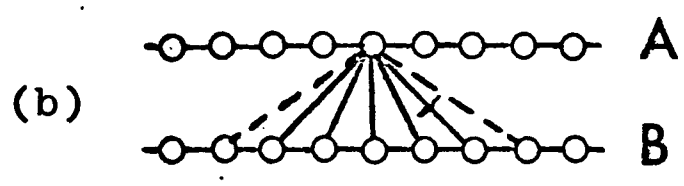


Fig. 2.1

(a) Two molecules, A and B, staggered by a distance s

(b) The relative positions of residues (open circles) in a small region of (a).

b) An interaction between two large hydrophobic residues (leucine, isoleucine, valine, phenylalanine or methionine) scores +1.

c) All other interactions score zero.

When the total interaction score is calculated for all possible intermolecular staggers the result is a so-called interaction curve. The computer programme used to calculate interaction curves is described in Appendix 1. In all the curves the ordinate is the total score divided by (2r + 1), where $\pm$ r is the interaction range.

iii) Results

The interaction curves obtained using the above scoring system and interaction ranges are shown in Fig. 2.2. When the intermolecular overlap is large (i.e. the stagger is small) there is a greater probability of finding many favourable interactions on a random basis than when the overlap is small. This explains the underlying negative slope common to both curves. Within the detailed fluctuations of the curves, however, there are local maxima at integral multiples of between 232 and 235 residues. The exact positions of the maxima are given in Table 2.1. Both the hydrophobic and the charge contributions have maxima in these regions. The corresponding staggers in Angstrom units can be obtained by multiplying the stagger in residues by the axial inter-residue rise in the collagen triple helix, as determined by Ramachandran (1967) using high angle X-ray diffraction. This latter value is 2.91 $\pm$ 0.01Å. The positions of the local maxima in the interaction curves are then integral multiples of 663 to 686Å. These limits correlate well with the value of 668 $\pm$ 2Å for the Hodge-Petruska D stagger

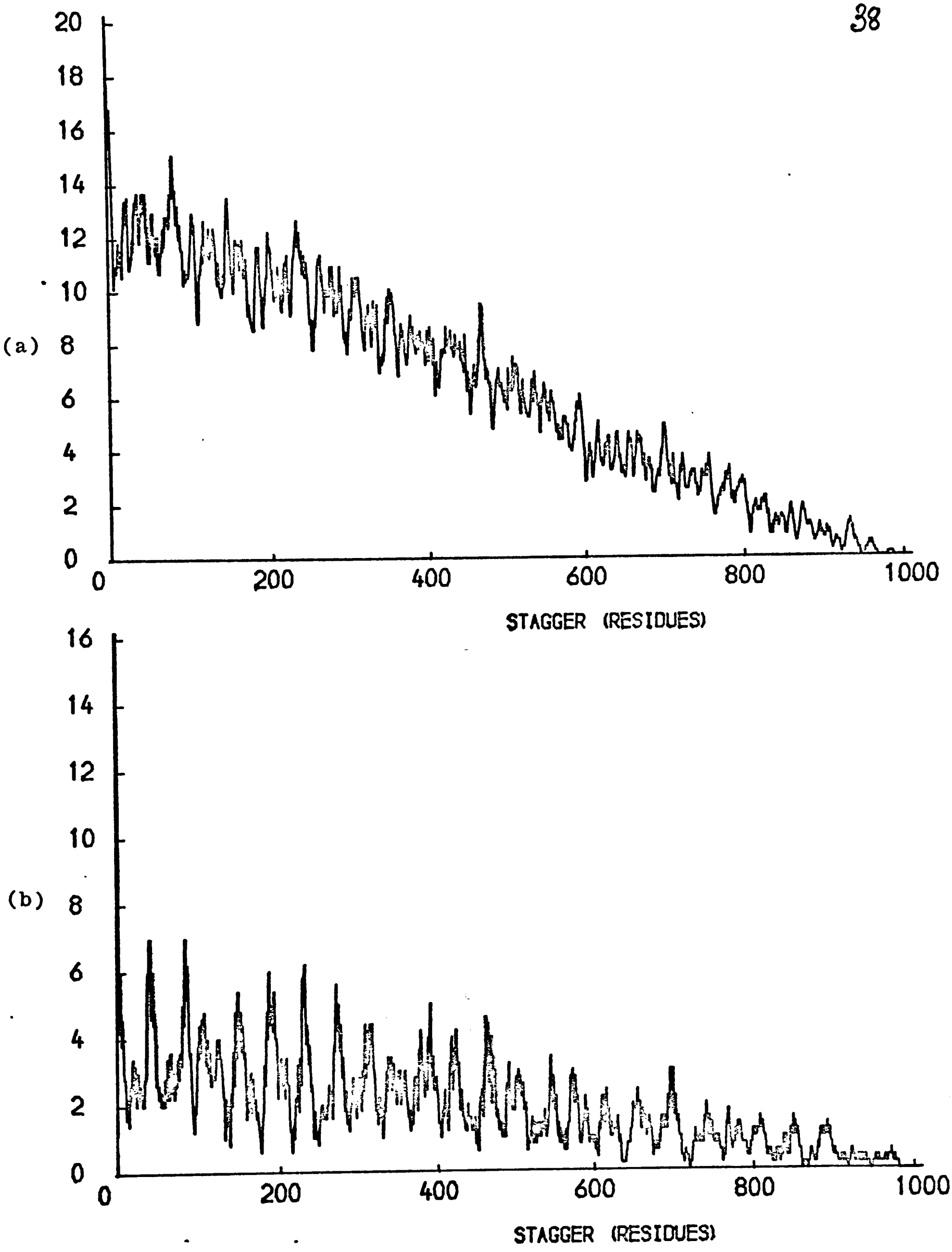


Fig. 2.2

(a) Interaction curve for electrostatic attractions between charged side-chains.

(b) Interaction curve for large hydrophobic side-chains.

TABLE 2.1 Positions of the n x D maxima in Fig. 2.2 (in residues)

Hydrophobic	Charged	Stagger
233	234	1D
464,466	468	2D
697,698,701	699,700	3D
-	933	4D

as determined by using low angle X-ray diffraction (Wray, 1972). The interaction peaks corresponding to multiples of the D stagger are not absolute maxima; there are other peaks which are higher. Nevertheless, the analysis has shown that the distribution of residues within the α1 sequence is such that favourable interactions can occur when molecules are staggered by integral multiples of D.

The statistical significance of the D related interaction peaks was assessed by comparing the results with those obtained using 20 random sequences. The latter were generated from the α1 sequences by randomising within positions corresponding to gly, X or Y in the repeating gly-X-Y triplet (see Fig. 1.2) but not between positions. The interaction curves were also normalised by dividing the score for each stagger by the corresponding molecular overlap. The peak values at intervals of D from the real sequence were 3 to 3.5 standard deviations from the mean, corresponding to probabilities of about 0.01 to 0.001 (Rattew, 1974; Hulmes et al., 1973).

In the above analysis electrostatic repulsions between like charges were not considered. When the calculation was repeated using the previous scoring scheme and interaction range but also scoring -1 for repulsions between either positive or negative charges, the curve shown in Fig. 2.3 was obtained. Clearly there are no significant peaks at integral D staggers. As pointed out earlier, however, the side chains are somewhat flexible and hence may be able to reduce the destabilising effect of repulsions by moving away from like charges.

Examination of the charge interaction curve in Fig. 2.2 yields an insight into the detailed distribution of charged residues within the sequence. The highest peak occurs when the linear sequences are in exact register. This result would be

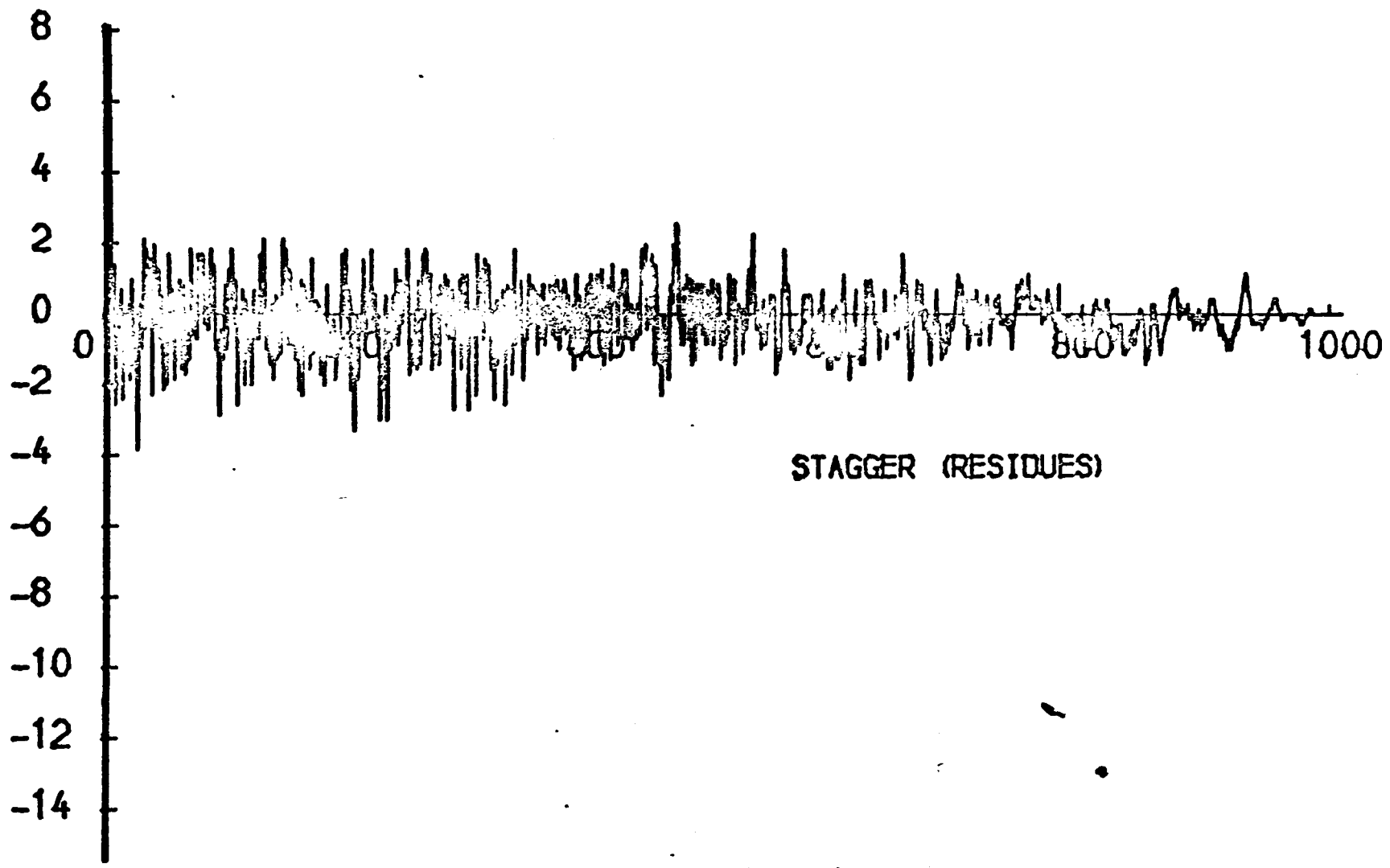


Fig. 2.3

Interaction curve for electrostatic attractions and repulsions between charged side-chains.

expected for the hydrophobic curve, where the interactions are between like residues, but it can only be explained for the charge curve if there is a strong tendency for oppositely charged residues to occur in pairs (separated by no more than the total allowed interaction range) within the sequence itself. This property also explains why there are favourable charge interactions at all integral multiples of D, and not just odd multiples, as would be expected without charge pairing (see Section 3 and also Rattew, 1974). In fact, of the 160 charged residues in the helical region of the αl sequence 100 occur as pairs of positively and negatively charged residues separated by no more than three residues. This compares with an average of 60 (standard deviation = 7) paired charges in the 20 random sequences. The paired and unpaired charges are listed in Tables 2.2 and 2.3 respectively.

It seemed worthwhile to investigate the possible role of this charge pairing in the sequence. The approach was to consider each charge pair as an electrostatic dipole and then investigate dipole-dipole interactions using the interaction curve analysis. When two oppositely directed dipoles from overlapping molecules were within a range of ± 3 residues, a score of +1 was given. No other interactions were considered. The result is shown in Fig. 2.4(a). The highest peak now occurs at a stagger of 235 residues. When the scoring scheme was extended to include a score of -1 for alignment of similarly directed dipoles the curve shown in Fig. 2.4(b) was obtained. Again the highest peak is at a stagger of 235 residues. The inclusion of dipole-dipole repulsions does not affect the position of the highest peak, unlike the effect of including charge repulsions in the original analysis. In the latter

TABLE 2.2 Paired charges in the rat/calf skin α1 sequence

plus-minus

ARG-ASP	ARG-GLU	LYS-ASP	LYS-GLU
144-146	183-186	50- 53	108-110
567-569	192-194	264-266	174-176
845-846	294-296	434-435	270-272
989-992	729-731	531-533	645-647
	737-740	564-566	722-723
	833-834	573-576	
	855-857	581-582	
	885-887	654-656	
		852-854	
		881-884	
		971-972	

minus-plus

ASP-ARG	ASP-LYS	GLU-ARG	GLU-LYS
384-386	96- 99	131-132	56- 59
620-621	372-374	308-309	288-290
	602-603	332-333	413-416
		347-350	476-479
		419-420	752-753
		452-453	800-803
		500-501	
		506-507	
		554-555	
		788-789	
		812-813	

The numbering begins at the N-terminal helical end.

TABLE 2.3 Unpaired charges in the rat/calf skin α1 sequence

ARG	LYS	ASP	GLU
9	87	54	29
42	219	147	32
62	252(18)	213	257(23)
66	327(93)	321(87)	287(53)
75	342(108)	488(20)	353(119)
90	360(126)	528(60)	437(203)
126	408(174)	639(171)	455(221)
134	681(213)	650(182)	482(14)
237(3)	726(24)	759(57)	510(42)
291(56)	915(213)		605(137)
315(81)	927(225)		623(155)
396(162)			830(128)
498(30)			840(138)
519(51)			947(11)
585(117)			
615(147)			
684(216)			
701(233)			
777(75)			
786(84)			
903(201)			
912(210)			
924(222)			
930(228)			
960(24)			
987(51)			

The numbering begins at the N-terminal helical end. Numbers in parentheses refer to positions in the D period (D is 234 residues).

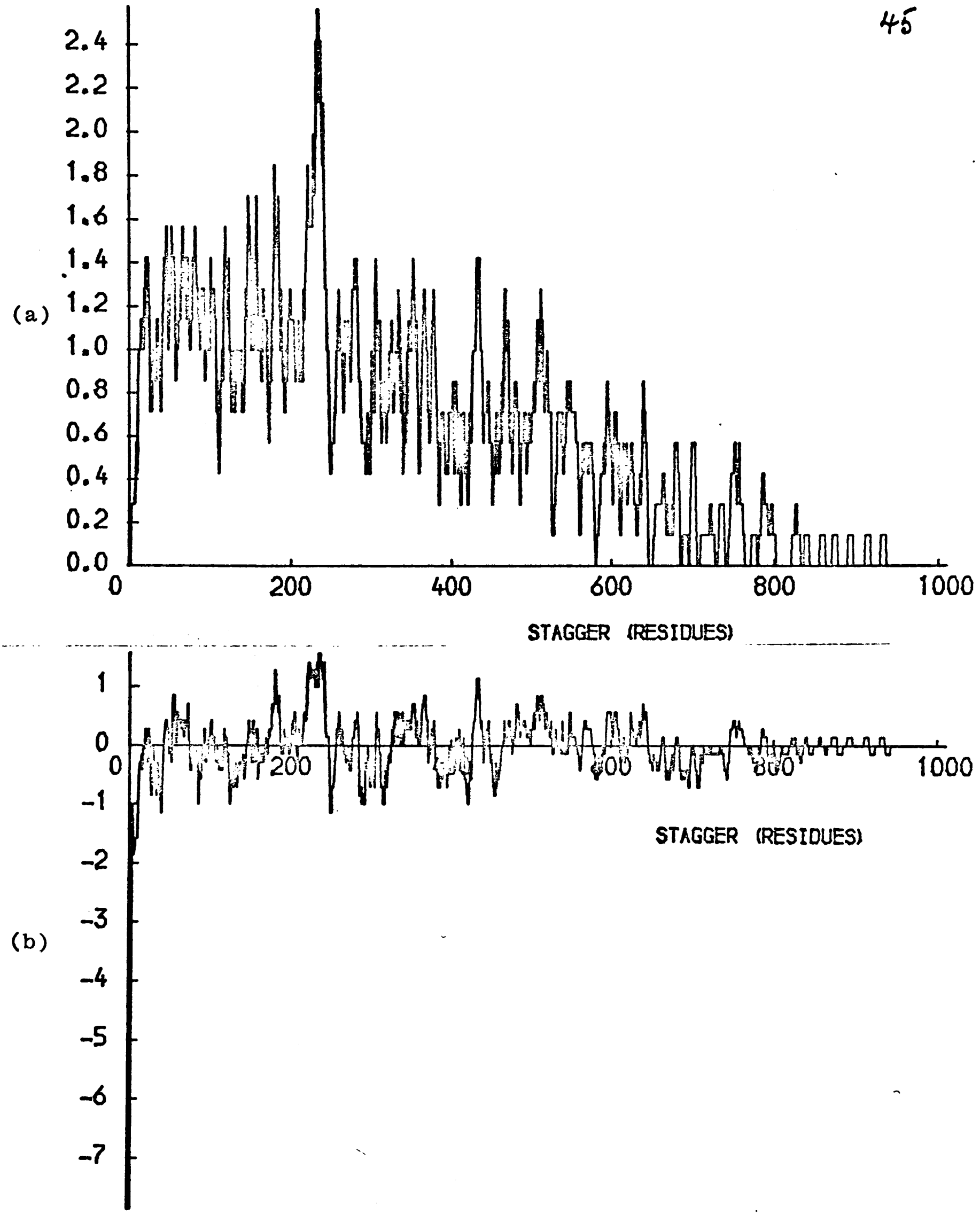


Fig. 2.4

(a) Interaction curve for attractions between oppositely directed dipoles.

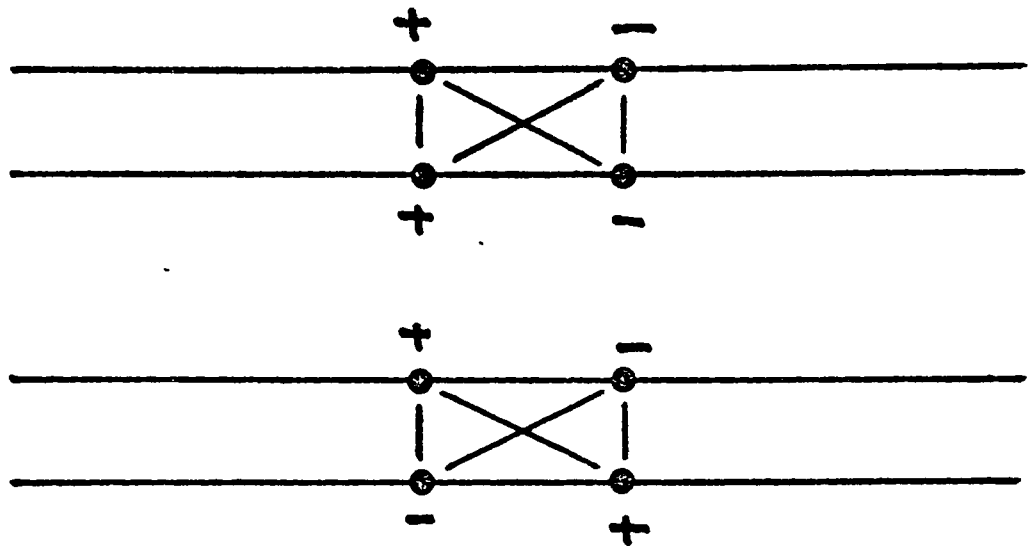
(b) Interaction curve for attractions and repulsions between dipoles.

approach the paired charges would not contribute much to the interaction curve since, when two charge pairs are exactly aligned, the net score would be zero, whatever their relative orientations, see Fig. 2.5. (The net score would oscillate very rapidly as the charge pairs move through the position of exact alignment. This explains the 'noisy' appearance of Fig. 2.3).

The interaction curves for the remaining unpaired charges were also calculated. Fig. 2.6(a) shows the result for electrostatic attractions only, and Fig. 2.6(b) shows the result for attractions and repulsions. The highest peaks are at staggers of 261 and 263 residues respectively, neither of them corresponding to the D stagger. The only indication of D stagger specificity in these curves is a peak at 469 residues (2D) which is the 5th highest in Fig. 2.6(a) and the second highest in Fig. 2.6(b).

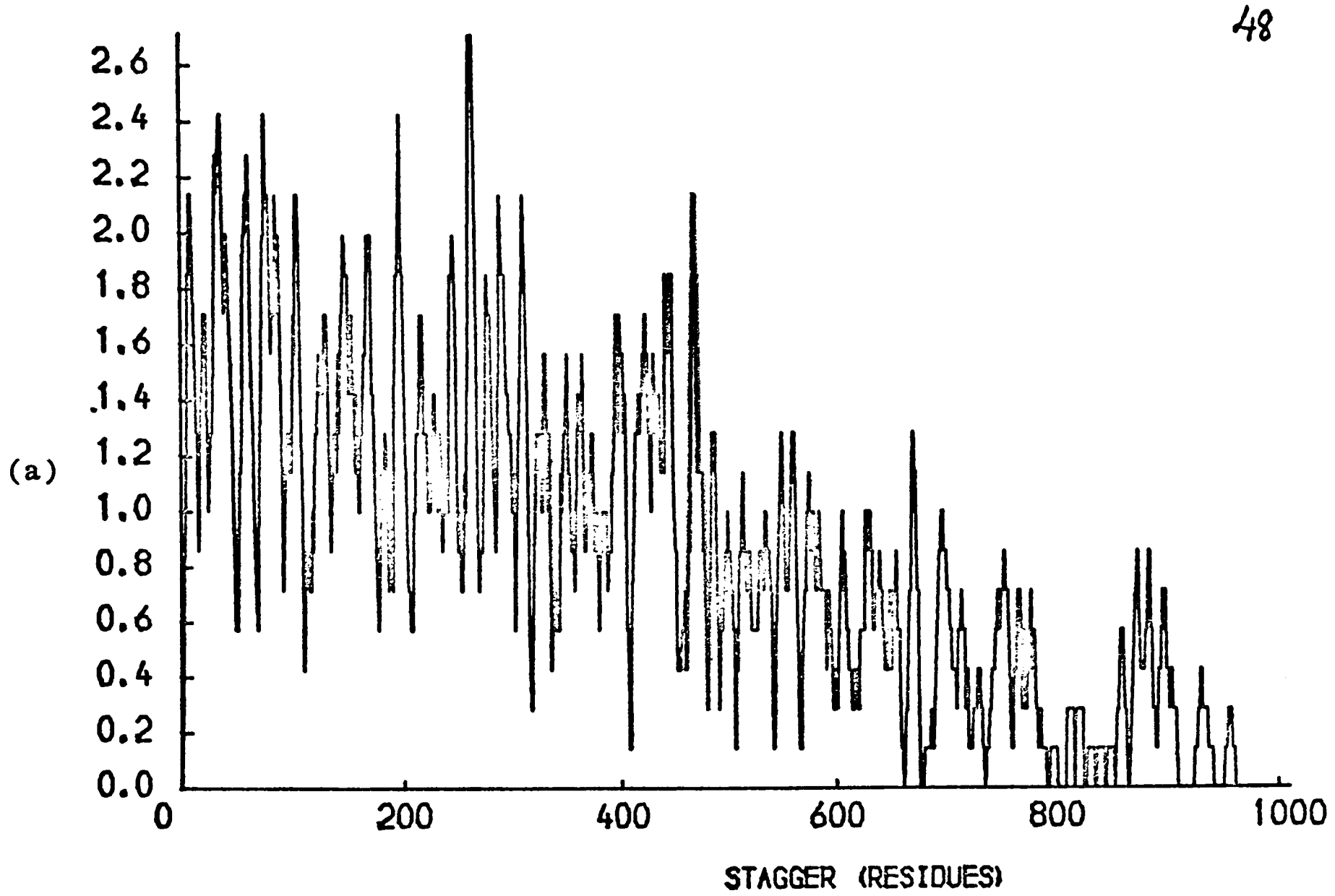
iv) Discussion

The most marked evidence for a D stagger specificity is seen in the interactions between those residues occurring in charge pairs. The interaction curve for electrostatic attractions only between all charged residues (Fig. 2.2(a)) shows peaks at all integral multiples of D. This is not the case in the dipole-dipole attractions in Fig. 2.4, where there is one major peak at 1D. As noted earlier the maximum hydrophobic interaction occurs at zero stagger, but there are also peaks at all integral multiples of D. It seems therefore that the specificity for a D staggered intermolecular arrangement is due to charged interactions, probably those involving charge pairs, whilst the hydrophobic interactions provide additional stability. The distributions of residues which give



Attractions	Repulsions
2	2
2	2

Fig. 2.5 Illustrating how dipole-dipole interactions would **not** be counted when **both** attractions and repulsions **are** considered between individual charged side-chains.



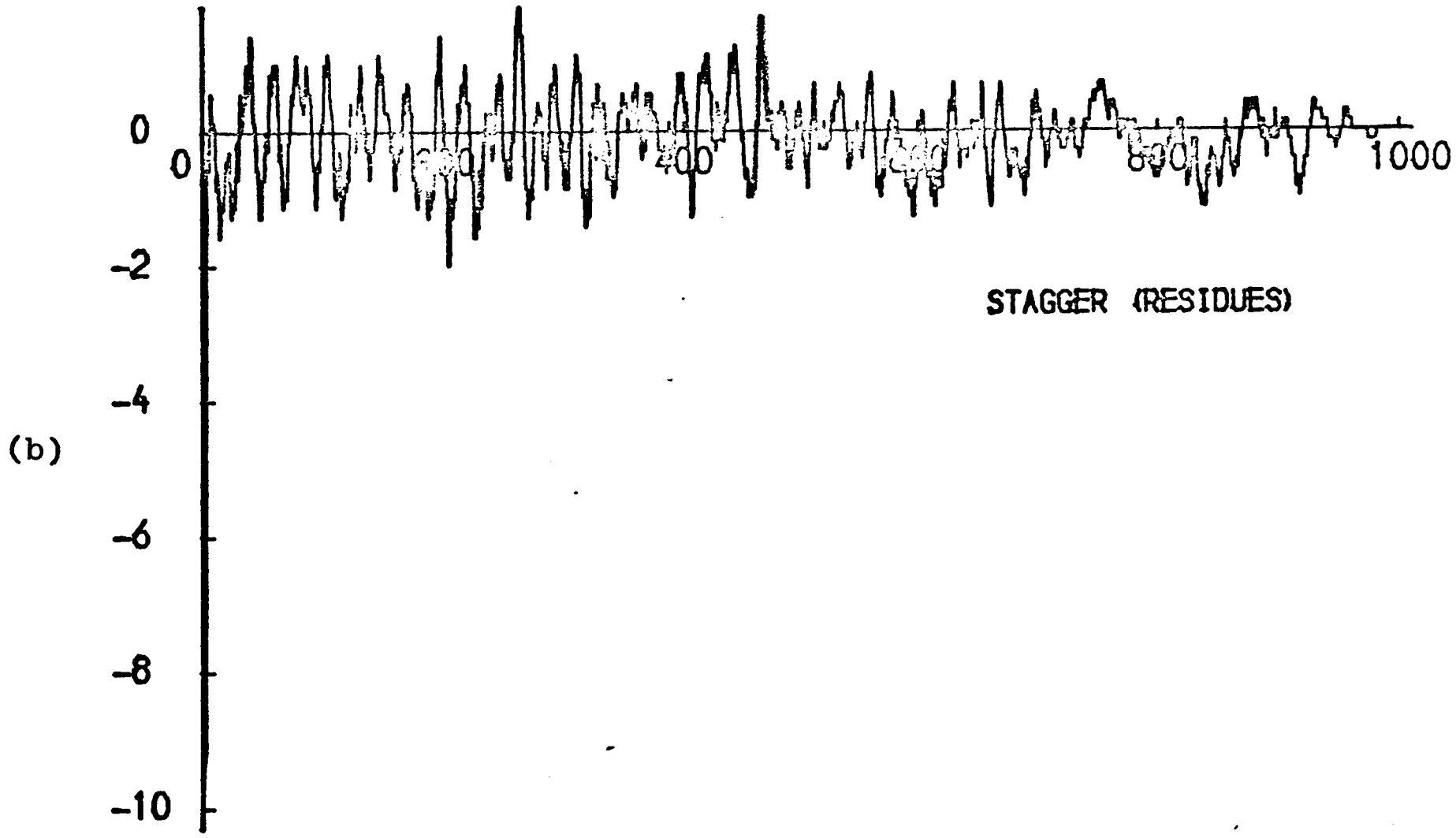


Fig. 2.6

(a) Interaction curve for attractions between unpaired, charged side-chains.

(b) Interaction curve for attractions and repulsions between unpaired, charged side-chains.

rise to these phenomena will be discussed in Section 4.

Clearly the interaction curve analysis is only in the crudest sense an energy calculation. The heights of the various peaks can only serve as a rough guide to the energy involved for each stagger. Furthermore the addition of two different kinds of interaction (i.e. hydrophobic and electrostatic) is not really valid without a knowledge of the energies involved.

The above interaction analysis was calculated for parallel molecules. When the interactions for anti-parallel molecules were calculated some peaks were higher than the main peaks for parallel molecules. Unlike the parallel interaction analysis, however, the main anti-parallel peaks for charge interactions were not coincident with those for hydrophobic interactions (see Fig. 2.7). This suggests that parallel alignment of molecules *in vivo* is due to the combined effect of both types of interaction.

In the next section a way of experimentally verifying the theoretically determined value for D is described.

3. Electron Microscopy and the Amino Acid Sequence

i) Introduction

When collagen molecules in solution are precipitated in the presence of adenosine triphosphate (ATP) structures are formed in which neighbouring molecules are aligned in parallel register. This structure was termed segment long spacing (SLS) collagen by Schmitt *et al*. (1953). When SLS is positively stained 58 parallel dark bands of varying width and separation can be observed (Bruns and Gross, 1973). By causing dimorphic growth of SLS on reconstituted native-type fibrils, Hodge and Schmitt (1960) were able to demonstrate a correspondence between

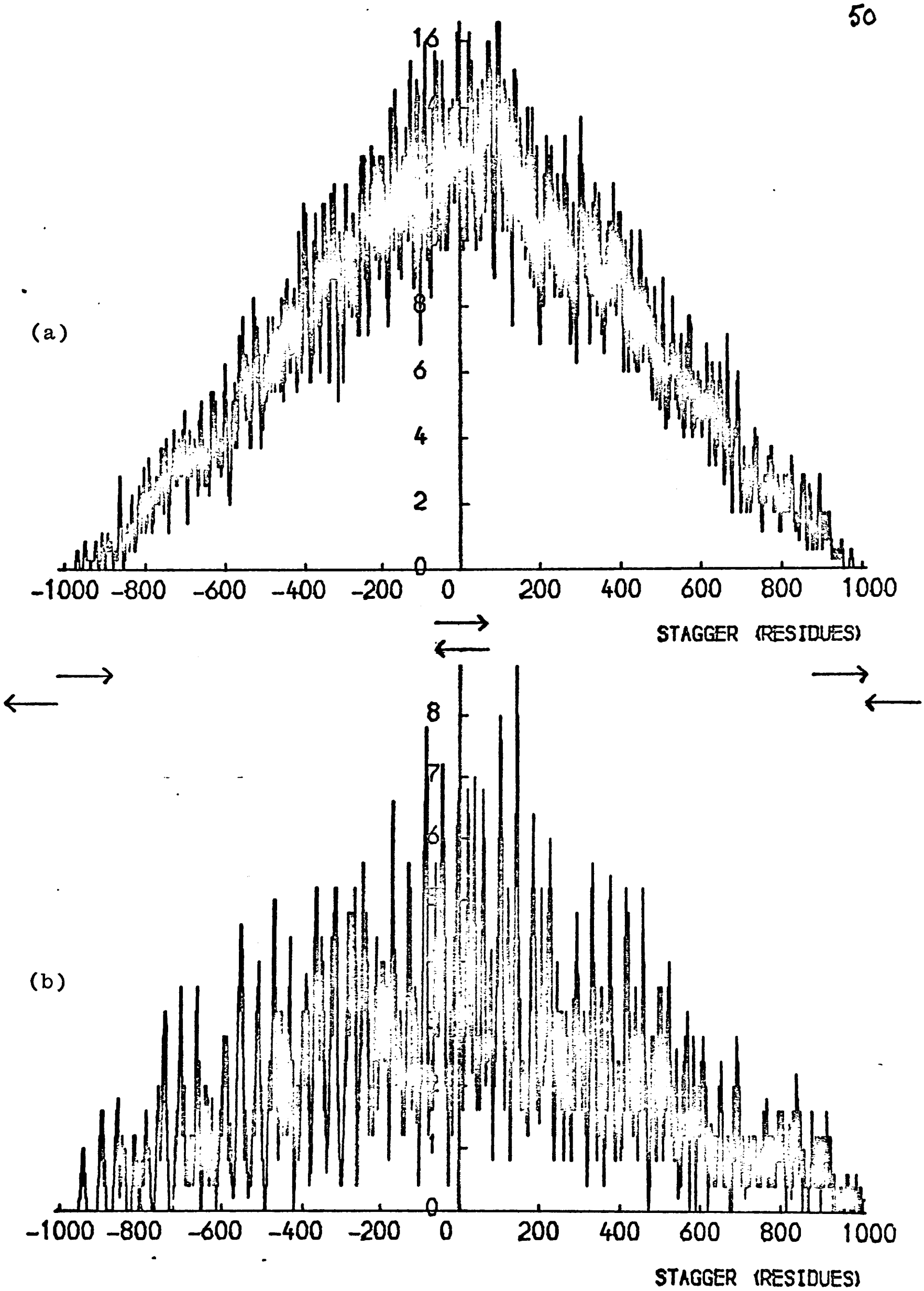


Fig. 2.7 Interaction curves for anti-parallel sequences.

The molecular arrangements at certain staggers are indicated by horizontal lines with arrowheads at the N-termini.

(a) Electrostatic attractions between all charged side-chains.

(b) Hydrophobic interactions.

the bands of SLS and the native-type banding pattern (see also Bruns and Gross, 1974). There is therefore a pseudo D periodicity in the SLS pattern, such that bands occur in the same positions in each D period but differ in width and intensity. Hodge and Schmitt (1960) also demonstrated the origins of the native-type positively stained banding pattern in the SLS pattern. By photographically superimposing the latter upon itself after repeated shifts of approximately one quarter of the molecular length (i.e. the segment length perpendicular to the bands) a low resolution agreement with the native banding pattern was observed. Kuhn and Zimmer (1961) carried out a similar analysis using multiple addition of SLS microdensitometer traces.

If the positively stained SLS banding pattern can be understood in terms of the amino acid sequence then the shift required to simulate the native pattern could be interpreted in terms of a number of residues. Hence there exists a possible way of checking the theoretical D stagger.

ii) Methods and Results

A method of correlating the sequence with the SLS pattern was suggested by von der Mark et al. (1970). They were able to diagrammatically represent the regions of high charge density in a 112 residue sequence from the α1-CB6 cyanogen bromide peptide from calf skin collagen. This was achieved by first assuming a constant residue translation along the molecular axis and then representing each charged region by a black block. By suitably adjusting the relative magnifications and comparing this diagram with the SLS banding pattern these authors were able to demonstrate a striking correspondence with one region of the SLS pattern.

With the completion of the rat/calf skin α1 sequence it was possible to extend the work of von der Mark et al. to the entire SLS pattern. Again a diagram was made of the distribution of charged residues assuming a constant residue translation. This time the computer linked graphical display equipment of the Oxford ICL 1906A Computing Service was used to achieve high accuracy. The programme was written so that the position of every Arg, Lys, Asp or Glu in the sequence was represented by a vertical line corresponding to a width of 10Å (assuming a length for the helical region of 2900Å). This width was chosen to make the resolution compatible with the electron microscope. (The subroutines used in this and subsequent banding pattern diagrams are described in the Oxford University Computing Service Manual, in the CUGHOST section). The resulting diagram is shown in Plate 2.1(b). Plate 2.1(a) is an electron micrograph of SLS collagen from calf skin positively stained with phosphotungstic acid and uranyl acetate. Both the interpretation of positive staining and the assumption of a constant residue translation are seen to be justified, at least to the limit of resolution of the electron microscope, by the correspondence between these patterns.

With the SLS-sequence correspondence it is possible to proceed to the native-type banding pattern. This was simulated from Plate 2.1(b) in an analagous way to that of Hodge and Schmitt (1960), though in a more quantitative manner. The method used was to make a multi-exposed negative by repeatedly photographing a diagram consisting of two copies of Plate 2.1(b) separated axially by 5D residues. A total of 5 exposures were made; the diagram was shifted axially by D from left to right between each exposure. The resulting negative contained a region in the centre of length 5D which showed a D periodic

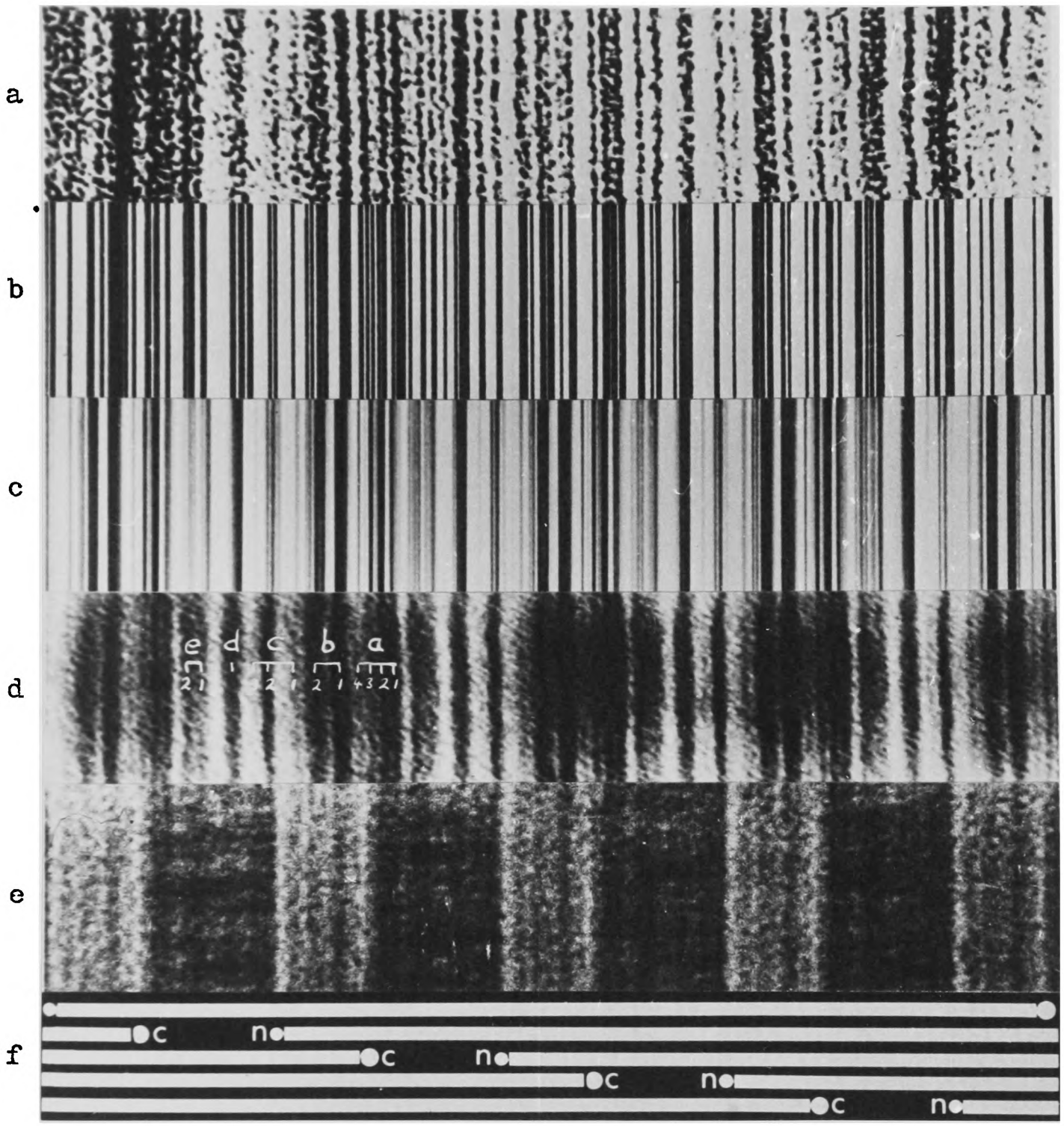


PLATE 2.1

a) Electron micrograph of positively stained SLS collagen (courtesy of Dr. K.A. Piez).
b) Simulation of (a) based on the α1 sequence (see text).
c) Simulation of the native-type positively stained banding pattern (see text).
d) Electron micrograph of a positively stained fibril from rat tail tendon (courtesy of Dr. J. Woodhead-Galloway).
e) Electron micrograph of a negatively stained fibril from rat tail tendon (courtesy of Dr. J. Woodhead-Galloway).
f) The Hodge-Petruska scheme for molecular packing.

banding pattern. The accuracy of this procedure was estimated to be within $\pm$2 residues. The chosen value for D was that suggested by the interaction curves in section 2, i.e. 234 residues. A print of the central region of the multi-exposed negative is shown in Plate 2.1(c), for comparison with a micrograph of a positively stained collagen fibril from rat tail tendon in Plate 2.1(d). The correspondence between these patterns provides an experimental verification of the theoretically determined D stagger.

iii) Discussion

The agreement between simulated and observed banding patterns is a sensitive function of D. This was demonstrated by Chapman and Hardcastle (1974) who showed that the visual agreement for D less than 231 or greater than 234 residues was so poor that the true value must be within this range. These authors also carried out a quantitative correlation analysis which yielded a value for D of 232.3 $\pm$ 0.5 residues. As Chapman and Hardcastle pointed out, however, the accuracy of this result is based on the assumptions that (a) the distribution of charged residues in the α2 chain is identical to that of the α1 chain and (b) the residue translation is uniform throughout. Both these assumptions have limited validity. The total number of charged residues in the rat or calf skin α2 chain is close to that in the α1. Of 250 α2 residues that have been sequenced (see Fietzek et al., 1974) approximately one third of the charged side chains undergo a non-charge substitution going from α1 to α2. If this is a general feature of the α2 chain, then, of the ~170 residues in the complete α2 sequence, roughly 55 (170/3) are located at different sites from the α1. Therefore in a molecule composed of

two α1 chains and one α2 chain the positions of ~450 (170 + 170 + 170 - 55) of the total ~500 (3 x 170) charged residues are known, i.e. 90%. With regard to the residue translation, measurements of the meridional width of the high angle X-ray reflection corresponding to this spacing indicates that it is only uniform for an average axial length of 90 to 140Å (Miller et al., 1975; Wray, 1972).

There is a third limit to the accuracy of any high resolution quantitative correlation analysis between simulated and observed banding patterns. This is due to the as yet unknown structure of the non-helical telopeptides at each end of the collagen molecule. In the α1 sequence there are 8 charged residues in the telopeptides. These would not make any striking difference to the appearance of Plate 2.1(c), as there are 160 charged residues in the helical region. They would however contribute to the kind of detailed analysis used by Chapman and Hardcastle (1974). If the entire α2 sequence were known and non-uniformity of the residue translation in the helical region could be understood, in terms of the effect on the distribution of positively stained bands, then a high resolution quantitative analysis of the native banding pattern may provide a method for determining the positions of the telopeptide charged residues. This could then give an insight into the conformation of the non-helical peptides.

The species differences between rat and calf skin are probably too few to affect the above analysis. The cyanogen bromide peptides are all of the same length and substitutions within the helical part of the molecule are rare. For example, there are only four differences in the 149 residues of α1-CB3

and they are all chemically conservative (W.T. Butler, private communication). Furthermore it is valid to compare patterns based on a rat/calf skin sequence with those observed in rat tail tendon since the α1 chains from rat skin and tendon are derived from the same gene (Bornstein, 1969).

In summary there are, at present, limits to the accuracy with which the D stagger can be estimated, in terms of the sequence, using electron microscopy. These limits can only be to a few residues as the matching of the sequence to the SLS pattern is so accurate. It is therefore possible to estimate the D stagger from electron microscopy to be between 231 and 234 residues. This is in excellent agreement with the theoretically determined value of 232 to 235 residues in section 2.

4. Patterns of Residues within the α 1 Sequence

i) Introduction

In this section the specificity for the D stagger in the hydrophobic and charged residues will be examined in terms of the distribution of those residues in the α1 sequence. First some general comments can be made.

Consider two parallel molecules, A and B, staggered by D, as shown in Fig. 2.8(a). If this interaction is favourable then a distribution of residues in molecule A must be complementary to another distribution in B. For hydrophobic interactions complimentarity would be the alignment of similar residues; for charged interactions it would be the alignment of either oppositely charged residues or oppositely oriented dipoles. If the molecules A and B are identical, the favourable D stagger interaction implies a D separation of complementary residues within each molecule. This is illustrated

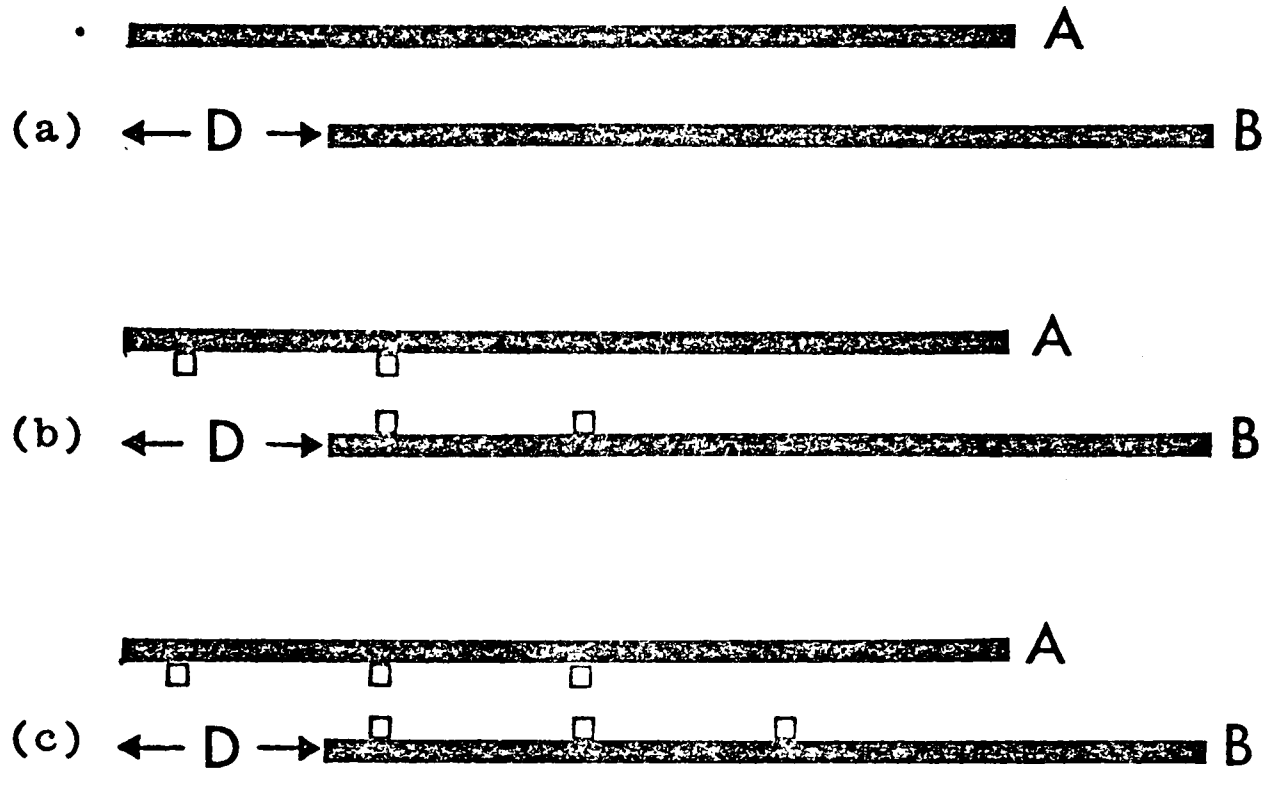


Fig. 2.8 Illustrating the way complementarity generates periodicities within the sequence (see text).

in Fig. 2.8(b) for the case of one interaction. There is then a situation in which two complementary residues are aligned and another two are isolated. For maximum efficiency in specifying the D stagger, these isolated residues should also be involved in favourable interactions. This requires the presence of another complementary residue in the sequence, as shown in Fig. 2.8(c). The equivalence of the molecules results in two more isolated residues. The process continues in this way until complementary residues are separated by D along the entire length of the molecule. The evolutionary implications of this principle have been discussed by Segrest and Cunningham (1973).

The above example shows only one interaction per D repeat but the same principle would hold for <u>any</u> distribution of interactions within D. On this principle, hydrophobic interactions would be expected to maximise at staggers of nD for all integral values of n from 0 to 4. Charge interactions would maximise only when n is odd. The observation that charge interactions are maximised at <u>all</u> integral multiples of D in Fig. 2.2(a), is evidence for charge pairing in the sequence (Section 1 and Rattew, 1974).

In summary, the above principle implies a D separation of complimentary groups of residues along the molecule. The only restriction on the pattern of residues within D is that it must not be such as to divide D into an integral number of n equivalent segments, in which case a D/n stagger would be specified.

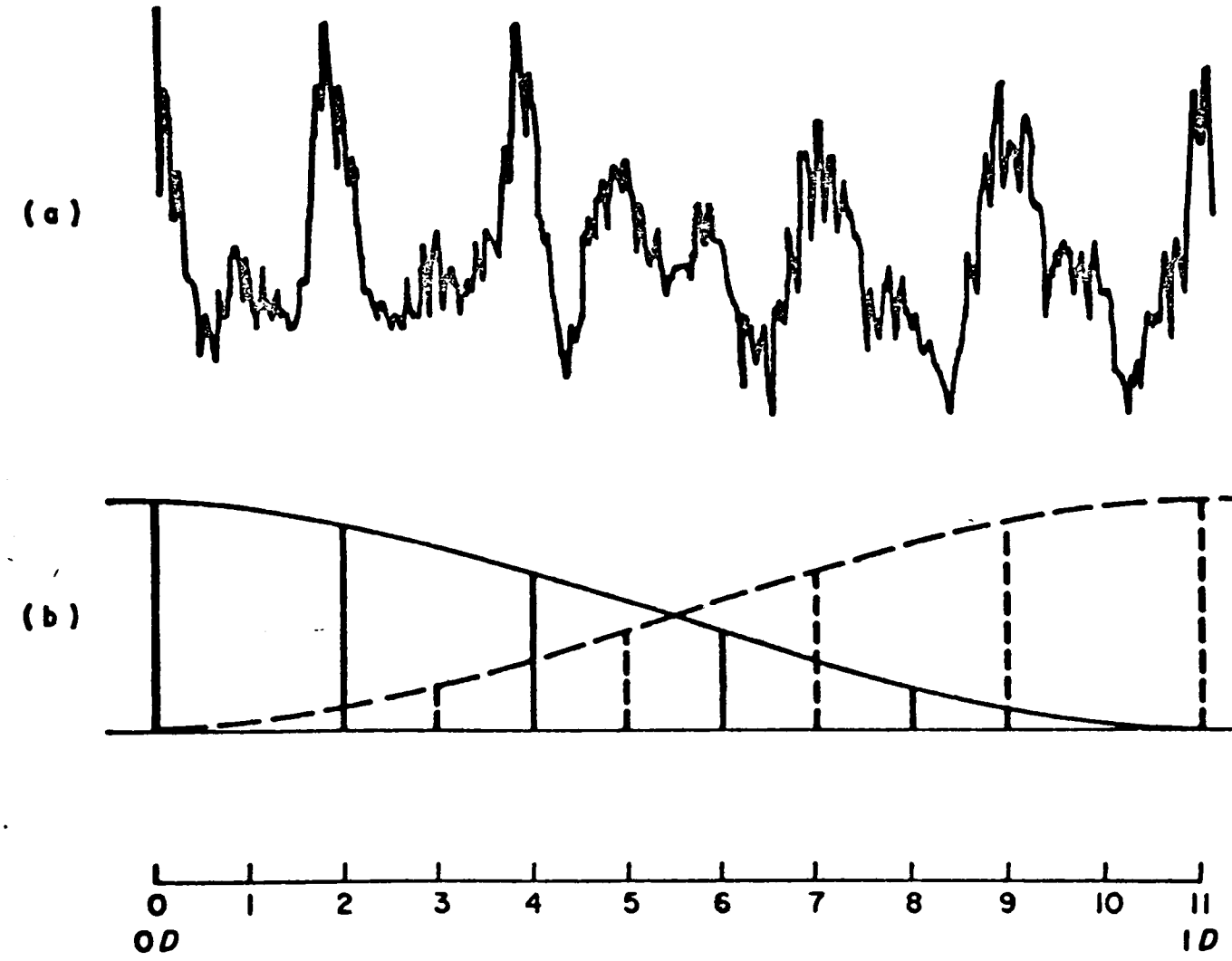
ii) <u>Hydrophobic residues</u>

The hydrophobic interaction curve in Fig. 2.2(b) shows regular fine structure. Between the peaks at multiples of D

there is a series of peaks at multiples of D/11. This is clear in the enlargement of the O to 234 residue region shown in Fig. 2.9(a). The pattern is repeated, though with smaller peaks, in every D segment. It can be considered as the sum of two oscillations of frequency 11/2D separated by D/11. Each oscillation is modified by a wave of frequency 1/2D. The 1/2D waves are 180^{o} out of phase. This is shown diagrammatically in Fig. 2.9(b).

Of the 62 hydrophobic residues in the helical region of the α1 sequence, 55 are involved in interactions in at least one of the 1D, 2D, 3D or 4D staggers. There is therefore a highly efficient use of residues and the principle of a D repeating pattern (Section 3(i)) must hold. Possible hydrophobic patterns can be inferred from Fig. 2.9. Since complementarity for hydrophobic interactions involves similar residues, the interaction curve is equivalent to an auto-correlation function. The inference of the repeating pattern is therefore similar to solving a Patterson map, in X-ray crystallography, for the position of heavy atoms. Fig. 2.9 suggests a repeat of 2D/11 exists. Since 2 is not divisible into 11 the 2D/11 repeat must start afresh every D. There would therefore be a discontinuity in the 2D/11 pattern every D equivalent to a shift of D/11. 2D/11 patterns separated by 2D would therefore be in phase; this is the origin of the 2D modulation of the interaction curve. A possible pattern within a D repeat is therefore $(2D/11)_5,(D/11)$. The auto-correlation function of this pattern is in fact Fig. 2.9(b).

If the positions of the hydrophobic residues are plotted along a line no repeating pattern is immediately apparent. However, it was considered that some of the residues might be involved in interactions more than others and the pattern was



g. 2.9. **(a) An enlarged portion (0–234 residues) of the hydrophobic interaction curve in Fig. 2.2. (b) An idealized representation of the curve showing the pattern of which it is composed.**

lost in the background. To bring out the suspected pattern the interactions between one molecule and two others separated by 5D, were counted for each residue at staggers of OD, 1D, 2D, 3D and 4D. Of the 62 residues, 7 were found to interact at only one stagger position (OD since all residues interact here), 17 at two positions, 18 at three positions, 17 at four positions and 3 at all five positions. A diagram was then drawn of the distribution of hydrophobic residues in the sequence, with each position indicated by a vertical line whose height was proportional to the number of nD interactions involving that residue. The result is shown in Fig. 2.10. Starting at the N-terminus there is a somewhat imperfect repeating pattern of $(2D/11)_4$, (D/11),(2D/11) . This is equivalent to the $(2D/11)_5$,(D/11) pattern inferred from Fig. 2.9 shifted by 2D/11.

As stated earlier, the principle of a repeating pattern in each D segment of the molecule does not impose any restriction on the nature of this pattern (except that it must not further divide the D period into an integral number of repeating segments). The observed $(2D/11)_4$,(D/11),(2D/11) pattern is, however, highly regular. The derivation of the principle of a D repeating pattern was, however, carried out with no regard to the true three dimensional packing of molecules. It seems likely that the hydrophobic pattern is related to the function of these residues in molecular packing. This is the case in tropomyosin, for example, where the repeating hydrophobic pattern every 7 residues has been shown to be related to the packing of chains in the molecular two stranded coiled coil (Crick, 1953; Stone et al., 1974). Furthermore an approximately 19½ residue repeat in tropomyosin may

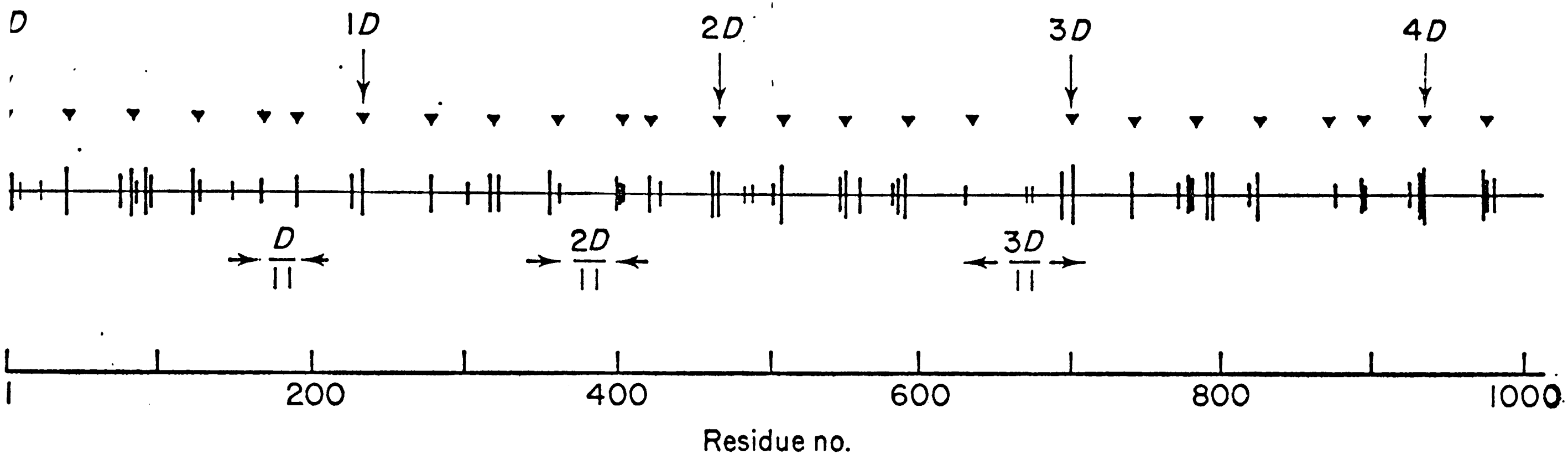


Fig. 2.10 . A representation of the locations of amino acids with large hydrophobic side chains in the α1 chain of collagen. The positions are indicated by lines whose length is proportional to the number of times the residue is involved in an interaction at staggers of $0D$, $1D$, $2D$, $3D$ and $4D$. The very short light lines indicate residues that interact at $0D$ only. The arrows show the patterns composed of intervals of $D/11$, $2D/11$ and $3D/11$. The residue positions are numbered beginning at the N-terminal end of the triplet region.

be related to the binding of this molecule to actin in the thin filaments of myofibrils (Miller, unpublished work; Parry, 1974; McLachlan and Stewart, 1975).

If the pattern of hydrophobic residues in collagen could be shown to be composed of a few simple periodicities then the analogy with tropomyosin may well be appropriate. Each component periodicity could have a different role in molecular packing (see discussion). This idea was tested by carrying out a Fourier analysis of the position of the α1 hydrophobic residues using the computer programme described in Appendix 2. The position of each hydrophobic residue was represented by a δ-function of unit height. The amplitudes and phases of the Fourier transform shown in Table 2.4 are those obtained when the sequence is convoluted with a one-dimensional lattice of spacing D. There are three large Fourier components, corresponding to wavelengths of D/5, D/6 and D/11. When these three waves are combined with the correct amplitudes and phases, the familiar $(2/11)_4$,(1/11),(2/11) pattern is regained, see Fig. 2.11. It should be noted that these strong Fourier components are real features of the hydrophobic distribution, i.e. not due to the density variation imposed by the gap/overlap phenomenon (Chapter 1), since the same features were apparent when only the first 936 residues (4D) were Fourier analysed.

It may be thought that this Fourier description of the pattern of hydrophobic residues is an unnecessary complication. However when other residues show similar periodicities, as described below, the analysis helps to form a more unified description of the α1 sequence.

TABLE 2.4 Fourier analysis of the positions of the hydrophobic* residues in the helical region of the α1 sequence Phase origin at residue 59.

Frequency (x 1/D)	Amplitude	Phase (°)
0	62.0	0
1	7.4	0
2	8.3	114
3	8.6	143
4	4.5	72
5	20.1	189
6	26.7	206
7	6.7	9
8	5.7	310
9	14.0	128
10	7.1	161
11	19.9	38
12	8.0	309
13	3.0	13
14	4.1	305
15	3.9	339
16	4.2	128
17	12.1	213
18	8.4	155
19	7.6	278
20	4.8	284
21	5.1	317
22	0.9	42
23	5.8	69
24	7.8	25
25	5.1	231

* valine, leucine, isoleucine, phenyalanine and methionine.

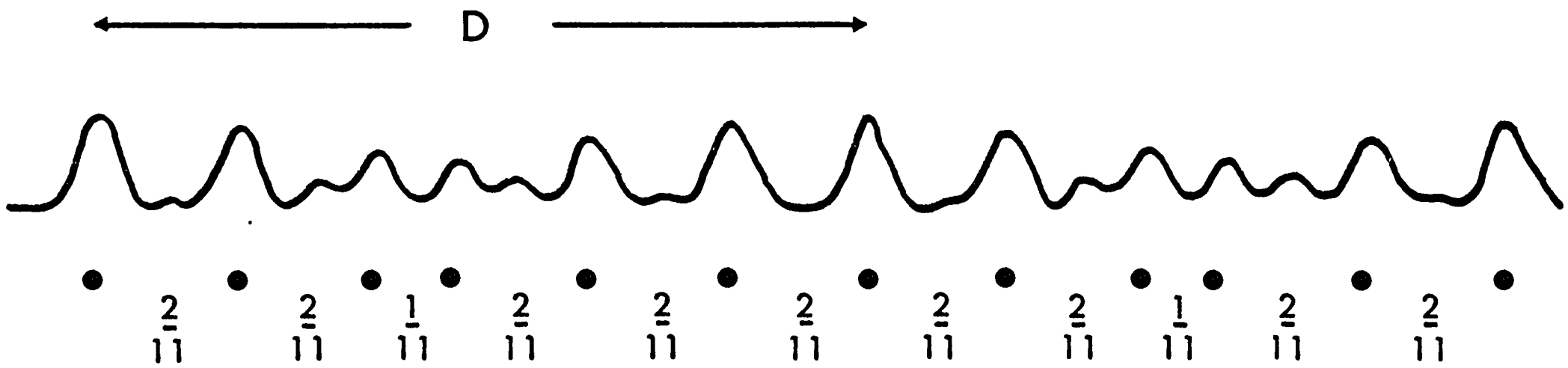


Fig. 2.11 Fourier synthesis of the pattern of large hydrophobic residues in the α1 sequence, using only the D/5, D/6 and D/11 components in Table 2.4.

iii) Charged residues

The charged interaction curve in Fig. 2.2 does not show a regular fine structure like the hydrophobic curve. This does not entirely preclude a regular charge pattern within D however, since charge complimentarity is such that the interaction curve is not equivalent to an auto-correlation function. In this case a more appropriate search for patterns is through the Fourier analysis. As for the hydrophobic residues above, the positions of charged residues were represented by δ-functions of unit height (with no regard to the sign of the charge) and then convoluted with a regular D lattice. The amplitudes and phases are shown in Table 2.5. Strong components now correspond to wavelengths of D/6 and D/21. Again these features are real properties of the charge distribution (see previous section). The phase of the D/6 wave is approximately 180^{o} out with respect to the corresponding hydrophobic wave.

The strong D/6 contribution is also shown in an optical diffraction pattern of an electron micrograph of a positively stained fibril, see Plate 2.2. The nearest maximum of the D/6 wave to the chosen phase origin corresponds to band b_1 in Plate 2.1(d). The source of the D/6 periodicity can be seen in the approximately equal separation of bands b_1, b_2, c_2, d, e and a.

The D/6 pattern is clear in the distribution of paired charges in the D period, as shown in Fig. 2.12. The positions of the dipoles in each D segment of the sequence are approximately equivalent, though their orientations are such that each segment is unique. Segments a and c or segments b and d are not equivalent. This is clear when groups of equally directed

TABLE 2.5 Fourier analysis of the positions of the charged* residues in the helical region of the α1 sequence. Phase origin at residue 59.

Frequency (x 1/D)	Amplitude	Phase (°)
0	160.0	0
1	17.4	36
2	14.4	312
3	17.2	264
4	8.0	216
5	16.8	270
6	30.3	38
7	15.4	280
8	16.6	30
9	20.0	314
10	20.9	354
11	14.6	246
12	13.4	283
13	15.4	326
14	17.5	181
15	9.8	257
16	14.8	231
17	4.7	27
18	14.3	41
19	17.7	98
20	13.5	129
21	26.4	172
22	17.2	198
23	7.0	75
24	11.9	299
25	14.5	49

* arginine, lysine, aspartic acid and glutamic acid.

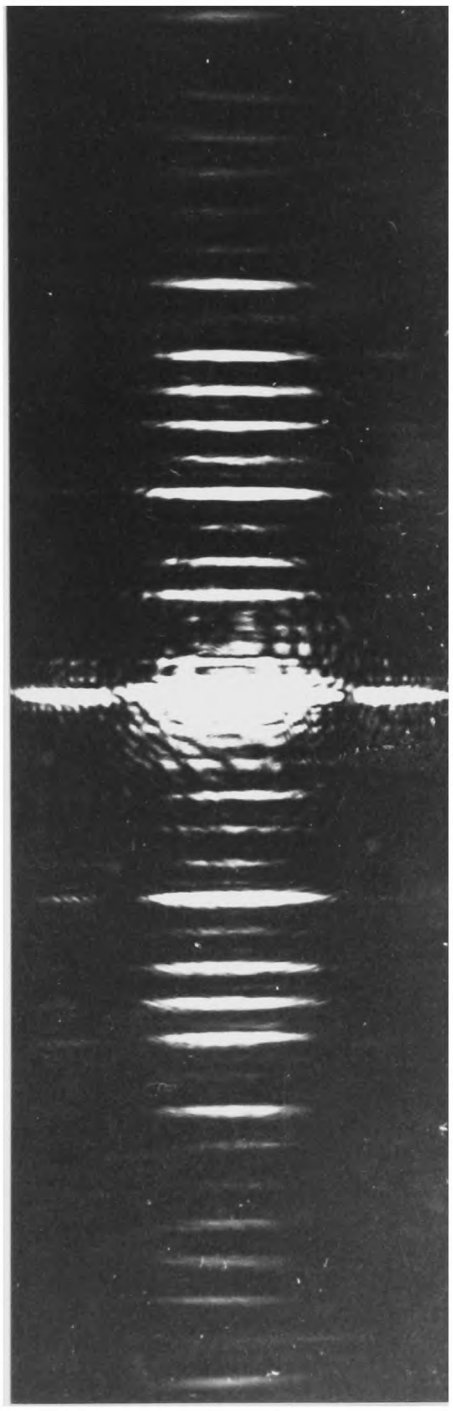


PLATE 2.2

Fraunhoffer diffraction pattern of Plate 2.1(d), showing the strong sixth order.

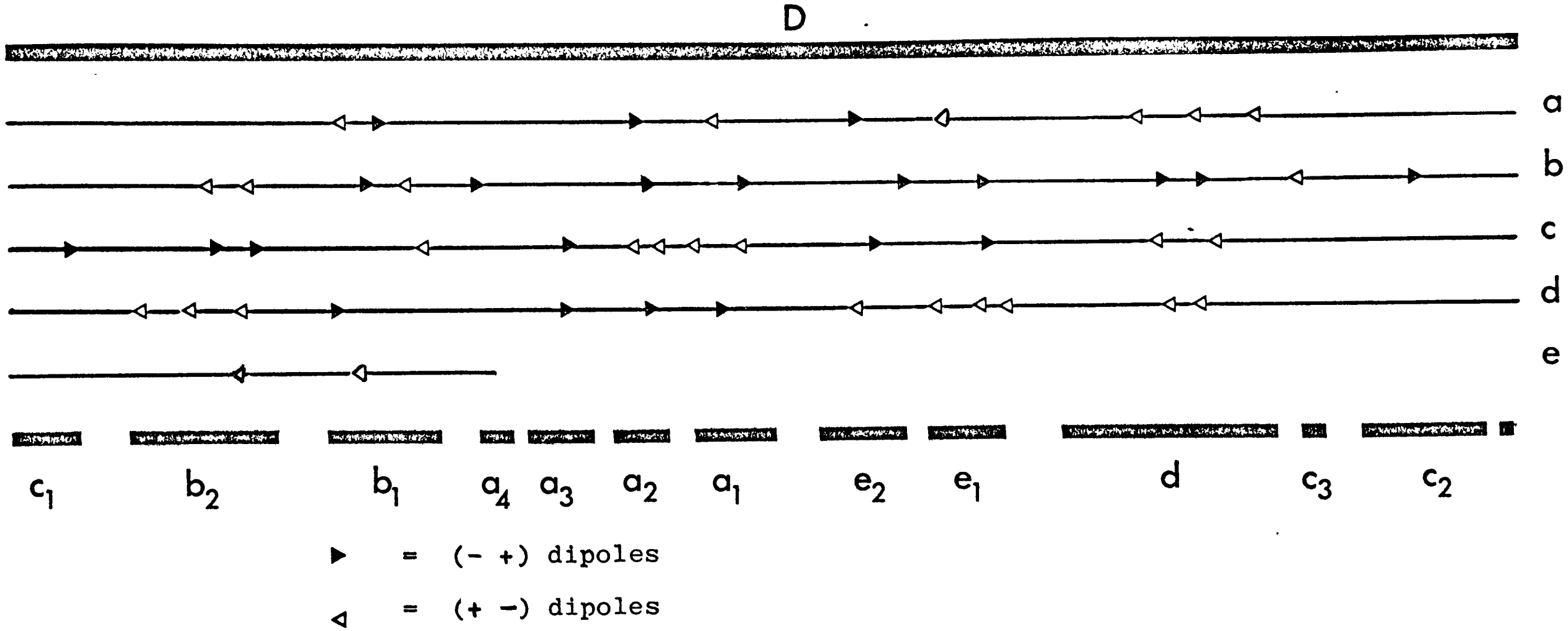


Fig. 2.12 The distribution of charged dipoles in the helical region of the α1 sequence. The sequence is represented by four 234 residue segments and one 75 residue segment. The positions of the positively stained bands are indicated.

dipoles are considered as a whole. The maximum efficiency principle of Section 3(i) is therefore not obeyed. The pattern is such that only the single D stagger is specified (Fig. 2.4). It seems that not all dipoles are involved in favourable interactions at 1D, but this is at the expense of removing any ambiguity with an alternative stagger of 3D. There is no major peak corresponding to a 3D stagger in Fig. 2.4.

iv) Other residues

There are therefore periodicities of D/5 and D/6 in the hydrophobic residues, and a D/6 periodicity in the charged residues which is out of phase with its hydrophobic counterpart. It was decided to look for any similar patterns in the remaining residues of the α1 sequence.

Autocorrelation functions for all the 21 different residues in the sequence were calculated by computing 'interaction curves' for each residue in turn (i.e. an auto-correlation function for aspartic acid, one for glutamine, one for histidine, and so on for all 21 different residues). In most cases too few residues were present in the sequence for the curves to show any evidence for periodicities. Of the four most abundant residues, i.e. glycine, proline, hydroxyproline and alanine, however, there was evidence of periodicities in glycine and proline. The glycine periodicity was of course the repeat every third residue, the proline periodicity, though, was a repeat of D/5, see Fig. 2.13. Fourier analysis of the proline positions using the procedure described in the previous sections also showed a strong contribution at a wavelength of D/5, see Table 2.6. The phase of the D/5 wave

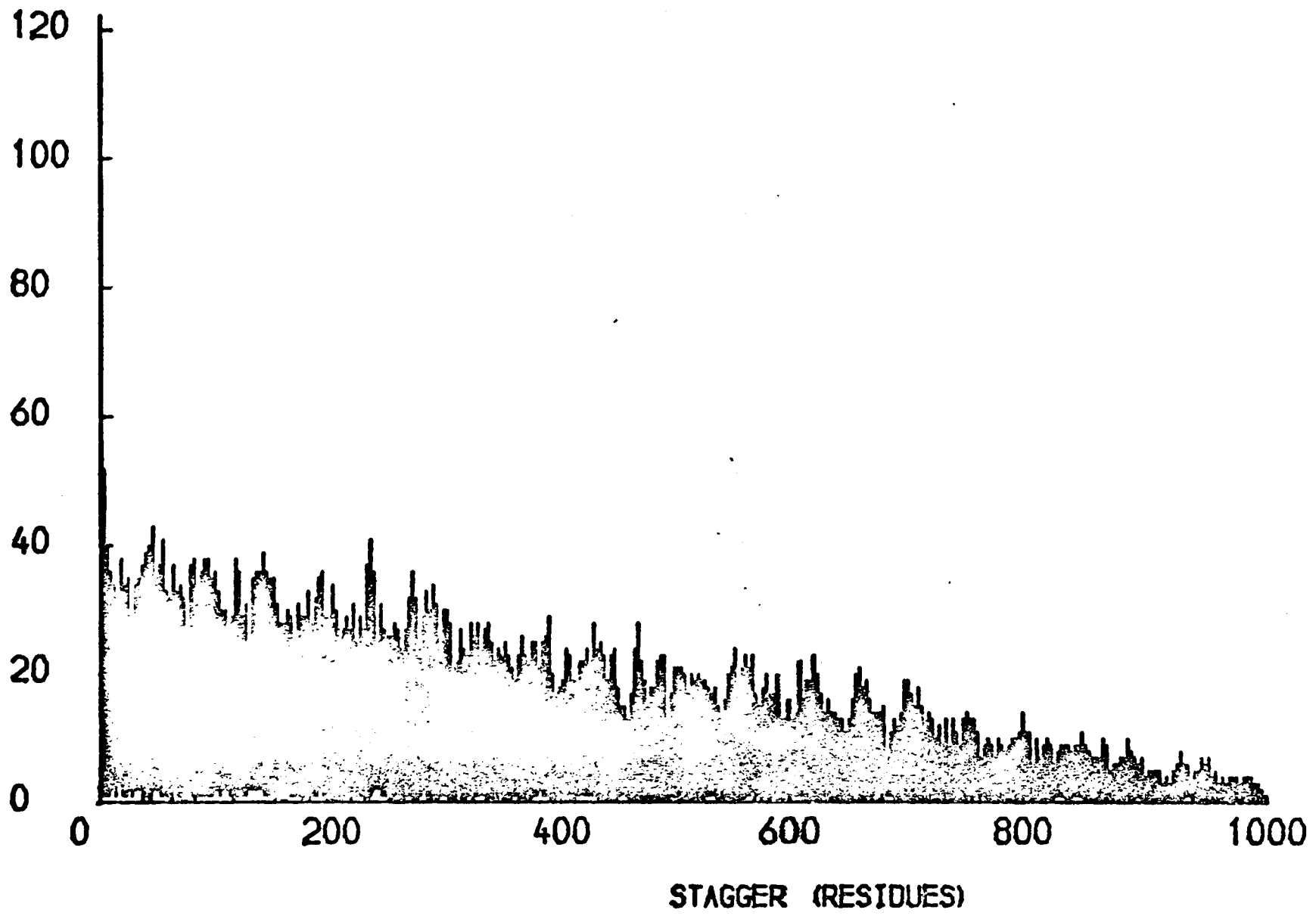


Fig. 2.13. Autocorrelation function for the positions of the proline residues in the helical region of the α1 sequence, showing the D/5 modulation.

TABLE 2.6 Fourier analysis of the positions of the proline residues in the helical region of the α1 sequence. Phase origin at residue 59.

Frequency (x 1/D)	Amplitude	Phase (°)
0	120.0	0
1	15.1	290
2	9.5	309
3	8.0	262
4	5.8	35
5	25.7	6
6	14.6	228
7	15.7	118
8	16.5	152
9	8.1	133
10	15.1	181
11	17.4	148
12	17.6	89
13	9.3	167
14	19.2	337
15	12.4	111
16	7.2	58
17	8.2	1
18	11.8	260
19	3.2	181
20	10.9	16
21	16.4	2
22	7.9	61
23	13.6	203
24	4.9	241
25	8.5	5

was 180° out with respect to the hydrophobic wave.

Fourier analysis of the alanine positions showed no strong periodicities, see Table 2.7. When the remaining polar residues (i.e. serine, threonine, asparagine, glutamine, tyrosine and histidine) were analysed as a whole again no strong periodicities were evident (Table 2.8). The only strong periodicity in the hydroxyproline positions was D/21, with a phase 125° out with respect to the charged D/21 periodicity (Table 2.9).

v) Discussion

A summary of the amplitudes and phases of the strong Fourier components in the α1 sequence is given in Table 2.10. As mentioned earlier, the possible significance of these distributions may be found in interactions involved in three-dimensional molecular packing within the collagen fibril. Segrest and Cunningham (1971, 1973) have discussed the possible existence of 'interacting edges' along the molecule, i.e. edges which arise when interacting residues in the α-chains are arranged in the triple helical structure. It is possible that the different periodicities in Table 2.10 may be such that each one is associated with a particular interacting edge. Molecules could then associate via these edges to produce possible hierarchies of structure, e.g. one set of edges for the formation of a microfibril and another set of edges for microfibril packing. The Fourier analysis may therefore separate patterns of interacting residues in the sequence in the same way that the triple helical structure separates residues into different interacting edges. By the nature of Fourier analysis the component periodicities must be continuous along

TABLE 2.7 Fourier analysis of the positions of the alanine residues in the helical region of the α1 sequence. Phase origin at residue 59.

Frequency (x 1/D)	Amplitude	Phase (°)
0	121.0	0
1	14.3	206
2	12.5	16
3	6.1	7
4	6.5	304
5	9.7	32
6	11.8	357
7	5.1	40
8	13.7	319
9	3.9	187
10	10.5	353
11	5.7	308
12	4.6	200
13	9.8	267
14	5.1	245
15	8.3	357
16	7.9	27
17	5.9	104
18	9.3	338
19	2.3	320
20	2.6	275
21	5.8	68
22	4.3	270
23	4.7	336
24	5.6	19
25	2.3	237

TABLE 2.8 Fourier analysis of the positions of the polar* residues in the helical region of the α1 sequence. Phase origin at residue 59.

Frequency (x 1/D)	Amplitude	Phase (°)
0	210.0	0
1	24.5	330
2	20.2	223
3	16.4	79
4	11.1	74
5	15.5	106
6	3.9	71
7	8.7	177
8	15.4	166
9	5.0	9
10	11.7	178
11	5.2	229
12	8.0	147
13	15.2	112
14	11.7	93
15	7.2	203
16	2.4	317
17	3.3	322
18	7.2	176
19	8.7	292
20	4.2	236
21	6.9	325
22	12.2	13
23	4.2	297
24	21.2	147
25	15.3	200

*serine, threanine, glutamine, asparagine, histidine, hydroxyproline.

TABLE 2.9 Fourier analysis of the positions of the hydroxyproline residues in the helical region of the α1 sequence. Phase origin at residue 59.

Frequency (x 1/D)	Amplitude	Phase (°)
0	113.0	0
1	12.9	352
2	12.9	281
3	7.8	74
4	9.9	57
5	6.7	108
6	11.8	217
7	5.8	116
8	8.9	178
9	6.1	100
10	10.8	145
11	8.6	139
12	2.4	219
13	13.2	115
14	12.2	95
15	1.3	322
16	2.5	88
17	9.7	301
18	5.1	218
19	2.5	38
20	7.2	63
21	17.8	317
22	10.3	16
23	5.9	267
24	12.5	171
25	5.2	157

TABLE 2.10 Principal Fourier components in the residue distributions within the α1 sequence.

Type	Amplitude	Phase	Wavelength
Hydrophobic	20.1	189	D/5
Proline	25.7	6	D/5
Hydrophobic	26.7	206	D/6
Charged	30.3	38	D/6
Hydrophobic	19.9	38	D/11
Charged	26.4	172	D/21
Hydroxyproline	17.8	317	D/21

the entire length of the molecule. This is unlike the 2D/11 repeat which involves a discontinuity every D period. It seems easier to think of interacting edge formation in terms of continuous periodicities along the sequence. The Fourier approach may therefore be useful in a study of possible three-dimensional interactions.

A feature that is common to all the periodic residue distributions is their approximate nature. The $(2/11_4,(1/11),(2/11)$ hydrophobic pattern in Fig. 2.10, for example, is not exact; neither is the D/6 repeat in the charge dipoles (Fig. 2.12). If a residue is displaced slightly from its 'ideal' position in the sequence, however, this would have a major effect on its azimuthal position in the molecule (see the radial co-ordinates of the Ramachandran-Kartha and Rich-Crick models given by Fraser and McRae (1973)). An interacting edge would therefore be difficult to observe in a molecular model based on the sequence. The unknown conformation of the side-chains would add to this uncertainty.

A further problem associated with a study of possible interacting edges is the α2 sequence. Whether edges are composed of contributions from all three chains or not, the α2 sequence is likely to introduce quite different interacting properties to a model based on three α1 chains.

When the α2 chain is fully sequenced and analysed in the same way as the α1, a study of the relation between the sequences and molecular packing should clarify the role of the various periodicities.

5. Summary

The α1 sequence of rat/calf skin collagen has been analysed for the origins of the D stagger. There were local

maxima in both the hydrophobic and charged interactions at all integral multiples of D (232 to 235 residues). When charge interactions were analysed in terms of dipoles there was marked specificity for a 1D stagger.

The one-dimensional analysis indicated that the hydrophobic residues are used efficiently at all nD staggers. This was correlated with an approximately repeating hydrophobic pattern in each D segment of the sequence. Dipoles are not used with maximum efficiency, but, although the positions of dipoles in each D segment are approximately the same, the orientations are such as to make each D segment unique. The approximate nature of the repeating D segments and the uniqueness imposed by the dipoles makes an explanation of the evolutionary origin of the sequence via gene duplication unlikely.

The theoretically determined value for the D stagger was correlated with the appearance of the electron microscopic positively stained banding pattern.

Strong Fourier components, of D/5, D/6, D/11 and D/21, in the distributions of various residue types within the α1 sequence were noticed. It was suggested that these could have a functional role in molecular packing in collagen fibrils. The relation of the sequence to the three-dimensional structure could be an exciting subject for future study.

CHAPTER 3

COLLAGEN POLYMORPHISM: ITS ORIGINS IN THE AMINO ACID SEQUENCE

1. Introduction

As mentioned in previous chapters, collagen displays an extensive polymorphism. Besides the 670Å periodically banded native-type fibrils found in most connective tissues, there have been reports of at least two other forms occurring *in vivo* (see Pérez-Tamayo (1972) and Banfield *et al.* (1973)). It has been suggested that these are natural forms of the *in vitro* molecular aggregates known as segment long spacing (SLS) and fibrous long spacing (FLS) collagen (Luse *et al.*, 1963). A full list of the known artificially induced and naturally occurring forms of collagen is given in Table 3.1. All of these structures show a characteristic banding pattern when stained and observed using the electron microscope and most of them, except the two forms of SLS, are fibrous.

The projection on to the fibre axis of the native 670Å (D) periodic structure is well understood in terms of the molecular arrangement and the amino acid sequence (see Chapter 2), though there is some doubt about the other structures. In particular the largest group, FLS, has been the subject of much research in which several workers have proposed different models to account for their observations. The two other groups, SLS and the D-periodic symmetric (DPS) structures, have also not yet been fully accounted for.

In this chapter an attempt is made to achieve a two-fold understanding of collagen polymorphism in terms of the amino acid sequence. First the molecular arrangements in projection

TABLE 3.1 *The observed polymorphic forms of collagen*

Name	Description	References
Native	Fibrils with periodic, asymmetric banding pattern after positive staining, repeating every 670 Å (*D*)	See Hodge (1967)
DPS	Same as native but with symmetric banding pattern	Kühn *et al.* (1964) Nemetschek (1965) Bard & Chapman (1968) Leibovich & Weiss (1970) Doyle *et al.* (1974a)
***D*/3 periodic**	Fibrils with periodic banding pattern repeating every 220 Å (*D*/3)	Fitton-Jackson (1953) Gross (1956) Porter & Pappas (1959) Kajikawa (1961)
Oblique striated	Tactoids composed of *D*-periodic polar sub-fibrils staggered with respect to nearest neighbours by about 90 Å	Kühn *et al.* (1964) Bruns *et al.* (1973) Doyle *et al.* (1974b)
***D*/6 periodic**	Tactoids with 110 Å (*D*/6) periodicity	Kühn *et al.* (1965)
SLS	Segments equal to the molecular length with asymmetric banding pattern after positive staining	Schmitt *et al.* (1953)
SLS symmetric	Same as SLS though with a symmetric banding pattern	Hormann & von Wilm (1964) Kühn *et al.* (1965) Chapman & Armitage (1972)
SLS fibrils	Fibrils with periodic, asymmetric banding pattern after positive staining, repeating every 2680 Å (4*D*)	Kühn (1969)
FLS	A class of fibrils all produced under similar conditions, each having a periodic, symmetric banding pattern after positive staining, repeating at a distance greater than 670 Å (*D*). See Table 2	Highberger *et al.* (1950) Fitton-Jackson & Randall (1953) Kühn & Zimmer (1961) Chapman & Armitage (1972)

on to the fibre axis are determined mainly by simulating and comparing banding patterns based on the sequence (as in Chapter 2, Section 2). Secondly the sequence is analysed to find the origins of the various structures in the distributions of interacting residues (as in Chapter 2, Section 1).

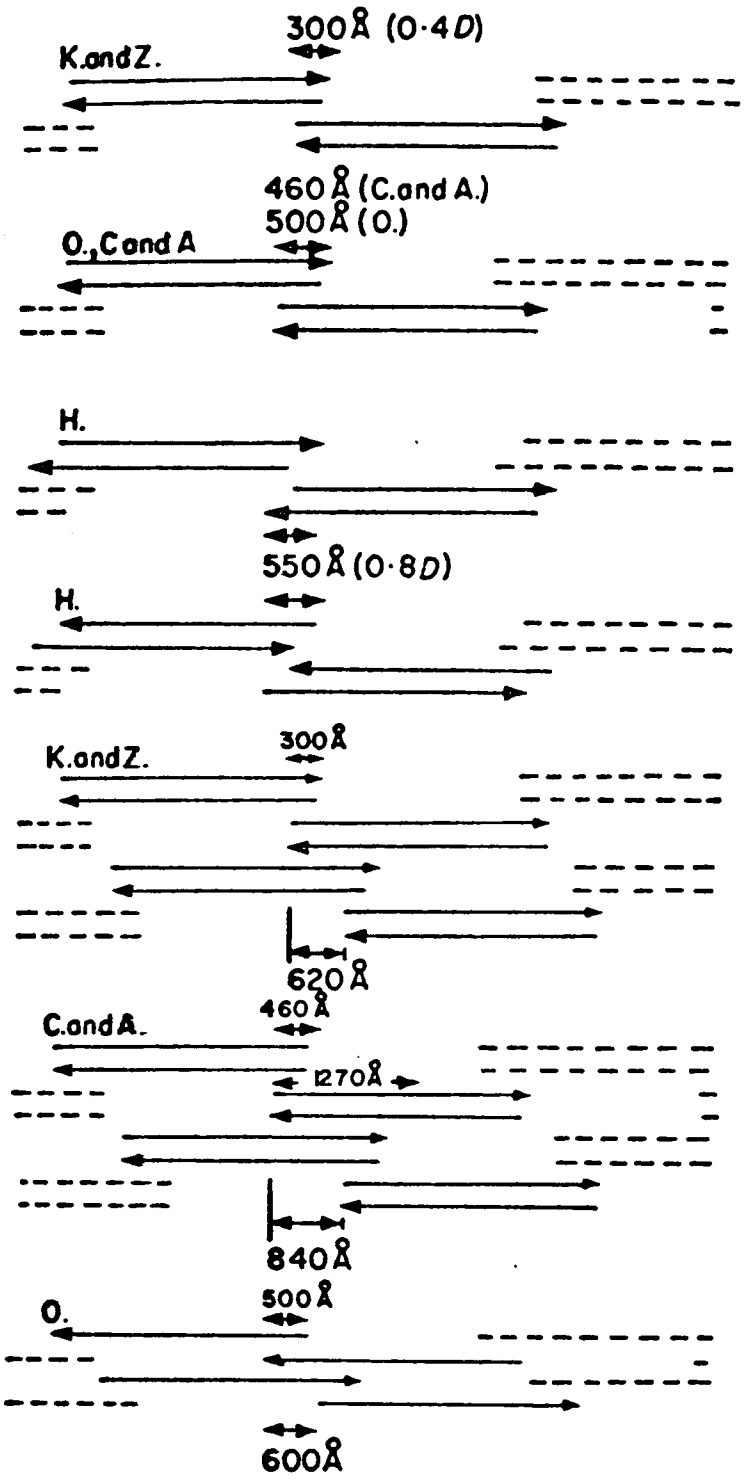
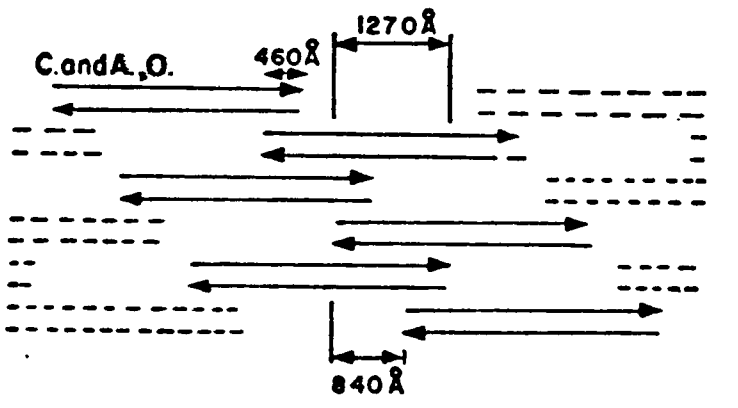
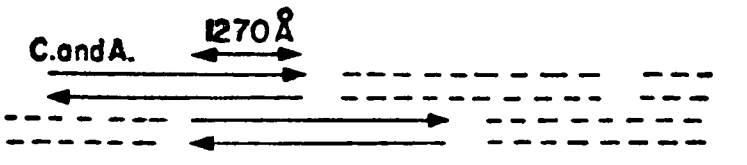
2. The Axially Projected Molecular Arrangements

i) Fibrous Long Spacing Collagen

Since the discovery by Highberger et al. (1950) of a long spacing (>670Å) form of collagen with a symmetric banding pattern many workers (Fitton-Jackson and Randall, 1953; Kühn and Zimmer, 1961; Olsen, 1963; Hodge, 1967; and Chapman and Armitage, 1972) have tried to understand the structure in terms of possible longitudinal displacements or staggers between molecules. Positive staining and heavy metal shadowing were used for the early observations. The former shows a banding pattern which indicates the location of charged amino acid side-chains in a structure, whereas the latter gives an impression of the surface relief. With these techniques it is difficult to know precisely where the regions of molecular overlap are in the structure (see Doyle et al., 1974a), though this was made possible with the introduction of negative staining (Brenner and Horne, 1959).

There are at least four different forms of fibrous long spacing collagen; FLS I, FLS II, FLS III and FLS IV. A description of their main features is given in Table 3.2 together with the suggested molecular arrangements. A variety of intermolecular staggers have been proposed, some of which are found in the native D-periodic structure (see the models of Hodge and of Kuhn and Zimmer).

TABLE 3.2 *Fibrous long spacing collagen*

Name	Description	Model
FLS I	Average periodicity ~2600 Å. Symmetric banding pattern with positive staining. One overlap band per period with negative staining.	
FLS II	Average periodicity ~2600 Å. Symmetric banding pattern with positive staining. Two overlap bands per period with negative staining.	
FLS III	Average periodicity ~900 Å. Symmetric banding pattern with positive staining. One overlap band per period with negative staining.	
FLS IV	Average periodicity ~1700 Å. Symmetric banding pattern with positive staining. One overlap band per period with negative staining.	

K. & Z., Kühn & Zimmer (1961*a*,*b*,*c*).
O, Olsen (1963).
H, Hodge (1967).
C. & A., Chapman & Armitage (1972).
In the diagrams the molecular length is assumed to be approximately 3000 Å.

To what extent do the various molecular models proposed reflect truly different structures? Is it, perhaps, that the techniques used do not permit an unambiguous solution to the problem? From careful analysis of the literature it seems that both possibilities are required. The most striking example of the first is the clear difference between the negatively stained pictures of FLS II published by Olsen (1963) and by Chapman and Armitage (1972). The unstained bands in both cases are the same width, though the ratio of the distance between the pair of unstained bands to the periodicity is approximately 1/4 for the former and 1/3 for the latter. As for the second possibility, it would be very difficult to distinguish, for example, between the Hodge type and Olsen-Chapman and Armitage type model for FLS I using negative staining alone. The overlap regions are the same width in both cases, and the small difference between the absolute values of the periodicity between the two models would probably be within the limits of error in measurement.

It must be remembered that all the models proposed for FLS are only one dimensional. They serve merely to explain structures in projection on to the fibre axis. Thus only staggers between molecules are relevant. The necessarily two dimensional scheme in which the models are presented does <u>not</u> imply any information about the lateral arrangement of molecules within the actual structures. Many alternative schemes could be presented which are equally valid, since they would share the same set of axial staggers. The models shown in Table 3.2, however, are all unique in this one dimensional sense, each one having a characteristic <u>set</u> of intermolecular staggers. Some <u>particular</u> staggers are common to at least two models.

Fortunately, there is now a method for determining which of these many proposals are in fact correct. This is possible using negative staining and the known amino acid sequence. Negative staining is used to give a small number of possible molecular arrangements, which are then tested by simulating the positive stained banding pattern from the sequence and comparing it to the observed distribution of charged groups.

a) FLS I

For FLS I, in the gap region, all the proposed molecular arrangements are identical apart from the two Hodge alternatives. This would seem, therefore, to be a good place to begin simulating the banding patterns.

The most detailed electron micrograph of positively stained FLS I was published by Kühn (1967). In this case the specimen was doubly stained with phosphotungstic acid (PTA) and uranyl acetate (UA), and here the banding pattern shows the distribution of both positively and negatively charged amino acid side chains. Photographic superposition of computer drawn banding patterns based on these residues (see Chapter 2, Section 2) has been used to simulate the three proposals for the gap region. By far the best agreement was found by superimposing two molecules in exact anti-parallel register (see Plate 3.1(a) and (b)). One of the most distinctive features of the pattern, the central dark band flanked by two white bands, was totally absent in the other two alternatives. The final banding pattern is very sensitive to the precise relation between the two molecules and so the result is accurate to about $\pm$ 10 residues. This result is in agreement with the conclusion of Kühn and Zimmer (1961) based on their addition of densitometer curves.

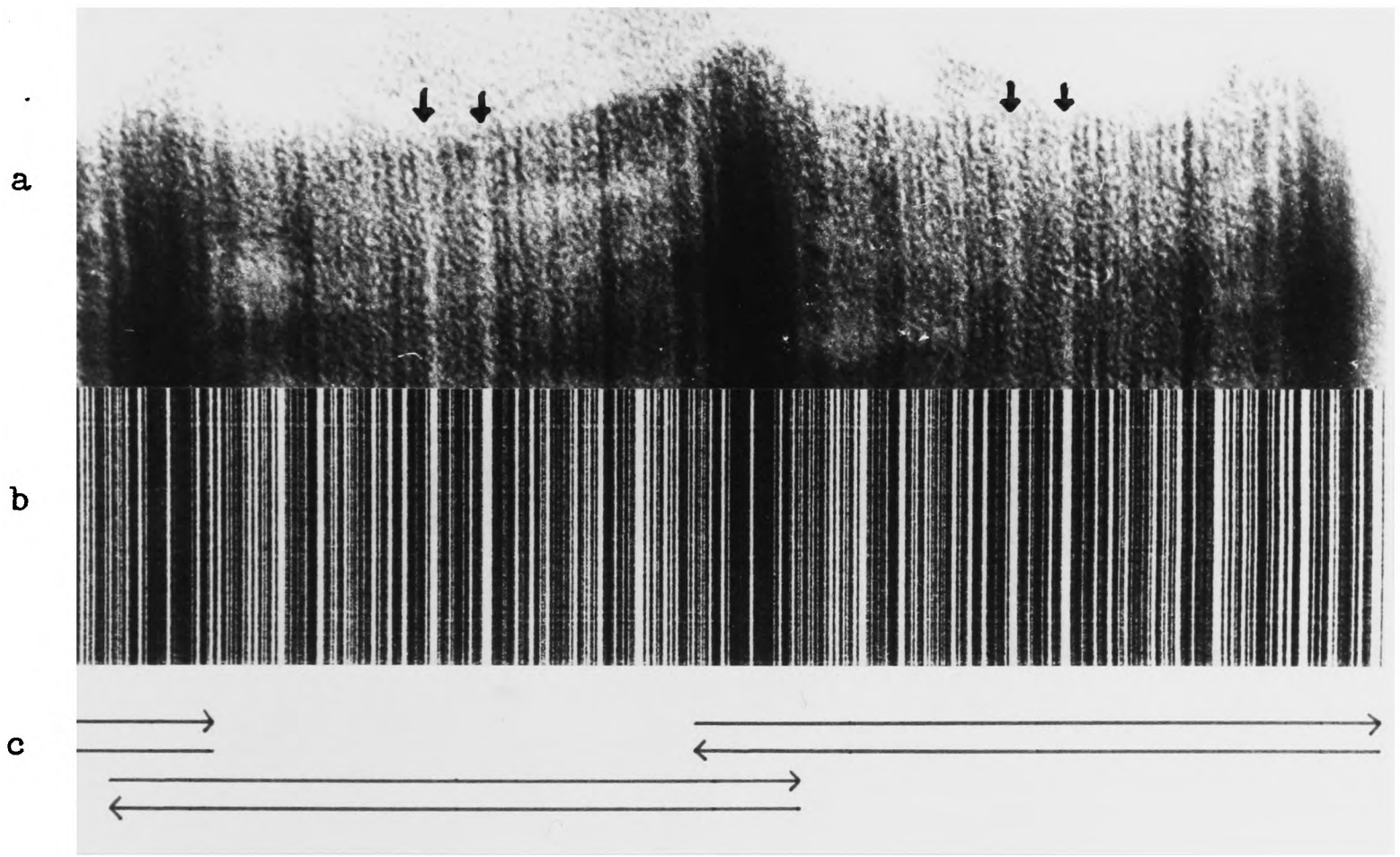


PLATE 3.1

a) Electron micrograph of FLS I, positively stained with phosphotungstic acid and uranyl acetate (courtesy of Prof. K. Kuhn). The distinctive pair of white bands, flanking the central dark band, is indicated (see text).

b) Simulation of (a), based on the α1 sequence and the molecular arrangement in (c).

c) One-dimensional model for FLS I.

Confirmation of the exactly anti-parallel arrangement was also obtained by simulating the negatively stained symmetric segment of Chapman and Armitage (1972). This was achieved by superposing a negatively stained polar SLS on itself anti-parallel (see Plates 3.2(a) and (b)). This expresses in a visual form the work of Olsen (1963) on the lining up of the so-called 'nodular thickenings' in SLS with FLS.

Though somewhat of a diversion from the main theme it is of interest to discuss the nature of these 'nodular thickenings' which show up as transverse white bands in the micrographs. The current interpretation of negative staining is that the stain fills up the depressions in the surface leaving the protruding features unstained. In the case of SLS these features would be the large amino acid side chains. At neutral pH, or thereabouts, however, it seems likely that there would be some electrostatic binding between charged side chains and the stain. When the stain is PTA, therefore, there is probably binding to the large, positively charged side chains of arginine and lysine. The only other large side chains to which the PTA would not be attracted would be those of phenylalanine, leucine, isoleucine, valine and methionine. Hence one would expect the positions of the nodular thickenings to correspond to the distribution of these residues. This banding pattern has been constructed from the sequence and the favourable comparison with SLS is shown in Plate 3.2(c) and (d).

Before returning to FLS I it is worth noting that the symmetric segments produced by Hormann and Von Wilm (1964) and Kühn et al. (1965) after hydroxylamine and trypsin treatment respectively, are probably other examples of collagen molecules being in exactly anti-parallel register. The positively

a

b

c

d

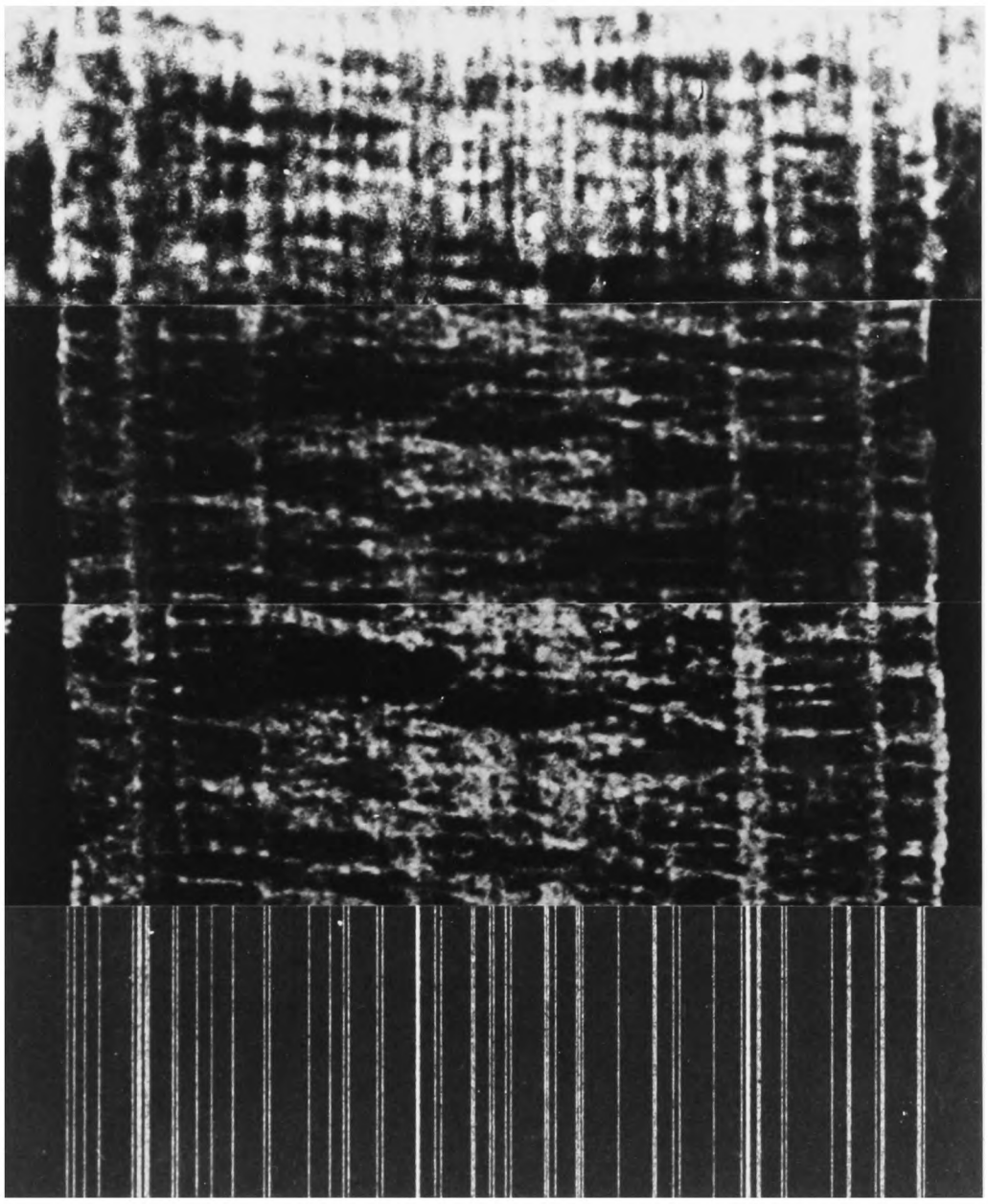


PLATE 3.2

a) Electron micrograph of a symmetrical segment of calf skin collagen, negatively stained with 2% sodium phosphotungstate (pH 7) (courtesy of Dr. J.A. Chapman).
b) Symmetrical banding pattern, produced by photographic superposition of two polar long spacing segments (see (c)) in exact anti-parallel register.
c) Electron micrograph of a polar long spacing segment of rat tail tendon, negatively stained with 1% phosphotungstic acid (courtesy of Prof. B.R. Olsen).
d) Distribution of the large hydrophobic residues in the α1 sequence (i.e. Phe, Met, Leu, Ile and Val).

stained banding patterns are strikingly similar to the FLS simulations.

Given the existence of an exactly anti-parallel relationship it is possible to determine the length of molecular overlap in FLS I. The procedure is first to adjust the relative magnifications between the observed and suggested gap region banding patterns until optimum agreement is found along the entire length. If this is done for two neighbouring regions the overlap may be calculated directly. Using this method the helical ends of the molecules have been found to be overlapping by 155 $\pm$ 15 residues (see Plate 3.1(a) and (b)), approximately 440Å. If one allows for the presence of telopeptides and their effect on the negatively stained pattern (see Doyle et al., 1974a) this value correlates well with the conclusion of Olsen and of Chapman and Armitage. It contradicts however the conclusion of Kühn and Zimmer based on their densitometer measurements.

b) FLS IV

FLS IV has only been described once before (Chapman and Armitage, 1972), and hence there is only one suggested molecular arrangement. Both positive staining and the sequence have been used to test this model.

Sometimes in the preparation of FLS a number of different forms are precipitated together. Occasionally compound structures are found which show features typical of at least two of the main forms (see Chapman and Armitage, 1972). The regions of transition from one type to another are very useful for understanding the molecular arrangement. One such region is shown in Plate 3.3. It shows a transition from FLS I to FLS IV. In the FLS I overlap region there are four clear dark, heavily stained lines.

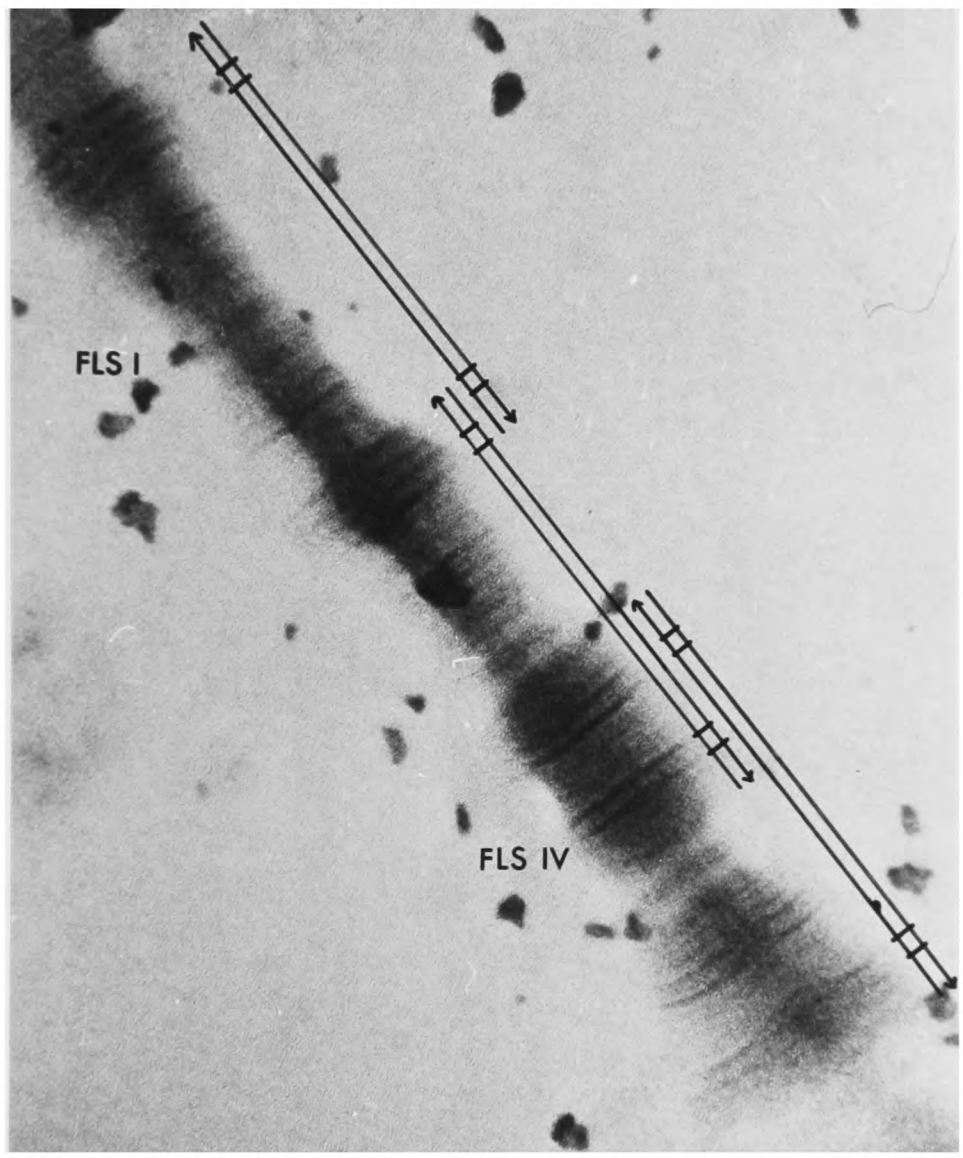


PLATE 3.3

Electron micrograph of a transition from FLS I to FLS IV, positively stained with 1% phosphotungstic acid (pH 7) (courtesy of Dr. J. Woodhead-Galloway). The specimen was prepared (by Dr. B. Brodsky-Doyle) according to Olsen's (1963) method, using lathyritic rat skin collagen. The accompanying diagram shows the axial arrangement of molecules, with markers indicating the origin of the heavily stained bands.

The diagram shows how these lines arise from the banding pattern of two molecules in exact anti-parallel register. As one follows through the transitional region to the FLS IV the lines are seen to separate into two widely spaced pairs. The diagram shows how this may be accounted for, the molecular overlap has become approximately three times larger than for FLS I. This compares well with the ratio of the overlap in FLS IV to that in FLS I measured by Chapman and Armitage.

As yet, specimens of positively stained FLS IV have not shown as much fine banding as, for example, in FLS I (Plate 3.1). Nevertheless measurements from Plate 3.3 together with those of Chapman and Armitage (1972) and reasonable assumptions about the molecular length do permit a calculation of the characteristic FLS IV helical overlap to be 425 ± 25 residues. A banding pattern using this value was constructed and the comparison with the relevant micrograph is shown in Plate 3.4.

c) FLS II

Kühn and Zimmer (1961b) showed, by addition of densitometer curves, that the positive stain banding pattern of FLS II can be constructed by superposition of two FLS I patterns staggered by about 600Å, a distance close to the native periodicity D. It was shown above that the correct model for FLS I is that of Olsen and of Chapman and Armitage. Thus the FLS II observed by Kühn and Zimmer and by Olsen must be accounted for by a combination of two Olsen/Chapman and Armitage FLS I models staggered by about 600Å.

The FLS II observed by Chapman and Armitage, however, was accounted for, by them, in terms of two superposed FLS I's staggered by 850Å, with intermolecular contacts of the FLS I and FLS IV type. Since many forms of FLS can be precipitated

a

b

c

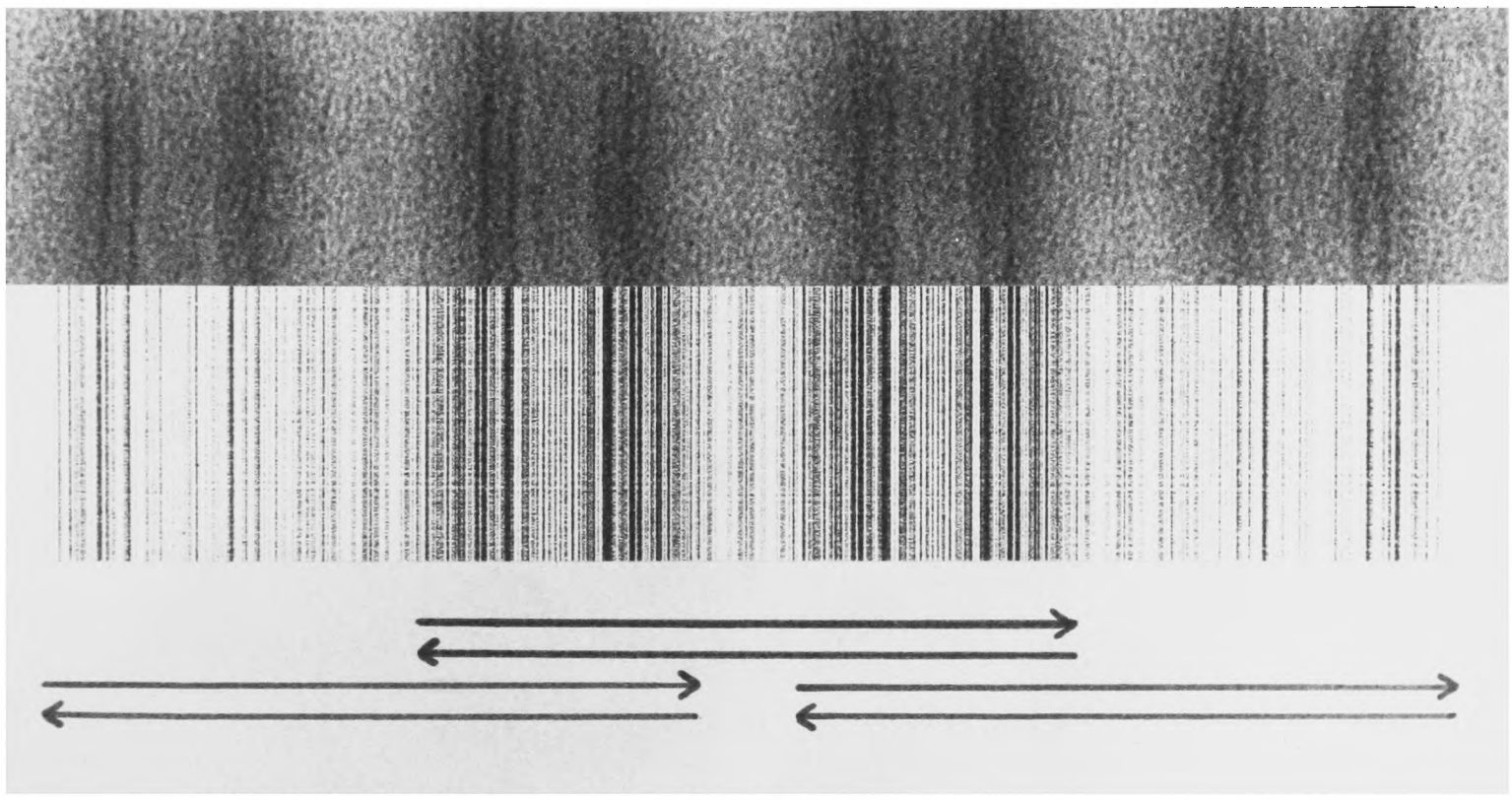


PLATE 3.4

a) Electron micrograph of FLS IV, positively stained with 1% phosphotungstic acid (courtesy of Dr. J. Woodhead-Galloway). The specimen was prepared (by Dr. B. Brodsky-Doyle) according to Olsen's (1963) method, using lathyritic rat skin collagen.
b) Simulation of (a), based on the distribution of Arg and Lys in the α1 sequence and the molecular arrangement in (c).
c) One-dimensional model for FLS IV.

together the existence of a structure having such a combination of contacts seems quite possible.

It is concluded, then, that there are two distinct forms of FLS II.

d) FLS III

Chapman and Armitage have also interpreted FLS III as a structure involving a combination of FLS I and FLS IV type contacts. There is certainly a very close relationship between FLS IV and FLS III, as shown in Plate 3.5. The characteristic molecular overlaps in FLS I and FLS IV, according to Chapman and Armitage (1972), are approximately 1/7 and 3/7 of the molecular length respectively. Hence the periodicity of FLS IV is approximately 2/3 that of FLS I. When both contacts are combined in FLS III the resulting structure has a periodicity equal to half that of FLS IV, or, equivalently, one third that of FLS I.

In summary, then the entire FLS series can be explained by appropriate combinations of four basic intermolecular overlaps:

A) Complete overlap of the helical regions of two anti-parallel molecules.

B) The characteristic FLS I helical overlap of 155 ± 15 residues ($\sim 440\text{Å}$).

C) The characteristic FLS IV helical overlap of 425 ± 25 residues ($\sim 1220\text{Å}$)

D) The intermolecular overlap of 800 ± 10 residues (i.e. a stagger of ~210 residues = 600Å).

(N.B. The conversion from residues to Angstrom units is achieved by assuming the residue translation along the axis of the collagen

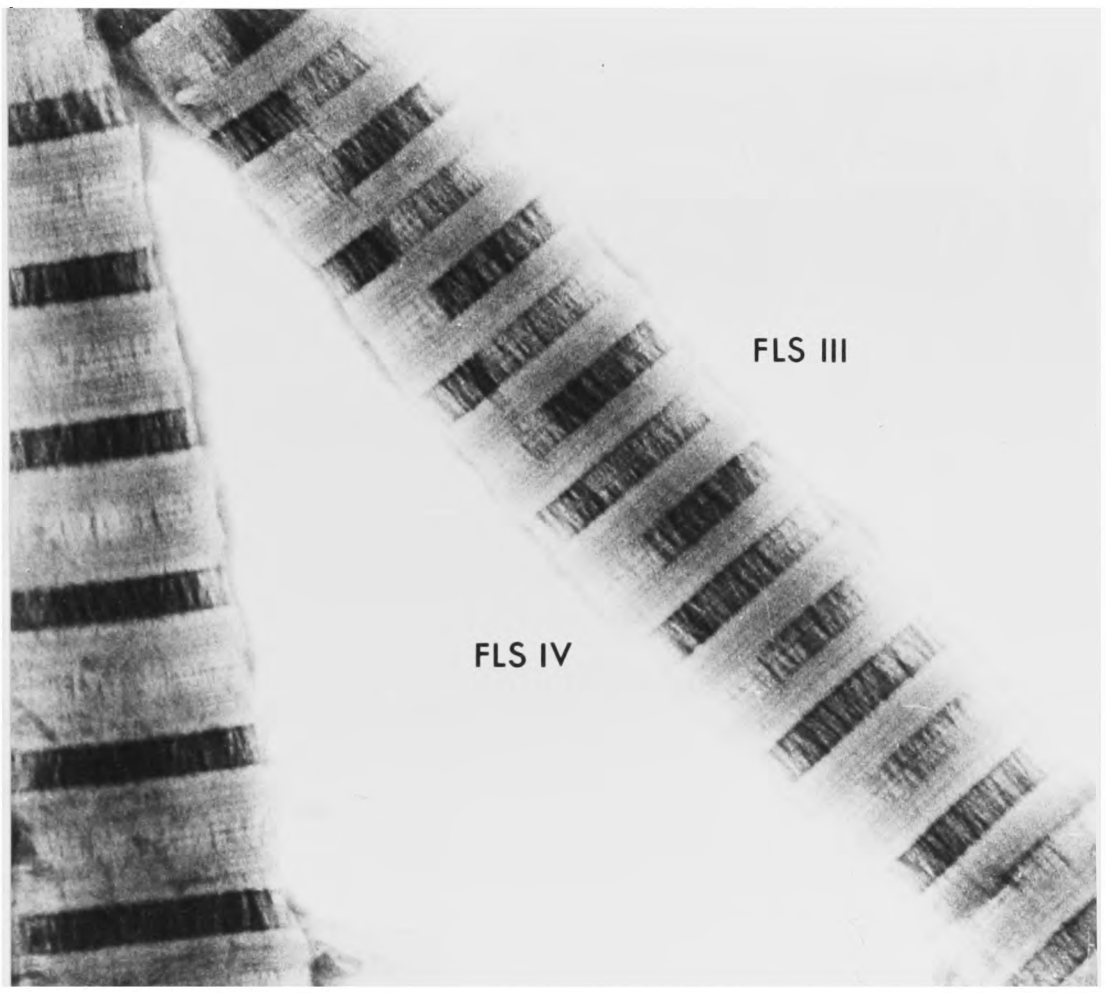


PLATE 3.5

FLS fibril showing a transition from FLS III to FLS IV, negatively stained with 2% phosphotungstic acid (pH 7) (courtesy of Dr. J. Woodhead-Galloway). The specimen was prepared (by Dr. B. Brodsky-Doyle) according to Olsen's (1963) method, using lathyritic rat skin collagen.

triple helix to be 2.86Å, see Hulmes et al. (1973)).

ii) D-periodic symmetric collagen (DPS)

This form has been described by several workers (Kühn et al., 1964; Nemetschek, 1965; Bard and Chapman, 1968; Leibowich and Weiss, 1970; Doyle et al., 1974a). It is always found in fibrillar or tactoidal aggregates, characterised by a repeating band pattern with the same D periodicity as in native collagen (670Å), though with a symmetric fine structure. It is precipitated at either low pH (~4.5) (Bard and Chapman; Doyle et al.) or after pepsin treatment (Kühn; Nemetschek; Leibowich and Weiss). The appearance in the electron microscope has been studied using both negative and positive staining. The negative stained patterns all show four bands per period, implying three densities of stain. The positive staining shows six heavily stained bands. All the patterns published are identical, except those of Doyle et al., which were obtained with cartilage collagen.

As the DPS positively stained banding pattern is centro-symmetric, and the collagen molecule is polarised, there must be an approximately equal number of molecules oriented in each direction along the fibril axis. The D periodicity is present, however, so it is likely that some native-type intermolecular staggers are present. The DPS negatively stained pattern could in fact be considered as the combination of two native-type patterns. This can be observed directly in those electron micrographs which show a two dimensional pattern rather than simple transverse banding, see Plate 3.6(a). The DPS tactoid is seen to be composed of native-type subfibrils of alternating polarity. This structure is less obvious in Plate 3.6(b) where the broad transverse white band corresponds to the region of

a

b

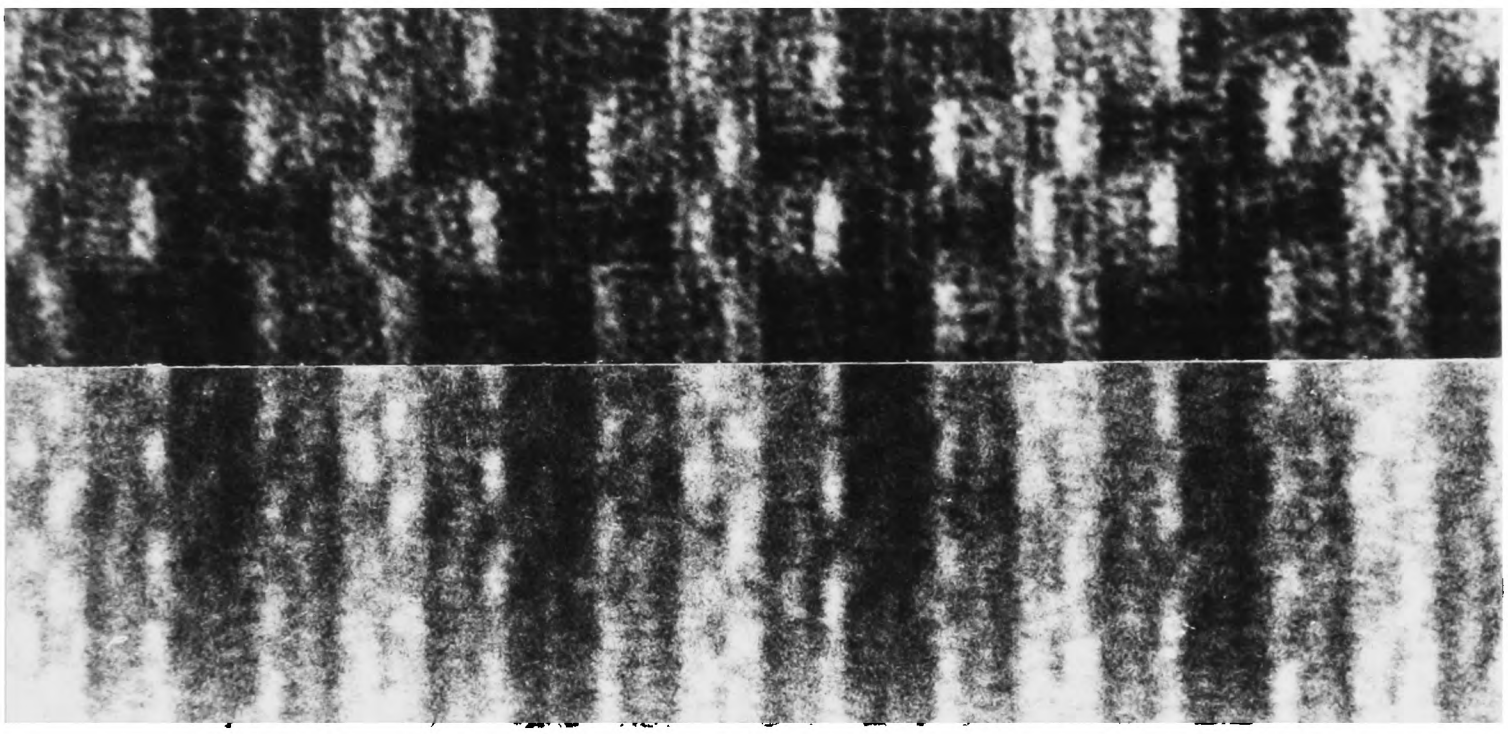


PLATE 3.6

Electron micrographs of tactoids precipitated from chick cartilage collagen (courtesy of Dr. J. Woodhead-Galloway). The specimens were prepared (by Dr. B. Brodsky-Doyle) according to the method of Bruns et al. (1973). The pattern in (a) is composed of two pairs of native-type subfibrils (width ~200Å), with alternating polarity.

proximity of either the N- or C-terminal ends of the native-type overlap regions. Plates 3.6(a) and (b) were obtained using chick cartilage collagen (Doyle et al., 1974a). For calf skin collagen the alternating subfibril structure has not been observed, though the interpretation of the simple negatively stained pattern is similar.

The problem of whether it is the N- or C-terminal ends of the native overlap regions which are closest together can best be solved by comparison of simulated and observed positively stained banding patterns. A necessary prerequisite for such a procedure is the equivalence of the native-type cartilage pattern to the simulated and observed patterns of Chapter 2. This is shown in Plate 3.7(a), (b) and (c). To investigate possible banding patterns for DPS two copies of Plate 3.7(c) were made on Xerox transparencies. The patterns were then overlayed anti-parallel on a light box and the superposition patterns for all possible relative shifts were observed. The cartilage DPS pattern is so distinctive that only one combination of native type patterns was possible, see Plate 3.7(e) and (f). This configuration corresponds to an anti-parallel N-terminal overlap of helical regions of 150 ± 5 residues. As the helical region contains 1011 residues and the D stagger is taken as 234 residues the helical overlap region is (1011- 4x234=) 75 residues, hence the cartilage DPS pattern corresponds to exact coincidence of helical C-termini. The schematic molecular arrangement is shown in Plate 3.7(g).

In skin DPS the configuration is slightly different from the cartilage case. Optimum agreement between simulated and observed patterns was found for an N-terminal helical overlap of 165 ± 5 residues, see Plate 3.8. It is interesting to note

a

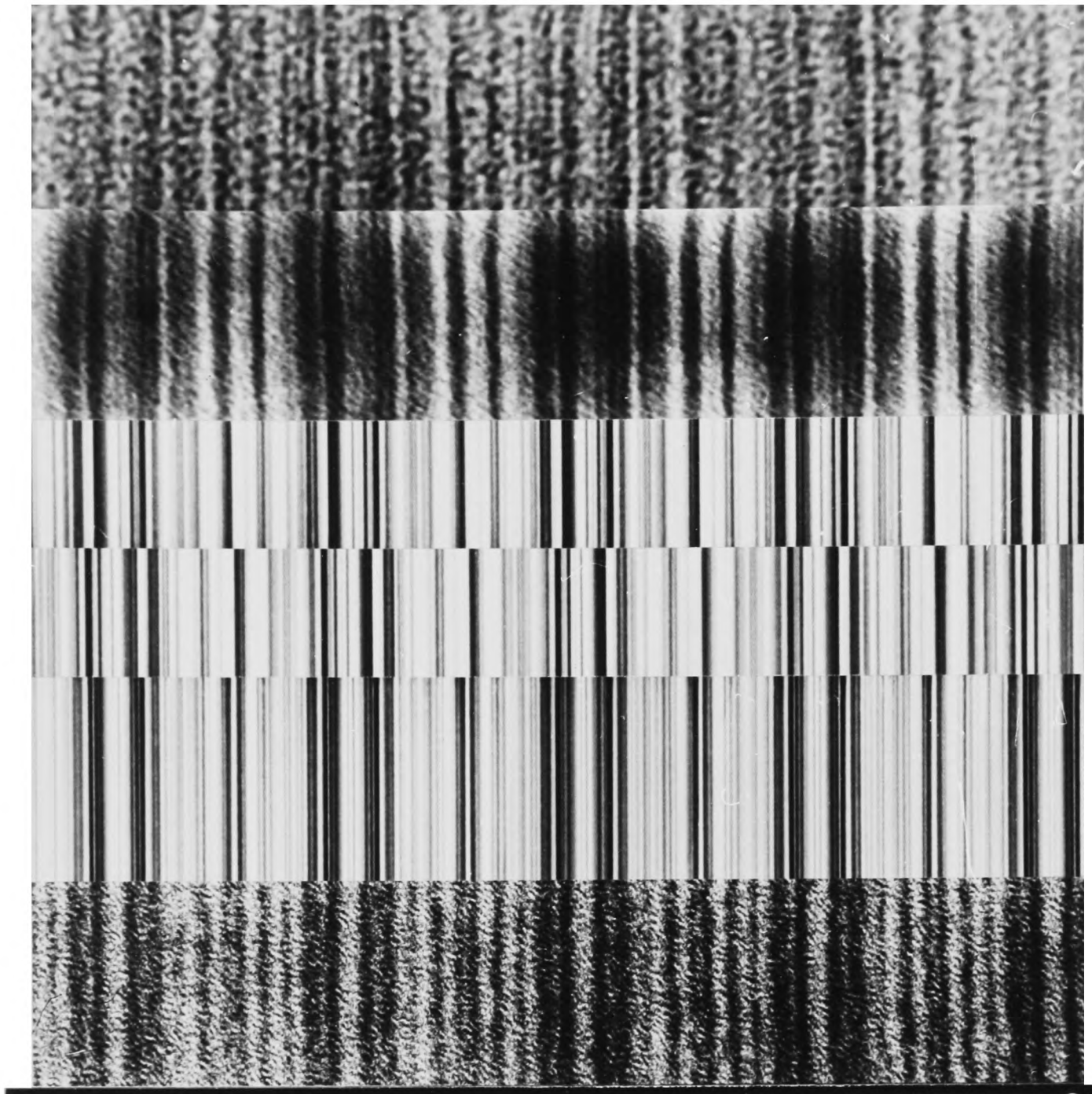
b

c

d

e

f



g

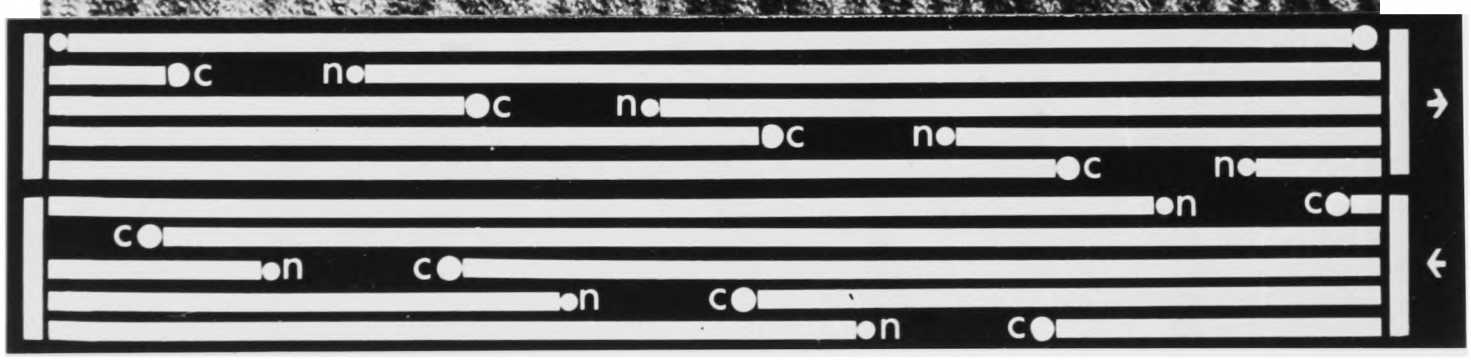


PLATE 3.7

a) Electron micrograph of heat precipitated chick cartilage collagen, positively stained.
b) As Plate 2.1(d).
c) As Plate 2.1(c).
d) As (c) but reversed in direction and staggered so that C-terminal helical ends are in contact (see text).
e) Photographic superposition of (c) and (d).
f) Electron micrograph of chick cartilage collagen, precipitated using the method of Bruns et al. (1973) and then positively stained.
g) Molecular arrangement corresponding to (e).

All micrographs courtesy of Dr. B. Brodsky-Doyle and Dr. J. Woodhead-Galloway.

a

b

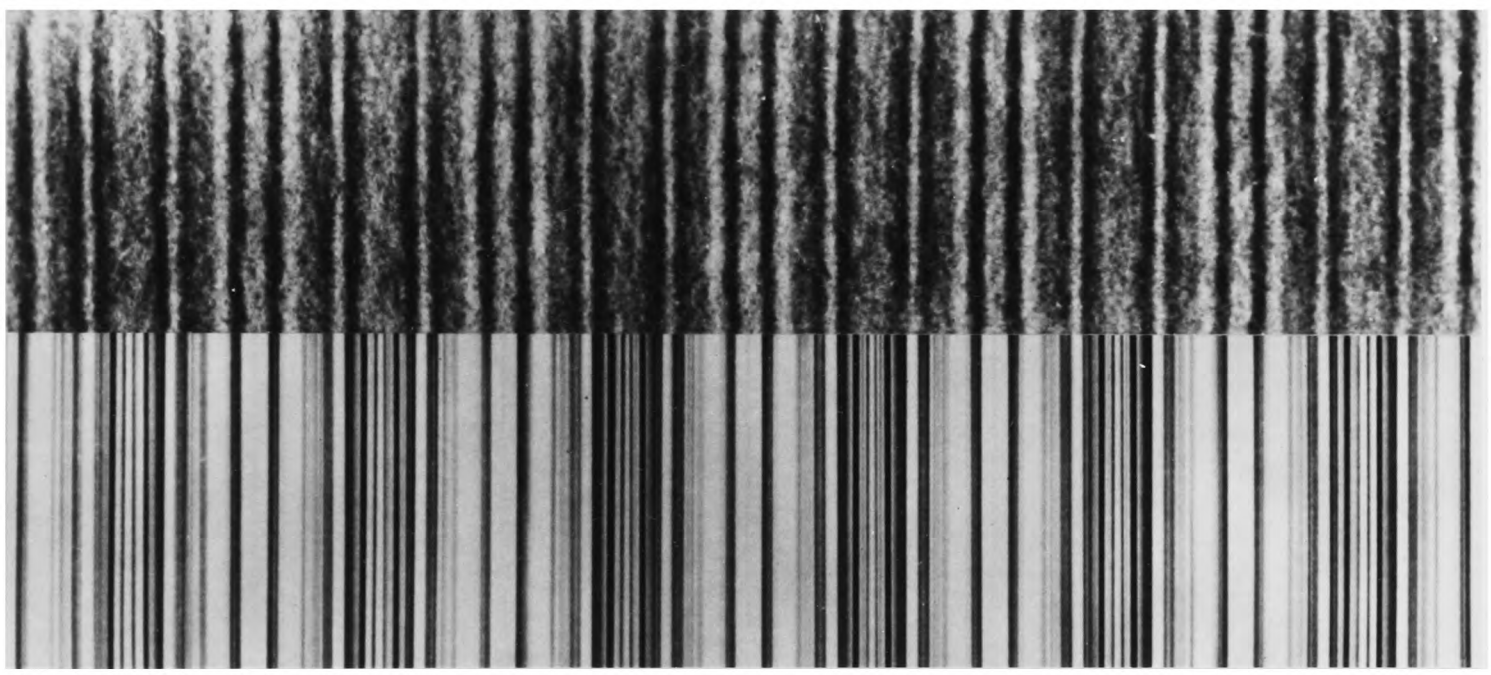


PLATE 3.8

a) Electron micrograph of a DPS fibril, formed from calf skin collagen and positively stained with phosphotungstic acid and uranyl acetate (courtesy of Prof. K. Kuhn).

b) Simulation of (a) using the α1 sequence (see text).

that though the difference in overlap between cartilage DPS and skin DPS is only 15 residues, the difference in charge distribution is quite marked (compare Plate 3.7(f) with Plate 3.8(a)).

iii) The obliquely striated and 110Å periodic forms

These forms of collagen have been reported by a number of workers, the former by Kühn et al. (1964), Bruns et al. (1973) and Doyle et al. (1974b), and the latter by Kühn et al. (1964,1965). They are each precipitated under similar conditions as for the DPS structures, so similar in fact, that compound structures have been observed showing transitions between two polymorphic forms (see Plate 3.9).

The obliquely striated structures have been interpreted by Bruns et al. (1973) to be tactoids composed of parallel D-periodic, polar subfibrils of equal width, each one staggered with respect to its nearest neighbours by approximately 90Å. The subfibril width in different tactoids from cartilage collagen varies over a considerable range, from 40 to 400Å (Doyle et al., 1974b). The stagger, however, is normally distributed about a mean of 91Å (in close agreement with Bruns et al. 1973), with a standard deviation of only 7Å. In spite of this variation these structures almost always show the characteristic oblique striations. There are exceptions, however. These occur when the stagger between subfibrils appears to be an integral sub-multiple of D (670Å), and the subfibril width is relatively small. In such cases the staining features which normally delineate the oblique striations are aligned perpendicular to the long axis of the tactoid, so giving the structure an apparent reduced periodicity equal to the stagger distance. The phenomenon is illustrated in Plate 3.10,

a

b

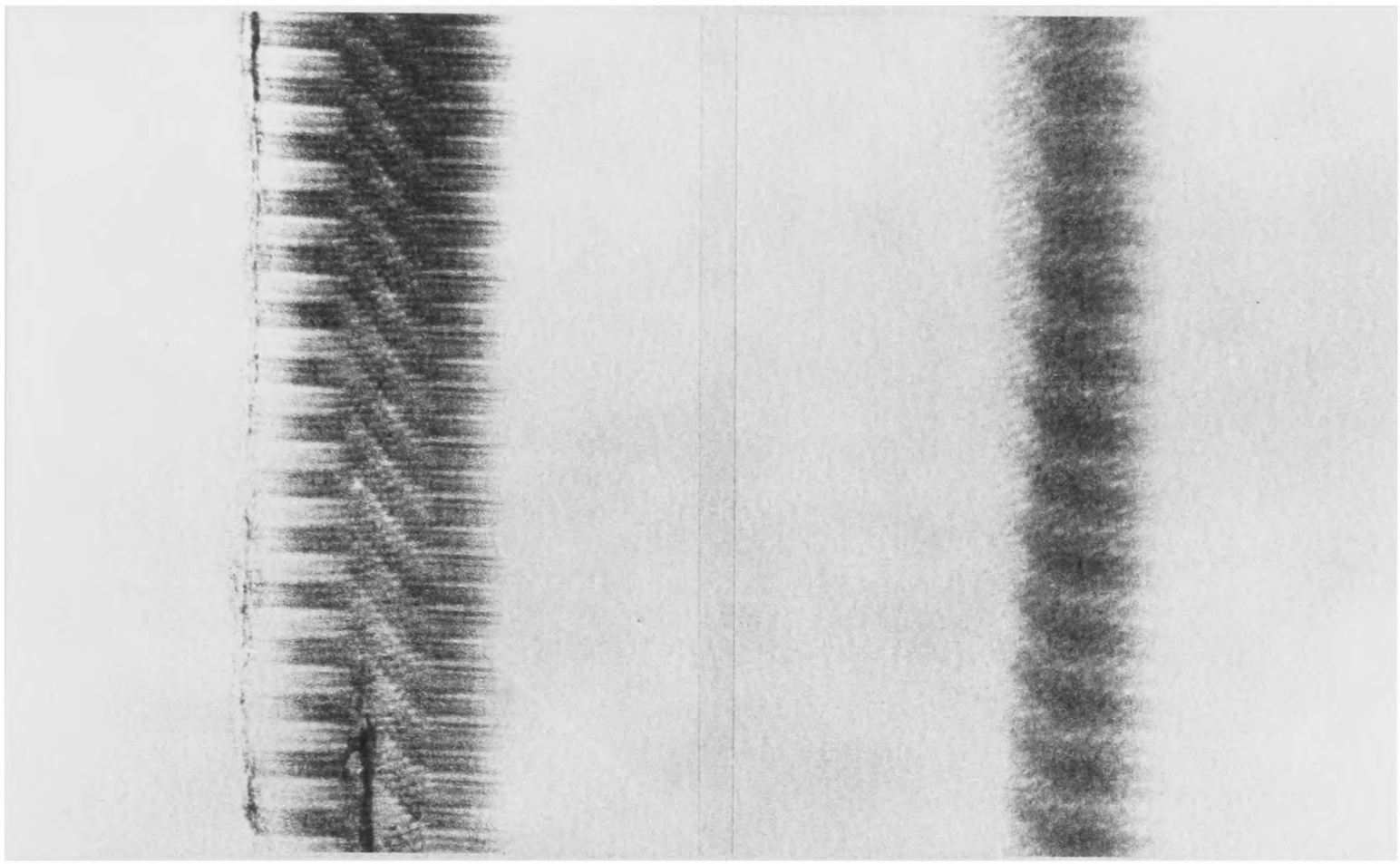


PLATE 3.9

Electron micrographs of structures obtained after precipitating chick cartilage collagen according to the method of Bruns et al. (1973). Both specimens were negatively stained with 3% uranyl acetate. Micrographs courtesy of Dr. J. Woodhead-Galloway and Dr. B. Brodsky-Doyle.

(a) A transition from the native to the oblique striated form.

(b) A transition from the DPS to the oblique striated form.



PLATE 3.10

Obliquely striated tactoid with pseudo periodicity of D/7. The specimen was prepared (by Dr. B. Brodsky-Doyle) according to the method of Bruns et al. (1973), and negatively stained with uranyl acetate. Micrograph courtesy of Dr. J. Woodhead-Galloway.

which shows an obliquely striated tactoid with a pseudo periodicity of D/7=96Å. This plate is also rather distinctive in that it shows two superimposed sets of oblique striations symmetrically disposed about the fibre axis. As the precise three dimensional structure of these tactoids is not known the exact significance of this observation is uncertain. It may be that the two sets of oblique striations are due to negative staining of the front and back of a helical structure. Whatever its cause, however, it may also enhance the apparent seventhing of the true D periodicity by a super-position effect.

When the stagger between neighbouring subfibrils is D/6 there would be an apparent periodicity of 111Å. It is suggested that this is the origin of the so-called 110Å periodic structure, it is merely a special case of the obliquely striated polymorphic form.

3. Origins of the Polymorphism in the Sequence

It is clear from the range of observed polymorphism that the way in which collagen molecules aggregate to produce ordered structures is critically dependent on many environmental factors such as pH and the presence of other macromolecules and ions (Gross, 1956). Hence it is likely that different properties of the sequence, under the various conditions, will become important in determining the overall minimum energy configuration. For the case of native-type aggregation, for example, it was shown in Chapter 2 that the n x 670Å stagger between neighbouring molecules has its origins, at least partly, in the pseudo-periodic distribution of hydrophobic and charged side chains. The success of the analysis showed that, for collagen, it is valid to consider the one dimensional structure in terms of the one dimensional distribution of amino acid residues along the

sequence (i.e. no account was taken of the azimuthal arrangement of the side chains).

i) DPS

It was suggested by Doyle et al. (1974a) that the origin of DPS and other anti-parallel structures may be in the presence of planes of mirror symmetry in the distribution of interacting residues. To decide which interactions might be important it is necessary to consider the chemical conditions in which these structures are precipitated.

D-periodic symmetric collagen appears to be produced either by precipitation at low pH (4.3 to 5.5) or after enzyme treatment. The pK's of the aspartic and glutamic acid side-chains are 3.9 and 4.3 respectively, hence these groups are largely undissociated at the lower limit of the above pH range. It therefore seems possible that, when the pH is low, electrostatic attractions between collagen molecules are inhibited and the ordered structures which are formed must be due to some other type of interaction. If these electrostatic forces are also affected by precipitating collagen at relatively high ionic strength, Gross (1956) has reported distinct changes in the ordered structures which result. He observed fibrils with an apparent 220Å periodicity after shadowing. It is possible that this is also a D-periodic symmetric structure since the distribution of the bulky terminal peptides has an approximate D/3 (220Å) periodicity (see Fig. 3.1 and Plate 3.6(a)). As an identical form of DPS collagen is also produced after enzyme treatment it seems likely that the same fundamental interactions are involved as in the previous cases. It is therefore suggested that the hydrophobic interactions are important for specifying DPS.

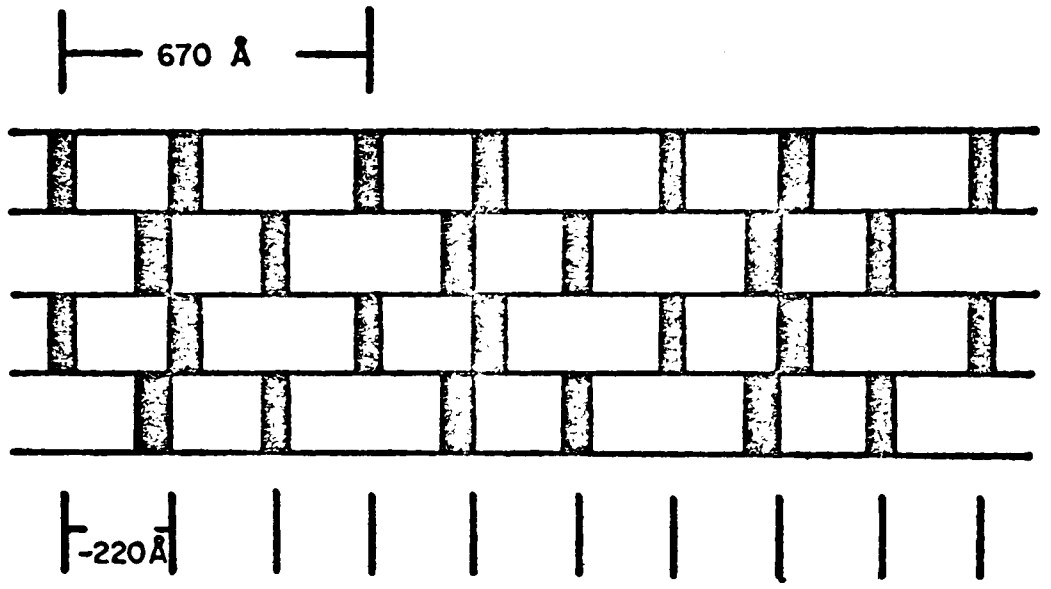


Fig. 3.1 . Diagram showing two pairs of anti-parallel native D-periodic sub-fibrils arranged as in Plate 3.6(a). The black bars show the locations of the bulky terminal peptides which occur approximately every 220 Å in projection on to the axis, thus giving the overall structure an apparent $D/3$ periodicity.

The planes of mirror symmetry in the pattern of large hydrophobic residues along the sequence are relatively easy to locate, as indicated in Fig. 3.2(a),(b) and (c). Using the computer programme in Appendix 1, the total number of hydrophobic interactions between two anti-parallel molecules for the full range of possible relative staggers were calculated. For each particular stagger a favourable interaction was scored whenever a large hydrophobic residue in one molecule was within a range of $\pm$ 2 residues from the adjacent residue in the other molecule. The resulting curve is shown in Fig. 3.3. It shows many peaks regularly placed at multiples of D/11, as expected from the known residue distribution (see Chapter 2), with the largest peak corresponding to an overlap of N-terminal helical ends of 866 residues. This situation does indeed correspond to the lining up of mirror planes in Fig. 3.2(c).

As must be the case, given the D-periodic distribution of hydrophobic residues, there are other peaks in the curve of Fig. 3.3 at multiples of 234 residues either side of the main peak. The fact that these are not all local maxima reflects the approximate way in which the hydrophobic residues conform to the $(\frac{2D}{11})_4$ $(\frac{D}{11})$ $(\frac{2D}{11})$ pattern. This D-periodic nature of the interaction curve must be true for all overlaps related by multiples of D. It is, therefore, possible to contract the curve somewhat by considering all D-related overlaps simultaneously. This may be done by adding to each overlap in the range 0 to 234 all interactions related by $\pm$ 234n (where n is an integer). The result is shown in Fig. 3.4. It is seen that the overall maximum occurs at an N-terminal helical overlap of 160 $\pm$ 234n residues. This value is in good agreement with that obtained quite independently from the simulation of the positively stained pattern (Section 2(ii)).

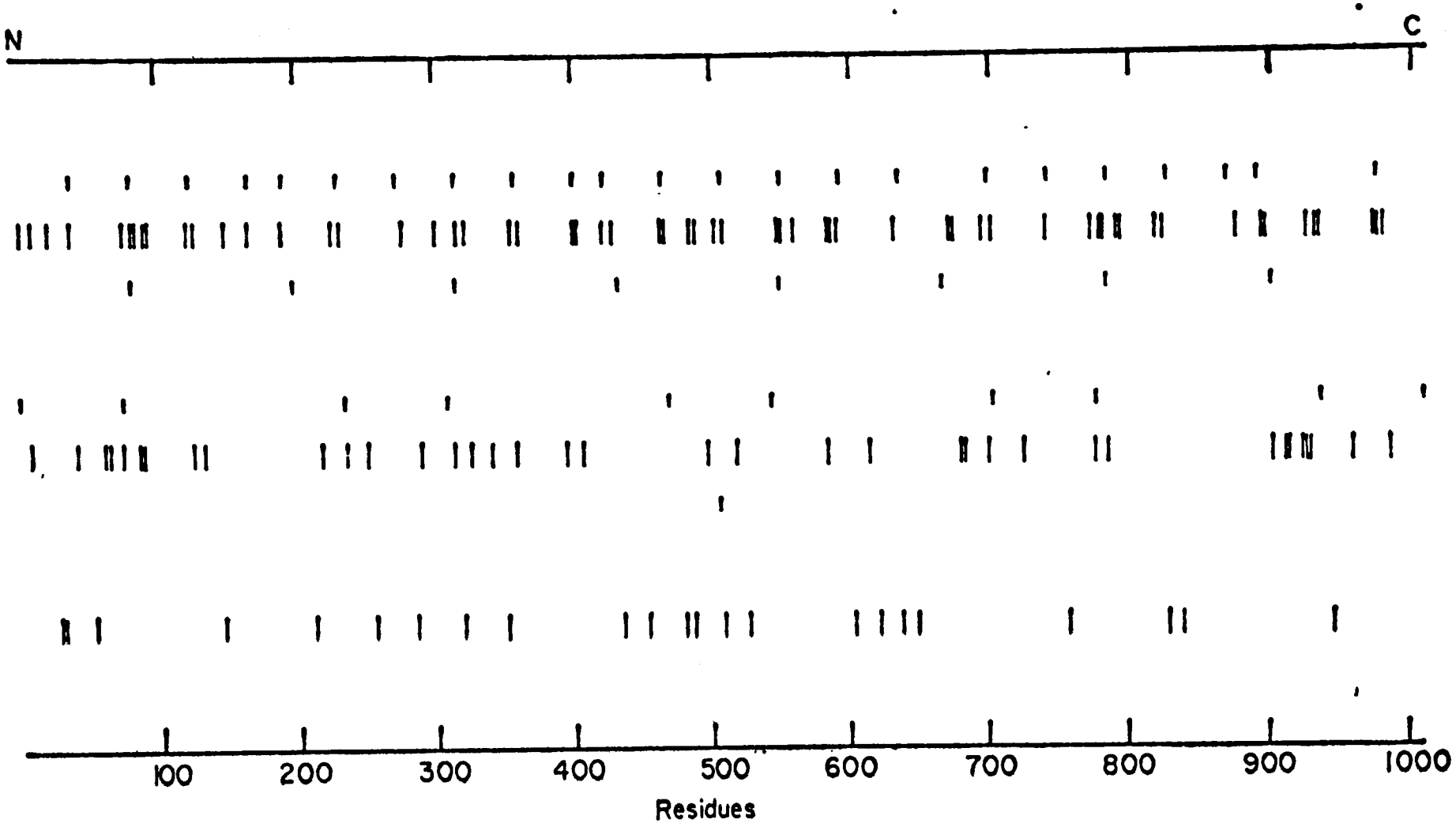


Fig. 3.2. Distributions in the 1011 residue helical region of the rat/calf skin collagen α1 amino acid sequence. The native 670 Å (*D*) periodicity is assumed to be 234 residues (from Hulmes *et al.*, 1973).

(a) Idealised distribution of large hydrophobic residues

(b) Actual distribution of large hydrophobic residues (i.e. Phe, Met, Leu, Ile and Val).

(c) Approximate local mirror planes in the patterns of Fig. 1(a) and (b), spaced periodically at intervals of *D*/2.

(d) Distribution of positions along the sequence which would correspond to both the N- and C-terminal helical ends of the so-called overlap region when molecules are assembled in a native fibril.

(e) Distribution of unpaired positively charged residues (i.e. Arg and Lys).

(f) The mid-point of the helical region.

(g) Distribution of unpaired negatively charged residues (i.e. Asp and Glu).

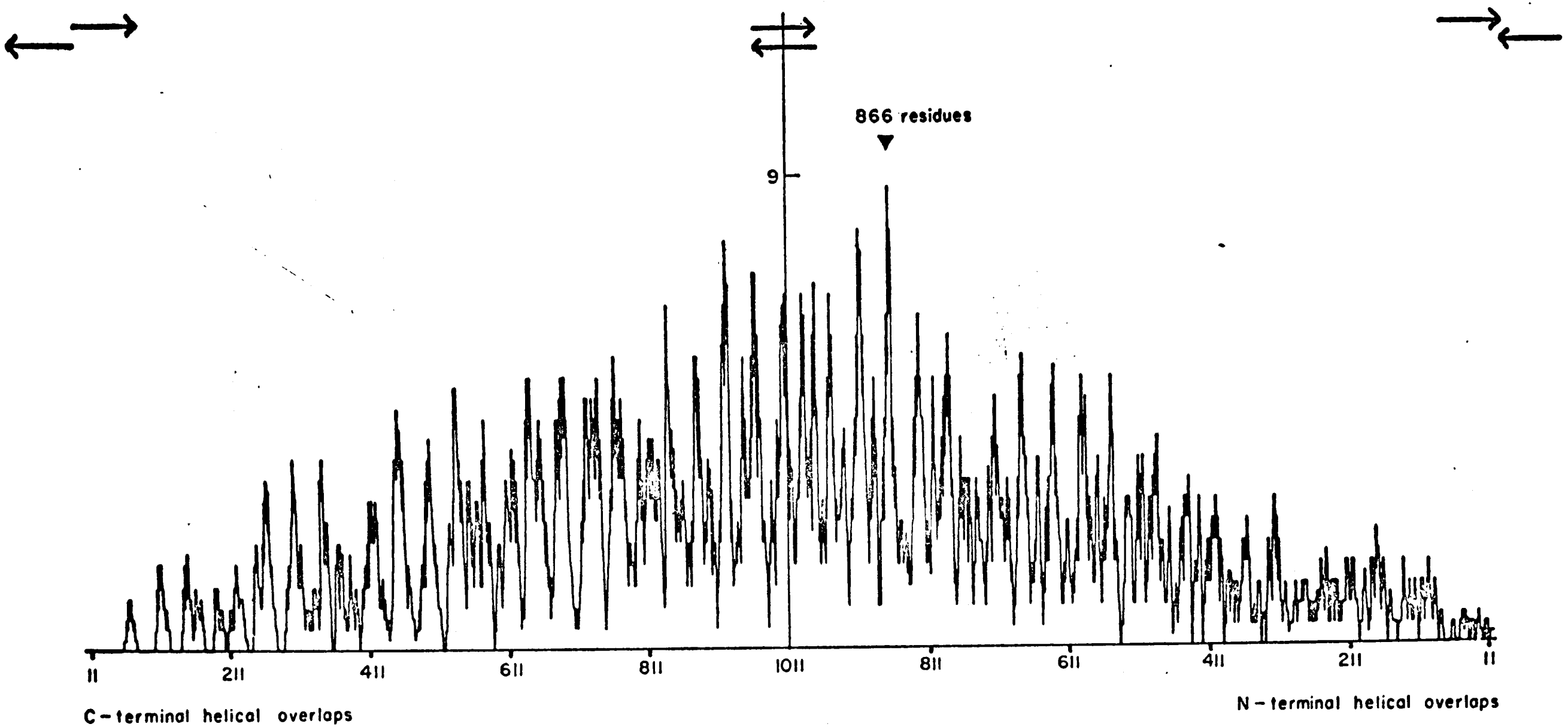


Fig. 3.3. Anti-parallel interactions between large hydrophobic residues of two α1 sequences. The molecular arrangements at selected overlaps are indicated. (A molecule is represented by a horizontal line with an arrowhead at the N-terminus.) For each overlap the total number of favourable interactions has been divided by the total allowed range (in this case 5 residues).

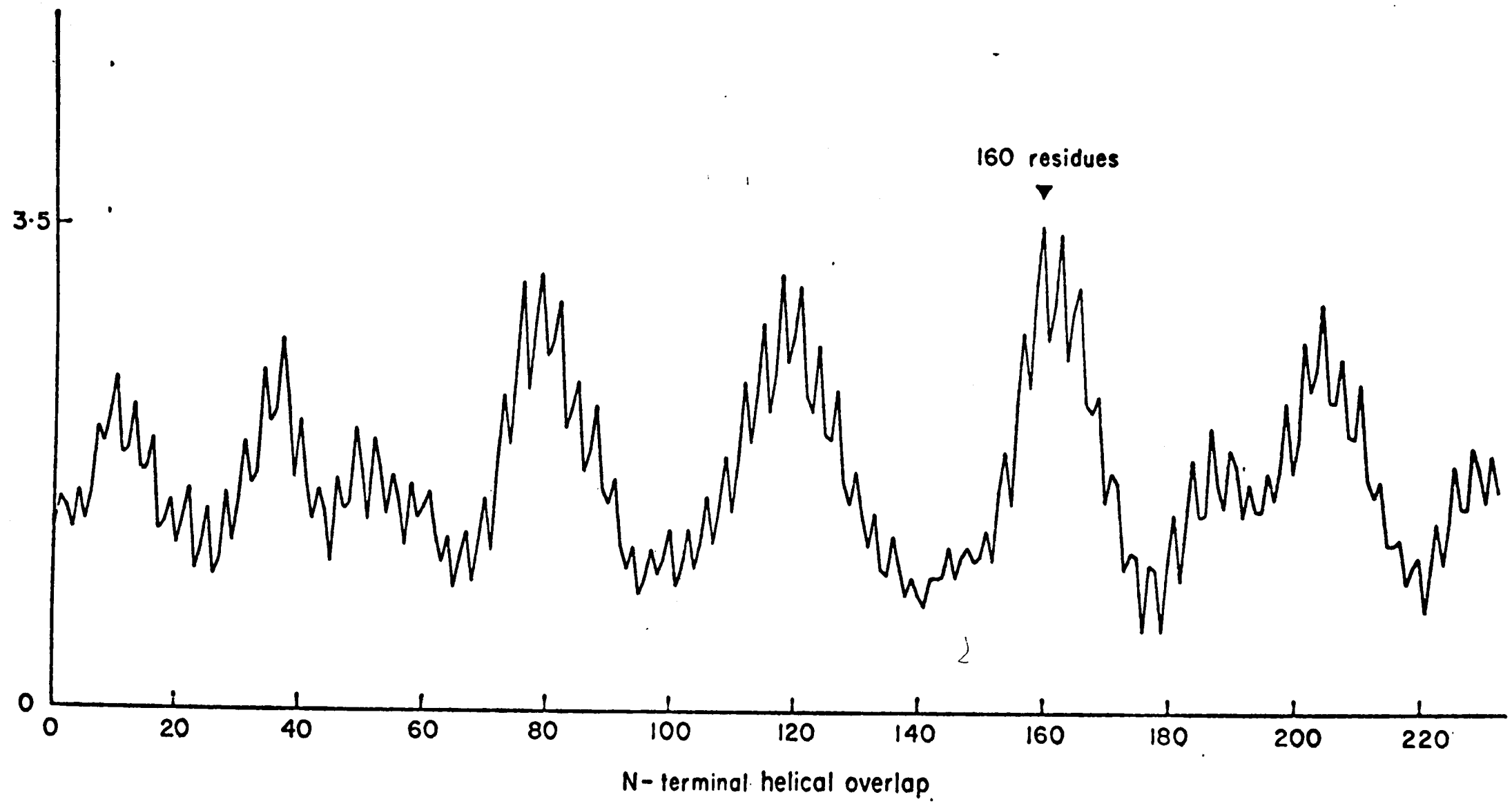


Fig. 3.4. Anti-parallel interaction curve constructed from Fig3.3 by adding together all overlaps related by $\pm 234n$ (where n is an integer). This curve is equivalent to the interaction curve between a pair of anti-parallel, axially projected, native D-periodic structures.

The above analysis has shown the origins of DPS for skin collagen in the rat/skin amino acid sequence. This sequence does not, however, explain the different overlap of 150 residues for cartilage collagen. This is perhaps not surprising since the sequences from skin and cartilage collagen may be quite different. The charged side chain distributions in SLS from the two types of molecule are, however, very similar (Stark et al., 1972). This supports the idea that DPS formation is not due to interactions between these charged side chains. If it were, there would be no difference in the structures using collage from either tissue. It is probable, therefore, that the difference in amino acid sequences between skin and cartilage collagen resides mainly in the distributions of large hydrophobic side chains.

ii) FLS

The formation of FLS is due to an interaction between collagen and other macromolecules. In most cases chondroitin sulphate has been used, though many other substances are capable of inducing FLS formation, see Dickinson (1969). All of them are relatively flexible polymers, many with a net negative charge. For serum glycoprotein, Gross et al. (1954) have estimated that 10-15% is present in FLS I. In the case of chondroitin sulphate, Dickinson (1969) has shown by bismuth nitrate staining that these molecules are located at fairly specific sites in the structure. Doyle et al. (1975) have also used this method for staining FLS IV, it shows two wide bands of chondroitin sulphate per period located at each end of the overlap region (see Plate 3.11(a)). An electrostatic interaction in solution between various muco-polysaccharides and proteoglycans with collagen has also been

a

b

c

d

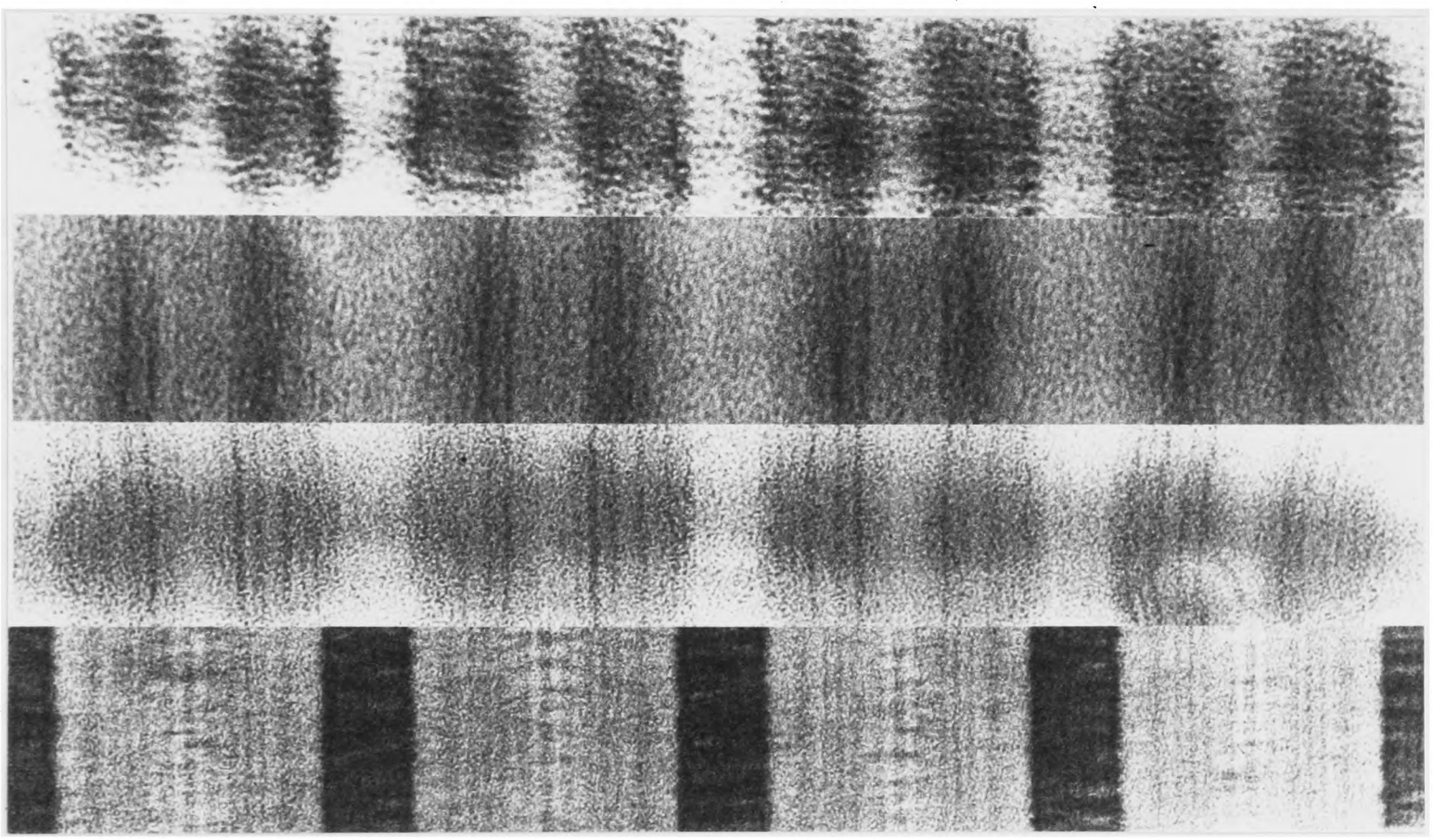


PLATE 3.11

Electron micrographs of FLS IV (courtesy of Dr. J. Woodhead-Galloway). All specimens were prepared (by Dr. B. Brodsky-Doyle) according to the method of Olsen (1963), using lathyritic rat skin collagen.

(a) FLS IV stained with saturated, unbuffered bismuth nitrate solution, showing the location of the chondroitin sulphate.

(b) FLS IV positively stained with 1% phosphotungstic acid (pH 3.4).

(c) FLS IV positively stained with 0.1% uranyl acetate (unbuffered).

(d) FLS IV negatively stained with 2% phosphotungstic acid (pH 7), showing clearly the gap and overlap regions.

demonstrated by Mathews (1965), Gelman and Blackwell (1973, 1974) and Obrink (1973).

A series of computer calculations on the anti-parallel interaction between two collagen sequences were carried out. There were no peaks corresponding to the FLS contacts for either hydrophobic interactions (Fig. 3.2) or electrostatic attractions between the charged side chains. This suggests that some account must be taken of the influence of the other macromolecules on the structure.

Electron microscopy provided a clue to the factors which may be important, by interpretation of the difference in appearance of FLS IV with different stains, see Plate 3.11(b) and (c). The distribution of positively charged side chains, as shown by staining with uranyl acetate, is quite different from the distribution of negatively charged side chains, as shown by staining with phosphotungstic acid. This sort of behaviour is not a feature of native collagen. There are two reasons for this: (1) there is intramolecular charge pairing (100 out of 160 charged groups can be considered as pairs of positively and negatively charged side-chains separated by no more than 3 residues); (2) there is intermolecular charge pairing (11 out of 12 bands per period are electrically almost neutral (see Doyle et al., 1974a)). Since the intramolecular charge pairing is always present in any collagen aggregate, the source of the stain-dependent appearance of FLS IV must lie in a tendency of the unpaired charges not to neutralise each other. Indeed it is possible that unpaired positive charges tend to line up in FLS due to a bridging effect via the periodically negatively charged chondroitin sulphate. This idea was tested by calculating the anti-parallel interactions, scoring for positive charge

alignment, for both the total and unpaired basic side chains. The range of interactions was $\pm$ 3 residues. The first curve does not have any special features at the FLS overlap positions; the second curve, however, is more interesting, as shown in Fig. 3.5(a). Though it looks at first sight to be rather 'noisy' there are low resolution maxima at both N and C terminal overlaps of about 150, 430, 780 and 1011 residues. The peaks are made clearer by allowing a much larger bridging range between positive charges, $\pm$ 30 residues rather than $\pm$ 3. The effect is shown in Fig. 3.5(b). The positions of the maxima correspond favourably to all the overlaps required to explain the FLS series, as listed at the end of Section 2(i).

So far it is not clear why the collagen molecules should assemble in FLS in an anti-parallel rather than parallel manner. A similar analysis as above was therefore calculated for two parallel molecules. The result is shown in Fig. 3.5(c). The exact centre of symmetry in this curve is a necessary result of the analysis. The striking similarity between this curve and Fig. 3.5(b), however, was not expected. Both the positions and heights of the peaks are almost identical. The only way this can be is that there must be an almost exact centre of symmetry in the distribution of unpaired positive charges within the molecule itself, see Fig. (3.2(e) and (f). This, of course, explains why the FLS structures are symmetric; the interacting residues are distributed approximately equally either side of the centre of the molecule, and hence it does not matter in which direction the molecules are oriented.

In contrast to the unpaired positive charge distribution, the distribution of unpaired negative charges does not have an approximate centre of symmetry, see Fig. 3.2(g). It may be, then, that repulsions between these negative charges when the

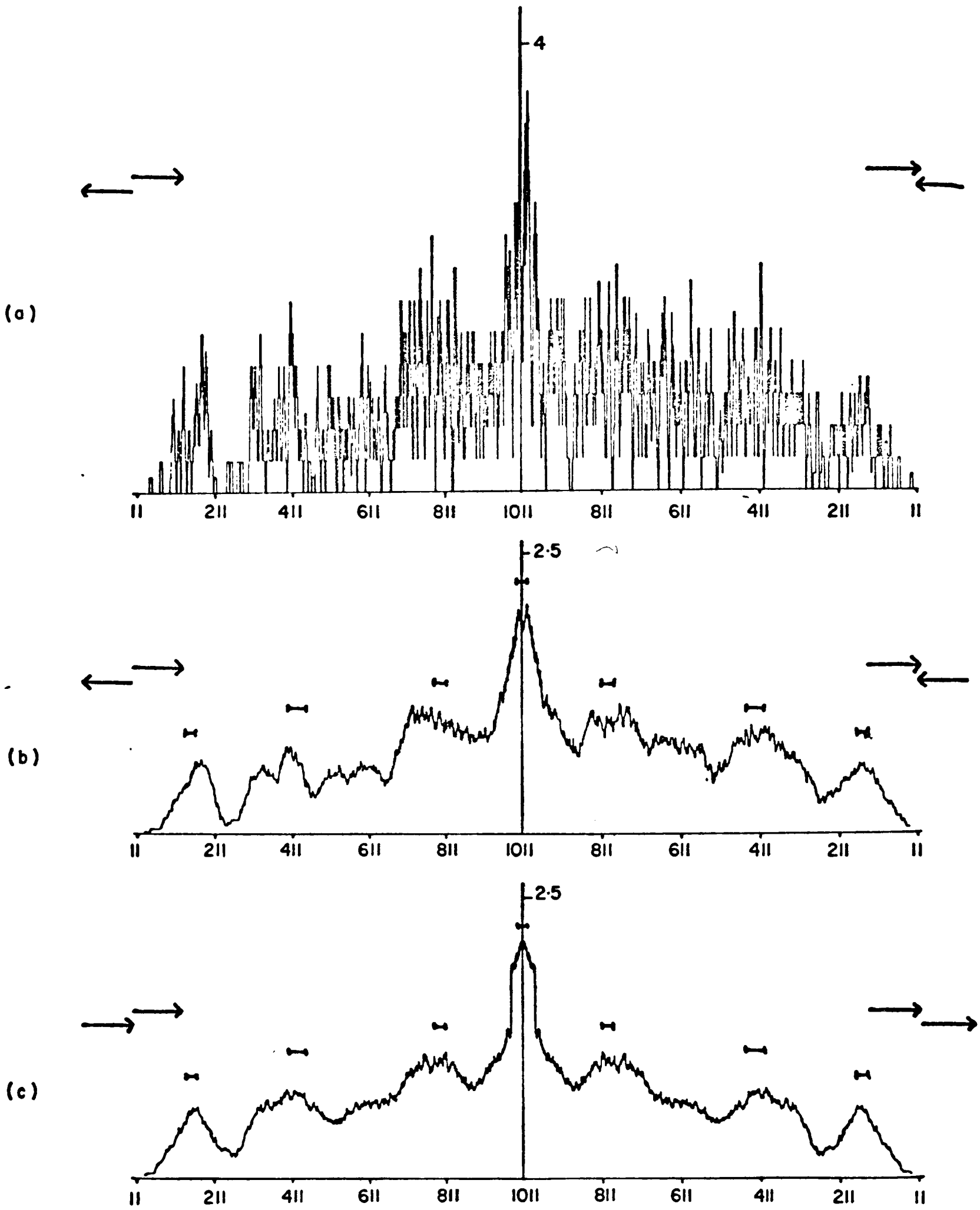


Fig. 3.5. (a) Anti-parallel interactions (i.e. scoring for charge alignment) between the unpaired positive charges of two *al* sequences. For each overlap the total number of interactions has been divided by the total allowed range (in this case 7 residues).

(b) As (a) except that the range of interactions is ± 30 rather than ± 3 residues. The total number of interactions from each stagger has therefore been divided by 61.

(c) Parallel interactions between unpaired positive charges with an allowed range of ± 30 residues.

In each graph the molecular arrangements at selected overlaps are indicated as in Fig. 3. Also in each case the abscissa is the helical overlap in residues. In (b) and (c) the experimentally determined limits of the FLS overlaps, as listed at the end of section 2(i), are indicated by short vertical bars separated by horizontal lines.

molecules are in parallel register may make the anti-parallel and in register situation more energetically favourable.

In summary, the origin of fibrous long spacing collagen lies in the distribution of unpaired positively charged side chains in the amino acid sequence. Adjacent molecules are held together by a bridging interaction between these charges using a variety of polyanions. The electrostatic attractions of the remaining positive charges must be effectively neutralised as a consequence of each one being paired with a negatively charged side chain within the sequence.

iii) SLS

The formation of polar SLS must be due to some sort of affinity between like amino acids. Such an affinity can be one of two kinds, either a direct hydrophobic interaction between large, uncharged side-chains, or an indirect bridging interaction between charged side-chains via some non-collagen molecule. Since ATP, or some other small negatively charged molecule (Bowden and Chapman, 1972; Bowden et al., 1968) is required for SLS formation, it is likely that this is used to bridge adjacent side-chains having a net positive charge (Kühn and Zimmer, 1961c). It is suggested that the combination of these two types of interaction specifies the formation of polar SLS.

The formation of symmetric SLS is less readily understood, unless some FLS inducing molecules are present together with the collagen; in which case the previous analysis for FLS would apply.

4. Discussion

i) Bridged interactions

It would seem that the principles governing the arrangement of collagen molecules in SLS and FLS are similar. In both cases, the interactions between collagen molecules are to some extent mediated by the presence of another type of molecule. For SLS this molecule is ATP, or a molecule of similar size and electrical properties (see Bowden and Chapman, 1972), and for FLS it is chondroitin sulphate or any other member of the list which appears in Dickinson (1969). These mediating molecules are similar in that they all have an affinity for positively charged amino acid side-chains: they differ in that it is the smaller molecules which are involved in SLS and the larger ones in FLS. To understand FLS formation it was necessary to invoke a large range over which positively charged side-chains could be bridged by the mucopolysaccharide. In the analysis presented here this range was up to 30 residues, or (30 x 2.86Å=) 90Å. Hence, according to this crude analysis, the distance between any two negative charges on an FLS inducing molecule must be at least 90Å. This is readily possible for a molecule of chondroitin sulphate with, for example, a molecular weight of the order of 10^4; such a molecule has a length of approximately 200Å with at least one negatively charged sulphate group approximately every 10Å. For SLS formation, on the other hand, this facility for a large bridging range is not available, since an ATP molecule is only about 10Å long. Hence ATP can only bridge positive charges between neighbouring collagen molecules which are relatively close together. The optimum situation is, as stated before, in the molecular arrangement for polar SLS, where all positive charges must be adjacent to corresponding charges on

neighbouring molecules. For polar SLS, then, positive charges are aligned much more precisely than in FLS. It is, therefore, the low resolution features of the collagen molecule which are important for FLS, and the high resolution features for polar SLS. The centre of symmetry in the distribution of unpaired positive charges in the sequence is only a low resolution feature. The length of the intermediate molecule, then, in both SLS and FLS, can be correlated with the degree of precision with which positive charges are aligned, and the degree of resolution which is important for determining the structure can have a marked influence on the minimum energy configuration.

Not all long, negatively charged polymers induce FLS formation, such an example is DNA (see Dickinson, 1969). This molecule is relatively rigid compared to some of those molecules which do induce FLS, for example certain gums and the chondroitin sulphates. For the bridging interaction discussed earlier to be most effective it would seem that flexibility of the FLS inducing molecules would be an advantage, since they would then be able to follow the perhaps contorted paths between positive charges. It is suggested that this is the explanation for the peculiar behaviour of DNA.

Dickinson (1969) has shown, by bismuth nitrate staining, that in FLS I the chondroitin sulphate is located predominantly in the overlap region. It is also located in the overlap region in FLS IV, at two specific sites per period (Plate 3.11(a)). These sites correspond to approximately 150 residue segments at each end of the collagen molecule, i.e. in the same regions which occupy the FLS I overlap. These observations correlate well with the distribution of unpaired positive charges used in the above analysis, see Fig. 3.2(e). There is a distinct

clustering of these residues at each end of the molecule in the regions which would appear in the FLS I overlap. This observation, then, is consistent with an interaction between chondroitin sulphate and the unpaired positive charges.

ii) Collagen-proteoglycan interaction in vivo

The relation of the above analysis of the *in vitro* interaction between mucopolysaccharide and collagen with the collagen-proteoglycan interaction *in vivo* is now considered. Fig. 3.2(d) shows the positions along the collagen sequence which would appear at the helical ends of the overlap region, when molecules are assembled in a normal fibril. The D stagger is assumed to be 234 residues, from Chapter 2. The distribution of unpaired positive charges approximates to this pattern fairly well. In particular there is a high density of positive charge 75 residues in from both the N and C terminal helical ends. The distribution, then, not only has a low resolution centre of symmetry, but also an approximate D periodicity. The effect of this is that, when the distribution of unpaired positive charges in a normal 670Å periodic fibril is calculated, there is a high density of positive charge at either side of the overlap region. Fig. 3.6 shows this distribution, together with the positions of the positive staining bands from Doyle *et al*. (1974a). The unpaired negative charge distribution is also shown, together with the approximate positions of the six unpaired charges in the terminal peptides. The high positive charge density is seen to coincide with bands $\underline{a}_2$, $\underline{a}_3$ and $\underline{a}_4$ near the C terminus and band $\underline{c}_2$ near the N terminus. Taken together $\underline{a}_2$, $\underline{a}_3$ and $\underline{a}_4$ have six net positive charges, $\underline{c}_2$ has four net positive charges (see Table 1 in Doyle *et al*., 1974a)). The rest of the D period is much more electrically neutral. It is,

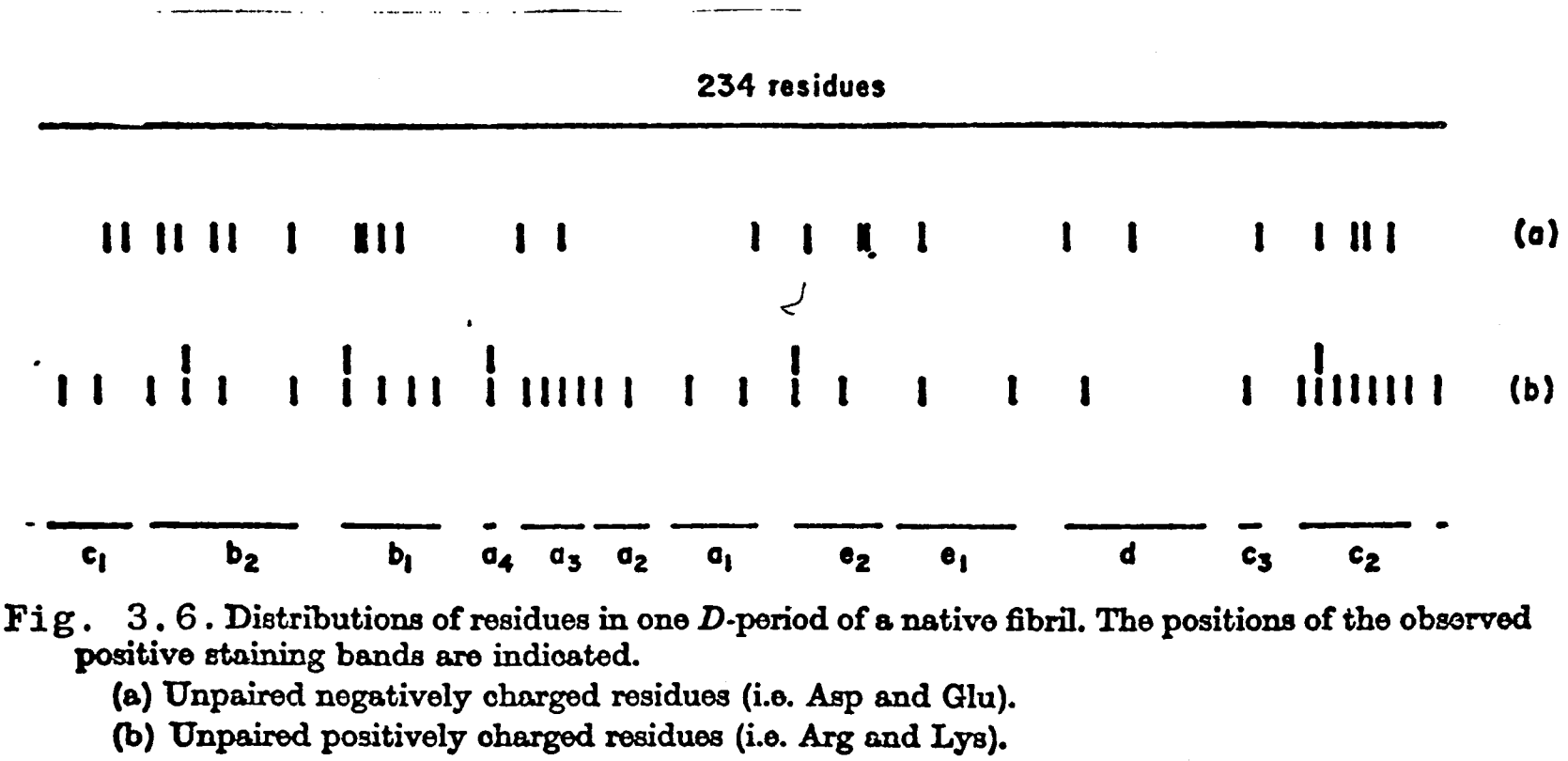


Fig. 3.6. Distributions of residues in one *D*-period of a native fibril. The positions of the observed positive staining bands are indicated.

(a) Unpaired negatively charged residues (i.e. Asp and Glu).

(b) Unpaired positively charged residues (i.e. Arg and Lys).

therefore, likely that the mucopolysaccharide component of the proteoglycan would be attracted to these two regions in the D period, with the highest affinity for the a band region.

Various workers have sought to locate the exact position of the proteoglycan with respect to collagen fibrils *in vivo*, using selective staining and electron microscopy (Serafini-Fracassini and Smith, 1966; Nakao and Bashey, 1972; Torp *et al*. (1975). Serafini-Fracassini and Smith, using bovine nasal cartilage, observed two bands of bismuth nitrate staining material per D period at the a and b_1 bands. Nakao and Bashey observed ruthenium red staining material on the surface of the fibrils from rabbit heart valve, between d and a_4 bands. Torp *et al*. have studied rat tail tendon, again ruthenium red stains material between the d and a bands (see Fig. 6 of their paper). Though these results are from different tissues, it is interesting that the a band is common to all observations.

The a band also seems to be important in the lateral aggregation of collagen fibrils in skin and tendon. Several workers have studied longitudinal sections of these tissues from various species (Gieseking, 1962; Borysko, 1963; Braun-Falco and Rupec, 1963). Two prominent features are observed - first, neighbouring fibrils can be either parallel or anti-parallel, and secondly, there is lining up of bands across the fibre. In particular the a band is often continuous across several fibres (Gieseking, 1962; Braun-Falco and Rupec, 1963). These phenomena have also been observed in rat tail tendon, as illustrated in Plate 3.12.

Strictly, one should only try to relate the above observations to the amino acid sequences for each tissue. The only

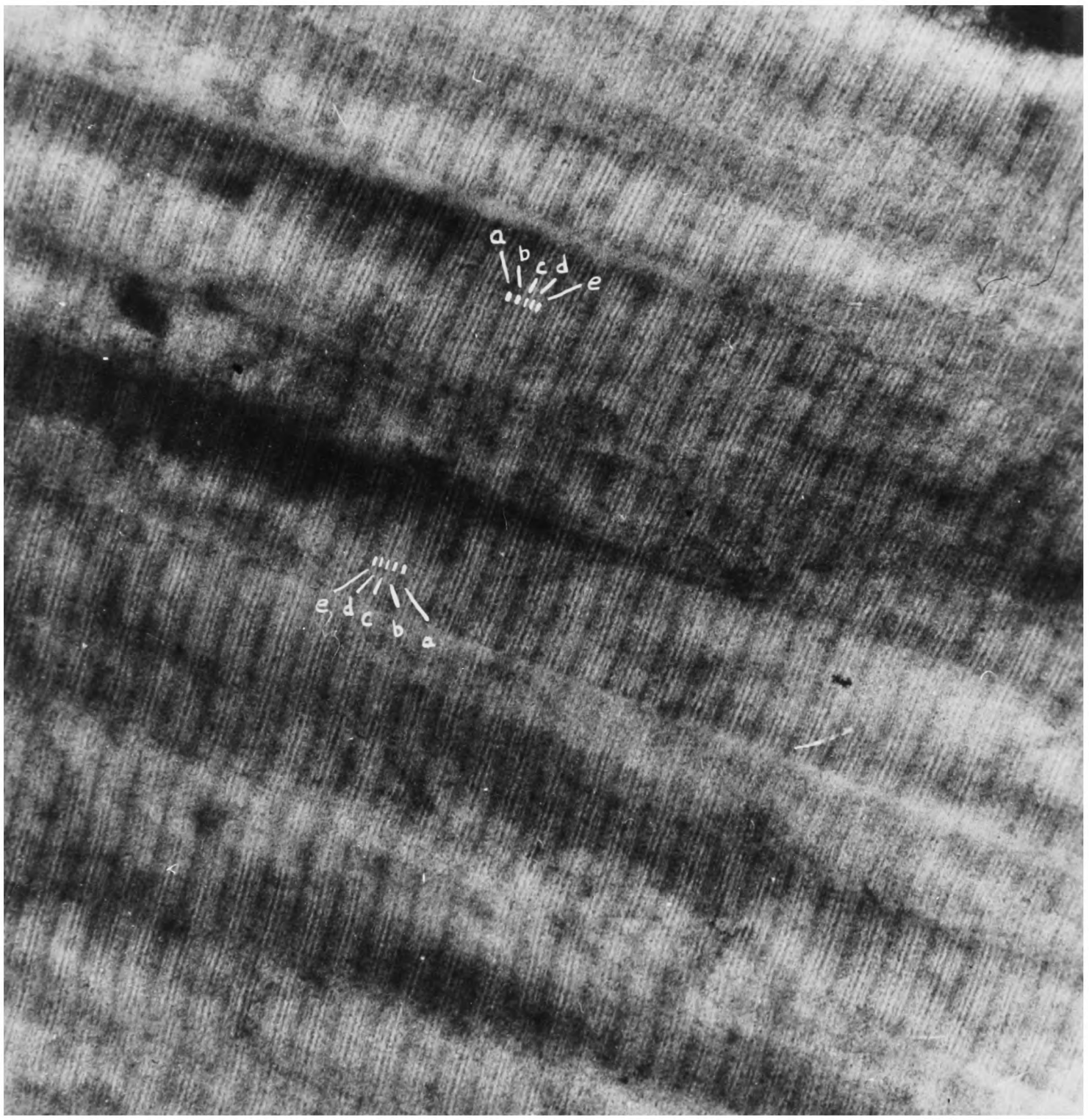


PLATE 3.12

Electron micrograph of a longitudinal section through rat tail tendon, stained with uranyl acetate and lead citrate (courtesy of Mrs. B.M. Luke). Both parallel and anti-parallel fibrils have the 'a' bands in register.

known sequence is for the α1 chain of rat/calf skin. Fortunately however, the band nomenclature of Hodge and Schmitt (1960) seems sufficient to describe the appearance of collagen fibrils from many tissues and species (see Bairati, 1972). Hence the charge distribution, at least is fairly conservative. The known sequence is certainly adequate to describe in detail the band pattern of rat-tail tendon (Chapter 2; Chapman, 1974; Chapman and Hardcastle, 1974). We can, therefore, say that the location of the mucopolysaccharide component of proteoglycan on the collagen fibrils, in the region of the a band, is very likely due to an interaction with the excess positive charge in this region. Furthermore, the lining up of the a bands between both parallel and anti-parallel collagen fibrils may well be due to an interfibrillar bridging interaction between this excess positive charge, via the negatively charged mucopolysaccharide side chains of the proteoglycans.

iii) Collagen morphology in the presence of other connective tissue components

In the preceding section the probable interaction of proteoglycans with collagen fibrils in connective tissue was discussed. Previously, however, it was shown that the mucopolysaccharide component of the proteoglycan can interact with collagen molecules to form FLS (i.e. fibrils with different periodicities from the normal 670Å). What are the factors which dictate the particular morphology of collagen aggregates in the presence of these other connective tissue components? Firstly, FLS is normally made with only the protein free mucopolysaccharide, so it might be suggested that the formation of either D periodic fibrils or FLS is dependent on the nature of the inducing molecule. This cannot be so, however, since FLS was also observed after

following the procedure of Olsen (1963) using the proteoglycan subunit (Hascall and Sajdera, 1969) instead of the free mucopolysaccharide (Doyle et al., 1975). Secondly, then, it is likely that the chemical conditions are the deciding factors. A number of workers (Keech, 1961; Mathews and Decker, 1968; Toole and Lowther, 1968) have studied the precipitates formed when collagen solutions are mixed with mucopolysaccharides or proteoglycan solutions at physiological pH and ionic strength. In all cases, D periodic fibrils were observed. When the two solutions were mixed at acid pH, however, and the pH was raised by addition of salt and subsequent dialysis against water (Olsen, 1963; Fitton-Jackson and Randall, 1953), FLS was formed. The initial pH, then, seems to be the important factor. When the pH is low the carboxyl side chains in the collagen molecule tend to be dissociated (see the section 3(i)), hence the electrostatic attraction between the molecules would be reduced. It may be, therefore, that, under these conditions, collagen molecules have a greater affinity for mucopolysaccharides or proteoglycans than for other collagen molecules; and the result is FLS.

iv) Some general observations on specificity

In Chapter 2, evidence was presented for the importance of dipole-dipole interactions in specifying the native D stagger. There are two advantages of this type of interaction. First, when collagen molecules self-assemble in the extracellular matrix, it is important that they should 'recognise' each other and not form abnormal FLS-like aggregates with the proteoglycans, as discussed in the previous paragraph. If collagen molecules interact with proteoglycans via isolated charges and with other collagen

molecules via dipoles, the protein-glycosaminoglycan and protein-protein recognition can be separated on the basis of the type of interaction. Paired charges in collagen would therefore be shielded from interaction with proteoglycan and free to interact with other paired charges. This may provide an evolutionary selection pressure for the apparent D stagger specificity in charged pairs. These considerations may be relevant to other multi-component self-assembling systems.

The second advantage of dipole-dipole interactions, which is not a property of simple isolated charge interactions, is that the relative orientation of interacting groups is specified. Thus if dipoles are oriented along the axis of fibrous protein molecules, interactions with similar molecules would serve to bring about either parallel or anti-parallel alignment. Dipoles, therefore, are vector quantities whereas isolated charges are scalars (Gerber et al., 1972).

The above analysis of collagen polymorphism has demonstrated the importance of the chemical conditions in governing the particular type of interactions involved. In the native case, both the interactions between large hydrophobic residues and the electrostatic interactions between charged residues are important (Chapter 2); for DPS it seems to be the interaction between large hydrophobic residues, and for FLS it is the bridged interactions between unpaired positively charged residues. Whatever the nature of the interaction, it is always, by definition, such as to bring together complementary regions of the molecule. Furthermore the mathematical operation which relates these regions must be either a simple translation or else a reflection through a local centre of complementarity. All the polymorphic forms listed in Table 3.1, with the

exception of SLS, show translational symmetry; also some forms (i.e. DPS, FLS, symmetric SLS) show centro-symmetry. In all cases, the mathematical operations of translation or reflection, which approximately relate complementary residue distributions within the molecule, are found as exact symmetry operations, relating equivalent regions of the molecular aggregates. In other words, when molecules are staggered by a distance ξ, there are at least two complementary distributions within each molecule, related by a translation of ξ (e.g. there are 4.4 equivalent D segments in the hydrophobic residues of the $\alpha 1$ sequence, but only 2 major D related distributions in the charged dipoles). When two molecules have an anti-parallel overlap of 2ζ, then complementary regions within each molecule must be related by a reflection through a point, distance ζ from the end involved in the overlap.

It is hoped that these ideas will be useful in understanding the self-assembly of many ordered fibrous aggregates.

CHAPTER 4

THE CORRELATION OF X-RAY AND NEUTRON DIFFRACTION PATTERNS OF NATIVE COLLAGEN WITH THE AMINO ACID SEQUENCE

1. Introduction

In previous chapters the analysis of electron microscopic banding patterns, in terms of the known α1 amino acid sequence, was used to solve the axially projected structures of the many forms of ordered aggregation of collagen molecules. In order to examine the specimens by electron microscopy, however, it was necessary to remove them from the conditions in which they were originally formed (either *in vivo* or *in vitro*). The specimens were treated with heavy metal stains and vacuum dried in the microscope. Other techniques, notably low angle X-ray or neutron diffraction, could, in principle, be used to study molecular packing in the original wet, unstained state. To give maximum structural information and sufficient diffracted intensity with these methods, however, it is necessary to have well oriented samples which are much larger than those used for electron microscopy. (A typical fibre from rat tail tendon used to give the low angle X-ray diffraction pattern in Plate 1.1(a), for example, contains approximately 10^6 of the fibrils shown in the electron micrograph in Plate 1.1(b)). The main disadvantage of the diffraction techniques is, of course, that the interpretation of diffraction patterns is less direct than the interpretation of electron micrographs. The intensities of the various parts of the diffraction pattern are related to the square of the modulus of the Fourier Transform of the specimen, whilst the relative phases of the Fourier components are not recorded. The nature of the molecular packing can therefore

only be inferred either by comparison of the observed diffraction pattern with a pattern calculated from a model, or else by obtaining the phases using a variety of crystallographic techniques (see, for example, Woolfson (1970)). If phases can be obtained, therefore, low angle diffraction techniques provide a means of investigating structure in the conditions in which the specimens were originally formed.

The Fourier transform of an object f(x,y,z) is given by

$$F(X,Y,Z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y,z)e^{2\pi i(Xx+Yy+Zz)}\,dxdydz$$

The variation of the transform along the Z-axis is therefore

$$F(o,o,Z) = \iiint f(x,y,z)e^{2\pi iZz}\,dxdydz = \int \sigma(z)e^{2\pi iZz}\,dz$$

where $\sigma(z)$ is the projection of f(x,y,z) on to the z-axis (i.e. the fibre axis). When corrected to give data proportional to the square of the modulus of the Fourier transform, therefore, the meridional intensities of an X-ray or neutron diffraction pattern of a collagen fibre depend solely on the variation in scattering length density in projection onto the fibre axis. The nature of the lateral packing of collagen molecules has no effect on the meridional intensities. The aim of the present chapter is to relate the observed meridional X-ray and neutron diffraction intensities from tendon to those expected on the basis of the amino acid sequence and axially projected structure described in Chapter 2. Besides being an extension of the structure-sequence theme developed in previous chapters, an understanding of the meridional diffraction to high resolution, in terms of the sequence, would represent a starting point for a number of interesting studies of biological problems of connective tissue. Such problems include:

a) the interaction of collagen fibrils and proteoglycans (see Chapter 3).

b) The action of certain enzymes on collagen fibrils, e.g. collagenase in tissue degradation, lysyl oxidase in cross-link formation, or procollagen peptidase in cleavage of the collagen precursor.

c) Genetic disorders, e.g. involving partial cleavage of the procollagen peptide (see Chapter 1).

d) Tissue specific or species specific collagens (see Marks et al., 1949).

All the above phenomena could involve localised changes in the electron (or neutron scattering length) density along collagen fibrils, by specific association of non-collagen molecules, by amino acid sequence changes or by the addition of extra peptides. The result would be a change in the meridional diffraction intensities which could be interpreted in terms of the rat tail tendon axially projected fibril structure. The advantage of the technique would be that it involves a minimal interference with the in vivo tissue structure. It would also be a method of localising changes detected by biochemical means.

2. Historical Background

The first evidence of a long range periodicity (i.e. greater than 200Å) in collagen fibres was obtained using low angle X-ray diffraction (Clark et al. (1935); Wyckoff et al. (1935); Clark and Schaad (1936); Corey and Wyckoff (1936)). As the diffraction patterns improved, the experimentally determined periodicity increased until Bear (1942) and Kratky and Sekora (1943) were able to index all observed meridional reflections, from dry beef achilles tendon and dry kangaroo tail tendon respectively, in terms of a fundamental periodicity of 640Å. This measurement was confirmed by the

direct observation of 640Å periodic fibrils in the electron microscope (Schmitt et al. (1942); Wolpers (1943)).

Before summarising the many attempts that have been made to interpret the low-angle meridional X-ray pattern, it is important to mention one fundamental assumption that is common to all the work. The assumption is that the diffraction effects are equivalent to those from a single collagen fibril, even though an X-ray specimen is composed of approximately 10^6 such fibrils. It has been assumed that the interference effects between fibrils are equivalent to those in gas diffraction, whereby the square of the modulus of the Fourier transform of the whole specimen is equal to the square of the modulus of the Fourier transform of an elementary unit (e.g. a gas molecule), multiplied by the total number of elementary units. The analogy with the gas is not quite appropriate, of course, since it is assumed that the fibrils in a collagen fibre are all parallel. The situation is more like a one-dimensional gas, where all axial shifts between neighbours are equally probable. The discussion in Chapter 3 showed, however, that there is evidence of lateral alignment between fibrils (possibly with the 'a' bands in register). Furthermore, there are equal numbers of fibrils oriented in each direction along the fibre axis (Braun - Falco and Rupec, 1963; D.A.D. Parry, private communication). If this alignment is perfect, the axially projected structure of a fibre would be the axially projected structure of a pair of anti-parallel fibrils, with 'a' bands in register, multiplied by the total number of fibril pairs in the fibre. Unlike the one-dimensional gas case, the Fourier transforms of the axially projected fibril pairs would now add together in phase. It seems unlikely that this alignment is so precise

throughout the fibre, however, in which case the inter-fibril interference effects would be intermediate between a one-dimensional crystal and a one-dimensional gas. The effect of this phenomenon on the intensities of the meridional Bragg reflections has not been fully discussed in the literature. There is some evidence, however, that the one-dimensional gas approximation is valid (see below).

The early interpretations of the low angle meridional X-ray reflections were concerned with two main aspects. First, there was the relation of the 640Å periodicity to the high angle 2.86Å reflection due to the residue translation in the triple helix and, secondly, there was the phenomena of the varying shapes of the low angle reflections in various humidities or in the presence of heavy metal stains and tanning agents. Bear (1944) showed that the spacing of the ~2.86Å reflection from tendon varies less with humidity, on a percentage basis, than the ~640Å spacing. A similar observation was made by Cowan et al. (1955b), though in this experiment the spacing of the high and low angle reflections were measured at constant humidity as a function of applied tension. The conclusions were that the 2.86Å reflection is associated with only the more highly ordered regions of the collagen fibril. This introduced the concept of ordered and disordered regions, an idea that was also used by Bear and his colleagues to explain the shape of the low-angle meridional reflections. When collagen fibres are fully hydrated, the observed low angle meridional reflections, though of varying intensities, all have the same shape perpendicular to the fibre axis. On drying, however, some reflections broaden and change shape. The broadening is approximately proportional to the index of the reflection.

Several theoretical models were calculated to explain these observations (Bear and Bolduan (1950a); Bear and Bolduan (1950b); Bolduan and Bear (1950); Bear (1951); Bear and Bolduan (1951); Bolduan and Bear (1951); Bear et al. (1951); Bolduan et al. (1951); Bear (1952); Rougvie and Bear (1953)). The conclusion was that the fibril is composed of ordered and disordered regions along the axis. The disordered regions were supposed to correspond to the dark bands seen in positively stained electron micrographs, whilst the ordered regions correspond to the unstained interbands. The need to invoke the concept of ordered and disordered regions was challenged by Tomlin and Ericson (1960), who interpreted the broadening of the meridional reflections in terms of a longitudinal shearing phenomenon. These authors did not, however, consider the evidence from the variation of the high and low angle X-ray spacings.

The next stage in the interpretation of the low angle meridional reflections was a number of attempts to relate fibril structure, as observed in the electron microscope, to the X-ray diffraction intensities. Kaesberg and Shurmann (1953) collected meridional data for wet and dry beef achilles tendon in the presence and absence of phosphotungstic acid (PTA). They then calculated Patterson profiles (see Woolfson, 1970) using the measured intensities. The Patterson for dry stained material showed some resemblance to a Patterson function based on the distribution of PTA positively stained bands in an electron micrograph of a single collagen fibril. If stained and air dried X-ray specimens are equivalent to the stained specimens observed in the electron microscope, then the result of Kaesberg and Shurmann indicates that the X-ray

meridional intensities from a fibre are indeed equivalent to those from a fibril. This problem of similating electron microscope conditions for X-ray work was later considered by Burge and Randall (1955) and Ericson and Tomlin (1959). Burge and Randall measured microdensitometer traces along the axis of unstained rat tail tendon fibrils observed in the electron microscope. They were then able to calculate an average density profile which, when Fourier transformed, gave a set of structure factor amplitudes for comparison with the observed X-ray diffraction amplitudes from an unstretched rat tail tendon fibre dried in vacuo. The agreement between the two sets of data was good, though it was necessary to introduce an exponential modulating function to reduce the amplitudes of the higher order calculated amplitudes. A similar experiment was carried out by Ericson and Tomlin (1959), who were able to obtain satisfactory agreement between observed and calculated amplitudes, without the need for a modulating factor, by measuring the X-ray meridional intensities from fibres dried at 200^{o}C (i.e. in conditions even more like those in an electron microscope). The conclusion from the work of Burge and Randall (1955) and Ericson and Tomlin (1959) is that it is appropriate, at least in the dry case, to consider the X-ray meridional intensities from a whole fibre as equivalent to those from a single fibril.

There have been two attempts (Bear and Morgan, 1957; Chandross and Bear, 1973) to relate the X-ray meridional intensities of unstained collagen fibres to the appearance of positively stained fibrils in the electron microscope. This approach can only be justified if the regions which take up positive stain are intrinsically electron dense, in which case an unstained fibril should show a similar band structure as a

stained fibril, though with reduced intensity. Early attempts to label individual amino acids with specific stains, and study the effects on the low angle meridional X-ray pattern (Bolduan et al., 1951; Bear, 1952), were interpreted as indicating that the disordered regions in fibrils (see previous paragraph), which were supposed to facilitate penetration of stain (and hence could be observed as bands 'a', 'b', 'c', 'd' and 'e' in the electron micrographs), were also the locations of charged amino acid residues. (This result was deduced without electron microscopy). Since the charged amino acids are larger than the smaller and less polar residues (Chandross and Bear, 1973; Bear, 1952) it was assumed by Bear and Morgan (1957) that stainable bands are intrinsically electron dense. It is shown in the next section that this assumption does have some validity.

Bear and Morgan (1957) obtained qualitative agreement between both wet and dry X-ray meridional intensities and the optical diffraction patterns of fibril models, the latter involving the combination of positively stained band structure and a background step function. For the wet case the background step width was 0.53D (D $\simeq$670Å), and for the dry case it was 0.78D. The significance of these background functions is immediately apparent in terms of the gap-overlap regions of the Hodge-Petruska model (see Chapter 1), where the overlap region is approximately 0.4D. This is reasonably close to the background step width in Bear and Morgan's model for wet collagen fibrils. The 0.78D background step width in the dry case could only be produced by a reduction of 8% in the intermolecular D stagger, if the step corresponds to the Hodge-Petruska overlap. Chandross and Bear (1973) have suggested that the 0.78D

step width does not arise from a change in the D stagger; they suggest it is due to a compaction of the gap region on drying. It is interesting to note that the optimal position of the background step function, with respect to the bands, in the fully refined model of Chandross and Bear (1973), corresponds to the position of the overlap region as deduced in Chapter 2. A quantitative comparison of the observed structure factor amplitudes of the first 15 orders from kangaroo tail tendon, with those calculated for the above wet and dry models, yielded R factors (see later) of 0.13 and 0.12 respectively. As Chandross and Bear (1973) pointed out, however, these models only take on any significance if the fluctuations in electron density can be related to the distribution of amino acids in the fibril. This is the purpose of this chapter.

There have been some attempts to deduce the axially projected collagen fibril electron density directly from the Patterson function obtained from the observed meridional intensities. For wet tendon, Kaesberg and Shurmann (1953), and later Tomlin and Worthington (1956), showed that the Patterson function is almost linear in falling from a peak at the origin to zero at 0.47D. Kaesberg and Shurmann (1953) interpreted this as due to an electron density function in the form of a D periodic step function, with the width of the step corresponding to 0.53D. Tomlin and Worthington chose a similar electron density, though with a step width of 0.47D. By Babinet's principle, the structure factor amplitudes of both models are identical. Tomlin and Worthington (1956) then used the phases from the step function model, together with the measured X-ray amplitudes, to produce a Fourier synthesis of the fibril electron density. As the periodic step function is centro-

symmetric, the phases introduce a centro-symmetry into the resulting Fourier synthesis. The Hodge-Petruska model would, of course, give a periodic step function in axial projection, and the gap/overlap characteristic is centro-symmetric. The collagen molecules are not centro-symmetric, however, so the fibril has a sense of polarity, as shown in Plate 2.1. It seems likely, then, that the electron density profile is centro-symmetric only at low resolution (see the next section).

A recent attempt to determine phases unambiguously, using heavy atom isomorphous derivatives, was made by Ellis and McGavin (1970). This study was mainly concerned with the axially projected structure of elastoidin, though some implications for beef achilles tendon fibrils were presented. Ellis and McGavin measured the low angle meridional X-ray intensities from moist elastoidin in the presence and absence of iodine, in the hope that the tyrosines in the protein would be specifically labelled, hence producing a change in relative intensities. Changes were in fact observed and the squares of the difference of the moduli of the measured structure factor amplitudes were used to calculate a so-called MD squared (modulus difference squared) synthesis. For a centro-symmetric structure, this function gives the Patterson function of the heavy atoms (Rossman, 1960; Ramachandran, 1964). The relative scaling of the two data sets was of course important; this was done by measuring the intensity of the direct beam relative to that of the strong third order, then equating direct beam intensities for each data set (see Ellis and McGavin, 1970). This procedure is only valid if all the samples used were of constant dimensions. Ellis and McGavin (1970) then deduced that the iodine was located in two bands per D period, separated by

0.4D. By assuming the iodine contribution made only small changes to the intensities from untreated material, and also that elastoidin was centro-symmetric, Ellis and McGavin were able to assign phases and hence produce a Fourier synthesis for native elastoidin. These authors then interpreted the difference Patterson for the intensities from moist native and iodine stained beef achilles tendon, obtained by Ericson and Tomlin (1959) and Tomlin and Worthington (1956), in terms of a similar two banded iodine distribution. Assuming a relative scaling for Tomlin and Worthington's data by equating first order intensities, Ellis and McGavin then assigned phases to the native beef achilles tendon. The resulting Fourier synthesis was similar to that obtained by Tomlin and Worthington (1956) by a different approach. In both cases the width of the (Hodge-Petruska) overlap region was 0.47D.

In summary, there are two assumptions that have been used to produce models for the axially projected electron density of collagen fibrils. The first was that positively stainable bands are intrinsically electron dense (Bear and Morgan, 1957; Chandross and Bear, 1973); the second was that the electron density is centro-symmetric (Tomlin and Worthington, 1956; Ericson and Tomlin, 1959; Ellis and McGavin, 1970). Both these assumptions can be tested by synthesising an electron density distribution based on the α1 amino acid sequence and the Hodge-Petruska model. Furthermore this distribution can be used to give a set of calculated structure factor amplitudes for comparison with the observed meridional data. These will be the topics of the next section.

3. Electron density distributions based on the amino acid sequence

i) Parameters used to describe model distributions

The Hodge-Petruska model, in axial projection, may be regarded as the convolution of a single collagen molecule with a one-dimensional lattice of spacing D (670Å). Each D period then contains five distinct molecular segments (see Fig. 1.1). The Fourier transform of this D periodic electron density distribution is equal to the transform of a single molecule sampled at intervals of 1/D. As a maximum of only 41 orders of the 670Å spacing have so far been quantified (see later), the resolution of low angle meridional X-ray diffraction pattern of rat tail tendon is limited to a spacing of (670/41≃) 16Å. A single amino acid residue, in projection on to the fibre axis, occupies a length of approximately 10Å, hence individual residues cannot be resolved using the present X-ray data. It was therefore considered appropriate to describe all models for the electron density in terms of residue positions rather than atomic positions. To calculate the Fourier transform of a model, each residue was treated as a δ-function with weight given by the total number of electrons in the residue. Hence not only the atomic form factors, but also the 'form factors' associated with the distribution of atoms in each residue, were approximated to 16Å resolution by a constant which depends on the number of electrons in the residue. To describe a model, therefore, two parameters, f_n (the scattering length) and z_n (the axial position) were used to describe each residue. These parameters are described below:

a) $\underline{z_n}$

It was assumed that, in the triple helical region of the collagen molecule, the residues on each α chain are equally separated along the entire length. The mutual positions of the three chains could therefore be described by the positions of their N-terminal glycine residues. Beginning with one α1 chain, the N-terminal glycine of the second α1 chain was staggered by one residue with respect to the N-terminal glycine of the first α1 chain, in the direction towards the C-terminus of the first chain. The N-terminal glycine of the α2 chain was further staggered, by one residue, in the same direction. This arrangement was only one of the three distinct possibilities (i.e. 1-1-2, 1-2-1, or 2-1-1).

The axial positions of the non-helical telopeptide residues were completely unknown, except of course that they were linked at the N- and C-termini of the helical regions. For the early models to be described, the assumption was made that the conformation of the telopeptides could be described by a single parameter r (r_N for the N-terminal telopeptide, r_C for the C-terminal telopeptide), which defined a uniform residue translation as a fraction of the triple helical translation. This simplification was abandoned for the later models, where non-homogeneous residue translations were used.

b) $\underline{f_n}$

It was convenient to include a term in f_n which was due to the scattering from the essentially water environment surrounding the molecules in a collagen fibril. In axial projection, the number of electrons in any small segment δz within the D period includes a contribution, p, from the protein

and a contribution, w, from the water. The total number of electrons in δz is therefore

$$s = p + w$$

The water contribution may be further expressed as

$$w = \rho\nu_w = \rho(\nu_f - \nu_p)$$

ρ is the electron density of water (0.33 electrons/Å^3)
ν_w is the volume of water in the fibril associated with the segment δz.
ν_f is the total volume of the fibril associated with the segment δz.
ν_p is the volume of protein in the fibril associated with the segment δz.

The total number of electrons in δz is therefore

$$s = p + \rho(\nu_f - \nu_p)$$

If the wet fibril diameter is constant throughout the D period the term $\rho\nu_f$ is constant. The variable part of the electron density is then

$$s' = p - \rho\nu_p$$

It follows that the effective number of electrons associated with the nth amino acid residue in the models to be described is given by

$$f_n = t_n - \rho\nu_n$$

t_n is the actual number of electrons in the residue
ν_n is the residue volume.

This method of allowing for water background was used by Fraser et al. (1965). The parameters ν_n where obtained from various published sets of residue volumes (see later).

The above equation, together with the z_n parameter, describes the contribution to the electron density from each residue in a water background. If other material (e.g. disaccharides or proteoglycans) are present, then this will affect the electron

density. Furthermore, not all the t_n values are known, as only certain parts of the α2 chain have so far been sequenced.

ii) The effect of varying the model parameters on the predicted meridional X-ray intensities

In the next few paragraphs the square of the modulus of the Fourier transforms of several electron density models are compared with the observed and corrected meridional X-ray intensities. This data was obtained from moist rat tail tendon held under tension over 0.15M sodium chloride (White, unpublished work) and corrected using the formula of Wang and Worthington (1975). The calculated and observed first order intensities were arbitrarily scaled to a value of 1000. Meridional Fourier transforms were calculated using the computer programme described in Appendix 2.

a) Amino Acid Residue Volumes

Three sets of residue volumes were obtained from the literature (Fraser et al., 1965; Goldsack and Chalifoux, 1972; Chothia, 1975). They are listed in Table 4.1, together with the total number of electrons in each residue and the net electrons after allowing for the water background.

To test the effect of using different volume sets, the calculated meridional intensities for an otherwise identical model were calculated. The details of the chosen model were:

3 α1 chains of 1052 residues
r_N =1) i.e. telopeptide residue translation same
r_C =1) as in triple helix
D = 234 residues

Table 4.2 shows the comparison with the observed intensities. It is clear that the positions of most of the strong orders are the same for the three volume sets, though the magnitudes with

TABLE 4.1 Three volume sets and the net electrons using the 3rd set

Amino Acid Residue	Number Electrons	VOLUME ($\AA^3$) 1st set (Fraser, McRae & Miller)	2nd set (Goldsack & Chalifoux)	3rd set (Chothia)	$(t_n - \rho v_n)$ for 3rd volume set
ARG	77	155.4	181.2	181.2	17.2
LYS	70	174.9	174.9	171.3	13.5
GLU	68	140.0	140.0	155.1	16.8
ASP	60	113.3	113.3	124.5	18.9
GLN	68	151.8	151.8	161.1	14.8
GLX	68	145.9	145.9	158.1	15.8
ASN	60	125.1	125.1	135.2	15.4
ASX	60	119.2	119.2	129.9	17.1
HIS	72	146.0	146.0	167.3	16.8
SER	46	90.8	90.8	99.1	13.3
THR	54	117.6	117.6	122.1	13.7
TYR	86	196.6	196.6	203.6	13.3
MET	70	166.2	166.2	170.8	13.6
PHE	78	192.8	192.8	203.4	10.9
VAL	54	140.5	140.5	141.7	7.2
LEU	62	167.2	167.2	167.9	6.6
ILE	62	167.2	167.2	168.8	6.3
PRO	52	130.2	122.2	129.3	9.3
HYP	60	141.7	126.0	133.1	16.1
GLY	30	60.3	60.3	66.4	8.1
ALA	38	87.0	87.0	91.5	7.8
HYL	78	-	-	175.1	20.2

N.B. Except for Arg, Pro and Hyp, the differences in residue volumes given by Fraser *et al.* and Goldsack and Chalifoux were so small ($<5\AA^3$) that they were neglected.

TABLE 4.2 The effect of using various volume sets

ORDER	OBSERVED INTENSITIES	1st Volume Set (from Fraser, McRae & Miller, 1965)	2nd Volume Set (from Goldsack, & Chalifoux, 1973)	3rd Volume Set (from Chothia, 1975)
1	1000	1000	1000	1000
2	27	7	11	9
3	109	108	109	106
4	11	35	9	10
5	54	154	76	87
6	17	124	49	61
7	21	103	42	48
8	6	18	4	6
9	71	23	10	8
10	19	28	4	7
11	20	115	69	64
12	31	19	8	7
13	2	90	44	47
14	4	57	24	34
15	4	49	29	33
16	5	32	14	14
17	11	12	21	19
18	4	23	3	4
19	6	28	15	23
20	34	39	17	17
21	24	122	31	36
22	11	16	1	2
23	0	5	2	2
24	0	3	6	6
25	14	32	12	16
26	5	24	21	24
27	11	37	18	21
28	1	2	3	0
29	9	9	1	0
30	16	31	20	15
31	1	0	0	0
32	2	27	2	1
33	2	11	3	3
34	17	10	10	10
35	4	19	18	23
36	6	3	0	1
37	4	11	5	9
38	3	4	4	4
39	4	9	1	2
40	5	17	7	12
41	11	2	2	2

respect to the first order are somewhat changed. The agreement between calculated and observed intensities is good at very low resolution (i.e. orders 1 to 5) and also for orders 20 and 21.

It should be said that, if the volume parameters are changed in some systematic way, then the effect on the predicted intensities can be striking. Table 4.3, for example, shows the effect of reducing the volumes of charged residues by 5% and increasing the volumes of hydrophobic residues by 5%. The biggest intensity changes are in the 6th and 11th orders, as might be expected from previous analysis of the sequence (Chapter 2).

b) Telopeptide residue translation (r_N and r_C)

The calculated intensities from three otherwise identical models with different r_N, r_C values are shown in Table 4.4. The details of the chosen model were:

2 α1 chains of 1052 residues
1 'pseudo α2' chain of 1024 residues (a pseudo α2 chain is defined as an α1 chain with 7 N-terminal and 21 C-terminal telopeptides removed, in accordance with the data on the length of the true calf skin α2 chain given by Fietzek et al. (1974)).
D = 234 residues.
3rd volume set (Chothia, 1975)

r_N, r_C = 0.524 corresponds to the residue translation in an α-helix (i.e. 1.5Å).

r_N, r_C = 1.224 corresponds to the residue translation in a β-sheet (3.5 Å).

Clearly the extended, β-sheet conformation in the telopeptides can be ruled out, as the agreement between calculated and observed intensities, even at very low resolution (i.e. orders 2 and 3), is so poor. The α-helical structure gives much better agreement for orders 1 to 10 and orders 20 to 40, though, as for the β-sheet structure, there is bad disagreement for orders 11 to 19.

TABLE 4.3 Effect of changing volumes of charged residues by -5% and hydrophobic residues by +5%

ORDER	OBSERVED INTENSITIES	BEFORE	AFTER
1	1000	1000	1000
2	27	9	10
3	109	106	163
4	11	10	4
5	54	87	100
6	17	61	7
7	21	48	39
8	6	6	1
9	71	8	8
10	19	7	0
11	20	64	23
12	31	7	3
13	2	47	41
14	4	34	24
15	4	33	24
16	5	14	7
17	11	19	17
18	4	4	5
19	6	23	9
20	34	17	5
21	24	36	19
22	11	2	0
23	0	2	3
24	0	6	8
25	14	16	9
26	5	24	18
27	11	21	15
28	1	0	0
29	9	0	0
30	16	15	14
31	1	0	1
32	2	1	1
33	2	3	1
34	17	10	7
35	4	23	13
36	6	1	1
37	4	9	2
38	3	4	2
39	4	2	0
40	5	12	12
41	11	2	3

TABLE 4.4 The effect of varying average residue translations in the N and C teleopeptides.

ORDER	OBSERVED INTENSITIES	$r_N = 0.524$ $r_C = 0.524$ (α-helix)	$r_N = 0.7$ $r_C = 0.7$	$r_N = 1.224$ $r_C = 1.224$ (β-sheet)
1	1000	1000	1000	1000
2	27	49	42	40
3	109	132	112	40
4	11	34	13	3
5	54	58	66	20
6	17	41	23	100
7	21	20	28	16
8	6	32	20	14
9	71	81	33	31
10	19	16	19	17
11	20	11	11	29
12	31	15	29	1
13	2	12	5	18
14	4	73	21	62
15	4	33	35	23
16	5	51	13	17
17	11	4	16	7
18	4	31	13	9
19	6	11	12	6
20	34	43	31	14
21	24	46	25	13
22	11	3	4	1
23	0	11	0	2
24	0	21	6	13
25	14	10	6	4
26	5	18	22	26
27	11	5	7	8
28	1	9	2	0
29	9	3	2	5
30	16	22	26	11
31	1	2	0	4
32	2	0	1	1
33	2	2	1	6
34	17	15	19	14
35	4	12	9	13
36	6	1	0	1
37	4	8	0	4
38	3	4	2	9
39	4	0	8	6
40	5	14	6	20
41	11	2	1	2

Table 4.4 shows that the calculated intensities of orders 20 to 41 are not very sensitive to the particular values of r_N and r_C. This is consistent with the idea that the telopeptide residue translation affects the low resolution features of the electron density (i.e. the gap/overlap step function) but does not alter the detailed electron density variation, which is dominated by the triple helical part of the structure.

A whole range of values for r_N and r_C were used to calculate predicted intensities. It seemed that the best overall agreement was with $r_N = 0.7$ and $r_C = 0.5$ (see column three of Table 4.5) though even then there is bad agreement for orders 7, 12, 14 and 15.

c) <u>The α2 sequence</u>

Not all of the α2 chain has so far been sequenced. A sequence has been compiled, however, which contains 40% of the rat/calf skin α1 sequence and 60% of the α2 (from Fietzek and Kuhn, 1975). As mentioned above, the α2 sequence contains 1024 residues. In the α2 regions of this composite α1/α2 sequence there are 227 non-identical amino acid substitutions compared to the corresponding regions of the α1 sequence. The composite α1/α2 sequence is henceforth referred to as the partial α2.

Table 4.5 shows the effect of using the known α2 residue substitutions on the calculated intensities from an otherwise identical model. The telopeptide residue translations are $r_N = 0.7$ and $r_C = 0.5$, and D = 234 residues. A disaccharide has been added to all three chains at position 87 of the helical region (see next paragraph). Clearly the effect of the amino acid substitution on the calculated intensities is very small. It would therefore be very difficult to use this technique to study sequence variation.

TABLE 4.5 The effect of using a partial α2 sequence

ORDER	OBSERVED INTENSITIES	Pseudo α2	Partial α2
1	1000	1000	1000
2	27	21	10
3	109	155	145
4	11	32	21
5	54	91	118
6	17	25	22
7	21	52	63
8	6	39	49
9	71	71	74
10	19	41	55
11	20	4	7
12	31	9	9
13	2	18	16
14	4	46	46
15	4	62	61
16	5	33	27
17	11	29	32
18	4	27	25
19	6	30	24
20	34	62	51
21	24	52	50
22	11	24	27
23	0	12	9
24	0	17	17
25	14	13	12
26	5	7	3
27	11	0	0
28	1	11	11
29	9	12	17
30	16	29	22
31	1	5	6
32	2	4	6
33	2	6	11
34	17	5	2
35	4	15	17
36	6	4	4
37	4	8	10
38	3	12	7
39	4	3	4
40	5	18	11
41	11	11	29

d) The disaccharide

As mentioned in Chapter 1, collagen is a glycoprotein. The carbohydrate is either galactose or glucosylgalactose covalently linked to hydroxylysine (Spiro, 1969). Rat tail tendon contains a very small amount of carbohydrate (0.4 units per α chain). In the rat skin α1 sequence there is only one carbohydrate, a glucosylgalactose at position 87 of the helical region (Butler, 1970).

Table 4.6 shows the effect of adding a glucosylgalactose (at position 87 in the helical region of all three chains) on the predicted meridional intensities from an otherwise identical model. The volume of the carbohydrate-hydroxylysine unit was estimated to be 524Å^3 (using the crystallographic data of Beevers and Hansen (1971)), hence giving a net 82.2 electrons above water background (compare with Table 4.1). Again the effect is to change magnitudes but not positions of the strong orders.

The details of the chosen model were:

2 α1 chains
1 partial α2 chain
$r_N = 0.7$
$r_C = 0.5$
$D = 234$ residues
3rd volume set (Chothia, 1975).

e) The D period

Table 4.7 shows the effect on the predicted intensities of varying the D stagger from 231 to 235 residues, on an otherwise identical model. The details of the chosen model were:

2 α1 chains
1 partial α2 chain
$r_N = 0.7$
$r_C = 0.5$
3rd volume set (Chothia, 1975).

TABLE 4.6 The effect of adding one disaccharide per α chain at position 87 of the helical region

ORDER	OBSERVED INTENSITIES	WITHOUT DISACCHARIDE	WITH DISACCHARIDE
1	1000	1000	1000
2	27	25	10
3	109	124	145
4	11	10	21
5	54	88	118
6	17	36	22
7	21	43	63
8	6	31	49
9	71	69	74
10	19	41	55
11	20	20	7
12	31	8	9
13	2	19	16
14	4	52	46
15	4	42	61
16	5	28	27
17	11	16	32
18	4	11	25
19	6	11	24
20	34	33	51
21	24	31	50
22	11	14	27
23	0	2	9
24	0	13	17
25	14	4	12
26	5	10	3
27	11	2	0
28	1	6	11
29	9	11	17
30	16	22	22
31	1	2	6
32	2	2	6
33	2	4	11
34	17	5	2
35	4	10	17
36	6	2	4
37	4	5	10
38	3	2	7
39	4	2	4
40	5	6	11
41	11	19	29

TABLE 4.7 The effect of varying the D stagger

ORDER	OBSERVED INTENSITIES	D = 231	D = 232	D = 233	D = 234	D =235
1	1000	1000	1000	1000	1000	1000
2	27	3	5	12	25	44
3	109	145	145	137	124	107
4	11	7	1	6	19	37
5	54	123	127	114	88	57
6	17	10	13	24	36	43
7	21	157	133	87	43	24
8	6	2	15	28	31	22
9	71	49	54	60	69	74
10	19	36	60	62	41	19
11	20	9	6	12	20	19
12	31	2	10	12	8	19
13	2	8	8	17	19	10
14	4	22	18	22	52	74
15	4	4	7	36	42	18
16	5	54	44	30	28	18
17	11	2	12	24	16	3
18	4	1	1	2	11	22
19	6	8	2	9	11	9
20	34	5	6	38	33	2
21	24	17	0	19	31	10
22	11	8	3	10	14	7
23	0	6	22	12	2	3
24	0	19	2	6	13	23
25	14	8	17	23	4	1
26	5	1	2	1	10	13
27	11	2	2	1	2	20
28	1	12	4	12	6	17
29	9	1	4	17	11	5
30	16	5	8	6	22	6
31	1	8	39	1	2	9
32	2	2	6	4	2	1
33	2	1	10	46	4	4
34	17	10	12	13	5	10
35	4	1	6	10	10	20
36	6	1	2	2	2	0
37	4	6	5	9	6	3
38	3	1	1	3	2	7
39	4	2	2	5	2	4
40	5	10	1	0	6	13
41	11	2	1	4	19	9

Clearly the intensities are very sensitive to the positions where the molecular transform is sampled by the D periodic lattice. This indicates that the molecular transform is a rapidly varying function. Note in particular that the 20th and 21st orders are both strong only for D staggers of 233 or 234 residues. This is a feature of all the models previously described (where D was taken to be 234 residues). In particular, the strength of orders 20 and 21 was independent of the chosen telopeptide residue translations. The previous paragraphs have shown that the two aspects of the electron density models, to which the calculated intensities are most sensitive, are the structure of the telopeptides and the D stagger. The dependence of the strength of the 20th and 21st orders on the D stagger, and their apparent independence of the telopeptide structure, implies that these two aspects can be separated in these higher resolution parts of the transform. Thus, if the true structure differs from the models solely in the structure of the telopeptides, then the agreement between calculated and observed 20th and 21st orders implies that the D stagger is probably between 232.5 and 234.5 residues.

The electron density model whose calculated meridional intensities are given in column four of Table 4.6 is shown in Fig. 4.1. The model was calculated using an inverse Fourier transform on the calculated amplitude and phase data, after multiplying the amplitudes by an artificial temperature factor to eliminate termination errors. This factor was such as to reduce the amplitude of order 41 to 1% of its original value. Fig. 4.1, therefore, shows the electron density model to a resolution of ~16Å.

It would, of course, be preferable to study the differences between the actual and model electron densities in real space,

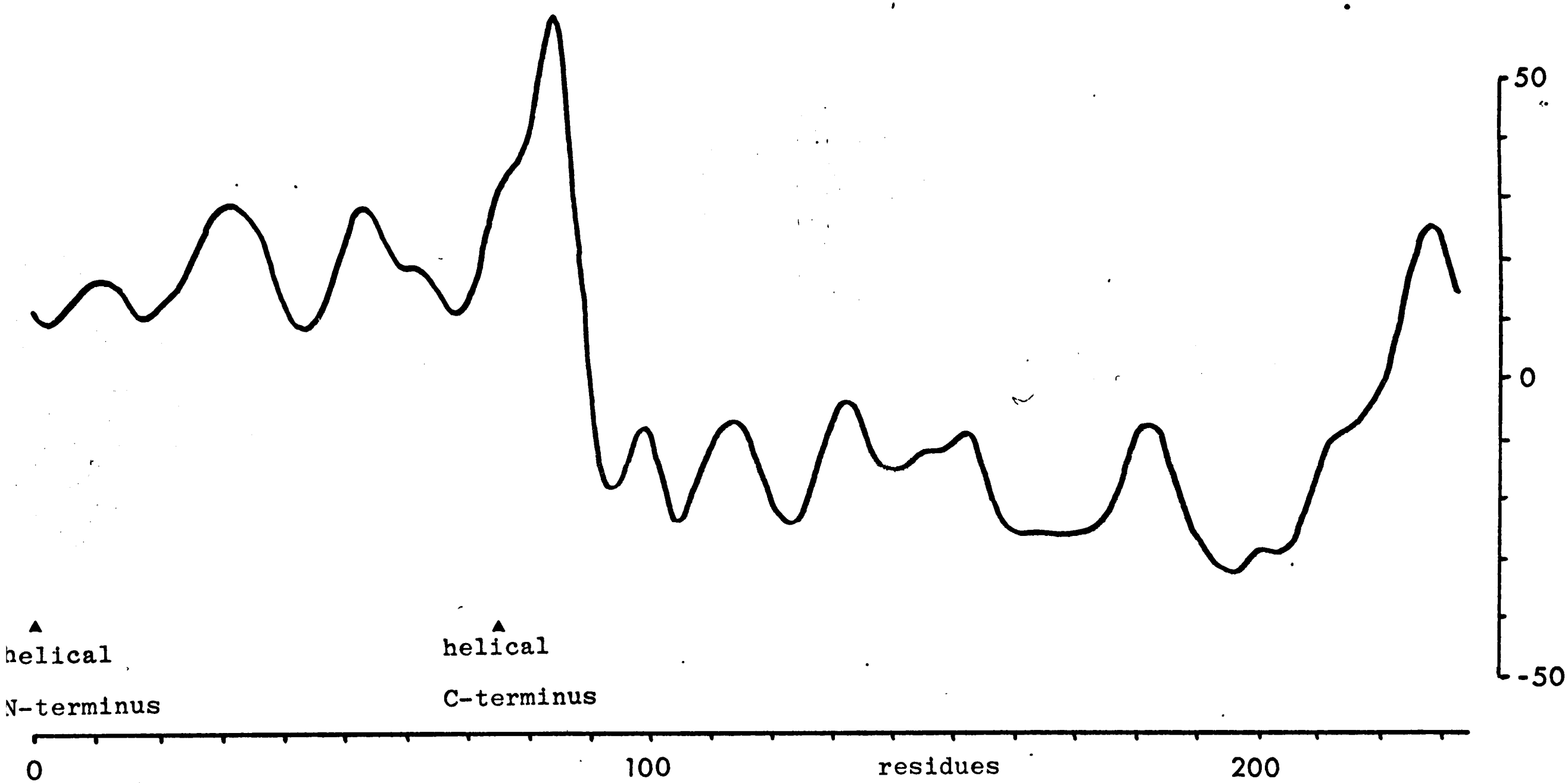


Fig. 4.1 Electron density variation (corrected for water background) for model C. The predicted meridional intensities are given in column four of Table 4.6.

rather than by comparison of meridional diffraction intensities. In principle this is not possible without knowledge of the true phases, though, if the differences are small, the following approximation can be made.

iii) Difference Fouriers

The electron density $\sigma(z)$ is related to its Fourier transform $F(Z)$ by:

$$\sigma(z) = C\int_{-\infty}^{\infty} F(Z)\, e^{-2\pi i z Z}\, dZ$$

$F(Z)$ contains both amplitude and phase information and hence may be rewritten as:

$$F(Z) = A(Z)e^{i\theta(Z)}$$

Hence

$$\sigma(z) = C\int_{-\infty}^{\infty} A(Z)e^{i(\theta(Z)-2\pi zZ)}\, dz$$

Let $\sigma(z)$ represent the model electron density, the true electron density is $\sigma'(z)$.

$$\sigma'(z) = C'\int_{-\infty}^{\infty} A'(Z)E^{i(\theta'(z)-2\pi zZ)}\, dz$$

In the above equations C and C' are constants.

The difference in electron density between the actual and model structure is $\Delta\sigma(z)$.

$$\Delta\sigma(z) = \sigma'(z)-\sigma(z)$$

$$= \int_{-\infty}^{\infty}\left[C'A'(Z)e^{i(\theta'(Z)-2\pi zZ)} - CA(Z)e^{i(\theta CZ)-2\pi zZ)}\right] dZ$$

$$= C\int_{-\infty}^{\infty}\left[\frac{C'}{C}A'(Z)e^{i(\theta'(Z)-2\pi zZ)} - A(Z)e^{i(\theta(Z)-2\pi zZ)}\right] dZ$$

The important unknowns in this equation are $\frac{C'}{C} = S$, a scale factor, and $\theta'(Z)$, the phases for the actual structure. If $\Delta\sigma(z)$ is not too large however the approximation $\theta(Z) = \theta'(Z)$ can be

used. This is the central assumption of the difference Fourier method, an approach to structural analysis which has been frequently used in X-ray crystallography (Blundell and Johnson, 1975). With this assumption the above equation can be rewritten,

$$\Delta\sigma(z) = C\int_{-\infty}^{\infty} \left[SA'(Z)-A(Z)\right] \; e^{i(\theta(Z)-2\pi zZ)} dZ$$

S is of course an important quantity. In the difference Fouriers illustrated below S is usually such as to equate calculated and observed first order amplitudes. This is justified below.

The first order amplitude and phase is dominated by the Hodge-Petruska gap/overlap step function. This is illustrated in Fig. 4.2, which shows the calculated amplitude variation in the unsampled molecular transform of three α1 chains with $r_N = 1$ and $r_C = 1$. The full curve shows the amplitude for an equivalent model in which all the residues were given an average scattering length. Such a model would produce a perfect step function when assembled with a D stagger, unlike the electron density in, for example, Fig. 4.1. The difference in amplitude, at the first order sampling position of the transforms (i.e. 1/D), is, however, less than 10%. This implies that if the actual structure and the model are similar in molecular length and total content then the assumption of equal first orders would not be unreasonable. In other words, local variations in electron density along the molecular length do not greatly affect the amplitude of the first order. Finally, in all the models involving variation of the telopeptide residue translations (Table 4.4) the first order amplitude varies by no more than 5% from the mean, even though there are greater changes in the higher orders. The amplitude of the first order is therefore

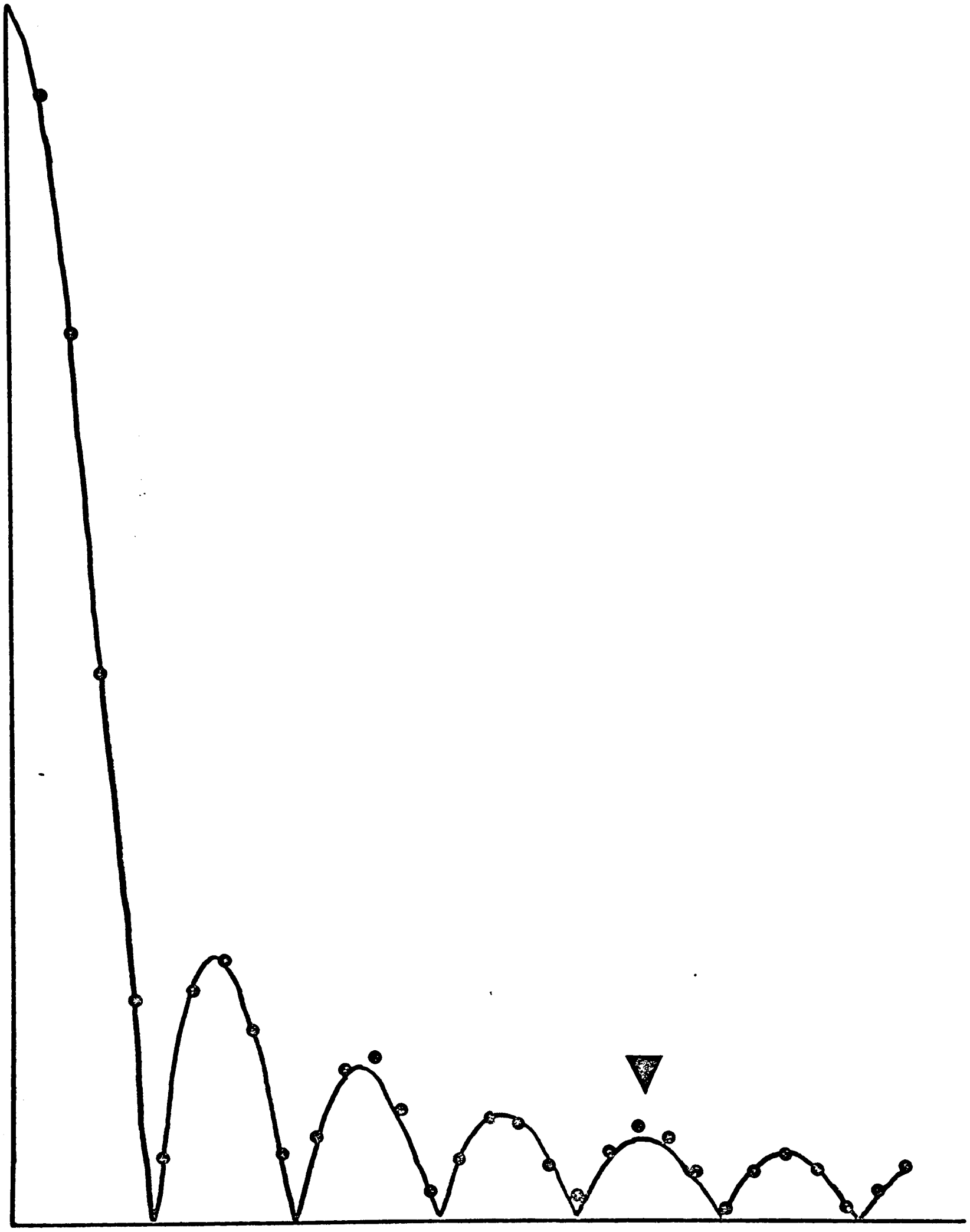


Fig. 4.2 The amplitude variation in the unsampled Fourier transform of a typical electron density model (●), compared with a model in which each residue is associated with an average scattering length. The first order sampling position is indicated.

almost independent of telopeptide conformation. The facts that the step function, which is common to all models, dominates the first order amplitude, and that this amplitude is unaffected by details of telopeptide structure, provide a justification for equating calculated and observed first order amplitudes in the difference Fouriers.

The above justification for the choice of scaling factor would not be valid if there is a contribution to the diffraction in the actual structure which is not present in the model (e.g. proteoglycans). The effect of such extra material is unpredictable.

Fig. 4.3 shows difference Fouriers obtained using observed amplitudes and the calculated amplitudes and phases from one model with two values of S, the scaling factor. In Fig. 4.3(a) the first order amplitudes were equated, but in Fig. 4.3(b) the observed amplitude was 90% of the calculated amplitude. The detailed features of the difference are retained, though, as expected, there is a slight modulation corresponding to a wavelength of D in Fig. 4.3(b). It is important to be aware of this slight uncertainty of the scaling factor in the following results.

a) Difference Fourier A

Fig. 4.4(a) shows the difference Fourier corresponding to the electron density model (A) whose predicted intensities are given in column 5 of Table 4.2 (i.e. three α1 chains, $r_N = r_C = 1$, D = 234 residues, 3rd volume set). As in all subsequent difference Fouriers the diagram shows the 'actual' minus the model electron density. Also, the calculated and observed first order amplitudes were equated and the difference amplitudes were multiplied by an artificial temperature factor.

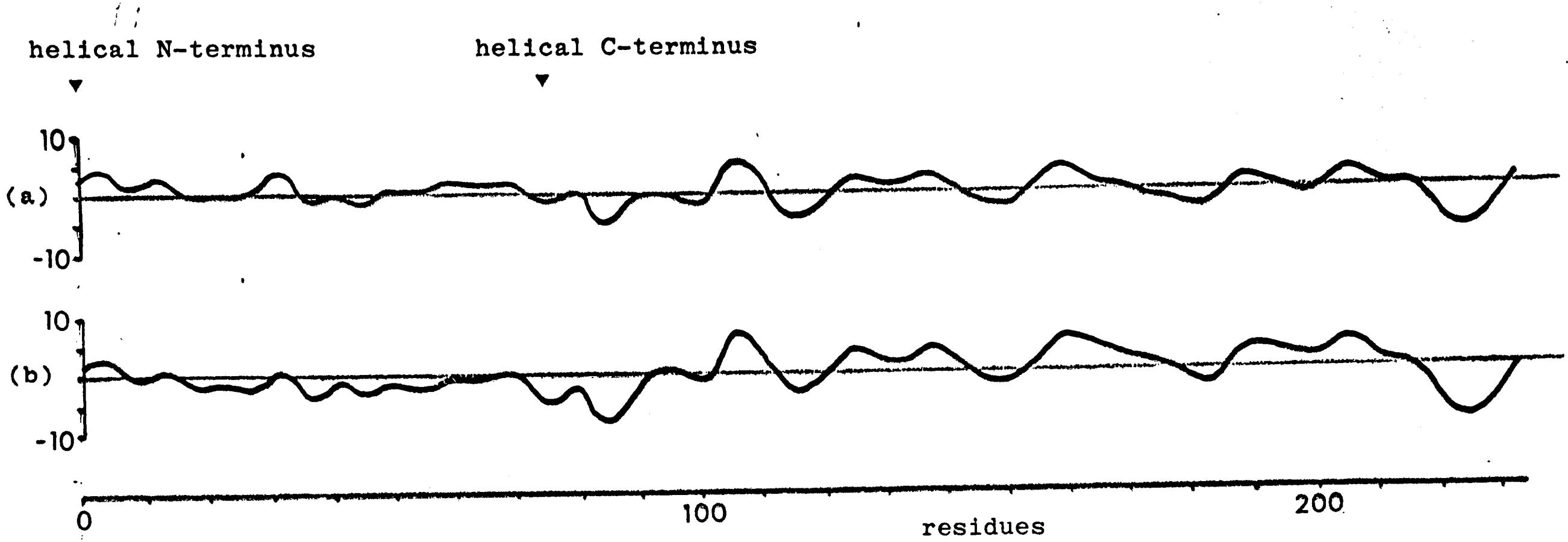


Fig. 4.3 Typical difference Fourier syntheses (i.e. 'actual' less expected electron density) using different scale factors (s).

(a) Observed and calculated first order amplitudes equated (i.e. s = 1).

(b) Observed first order amplitude = 0.9 x calculated first order amplitude (i.e. s = 0.9).

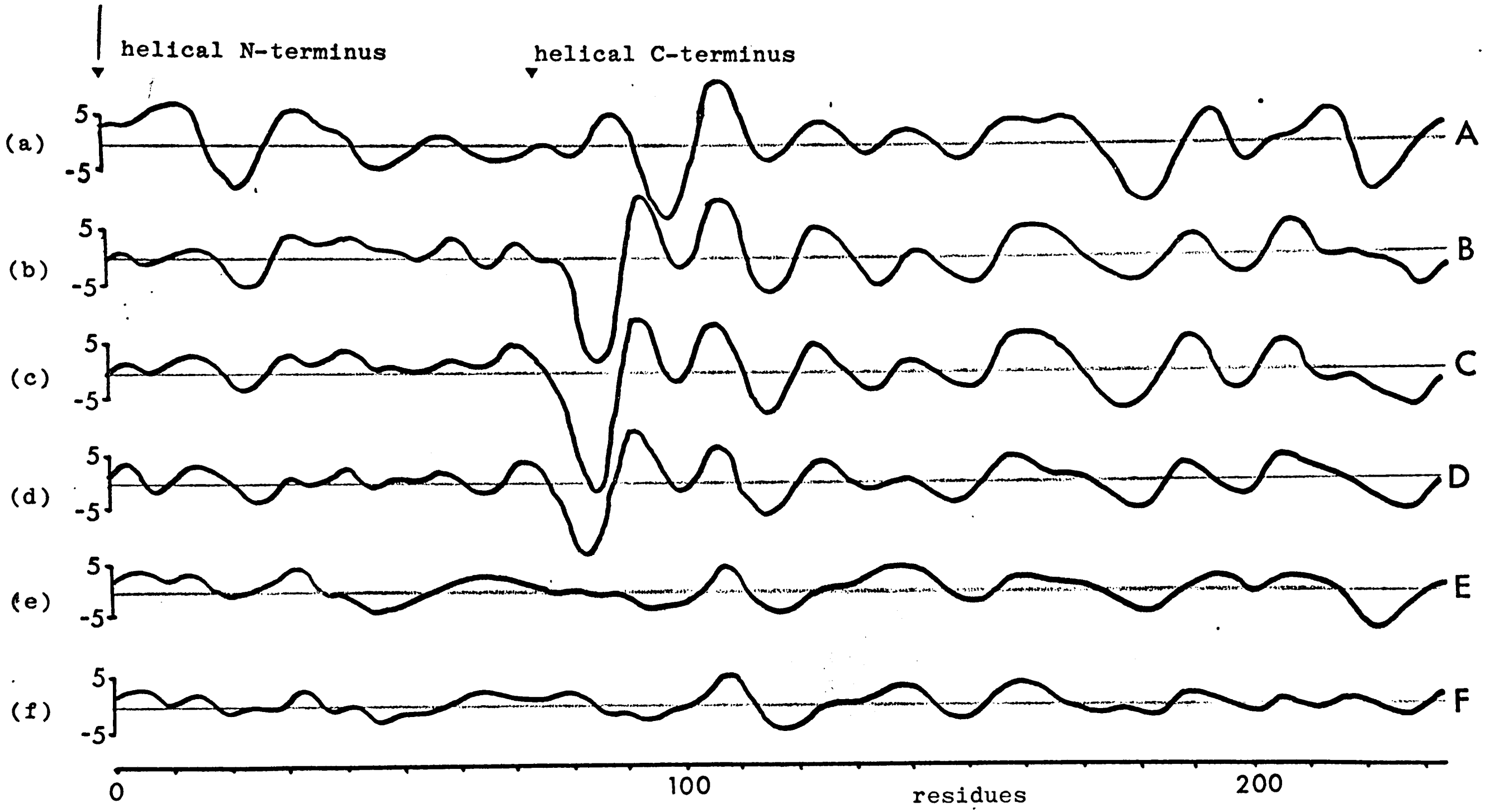


Fig. 4.4 Difference Fourier synthesis (i.e. 'actual' less expected electron density) for six models (see text).

to eliminate the Fourier termination errors.

Fig. 4.4(a) indicates that model A has too much density near positions, 25, 100, 180 and 220 and not enough density near positions 10, 30, 90, 105, 160, 190 and 210.

b) Difference Fourier B

Fig. 4.4(b) shows the difference Fourier corresponding to column 3 of Table 4.5 (i.e. 2 α1 chains, 1 pseudo α2 chain, $r_N = 0.7$, $r_c = 0.5$, D = 234, 3rd volume set , added disaccharide). There is now too much density in the model near position 85 and not enough near position 90. This is consistent with three changes that were made in the region of the C-terminal telopeptide in going from model A to model B. These charges are:

(1) addition of a disaccharide on all three chains at position 87.

(2) removal of 21 residues due to modifying one α1 chain to a pseudo 2 chain.

(3) relative compression ($r_c = 0.5$) of the remaining C-terminal telopeptide.

The shift in the negative peak from position 220 in Fig. 4.4(a) to position 230 in Fig. 4.5(b) is also consistent with the loss of 7 N-terminal telopeptide residues in going from α1 to pseudo α2 and the relative compression ($r_N = 0.7$) of the remaining telopeptide residues.

The difference Fouriers for models A and B are not identical in the non-telopeptide regions, even though the models themselves are identical. This apparent contradiction is due to the assumption that actual and model phases are equal (see later).

c) Difference Fourier C

Fig. 4.4(c) shows the difference Fourier for model C. This model corresponds to the electron density of Fig. 4.1, and to the predicted intensities of column four of Table 4.6. The model differs from model B in the replacement of a pseudo α2 by a partial α2. Clearly there is little effect on the difference Fourier.

d) Difference Fourier D

Model D, obtained from model C after removing the disaccharides, and hence corresponding to column 3 of Table 4.6, gives the difference Fourier shown in Fig. 4.4(d). The main change is the reduced height of the negative peak at position 87, as expected.

e) Difference Fouriers E and F

The main difference peak in Fig. 4.4(d) occurs in the middle of the C-terminal telopeptide region near position 85. The fact that this is one of the most uncertain parts of the axially projected structure, and hence would be expected to show differences against the true structure, suggests that the difference Fourier approach is valid for this problem.

In model D the C-terminal telopeptide occupies the region 75 to 88. The difference Fourier indicates that there is too much density in the model near position 85 and too little density near positions 70 and 90. All three peaks can be removed by withdrawing the restriction that the telopeptide residue translation must be constant. The C-terminal telopeptide can then be given an axially projected structure in which there is a low density region near position 85 with high density regions near positions 70 and 90. An electron density model is then

produced whose difference Fourier is shown in Fig. 4.4(e). The positions of the residues in the C-terminal telopeptide are given in Table 4.8, and the predicted meridional intensities are shown in Table 4.9. The overall agreement between calculated and observed intensities is now very good for this model, particularly in the region of orders 11 to 19, where previous models showed very poor agreement.

The main peak in the difference Fourier for model E is in the N-terminal telopeptide region. The model has too much density near position 220. Unlike the C-terminal negative peak in Fig. 4.4(d) however, this N-terminal negative peak is not flanked by positive peaks. There is therefore no way of redistributing the N-terminal telopeptide residues to remove this peak. The peak can be removed, however, if four residues are eliminated from the N-terminus of each $\alpha 1$ chain. The difference Fourier for this model (F) is shown in Fig. 4.4(f), and its predicted meridional intensities are given in Table 4.9. There is unfortunately no chemical evidence for the loss of four residues at the $\alpha 1$ N-terminus for rat tail tendon, so model E must be regarded as the best electron density model which is consistent with the chemical data.

The electron density for model E is shown in Fig. 4.5. The positions of the 25 C-terminal telopeptides are given in Table 4.8; the 4 C-terminal $\alpha 2$ telopeptides are defined by $r_c = 0.5$; the N-terminal telopeptides are defined by $r_N = 0.9$; the D period is 233.5; the partial $\alpha 2$ chain and 3rd volume set were used and there was no added disaccharide. The parameters for model F were identical except that 4 residues were eliminated from the $\alpha 1$ N-terminus.

Before discussing this electron density model in the light of other attempts to solve the axially projected structure it is

TABLE 4.8 The positions* of the α1 C-terminal telopeptide residues in model E

Residue	Position 1st chain	Position 2nd chain
SER	77	78
GLY	78	79
GLY	78	79
TYR	79	80
ASP	79	80
LEU	80	81
SER	81	82
PHE	82	83
LEU	83	84
PRO	84	85
GLN	85	86
PRO	85	86
PRO	86	87
GLN	86	87
GLN	87	88
GLX	88	89
LYS	89	90
ALA	90	91
HIS	91	92
ASP	92	93
GLY	92	93
GLY	93	94
ARG	93	94
TYR	94	95
TYR	95	96

* in residues from the helical N-terminus of the first α1 chain

TABLE 4.9 Effect of removing 4 residues at N terminus of α1

Order	Observed Intensities	Model E	Model F (i.e. Model E less 4 residues at the N-terminus of the α1 chains)
1	1000	1000	1000
2	27	15	25
3	109	104	89
4	11	2	8
5	54	85	66
6	17	21	12
7	21	60	40
8	6	7	16
9	71	32	34
10	19	18	13
11	20	32	22
12	31	24	37
13	2	12	6
14	4	11	11
15	4	11	19
16	5	5	7
17	11	13	21
18	4	10	4
19	6	8	7
20	34	20	30
21	24	27	17
22	11	4	9
23	0	2	1
24	0	12	7
25	14	6	3
26	5	22	21
27	11	5	2
28	1	10	6
29	9	27	25
30	16	30	39
31	1	14	8
32	2	2	5
33	2	21	15
34	17	19	20
35	4	16	13
36	6	4	2
37	4	4	3
38	3	1	1
39	4	5	7
40	5	4	2
41	11	18	18

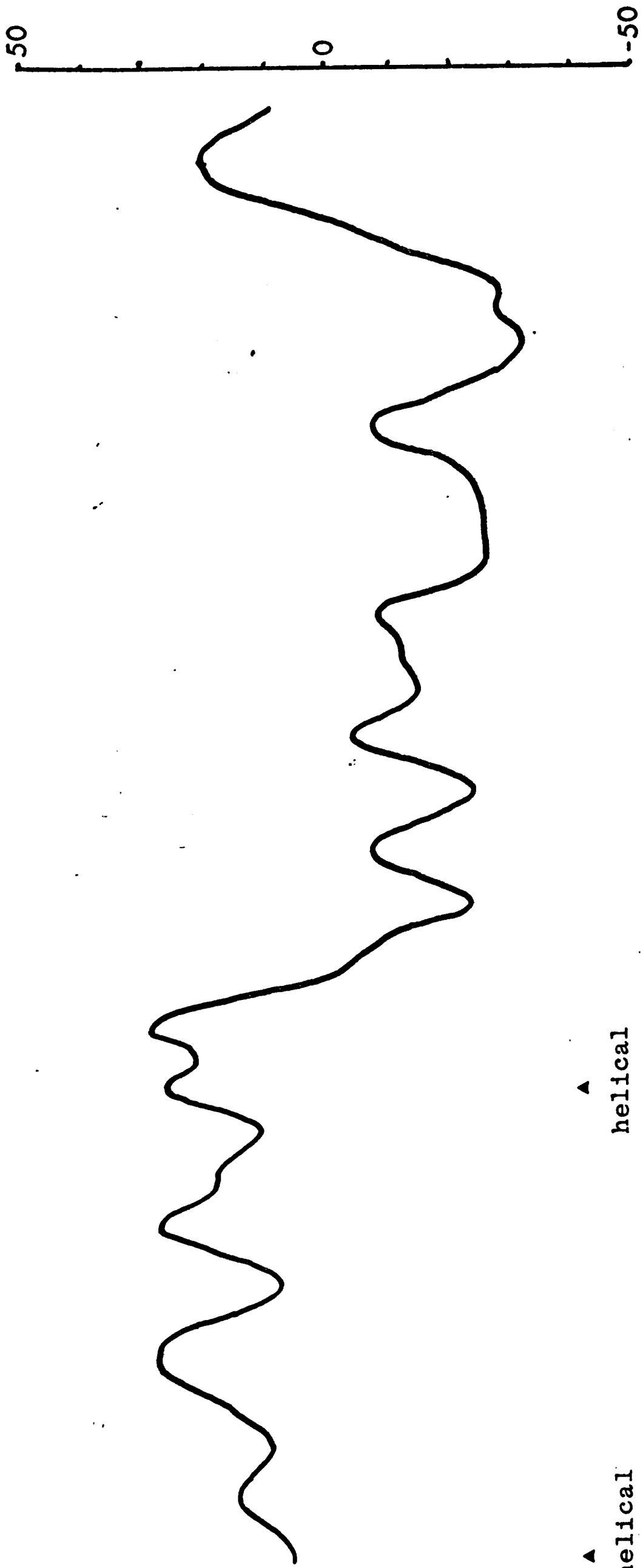


Fig. 4.5 Electron density variation (corrected for water background) for model E. The predicted meridional intensities are given in column three of Table 4.9.

necessary to mention briefly an attempt to refine an electron density model based on the sequence.

iv) Refinement

A Bayesian least squares refinement programme has been written by French (1975). This programme was used to optimise the agreement between the 41 calculated and observed meridional intensities by making slight adjustments to 26 parameters (23 residue volumes, r_N, r_C and D). The uncertainty in each parameter was used by the programme to guide the path of refinement. After eleven cycles the programme converged to give the calculated intensities shown in Table 4.10. The initial intensities were from model C (i.e. Fig. 4.1, column four of Table 4.6). The parameters before and after refinement are given in Table 4.11.

The excellent agreement between observed and refined calculated intensities would represent an interpretible solution to the axially projected electron density if the refined parameters are acceptable. Unfortunately, for some parameters, this is not the case. The refined glycine volume, for example, is 25% less than the volume of the residue in a crystal of the amino acid (Chothia, 1975). The refined volumes for Asx and Glx would be expected to be similar to either Asp, Asn or Glu, Gln respectively, but this is not observed. Finally the refined volume for hydroxylysine is almost 50Å^3 less than that of lysine; this again seems unacceptable. The conclusion, therefore, is that some of the parameters used to describe the refined electron density model are not physically meaningful. They are therefore inappropriate to describe the structure. Nevertheless, it is possible that the refined electron density may still be correct if this structure can be formed by changing

TABLE 4.10 Intensities before and after refinement

ORDER	OBSERVED INTENSITIES	BEFORE REFINEMENT	AFTER REFINEMENT
1	1000	1000	1000
2	27	10	37
3	109	145	97
4	11	21	6
5	54	118	58
6	17	22	21
7	21	63	18
8	6	49	8
9	71	74	48
10	19	55	17
11	20	7	24
12	31	9	28
13	2	16	2
14	4	46	6
15	4	61	12
16	5	27	7
17	11	32	19
18	4	25	8
19	6	24	7
20	34	51	37
21	24	50	25
22	11	27	15
23	0	9	3
24	0	17	3
25	14	12	12
26	5	3	6
27	11	0	8
28	1	11	2
29	9	17	11
30	16	22	17
31	1	6	5
32	2	6	4
33	2	11	4
34	17	2	19
35	4	17	4
36	6	4	9
37	4	10	4
38	3	7	3
39	4	4	4
40	5	11	10
41	11	29	7

TABLE 4.11 The effect of refinement

AMINO ACID RESIDUE	VOLUME BEFORE REFINEMENT	VOLUME AFTER REFINEMENT
GLY	66.4	45.4
ALA	91.5	78.7
VAL	141.7	139.4
LEU	167.9	182.9
ILE	168.8	179.8
PHE	203.4	197.0
PRO	129.3	119.2
HYP	133.1	127.4
SER	99.1	86.4
THR	122.1	123.3
MET	170.8	140.6
ASN	135.2	127.7
GLN	161.1	150.8
ASP	124.5	130.0
GLU	155.1	140.0
TYR	203.6	194.1
HIS	167.3	167.5
LYS	171.3	176.7
HYL	175.0	127.7
ARG	181.2	161.2
ASX	129.9	186.7
GLX	158.1	136.3
Hyl + disaccharide	524.0	794.0
D STAGGER	234	233.5
r_N	0.7	0.762
r_C	0.5	0.670

some other aspects of the original structure in a realistic way. To examine this possibility it is necessary to study the electron density before and after refinement. This is shown in Fig. 4.6. There are two main changes:

(1) The refined step height is larger. This is due to the increased weighting of the glycine residues.

(2) There is a localised change in the region of the C-terminal telopeptide.

The increased step height is not important. This could be obtained from the original electron density by multiplying by some constant. This would not change the relative meridional intensities.

The localised change in the C-terminal telopeptide region after refinement corresponds exactly to the large peaks in the difference Fourier for the initial model (model C, Fig. 4.4(c)). Thus, in the same way that the difference Fourier peaks were eliminated, the refined electron density can be produced from model C by suitable choice of C-terminal telopeptide structure and removal of the disaccharides (resulting in model E). It is encouraging that two quite independent techniques, reciprocal space refinement and real space difference Fouriers, both indicate that the initial electron density model differs from the actual structure mainly in the conformation of the C-terminal telopeptide.

v) Discussion

Model E gives a set of phases which may be combined with the observed amplitudes to give a 'true' electron density distribution for an axially projected wet collagen fibril. The

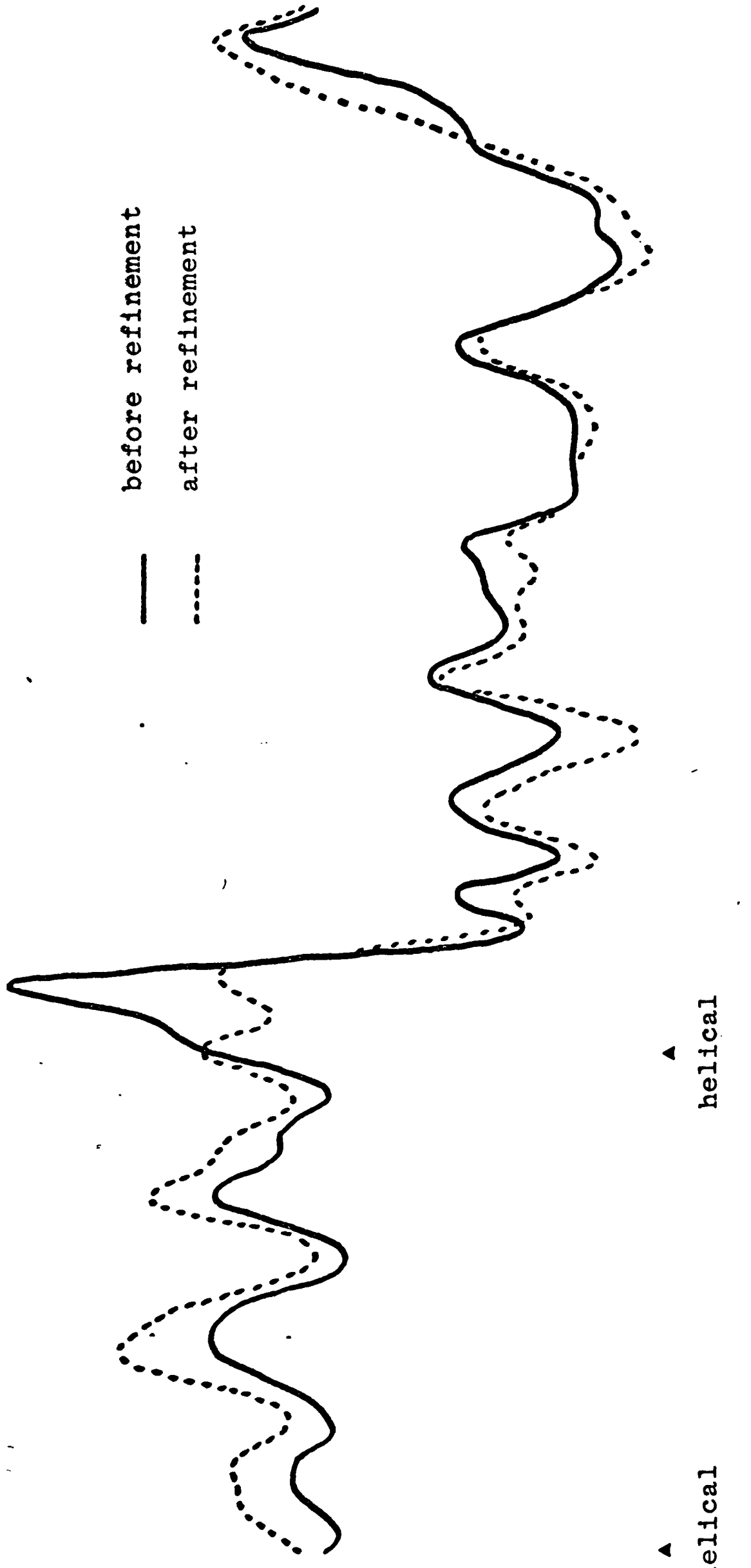


Fig. 4.6 Electron density variation before (model C) and after refinement (see text).

criteria for the validity of these model phases are:

(1) physical and chemical plausibility of the model.

(2) agreement of calculated and observed intensities or, equivalently, a zero difference Fourier throughout the D period.

Model E (Fig. 4.5) satisfies the first criterion. The difference Fourier, however, is not zero; though the maximum differences (Fig. 4.4(e)) are only 10% of the step height. (The electron density maps and difference Fouriers are on the same scale). If these differences could be accounted for in physically meaningful terms, and their rectification causes negligible change in the phases, then the phases from model E should be valid for the true structure (by the above criteria). Evidence that the phases would be essentially unchanged is given below.

Consider models E and F, differing only in the presence and absence of 4 α1 N-terminal telopeptide residue. Their difference Fouriers with the observed amplitudes are given by

$$\Delta\sigma_E(z) = \int \left[A_E(Z)-A_o(Z)\right] \; e^{i(\theta_E(Z)-2\pi zZ)} dZ$$

$$\Delta\sigma_F(z) = \int \left[A_F(Z)-A_o(Z)\right] \; e^{i(\theta_F(Z)-2\pi zZ)} dZ$$

The difference between these difference Fouriers is given by

$$\Delta\sigma_E(z)-\Delta\sigma_F(z) = \int A_o(Z) \left[e^{i\theta_F(Z)} - e^{i\theta_E(Z)}\right] e^{-i2\pi zZ} dZ$$

$$+ \int \left[A_E(Z)e^{i\theta_E(Z)} - A_F(Z)e^{i\theta_F(Z)}\right] e^{-i2\pi zZ} dZ$$

The second term in this equation is merely the difference between models E and F, i.e. the loss of 4 α1 N-terminal telopeptides. The first term arises from the assumption that real and model phases are identical. The fact that the difference between the

difference Fouriers for models E and F does indeed correspond to simply the difference between the models (see Fig. 4.4(e) and (f)), indicates that the first term in the above equation is negligible, and therefore $\theta_F(Z) \simeq \theta_E(Z)$. Thus the rectification of the difference in the N-terminal telopeptide region causes a negligible change in the phases. (This is in fact observed, see Table 4.12). Unfortunately chemical evidence does not support the loss of 4 α1 N-terminal telopeptides; nevertheless if this difference can be rectified in some physico-chemically acceptable manner then the above argument is still appropriate. The implication of the argument is that, if the other small differences throughout the rest of the D period can be explained in acceptable terms, then the effect on the calculated phases should be negligible.

There are a number of ways in which the difference Fourier peaks could be eliminated. One possibility is that inhomogeneities in the triple helical residue translation may cause a sufficient redistribution in electron density. Furthermore amino acids of the same type may exclude different amounts of water depending on their location in the fibril. Small quantities of proteoglycans or inorganic ions may be specifically bound. Finally the assumption of a constant 'residue scattering factor' (Section 3(i)) is probably only valid at low resolution.

It must be stressed that the above criterion of a physically plausible structure with the observed meridional intensities is not a certain way of determining the true phases. Several models may satisfy this criterion. Nevertheless the electron density obtained using observed amplitudes and phases from model E is shown in Fig. 4.7. This may be compared with the results of Tomlin and Worthington (1956), Ellis and McGavin (1970) and

TABLE 4.12 Phases (^{o}) for models E and F referred to an origin 44.5 residues from the helical N-terminus

ORDER	MODEL E	MODEL F
1	5	7
2	320	332
3	156	163
4	214	204
5	326	332
6	232	245
7	103	101
8	192	198
9	209	195
10	271	289
11	169	180
12	187	194
13	341	355
14	153	129
15	283	294
16	321	292
17	104	115
18	216	202
19	223	248
20	279	273
21	2	7
22	100	91
23	202	247
24	112	123
25	331	315
26	193	181
27	223	246
28	52	36
29	83	93
30	141	135
31	193	192
32	322	310
33	50	55
34	36	46
35	167	159
36	127	129
37	98	116
38	350	21
39	11	18
40	126	122
41	318	311

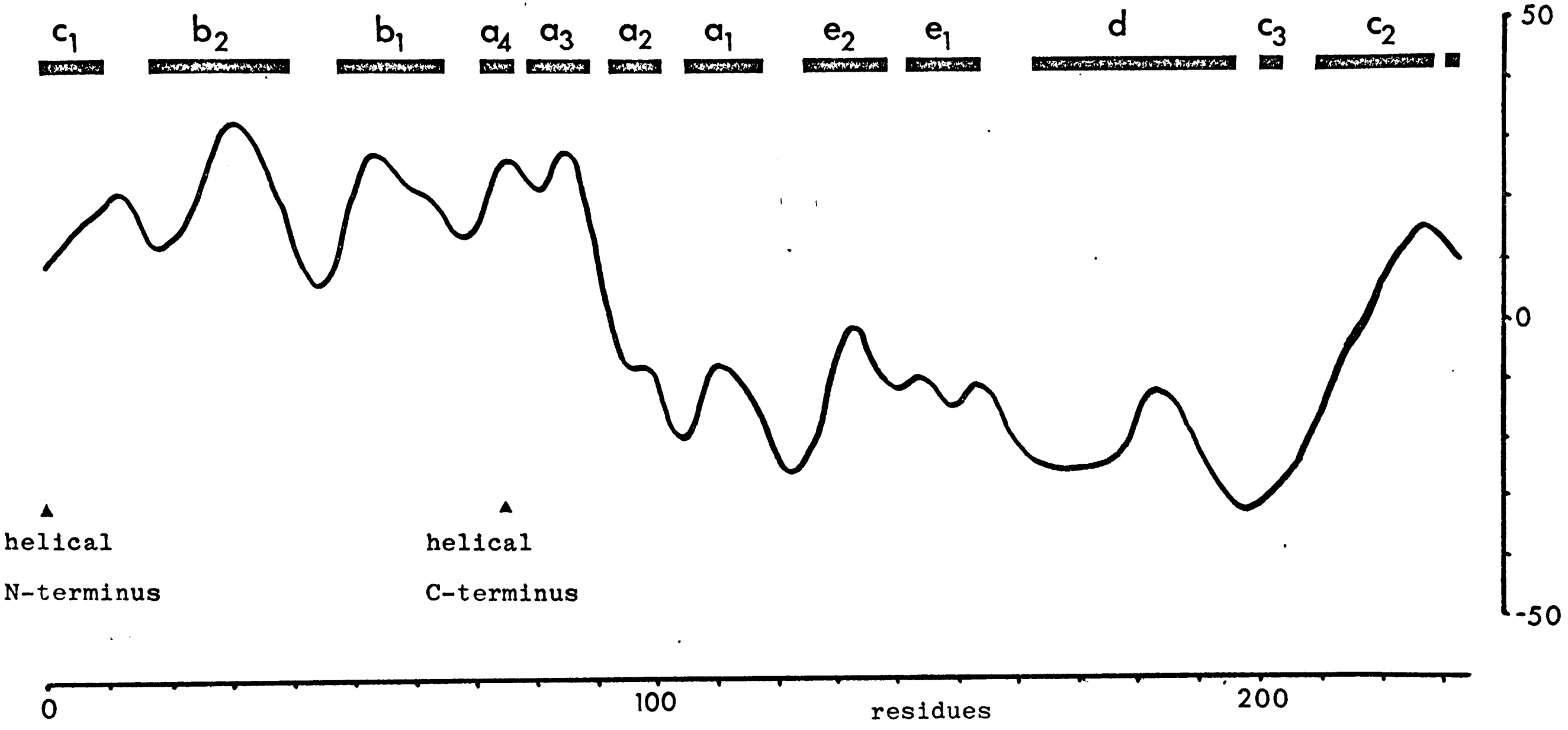


Fig. 4.7 Fourier synthesis of the axially projected electron density of a collagen fibril from moist rat tail tendon. The observed amplitudes were used with the calculated phases from model E. The positions of the positive staining bands are indicated.

Chandross and Bear (1973).

As mentioned earlier, Bear and Morgan (1957) and Chandross and Bear (1973) assumed that the positively stained bands of collagen fibrils which can be observed with the electron microscope, correspond to intrinsically electron dense regions. Fig. 4.7 shows the positions of these bands. Comparison with the present electron density model shows that the assumption is valid for the wet case; each peak in the background corrected electron density coincides with the centre of a positively stainable band. The model of Chandross and Bear (1973) differs from the present model mainly because they omitted bands 'a', and 'd'. The present model is an improvement on that of Chandross and Bear (1973), as the electron density can now be related to the amino acid sequence and molecular structure.

The so-called R factor (defined as $(\Sigma_n ||F_{no}| - |F_{nc}|| / \Sigma_n |F_{no}|)$, where F_{no} and F_{nc} are the observed and calculated structure factor amplitudes respectively) is 0.30 for model E. This compares with 0.21 for Tomlin and Worthington's (1956) model and 0.13 for Chandross and Bear's (1973) model. Direct comparison is not very useful however as the present work was based on 41 orders rather than 25 for Tomlin and Worthington and 15 for Chandross and Bear. Furthermore, unlike White's (1975) data, the intensities of Tomlin and Worthington are all very weak after the 12th order.

The step width in the present model (Fig. 4.7), defined using the intersections of the overlap region with the horizontal axis, is 0.47, in excellent agreement with the simple step function model of Tomlin and Worthington (also 0.47). The step width in Chandross and Bear's model is 0.455. It was shown

earlier that a step width of 0.53 (Kaesberg and Shurmann, 1953), which can be obtained by extending the telopeptide residue translation, gives very poor agreement with the observed intensities (Table 4.4).

The extent to which the present electron density model is centro-symmetric can be ascertained from the calculated phases. It is important to choose an appropriate phase origin. In Table 4.13 the calculated phases, referred to an origin in the exact centre of the overlap region, of the Tomlin and Worthington (1956), Chandross and Bear (1973) and present models are compared. (The Chandross and Bear model was also inverted through the phase origin to make the polarity of the band pattern compatible with the present model). All the phases are in approximate agreement for the first three orders. They are therefore close to either 0^{o} or 180^{o}. From the fourth order onwards most of the phases from model E do not correspond to a centro-symmetric structure; this is also true of the Chandross and Bear model from the fifth order onwards. Tomlin and Worthington's (1956) assumption of centro-symmetry is, therefore, only compatible to an electron density model based on the sequence (model E) to a resolution of (670/3≃) 220Å.

No attempt has been made in the present study to examine the dry axially projected electron density of collagen fibrils. This is because there is evidence of a considerable loss of order on drying (see Bear (1952)), which probably causes a deformation in the axially projected structure. The micro-densitometer traces of unstained fibrils in the electron micro-scope (Burge and Randall, 1955) bear no resemblance to an approximately equal length gap and overlap region, which would be expected for a model based on the sequence (i.e. an

TABLE 4.13 A comparison of the phases for three axially projected electron density models for tendon. Phase origins at the exact centre of the overlap regions.

	PHASES (degrees)			Amplitudes
ORDER	Tomlin & Worthington (1956)	Chandross & Bear (1973)	Model E	Model E
1	0	10	15	11.773
2	0	314	335	1.439
3	180	140	179	3.795
4	180	356	246	0.524
5	0	67	4	3.434
6	0	227	278	1.678
7	180	192	157	2.885
8	180	287	253	1.003
9	0	348	278	2.094
10	0	271	348	1.588
11	180	266	254	2.111
12	180	243	279	1.844
13	0	11	81	1.304
14	0	63	261	1.228
15	180	240	38	1.208
16	180	-	84	0.801
17	180	-	235	1.354
18	0	-	354	1.173
19	0	-	9	1.071
20	180	-	73	1.642
21	180	-	163	1.919
22	0	-	269	0.712
23	0	-	119	0.564
24	180	-	297	1.286
25	180	-	163	0.929
26	-	-	33	1.763
27	-	-	71	0.827
28	-	-	267	1.142
29	-	-	306	1.915
30	-	-	11	2.042
31	-	-	71	1.385
32	-	-	208	0.537
33	-	-	284	1.697
34	-	-	297	1.619
35	-	-	76	1.503
36	-	-	44	0.776
37	-	-	23	0.778
38	-	-	282	0.433
39	-	-	311	0.816
40	-	-	74	0.733
41	-	-	274	1.552

approximately 234 residue D stagger and a constant triple helical residue translation). When fibrils are negatively or positively stained, however, the electron micrographs are entirely consistent with these assumptions (Chapter 2). It may be that staining preserves the structure in some way. Finally, the dry meridional intensities are difficult to interpret as they are a critical function of the state of tension of the specimen (Cohen, 1954).

In summary, the best phases that have been obtained for the wet, axially projected electron density, which are compatible with the amino acid sequence, are those of model E (Table 4.13). The electron density obtained by combining these phases with the observed and corrected meridional intensities from wet, rat tail tendon is shown in Fig. 4.7.

4. Neutron Diffraction from Collagen

i) Introduction

One of the most attractive features of neutron diffraction for studying biological material is the possibility of 'contrast variation' (see Schoenborn, 1975). By varying the ratio of D_2O to H_2O in a sample, the contrast between certain parts of that sample and the water background may be minimised. If a biological system contains components which differ enough in their scattering length densities, these components may be individually 'contrasted out'. Baldwin et al. (1975), for example, have shown that the range of scattering length densities for D_2O/H_2O mixtures is sufficient to match the average histone scattering length density at 37.5% D_2O, and to match the average nucleic acid scattering density at 63.5% D_2O. It must be said at the outset, however, that these calculations refer to average

scattering lengths. Proteins and nucleic acids do not have homogeneous scattering length densities. These internal inhomogeneities can only be detected, however, at a certain resolution. It may well be that, at the resolution at which Baldwin et al. (1975) were working, the internal inhomogeneities were insignificant compared to the large difference in average scattering length densities between histones and nucleic acids. Nevertheless it is important to bear in mind when interpreting contrast variation experiments that there is a limit to the resolution at which the components of a biological structure may be treated as homogeneous. This limit increases (in reciprocal space) as the sizes of the components decrease.

There are a number of particular advantages for the use of neutron diffraction in the study of molecular packing in connective tissue. First, as mentioned in Chapter 1, connective tissues are multi-component systems (proteins, polysaccharides, minerals and water). It may be possible to use contrast variation to study the structural relationships between these components. More specifically, even though glycosaminoglycans make up only 0.5% of the dry weight of tendon (Meyer et al., 1956), it may be possible to study the effect of collagen-proteoglycans interaction on the axially projected structure (see Chapter 3 and previous sections of this Chapter). Secondly the negative neutron scattering length of hydrogen means that the scattering length density of H_2O is also negative. The scattering length density of D_2O is positive. There is thus a particular D_2O/H_2O (8.7% D_2O) mixture for which the average scattering length density is zero. Collagen fibres may therefore be examined in their native wet state but without the water (i.e. 8.7% D_2O/91.3% H_2O by volume) contributing to the low angle

diffraction. In the model calculations described in the previous X-ray section, therefore, the residue volume parameters (and their associated uncertainty) may be ignored for neutron diffraction in 8.7% D_2O.

The third advantage of neutron diffraction for collagen work is that, as neutron scattering lengths are independent of X-ray scattering lengths, the same model may be used to give two sets of predicted meridional intensities for comparison with experimental observations. The atomic scattering factors for X-rays would simply be replaced by neutron scattering lengths. In this section the predicted neutron meridional intensities for rat tail tendon, based on model E of the previous section, are compared with the observed intensities in a range of D_2O/H_2O mixtures.

ii) Data collection

The low angle neutron diffraction apparatus, D11, at ILL, Grenoble, was used to collect the meridional intensities from rat tail tendon in a range of D_2O/H_2O mixtures, from 0 to 80% D_2O. The data was recorded initially on a 64 by 64 area detector. The intensity of the unscattered beam was recorded for each experiment (see later) and the integrated intensities above background were obtained using a specially written computer programme. The finite spread of wavelengths ($\pm$ 4% at 7Å) caused a broadening of the meridional reflections which increased with diffraction order. There was also a broadening due to beam divergence. Both these factors limited the resolution to eight orders of the 670Å periodicity.

As only 48 hrs were available on D11, and as measurements were taken for nine D_2O/H_2O mixtures, in samples that had been equilibrated for 8 to 24 hrs, it was not possible to collect all

the data on one sample. Four samples, A, B, C and D were used and the measurements were made in three stages. At the beginning of the experiment data was collected for all samples in H_2O (phosphate buffered to pH 7). The four specimens were then left to equilibrate in 10%, 30%, 50% and 70% D_2O respectively. After 24 hrs, further diffraction data was collected and the specimens were re-equilibrated in 20%, 40%, 60% and 80% D_2O respectively. (The pD of the D_2O was phosphate buffered to pD 7). The meridional intensities were then corrected, for variations in sample size and absorption in different D_2O/H_2O mixtures, in the following way.

a) Sample size and absorption corrections

The measured intensity of a particular meridional reflection may be expressed as:

$$I' = I \,.\, N.a.c. \qquad (1)$$

where $I' = I'(n,s,p)$

- n is the diffraction order
- s is a set of parameters describing the size, shape and disorientation of the specimen
- p is the number of deuterium atoms in the specimen expressed as a percentage of the total number of exchangeable hydrogens.
- N is the total number of fibrils in a specimen, and hence is a function of s (i.e. $N = N(s)$). It is assumed that the diffracted intensities from all the fibrils simply add together.
- a is an absorption corrected, which is a function of n,s, and p (i.e. $a = a(n,s,p)$.
- c is a correction factor which incorporates Lorentz, polarisation, disorientation factors, etc. $c = c(n,s)$.
- I is the square of the structure factor amplitude per fibril. $I = I(n,p)$. This is the quantity that is required from the data.

From equation (1),

$$I(n,p) = \frac{I'(n,s,p)}{N(s)\ a(n,s,p)\ c(n,s)} \qquad (2)$$

Thus, for a given n and p, n_1 and p_1,

$$\left(\frac{I'}{Nac}\right)_{S_A} = \left(\frac{I'}{Nac}\right)_{S_B} = \left(\frac{I'}{Nac}\right)_{S_C} = \left(\frac{I'}{Nac}\right)_{S_D} \qquad (3)$$

for samples A, B, C and D. From equation (2), the number of fibrils in samples B, C and D relative to the number in sample A may be calculated.

$$\text{i.e.} \quad \frac{N(s_A)}{N(s)} = \frac{I'(n_1,s_A,p_1)\ a(n_1,s,p_1).\ c(n_1,s)}{I'(n_1,s,p_1)\ a(n_1,s_A,p_1).\ c(n_1,s_A)} \qquad (4)$$

Assuming the disorientation in all samples is the same, i.e.

$$c(n,s) \simeq c(n)$$

and that the dependence of a on n may be neglected for low angle diffraction, i.e.

$$a(n,s,p) \simeq a(s,p)$$

equation (4) may be re-written,

$$\frac{N(s_A)}{N(s)} = \frac{I'(n_1,s_A,p_1).\ a(s,p_1)}{I'(n_1,s,p_1).\ a(s_A,p_1)} \qquad (5)$$

Assuming I(o,s,p) is negligible compared to the unscattered beam intensity, B, i.e.

$$I'(o,s,p) = B.a(s,p) \qquad (6)$$

the absorption factors in equation (5) may be written as

$$\frac{a(s,p_1)}{a(s_A,p_1)} = \frac{I'(o,s,p_1)}{I'(o,s_A,p_1)}$$

$$\therefore \quad \frac{N(s_A)}{N(s)} = \frac{I'(n_1,s_A,p_1).\ I'(o,s,p_1)}{I'(n_1,s,p_1).\ I'(o,s_A,p_1)} \qquad (7)$$

The average value of this quantity may be calculated for all values of n, hence

$$\left\langle \frac{N(s_A)}{N(s)} \right\rangle = \frac{I'(o,s,p_1)}{I'(o,s_A,p_1)} \left(\sum_{n=1}^{n_{max}} \frac{I'(n,s_A,p_1)}{I'(n,s,p_1)} \right) \cdot \frac{1}{n_{max}} \qquad (8)$$

This quantity can therefore be calculated from the data for all samples in one particular D_2O/H_2O mixture. It is convenient to make these measurements in H_2O.
From equation (2)

$$I(n,p).N(s_A).c(n) = \frac{I'(n,s,p).}{a(s,p)} \frac{N(s_A)}{N(s)}$$

But, from equation (6)

$$a(s,p) = \frac{I'(o,s,p)}{B}$$

$$\therefore \quad I(n,p). \frac{N(s_A).c(n)}{B} = \frac{I'(n,s,p).}{I'(o,s,p)} \frac{N(s_A)}{N(s)}$$

$$= \frac{I'(n,s,p).}{I'(o,s,p)} \frac{I'(o,s,p_1).}{I'(o,s_A,p_1)} \left(\sum_{n=1}^{n_{max}} \frac{I'(n,s_A,p_1)}{I'(n,s,p_1)} \right) \cdot \frac{1}{n_{max}}$$

from equation (8).

The product $I(n,p).c(n)$ is therefore given, on an arbitrary scale, by

$$I(n,p).c(n) = \frac{I'(n,s,p).}{I'(o,s,p)} I'(o,s,p\ ). \left(\sum_{n=1}^{n_{max}} \frac{I'(n,s_A,p_1)}{I'(n,s,p_1)} \right) \qquad (9)$$

The square root of this quantity is a measure of the structure factor amplitude per fibril as a function of the diffraction order and percentage deuteration. Since it was assumed that the amount of disorientation was the same for all samples, equation (9) represents the meridional intensity data corrected for variations in sample size, shape and absorption.

iii) Results

The square root of $I(n,p).c(n)$, as obtained from the

experimental data and equation (9), is given in Table 4.14, for the first eight orders and 0 to 80% D_2O. The second order amplitude is probably an overestimate as the first order intensity was so strong that there was some overlapping with the second order. Fig. 4.8 shows the amplitude variation for the strong first, third, fifth and sixth orders.

There are three striking features about these results. First, all the amplitudes go through a minimum between 25% and 50% deuteration. Secondly, away from the minima the amplitudes seem to vary linearly with deuteration. Thirdly, though most of the amplitudes are approximately equal at 0% and 80% deuteration, the sixth order shows a two-fold increase.

iv) Calculated Neutron Structure Factor Amplitudes

In the previous section the meridional X-ray structure factors were calculated for axially projected electron density models based on the amino acid sequence. The models were defined using two parameters for each residue (section 3(i)), z_n, the position, and f_n, the net scattering length above water. The net scattering length was defined by the following equation

$$f_n = t_n - \rho\nu_n$$

where t_n was the total number of electrons in the nth residue, ν_n was the volume and ρ was the electron density of water. For the calculation of neutron diffraction structure factors the above equation is replaced by

$$f_n = C_n + H_n b_D r + H_n b_H(1-r) - \nu_n \frac{(b_o + 2rb_D + 2(1-r)b_H)}{\nu_w}$$

where C_n is the total elastic neutron scattering length of the unexchangeable atoms in the residue
H_n is the number of exchangeable hydrogens in the residue
b_D is the elastic neutron scattering length of deuterium

TABLE 4.14 Square roots of the observed meridional neutron reflections from wet rat tail tendon, corrected for sample variation and absorption (see text).

n	1	2	3	4	5	6	7	8
% D_2O								
0	45.0	11.4	14.8	5.6	7.8	5.4	3.4	4.9
10	36.5	6.0	14.3	3.2	7.4	4.7	3.1	4.4
20	25.7	4.4	9.9	1.6	5.4	2.7	2.5	3.5
30	18.8	5.0	5.2	1.6	3.3	2.2	0.7	2.1
40	12.1	5.7	4.2	1.9	2.4	3.1	2.3	2.8
50	15.1	9.1	6.3	2.0	2.5	4.7	3.2	2.2
60	24.6	12.9	8.8	4.3	4.5	6.6	4	3.8
70	29.5	13.7	11.4	5.8	5.9	8.7	3.9	4.2
80	42.9	18.0	15.5	7.6	8.1	10.7	4.7	5.0

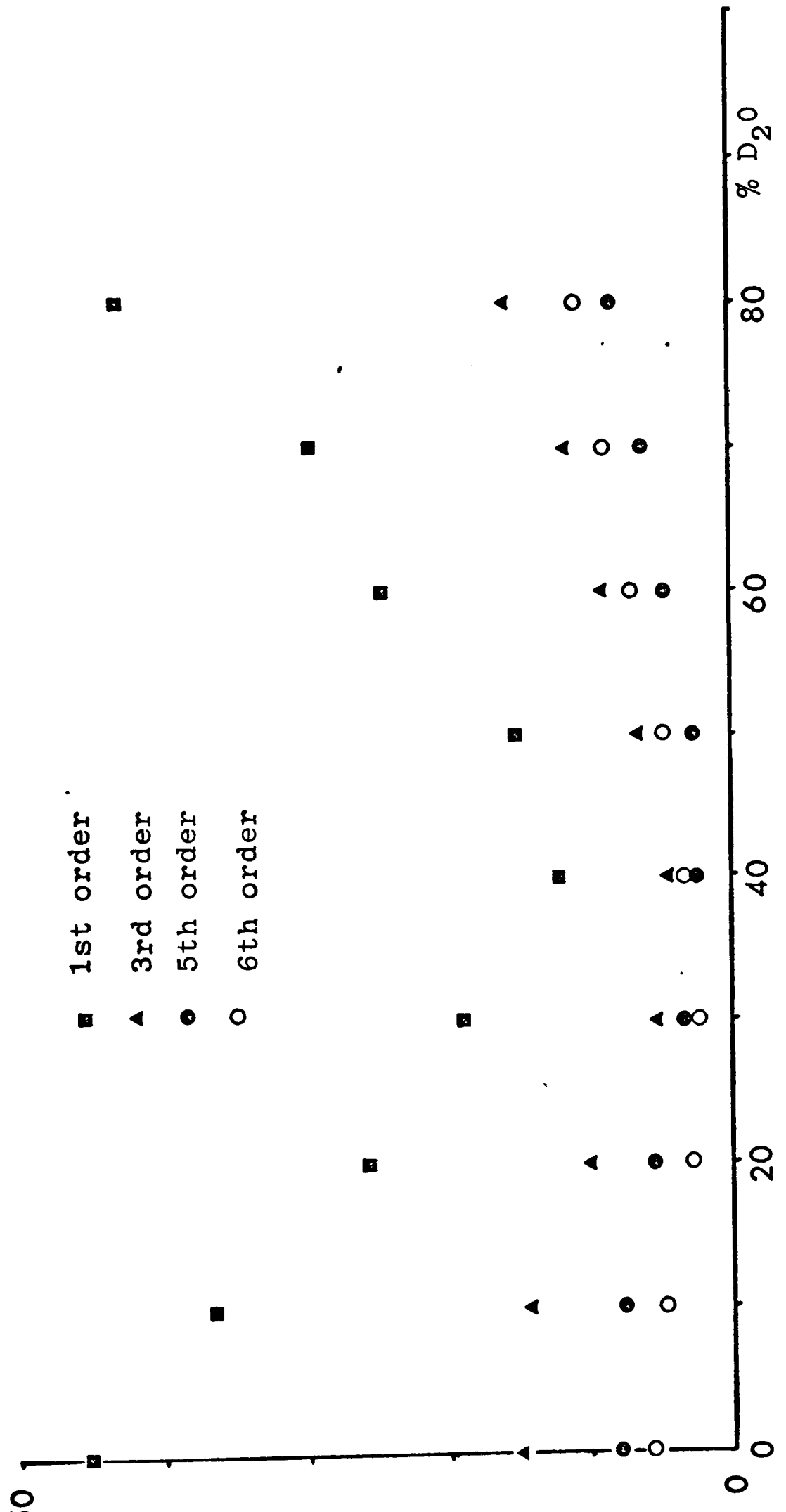


Fig. 4.8. Observed amplitude variation, of the meridional reflections, for rat tail tendon in a range of D_2O/H_2O mixtures.

b_H is the elastic neutron scattering length of hydrogen
b_O is the elastic neutron scattering length of oxygen
ν_w^o is the volume of a water molecule (30 $Å^3$)
r^w is the fractional deuteration.

The first three terms on the right hand side of the above equation represent the average scattering length of the nth residue in the unit cell, after the sample has been equilibrated with a water solution containing r% D_2O and (100-r)% H_2O (by volume). Similarly, the fourth term represents the background correction from water molecules with the average scattering length for that particular D_2O/H_2O ratio. It has been assumed that the sample was equilibrated with a large excess of water, so that the number of exchangeable hydrogen atoms in the protein was insignificant compared to the number in the surrounding water.

The above equation can be re-expressed as the combination of a constant and variable part, i.e.

$$f_n = \left[C_n + H_n b_H - \frac{\nu_n(b_o + 2b_H)}{\nu_w}\right] + r\left[(b_D - b_H)(H_n - \frac{2\nu_n}{\nu_w})\right]$$

$$= K_n + r\,S_n$$

where K_n is the H_2O corrected scattering length of a fully hydrogenated residue and S_n is the difference scattering length term for a fully deuterated residue in D_2O.

When the axially projected positions z_n, of all the residues in the unit cell are given, the meridional structure factors are,

$$F_l = \sum_{n=1}^{N} (K_n + rS_n)e^{2\pi i l z_n/D}$$

where N is the total number of residues in the unit cell and F_l is the structure factor for the lth order. (N.B. Each residue is represented by a δ-function).

F_1 is therefore the sum of two complex components, i.e.

$$F_1 = P_1 + r\,Q_1$$

where

$$P_1 = \sum_{N=1}^{N} K_n\, e^{2\pi i l z_n/D} = p e^{i\theta_p}$$

and

$$Q_1 = \sum_{N=1}^{N} S_n\, e^{2\pi i l z_n/D} = q\, e^{i\theta_q}$$

This equation is illustrated in the form of an Argand diagram in Fig. 4.9. Clearly the modulus of F_1 is given by,

$$|F_1| = |(p^2 + r^2q^2 + r2pq\cos(\theta_p - \theta_q))^{\frac{1}{2}}|$$

$$= |(p^2 + r^2q^2 + r2pq\cos\Delta\theta)^{\frac{1}{2}}|$$

$|F_1|$ can only go through a minimum if $\cos\Delta\theta$ is negative (i.e. $90^o < \Delta\theta < 270^o$). Furthermore the minimum will only be in the physically obtainable range of r = 0 to r = 1 if the projection of p on to q is less than q (i.e. $-p\cos\Delta\theta < q$). If these conditions are satisfied the value of $|F_1|$ at its minimum is given by $p\sin\Delta\theta$ (see Fig. 4.9). As p is known from the value of $|F_1|$ at r = 0 the height of the minimum gives $\sin\Delta\theta$. Finally q can be calculated from the value of r at the minimum by using the equation,

$$r_{min} = -p\cos\Delta\theta \quad \text{(see Fig. 4.9).}$$

Thus, if there is a minimum in the variation of the square root of the experimentally determined intensities, both the relative amplitudes and relative phases of the constant and r-depenedent parts of the structure factor can be determined. Therefore if the absolute phases of either the constant or the variable part can be determined by, for example, Patterson techniques, the phases of the remaining part can be calculated directly.

The best model, based on the sequence, for the axially

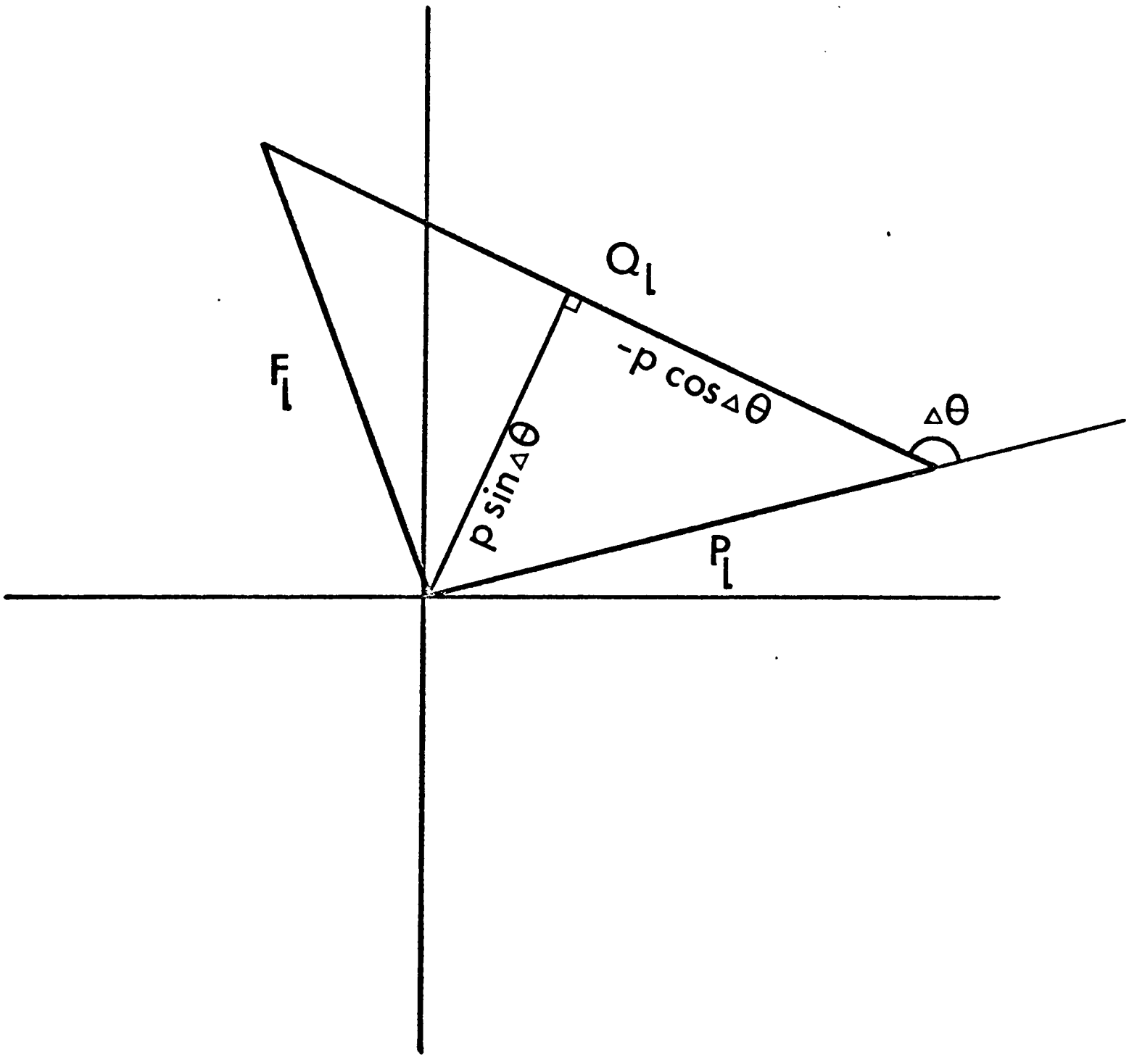


Fig. 4.9 Argand diagram showing the contribution to the structure factor of the lth order (F_l). P_l is independent of the fractional deuteration, whereas the modulus of Q_l varies from zero to q_l (see text).

projected electron density of rat tail tendon fibrils was shown in Fig. 4.7. The parameters used to define this model (model E) were given on page 161. By replacing the net number of electrons per residue by the net neutron scattering length defined above, it was possible to calculate the predicted amplitude variation of the first eight meridional neutron reflections from rat tail tendon. The number of exchangeable hydrogens per residue was taken as the number of hydrogens not covalently linked to carbon (Hvidt and Nielsen, 1966), minus the number involved in hydrogen bonds within the collagen triple helix (see Brodsky-Doyle et al. (1975) for references). This latter number was calculated assuming that glycine is the only hydrogen bond donor (i.e. the Rich-Crick Model II, see Chapter 1). The results of this structure factor calculation are shown in Table 4.15. The expected amplitude variation of the first, third, fifth and sixth orders is illustrated in Fig. 4.10.

v) Discussion

There is clearly good general agreement between the observed (Table 4.14, Fig. 4.8) and expected (Table 4.15, Fig. 4.10) amplitude variations. All the curves have minima and they are approximately linear away from the minima. The linearity suggests that the phase difference between the constant and r-dependent parts of the structure factor is approximately 180^{o}. If this were the case the expression for the modulus of the structure factor, given in the previous section, would simplify to

$$|F_1| = |(a^2 + r^2b^2 - r2ab)^{\frac{1}{2}}|$$

$$= |a - rb|$$

TABLE 4.15 Predicted amplitudes of the meridional neutron reflections from wet rat tail tendon. The data has been scaled so that the observed and predicted 0% D_2O, first order amplitudes are equal.

n	1	2	3	4	5	6	7	8
% D_2O								
0	45	1.9	17.1	2.4	14.6	5.1	9.9	5.1
10	34.7	1.7	13.4	2.1	11.4	3.7	8.8	4.6
20	24.5	1.7	9.8	2.6	8.1	5.2	7.8	4.2
30	14.3	1.7	6.6	3.6	4.9	8.1	6.9	3.9
40	4.4	1.8	4.6	4.8	1.8	11.3	6.0	3.9
50	6.8	2.0	5.3	6.1	1.9	14.7	5.2	4.2
60	16.9	2.2	8.0	7.4	5.0	18.2	4.6	4.6
70	27.1	2.5	11.5	8.8	8.2	21.7	4.2	5.1
80	37.4	2.8	15.1	10.1	11.4	25.2	4.1	5.8
90	47.7	3.2	18.8	11.5	14.6	28.7	4.2	6.5
100	57.9	3.5	22.6	12.9	17.8	32.2	4.7	7.3

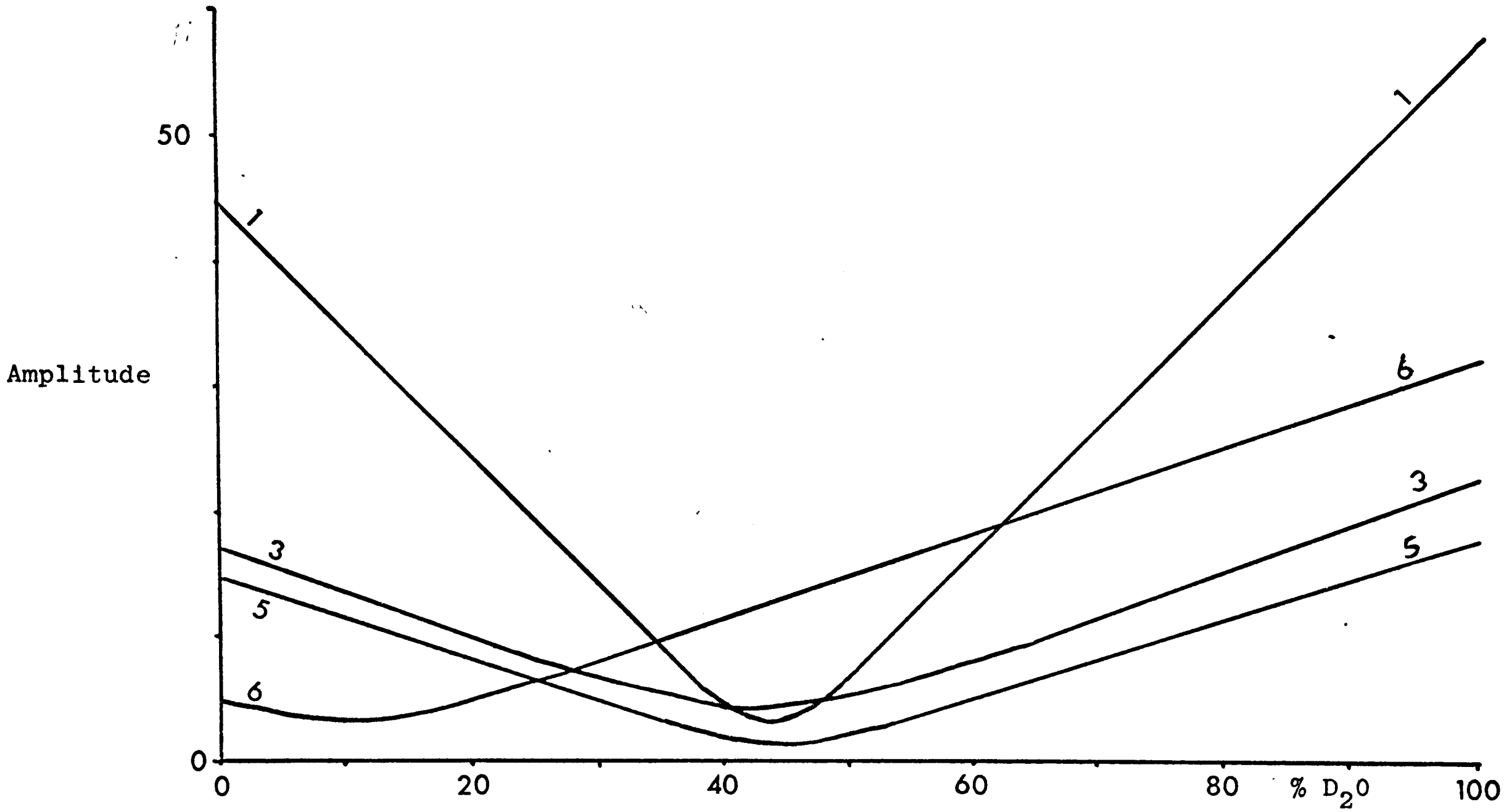


Fig. 4.10. Predicted amplitude variation, of the meridional reflections, for rat tail tendon in a range of D_2O/H_2O mixtures. The first order, 0% D_2O amplitude was arbitrarily equated to the observed amplitude.

As the first order amplitude is dominated by the Hodge-Petruska gap/overlap step function (section 3(iii)) the origin of the variation for this order can be seen clearly in the water background corrected neutron scattering length density distributions calculated for the model in H_2O and D_2O. These are shown in Fig. 4.11. Clearly the step function is inverted due to the filling of the gap region by D_2O, which has a higher average scattering length density than the fully deuterated protein. From Figs 4.8 and 4.10 the step function is minimised at approximately 45% deuteration.

The fact that the observed structure factor amplitudes do not all go through a minimum at the same percentage deuteration illustrates the limitations of the contrast matching technique discussed in the introduction to this section. Preliminary calculations show that the average neutron scattering length density of glycosaminoglycans corresponds to a D_2O/H_2O mixture of approximately 25% D_2O. This is close to the positions of the minima of some of the orders in Fig. 4.8. It would therefore not be appropriate to consider collagen as homogeneous to a resolution of the 8th order, when contrast variation is used to study the structural arrangement of glycosaminoglycans and collagen in connective tissues. A more detailed understanding of the contribution from collagen alone is a necessary prerequisite of such a study.

A most satisfactory feature of the calculated amplitude variation (Fig. 4.10) is the rise in the sixth order with increasing deuteration, as observed. This feature can be related directly to the D/6 periodicities discussed at the end of Chapter 2. Table 4.16 shows the contributions to the model's sixth order from the large hydrophobic residues (val,

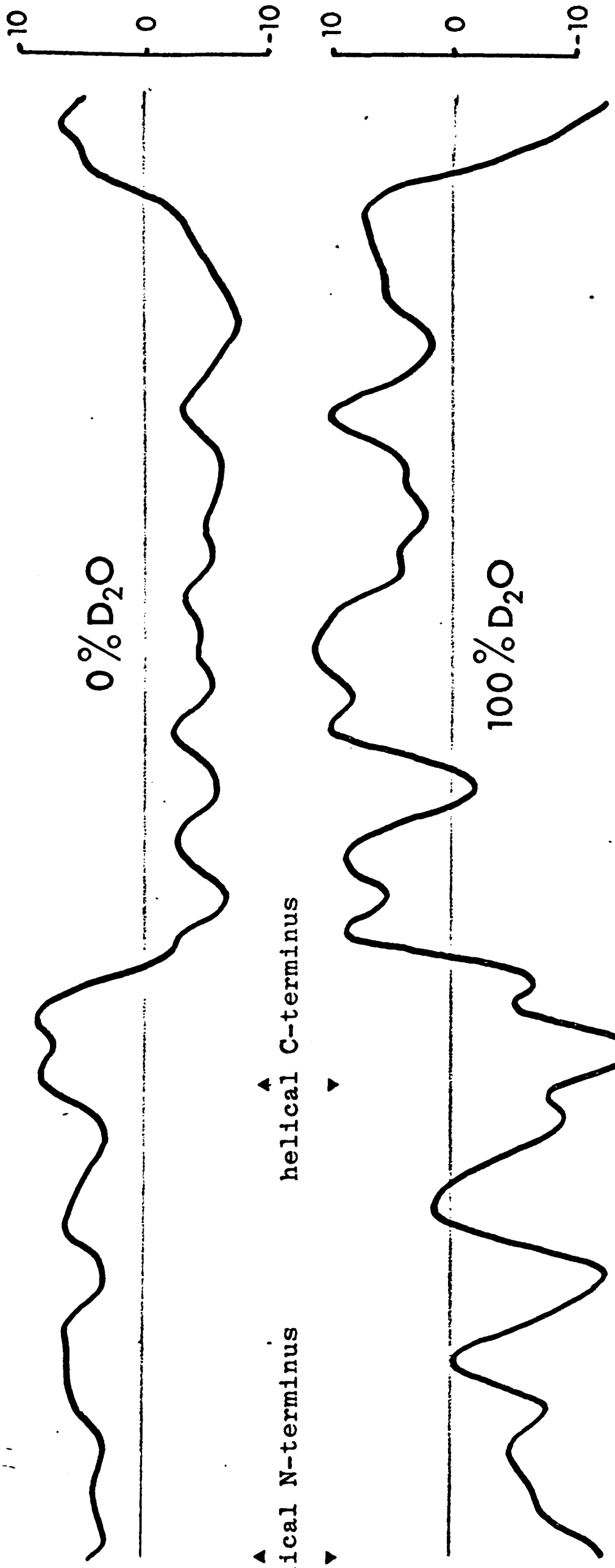


Fig. 4.11 Predicted neutron scattering length variations, for the axially projected structure of collagen fibrils, in H_2O and in D_2O.

TABLE 4.16 Contributions to the structure factor for the sixth order. The amplitudes are given in arbitrary units and the phases (referred to an origin at residue position 44.5 of the helical region, i.e. the approximate centre of the overlap) are in brackets.

	Percentage deuteration	
	0%	100%
hydrophobic residues	218 (355°)	593 (175°)
charged residues	289 (200°)	338 (6°)
the rest	112 (26°)	270 (195°)
total	82 (305°)	516 (178°)

met, ile, leu, phe), the charged residues (asp, glu, arg, lys, hyl) and the rest (asn, asx, gln, glx, ser, thr, tyr, pro, gly, ala, his, hyp) for 0% and 100% deuteration. For 0% deuteration the calculated sixth order is weak because the contribution from the hydrophobic residues and the rest is approximately equal and out of phase with the charged residue contribution. This balance is upset for 100% deuteration mainly by the approximately three-fold increase in the amplitude of the hydrophobic contribution. The charged contribution increases only slightly whilst the contribution from the rest increases by a factor of 2.5. As a result of these charges the total structure factor amplitude for the sixth order is 100% D_2O is large.

It is not possible at present to use the difference Fourier techniques of the previous X-ray section for analysis of the neutron diffraction data. There are two reasons for this. First it is necessary to determine the appropriate correction factor for neutron data analagous to that of Wang and Worthington (1975) for X-rays. This factor must take account of the large wavelength spread of the neutron source. Secondly, the present data must be extended to the higher orders of diffraction. This second limitation has been remedied very recently by Brodsky-Doyle et al. (1975) and by Miller, Timmins and White (unpublished), who have collected meridional diffraction data for rat tail tendon in various D_2O/H_2O mixtures on the D16 instrument at ILL, Grenoble. Preliminary inspection of this data indicates that the agreement between calculated and observed intensities for 100% deuteration is much better than the agreement for 0% deuteration (see Table 4.17). A possible explanation of this can be seen in the predicted

TABLE 4.17 Comparison of the predicted meridional intensities, for rat tail tendon in 100% D_2O, with the observed intensities. Both data sets are on arbitrary scales.

ORDER	PREDICTED INTENSITIES	OBSERVED INTENSITIES
1	1000	not measurable
2	4	not measurable
3	152	7300
4	49	800
5	95	1300
6	309	4000
7	7	200
8	16	500
9	55	660
10	23	150
11	52	720
12	25	110
13	16	150
14	46	100
15	14	80
16	36	210
17	49	-
18	7	40
19	8	150
20	62	150
21	67	150
22	19	100
23	21	30
24	0	40
25	18	20

neutron scattering length distributions of Fig. 4.11. For 0% deuteration the local fluctuations in the curve relative to the height of the step function are much smaller than those for 100% deuteration. Hence it would be expected that the Fourier transform of the 0% deuteration curve would be highly dependent on the precise shape of the step function, which is itself a critical function of the telopeptide structure. This may not be the case for 100% deuteration as there is so much inhomogeneity in the density of the main, helical parts of the structure. The telopeptide contribution may therefore be relatively weak in this case. It may be, therefore, that the structure in D_2O is dominated by the more certain helical region and this is why the agreement between calculated and observed intensities is good. Such an explanation is consistent with the fact that the net scattering lengths of the large hydrophobic residues, which occur predominantly in the helical region, are dominant in 100% D_2O (see Table 4.18).

5. Conclusion

In Chapter 2 the amino acid sequence of collagen was correlated with the appearance of native fibrils in electron micrographs. The present chapter has been an extension of this theme to the correlation of the sequence with the meridional X-ray and neutron diffraction patterns. It has been possible to devise a model, based on the sequence and the Hodge-Petruska scheme in axial projection, which gives an R-factor of 0.3 with the X-ray data. This model also explains many features of the neutron diffraction observations.

So far the interpretation of the neutron diffraction patterns has been heavily dependent on the model devised to fit the X-ray data. In future it is hoped that the neutron data

TABLE 4.18 Effective neutron scattering lengths for amino acid residues in 0% D_2O and 100% D_2O. Residue volumes from Chothia (1975).

Amino Acid Residue	Effective scattering lengths	
	0% D_2O	100% D_2O
HYP	3.53	-4.51
ASP	4.16	-2.31
ASN	4.21	-1.98
ASX	4.19	-2.15
THR	2.81	-3.50
SER	2.77	-1.96
GLU	4.25	-4.32
GLU	4.26	-3.70
GLX	4.25	-4.01
PRO	2.93	-5.93
GLY	2.10	-1.42
ALA	2.15	-3.10
VAL	2.25	-6.44
MET	2.73	-7.96
ILE	2.31	-8.23
LEU	2.30	-8.18
TYR	5.84	-6.06
PHE	5.25	-7.66
LYS	2.88	-5.78
HIS	5.32	-4.09
ARG	4.82	-2.48

can be quantified and corrected so that it is possible to compare predicted and observed structure factor amplitudes using difference Fourier and difference Patterson techniques. The neutron data could then be used to suggest alterations in the X-ray model and vice-versa.

Ultimately the electron density model based on the sequence (model E) must be compared with a model deduced directly from X-ray diffraction data and the results of the heavy atom, isomorphous replacement technique. An attempt to solve the structure using these latter methods is currently in progress (White, unpublished work).

A lot of effort has been devoted to interpreting the X-ray and neutron diffraction patterns in terms of the sequence. The justification for this is that, if the diffraction effects of rat tail tendon can be understood in such a way, then there is the possibility of being able to interpret the diffraction patterns from a variety of connective tissues. Furthermore, these tissues could be examined in many metabolic states (e.g. age variation, tissue degradation).

As mentioned earlier, electron microscopy has the disadvantage that specimens are not examined in their natural states. X-ray and neutron diffraction suffer much less from this limitation; their main drawback is the phase problem. The main concern of this chapter has been the determination of phases for rat tail tendon. It is suggested that this phase information, in conjunction with the crystallographic 'difference' techniques, could widen the scope of X-ray and neutron diffraction in the study of connective tissue.

APPENDIX 1

The following computer programme calculates interaction curves for any given amino acid sequence. The input and output is described in the comments. On the Oxford ICL 1906 A computer, the programme calculates one interaction curve in approximately 35 seconds. It requires 37K of core.

```
DØCUMENT     INTERACTIØNS

*FØRTRAN
       LIST
       PRØGRAM(FURT)
       INPUT 2=CR0
       INPUT 3=CR1
       INPUT 4=CR2
       INPUT5=CR3
       INPUT 8=CR4
       ØUTPUT6=LP0
       ØUTPUT7=LP1
       CØMPRESS INTEGER AND LØGICAL
       TRACE 1
       END
       MASTER INTERACT
       DIMENSIØN NSEQ(1250),MX(5,5),NSEQ2(1250),Y(2500),IX(2500)
      1,SUMN(1250),SUMC(1250),X(2500),IPLUS(50),MINUS(50),Y2(2500),
      2IAA(23),IAAT(23),LABEL(20)
C
C
C      THIS PRØGRAM CALCULATES INTERACTIØNS BETWEEN TWØ
C      IDENTICAL 'ØNE DIMENSIØNAL' PRØTEIN SEQUENCES
C
C      THE PARAMETERS ARE:
C
C      NRES : THE NUMBER ØF RESIDUES IN THE SEQUENCE
C
C      IT1 : IF IT1=0 THE ENTIRE SEQUENCE ØF RESIDUES IS READ
C            ØN INPUT CHANNEL 3 INTØ THE ARRAY 'NSEQ'. TWØ ØTHER
C            ARRAYS, 'IAA' & 'IAAT', ARE ALSØ READ ØN CHANNEL 4.
C            THESE ARE USED TØ CØNVERT THE SEQUENCE INTØ A LINEAR
C            NUMERICAL ARRAY, ALSØ CALLED 'NSEQ', IN WHICH THE
C            INTERACTIØN PRØPERTY ØF EACH RESIDUE TYPE IS
C            ASSOCIATED WITH A PARTICULAR NUMBER.
C
C            IF IT1=1 THE PØSITIØNS ØF THE 'IP' UNPAIRED PØSITIVE
C            CHARGES (ARRAY 'IPLUS') AND THE PØSITIØNS ØF THE
C            'IM' UNPAIRED NEGATIVE CHARGES (ARRAY 'MINUS') ARE READ
C            ØN INPUT CHANNEL 5.
C
C            IF IT1=2 THE PØSITIØNS ØF THE 'IP' N-TERMINALLY DIRECTED
C            CHARGE DIPØLES (ARRAY 'IPLUS') AND THE PØSITIØNS ØF THE
C            'IM' C-TERMINALLY DIRECTED CHARGE DIPØLES (ARRAY 'MINUS')
C            ARE READ ØN INPUT CHANNEL 8.
C
C      IT2 : IF IT2=0 THE INTERACTIØNS ARE CALCULATED FØR PARALLEL
C            SEQUENCES.
C
C            IF IT2=1 THE INTERACTIØNS ARE CALCULATED FØR
C            ANTI-PARALLEL SEQUENCES.
C
C      IT3 : IF IT3=0 THE INTERACTIØNS FØR ALL STAGGERS ARE DISPLAYED
C            GRAPHICALLY AND WRITTEN INTØ TWØ NUMERICAL ØUTPUT
C            CHANNELS.
C
C            IF IT3=1 THE INTERACTIØN CURVE IS CØMPRESSED BY
C            ADDING TØGETHER ALL INTERACTIØNS RELATED BY
C            STAGGERS ØF N*'ND'. THESE RESULTS ARE THEN DISPLAYED
C            GRAPHICALLY AND WRITTEN INTØ TWØ ØUTPUT CHANNELS.
C
C      LS : THE ALLØWED INTERACTIØN RANGE IS PLUS ØR MINUS 'LS'
C           RESIDUES.
```

```
C
      READ(2,105)NRES,IT1
  105 FØRMAT(2I0)
      NL1=NRES-1
      NP1=NRES+1
      RNL1=NL1
C
C     DEPENDING ØN THE VALUE ØF IT1 THE APPRØPRIATE ARRAYS
C     ARE READ IN
C
      IF(IT1.GT.0)GØ TØ 1000
      READ(4,102)(IAA(K),IAAT(K),K=1,23)
  102 FØRMAT(A3,1X,I1)
      READ(3,103)(NSEQ(K),K=1,NRES)
  103 FØRMAT(20(A3,1X))
      DØ 70 J=1,NRES
      DØ 70 K=1,23
      IF(NSEQ(J).NE.IAA(K))GØ TØ 70
      NSEQ(J)=IAAT(K)
   70 CØNTINUE
      GØ TØ 1001
C
C
1000  CØNTINUE
      IF(IT1.EQ.1)M=5
      IF(IT1.EQ.2)M=8
      READ(M,105)IP,IM
      READ(M,100)(IPLUS(N),N=1,IP)
  100 FØRMAT(I0)
      READ(M,100)(MINUS(N),N=1,IM)
      DØ 50 N=1,NRES
   50 NSEQ(N)=3
      DØ 51 N=1,IP
      M=IPLUS(N)
   51 NSEQ(M)=1
      DØ 52 N=1,IM
      M=MINUS(N)
   52 NSEQ(M)=2
1001  CØNTINUE
C
C
C     ØNCE 'NSEQ' IS CREATED ACCØRDING TØ THE VALUE ØF IT1 THE
C     PRØGRAM CAN CYCLE THRØUGH 'NNN' SETS ØF VALUES FØR
C     IT2,IT3,LS AND THE SCØRE MATRIX
C
C
      READ(2,100)NNN
      DØ 999 LMN=1,NNN
      READ(2,109)LABEL
109   FØRMAT(20A4)
      READ(2,108)IT2,IT3,LS
108   FØRMAT(3I0)
C
C     THE ARRAY 'MX(I,J)' SPECIFIES THE INTERACTIØN SCØRE BETWEEN
C     RESIDUE TYPE 'I' AND RESIDUE TYPE 'J'.
C
      READ(2,100)NUM
C
C
      DØ 1 I=1,5
      DØ 1 J=1,5
    1 MX(I,J)=0
C
C     THE 'NUM' NØN-ZERØ SCØRES ARE READ IN ØN CHANNEL 2.
C
      DØ 2 K=1,NUM
      READ(2,108)INA,INB,IS
      MX(INA,INB)=IS
    2 CØNTINUE
C
C     AN ARRAY 'NSEQ2' IS CREATED WHICH IS EITHER THE SAME AS 'NSEQ'
C     (IT2=0) ØR 'NSEQ' WITH REVERSED PØLARITY (IT2=1)
C
      IF(IT2.EQ.0)GØ TØ 1002
      DØ 53 K=1,NRES
   53 NSEQ2(NP1-K)=NSEQ(K)
      GØ TØ 1003
1002  CØNTINUE
      DØ 54 K=1,NRES
   54 NSEQ2(K)=NSEQ(K)
1003  CØNTINUE
C
```

```
C
CC      THE BASIC ZERØ RANGE INTERACTIØN CURVE IS CALCULATED HERE.
C
        M=0
        DØ 3 IS=M,NL1
        SUMN(IS+1)=0.0
        SUMC(IS+1)=0.0
        ID=NRES-IS
        DØ 4 I=1,ID
        ISI=IS+I
        IA=NSEQ2(I)
        IO=NSEQ(ISI)
        SUMC(IS+1)=SUMC(IS+1)+MX(IA,IO)
        IA=NSEQ(I)
        IO=NSEQ2(ISI)
        SUMN(IS+1)=SUMN(IS+1)+MX(IA,IO)
    4 CØNTINUE
    3 CØNTINUE
        DØ 54 N=1,NRES
   54 Y(N)=SUMC(NP1-N)
        DØ 55 N=1,NL1
   55 Y(N+NRES)=SUMN(N+1)
C
CC      THE INTERACTIØN RANGE IS INTRØDUCED HERE.
C
        LS3=NRES+NL1-(2*LS)
        LS4=(LS3+1)/2
        LS5=(2*LS)+1
        DØ 57 N=1,LS3
        NLS=N+LS
        Y2(N)=Y(NLS)
        IF(LS.EQ.0)GØ TØ 57
        DØ 58 K=1,LS
   58 Y2(N)=Y2(N)+Y(NLS-K)+Y(NLS+K)
        Y2(N)=Y2(N)/FLØAT(LS5)
   57 CØNTINUE
C
CC      THE INTERACTIØN CURVE IS EITHER WRITTEN STRAIGHT AWAY
C       (IT3=0) ØR FURTHER PRØCESSED (IT3=1)
C
        IF(IT3.EQ.0)GØ TØ 1004
        READ(2,106)ND
  106 FØRMAT(I0)
        NR2=NRES+NRES
        IF=(NR2/ND)+1
        NI=IF*ND
        NI1=NI+1
        RND=ND
        DØ 71 L=1,ND
        X(L)=L-1
   71 IX(L)=L-1
        LS6=LS+1
        LS7=LS3+LS6
        LS8=NI-LS7
        LS9=LS8+1
        LS10=NI-LS6
        LS11=LS10+1
        DØ 60 L=1,LS8
   60 Y(L)=0.0
        DØ 61 L=LS9,LS10
   61 Y(L)=Y2(L-LS8)
        DØ 62 L=LS11,NI
   62 Y(L)=0.0
        DØ 63 L=1,ND
        Y2(L)=0.0
        M1=NI1-L
        DØ 64 M=1,IF
        M2=(M-1)*ND
   64 Y2(L)=Y2(L)+Y(M1-M2)
   63 Y2(L)=Y2(L)/FLØAT(IF)
        GØ TØ 1005
C
C
1004  CØNTINUE
        DØ 59 N=1,LS3
        X(N)=N-LS4
   59 IX(N)=N-LS4
1005  CØNTINUE
C
```

```
C
C       GRAPHICAL OUTPUT
C
        CALL PSPACE(0.1,0.7,0.5,0.9)
        CALL CTRMAG(10)
        IF(IT3)1006,1006,1007
1006    LIM=LS3
        X2=RNL1
        GO TO 1008
1007    CONTINUE
        LIM=ND
        LIMO=1
        X2=RND
        X1=0.0
        GO TO 1009
1008    CONTINUE
        IF(IT2.EQ.1)GO TO 1010
        X1=0.0
        LIMO=LS4
        GO TO 1009
1010    CONTINUE
        X1=-RNL1
        LIMO=1
1009    CONTINUE
        YM=0.0
        YMN=0.0
        DO 80 N=1,LIM
        Z=Y2(N)
        IF(Z.GT.YM)YM=Z
        IF(Z.LT.YMN)YMN=Z
   80   CONTINUE
        CALL MAP(X1,X2,YMN,YM)
        CALL AXES
        CALL PTPLOT(X,Y2,LIMO,LIM,-16)
        H1=(5.0*YMN-YM)/4.0
        H2=(9.0*YMN-YM)/8.0
        XM=X2/2.0
        CALL PLOTCS(X1,H1,LABEL,80)
        CALL PLOTCS(XM,H2,'STAGGER (RESIDUES)',18)
        CALL FRAME
C
C
C       NUMERICAL OUTPUT
C
        WRITE(6,202)LABEL
        WRITE(6,200)(IX(I),Y2(I),I=LIMO,LIM)
        WRITE(7,201)(Y2(N),N=LIMO,LIM)
        WRITE(7,203)LABEL
202     FORMAT(1H1,20A4)
203     FORMAT(1H ,20A4)
  201   FORMAT(10(1X,F7.3))
  200   FORMAT(6(/X,I5,1X,F7.3))
C
C
C
999     CONTINUE
        CALL GREND
        STOP
        END
        FINISH
*END LIB CUGHOST
****
```

APPENDIX 2

The following computer programme calculates the Fourier Transforms of models for the axially projected structure of collagen fibrils. The input and output is described in the comments. On the Oxford ICL 1906 A computer, the programme calculates 40 orders of a transform in approximately 15 seconds. It requires 13K of core.

```
DØCUMENT     X-TRANS

*FØRTRAN
       LIST
       PRØGRAM(FØRT)
       INPUT 5=CR0
       INPUT 1=CR1
       INPUT 2=CR2
       INPUT 3=CR3
       INPUT 4=CR4
       ØUTPUT 6=LP0
       TRACE 0
       END
C
CC
CC     MAIN PRØGRAM SEGMENT
C
       MASTER FT
       DIMENSION LABEL(20)
       CØMMØN /DAT1/NSEQ(1200),IAA(23),AAV(23),SL(23),NSEQ2(1200),
      1IW,NRES,NRESB,NA1A,NA1B,NA2,NAA(100)
       CØMMØN /DAT2/F(1202),F2(100)
       CØMMØN /DAT3/FTMØD(100),FIPHASE(100),DP,PHASEØR
       CØMMØN /DAT4/FTINT(100),LMAX
       CØMMØN /DAT5/MØR
       CØMMØN /DAT6/AAP(100)
C
CC
C      INPUT DATA
CCCC
```

```
C
      READ(1,100)LABEL
      READ(1,101)NRES,NRESB,NA1A,NA1B,NA2,DP,PHASEØR,IW,LMAX
      READ(2,102)(NSEQ(K),K=1,NRES)
      READ(5,102)(NSEQ2(K),K=1,NRESB)
      READ(3,103)(IAA(K),SL(K),K=1,23)
      READ(4,104)(AAV(K),K=1,23)
1000  READ(1,105,END=999)IN1,IC1,IN2,IC2,SQ1,SQ2,BG
      READ(1,106)MØR
      IF(MØR.EQ.0)GØ TØ 2000
      READ(1,107)(NAA(K),AAP(K),K=1,MØR)
2000  CØNTINUE
C
      WRITE(6,200)LABEL
      WRITE(6,201)LMAX
      WRITE(6,202)DP
      WRITE(6,203)PHASEØR
      IF(IW.EQ.1)WRITE(6,214)BG
      WRITE(6,209)IN1,IC1,IN2,IC2
      WRITE(6,210)SQ1,SQ2
      IF(MØR.EQ.0)GØ TØ 2001
      WRITE(6,212)
      WRITE(6,211)(NAA(K),AAP(K),K=1,MØR)
2001  CØNTINUE
      WRITE(6,213)NA1A,NA1B,NA2
C
      CALL CØDE(BG,IN1,IC1,IN2,IC2)
C
      CALL TRANSFØRM(SQ1,SQ2)
C
C
C     ØUTPUT DATA
C
C     FTMØD : AMPLITUDES
C     FTPHASE : PHASES
C     FTINT : INTENSITIES
C
      WRITE(6,204)
      WRITE(6,205)(K,FTMØD(K),FTPHASE(K),FTINT(K),K=1,LMAX)
C
      CALL SCALE
C
      WRITE(6,200)LABEL
      WRITE(6,206)
      WRITE(6,207)
      WRITE(6,208)(K,FTINT(K),K=1,LMAX)
      GØ TØ 1000
999   CØNTINUE
C
C
  100 FØRMAT(20A4)
  101 FØRMAT(5I0,2F0.0,2I0)
  102 FØRMAT(20(A3,1X))
  103 FØRMAT(A3,1X,F6.2)
  104 FØRMAT(4X,F5.1)
  105 FØRMAT(4I0,3F0.0)
106   FØRMAT(I0)
107   FØRMAT(A3,F0.0)

  200 FØRMAT(1H1,20A4,//)
  201 FØRMAT(1H ,'DATA FØR THE FIRST ',I2,' ØRDERS',//)
  202 FØRMAT(1H ,'0 PERIØD=',F6.2,' RESIDUES',//)
  203 FØRMAT(1H ,'PHASE ØRIGIN AT RESIDUE NUMBER ',F6.2,//)
  204 FØRMAT(1H ,'ØRDER',10X,'AMPLITUDE',6X,'PHASE',10X,'INTENSITY',/)
  205 FØRMAT(1X,I4,10X,E11.4,5X,F6.1,9X,E11.4)
  206 FØRMAT(1H ,/,' INTENSITIES SCALED TØ FIRST ØRDER=1000',/)
  207 FØRMAT(1H ,'ØRDER',10X,'INTENSITY',/)
  208 FØRMAT(1X,I4,11X,F10.2)
  209 FØRMAT(1H ,'RESIDUES REMØVED:',I4,' AT A1N',I4,
     1' AT A1C',I4,' AT A2N',I4,' AT A2C',//)
  210 FØRMAT(1H ,'N-TRANS=',F5.3,'   C-TRANS=',F5.3,//)
  212 FØRMAT(1H ,'EXTRA SCATTERERS',/,' RESIDUE   PØSITIØN')
  211 FØRMAT(4X,A3,5X,F6.2)
  213 FØRMAT(1H ,'ALPHA 1 STARIS',2I4,'     ALPHA 2 STARIS',I4,/)
  214 FØRMAT(1H ,'ELECTRØN DENSITY ØF BACKGRØUND = ',F5.3,//)
      STØP
      END
C
```

```
C
      SUBROUTINE CODE(BG1,NN1,NC1,NN2,NC2)
C     THIS SUBROUTINE CALCULATES THE ARRAYS F AND F2.
CC    THESE CARRY THE SCATTERING LENGTHS (WATER CORRECTED IF
CC    DESIRED) OF THE HOMOGENEOUS TRANSLATION PART AND THE -
CC    INHOMOGENEOUS PART RESPECTIVELY.
C
      COMMON /DAT1/NSEQ(1200),IAA(23),AAV(23),SL(23),NSEQ2(1200),
     1IW,NRES,NRESB,NA1A,NA1B,NA2,NAA(100)
      COMMON /DAT2/F(1202),F2(100)
      COMMON /DAT5/MOR
      NC1=NRES-NC1
      NC2=NRESB-NC2
      DO 2 K=1,1202
    2 F(K)=0.0
      IF(IW.EQ.1)EDW=BG1
      IF(IW.EQ.0)EDW=0.0
      IF(MOR.EQ.0)GO TO 5
      DO 8 N=1,MOR
      DO 4 L=1,23
      IF(NAA(N).NE.IAA(L))GO TO 4
      F2(N)=SL(L)-(AAV(L)*EDW)
    4 CONTINUE
    8 CONTINUE
    5 CONTINUE
      DO 1 N=1,NRES
      IF(N.LE.NN1.OR.N.GT.NC1)GO TO 1
      DO 3 L=1,23
      IF(NSEQ(N).NE.IAA(L))GO TO 3
      A=SL(L)-(AAV(L)*EDW)
      F(N+NA1A)=F(N+NA1A)+A
      F(N+NA1B)=F(N+NA1B)+A
    3 CONTINUE
    1 CONTINUE
      DO 6 N=1,NRESB
      IF(N.LE.NN2.OR.N.GT.NC2)GO TO 6
      DO 7 L=1,23
      IF(NSEQ2(N).NE.IAA(L))GO TO 7
      A=SL(L)-(AAV(L)*EDW)
      F(N+NA2)=F(N+NA2)+A
    7 CONTINUE
    6 CONTINUE
      RETURN
      END
C
```

```
C
      SUBRØUTINE TRANSFØRM(SQF1,SQF2)
C     THIS SUBRØUTINE CALCULATES THE FØURIER TRANSFØRM
C     ØF THE MØDEL, WITH DUE REGARD TØ TELØPEPTIDE PARAMETERS.
C
      CØMMØN /DAT2/F(1202),F2(100)
      CØMMØN /DAT3/FTMØD(100),FTPHASE(100),DP,PHASEØR
      CØMMØN /DAT4/FTINT(100),LMAX
      CØMMØN /DAT5/MØR
      CØMMØN /DAT6/AAP(100)
      DIMENSIØN B(100),C(100)
      TP=6.283185307/2
      DØ 1 L=1,LMAX
      B(L)=0.
      C(L)=0.
      TPK=(TP*FLØAT(L))/DP
      DØ 2 M=1,1202
      FM=M
      IF(M.GT.16)GØ TØ 10
      MD=16-M
      FM=16-(MD*SQF1)
      GØ TØ 11
   10 CØNTINUE
      IF(M.LT.1028)GØ TØ 11
      MD=M-1028
      FM=1028+(MD*SQF2)
   11 CØNTINUE
      TPKM=TPK*(FM-PHASEØR)
      B(L)=B(L)+(F(M)*CØS(TPKM))
      C(L)=C(L)+(F(M)*SIN(TPKM))
    2 CØNTINUE
      DØ 3 J=1,MØR
      FM=AAP(J)
      TPKM=TPK*(FM-PHASEØR)
      B(L)=B(L)+(F2(J)*CØS(TPKM))
      C(L)=C(L)+(F2(J)*SIN(TPKM))
    3 CØNTINUE
      A=(B(L)**2)+(C(L)**2)
      FTINT(L)=A
      FTMØD(L)=SQRT(A)
      D=ARCTAN(C(L),B(L))
      FTPHASE(L)=(D*360.)/TP
    1 CØNTINUE
      RETURN
      END
C
C
      FUNCTIØN ARCTAN(CC,BB)
      PI=3.1415926536
      IF(BB.GT.0.0.AND.CC.GT.0.0)E=ATAN(CC/BB)
      IF(BB.LT.0.0.AND.CC.GT.0.0)E=ATAN(CC/BB)+PI
      IF(BB.EQ.0.0.AND.CC.GT.0.0)E=PI/2.0
      IF(BB.EQ.0.0.AND.CC.LT.0.0)E=PI*1.5
      IF(BB.GT.0.0.AND.CC.LT.0.0)E=ATAN(CC/BB)+(2.0*PI)
      IF(BB.LT.0.0.AND.CC.LT.0.0)E=ATAN(CC/BB)+PI
      IF(BB.GE.0.0.AND.CC.EQ.0.0)E=0.0
      IF(BB.LT.0.0.AND.CC.EQ.0.0)E=PI
      ARCTAN=E
      RETURN
      END
C
C
      SUBRØUTINE SCALE
C     THIS SUBRØUTINE SCALES THE CALCULATED INTENSITIES
C     TØ A FIRST ØRDER VALUE ØF 1000.
C
      CØMMØN /DAT4/FTINT(100),LMAX
      SCAL=1000./FTINT(1)
      DØ 1 I=1,LMAX
    1 FTINT(I)=FTINT(I)*SCAL
      RETURN
      END
C
C
      FINISH
*END
****
```

REFERENCES

Anfinsen, C.B., Haber, E., Sela, M. & White, F.H. (1964). Proc. Nat. Acad. Sci. (Wash.) 47, 1309-1314.

Anfinsen, C.B. (1972). Biochem. J. 128, 737-749.

Anseth, A. & Fransson, L.A. (1970). Acta. Opthal. 48, 455-464.

Astbury, W.T. (1933). Trans. Faraday Soc. 29, 193-211.

Bairati, A. (1972). Bolletino de Zoologica 39, 206-248.

Baldwin, J.P., Boseley, P.G. & Bradbury, E.M. (1975). Nature 253, 245-249.

Banfield, W.G., Lee, C.K. & Lee, C.W. (1973). Arch. Pathol. 95, 262-266.

Bard, J.B.L. & Chapman, J.A. (1968). Nature 219, 1279-1280.

Bear, R.S. (1942). J. Amer. Chem. Soc. 64, 727.

Bear, R.S. (1944). J. Amer. Chem. Soc. 66, 1297-1305.

Bear, R.S. & Bolduan, O.E.A. (1950a). Acta Crys. 3, 230-235.

Bear, R.S. & Bolduan, O.E.A. (1950b). Acta Crys. 3, 236-241.

Bear, R.S. (1951). J. Amer. Leather Chem. Assoc. 46, 438-445.

Bear, R.S. & Bolduan, O.E.A. (1951). J. Appl. Phys. 22, 191-198.

Bear, R.S., Bolduan, O.E.A. & Salo, T.P. (1951). J. Amer. Leather Chem. Assoc. 46, 107-123.

Bear, R.S. (1952). Adv. Prot. Chem. 7, 69-160.

Bear, R.S. & Morgan, R.S. (1957). In 'Connective Tissue' (ed. R.E. Tunbridge), pp.321-333. Blackwell, Oxford.

Beevers, C.A. & Hansen, H.N. (1971). Acta. Crys. B27, 1323-1325.

Bello, J., Harker, D. & De Jarnette, E. (1961). J. Biol. Chem. 236, 1358-1360.

Blundell, T. & Johnson, L.N. (1975) 'Protein Crystallography', Academic Press.

Boedtker, H. & Doty, P. (1956). J. Amer. Chem. Soc. 78, 4267-4280.

Bolduan, O.E.A. & Bear, R.S. (1950). J. Polymer. Sci. 5, 159-168.

Bolduan, O.E.A. & Bear, R.S. (1951). J. Polymer. Sci. 6, 271-284.

Bolduan, O.E.A., Salo, T.P. & Bear, R.S. (1951). J. Amer. Leather Chem. Assoc. 46, 124-137.

Bornstein, P. (1967). Biochemistry 6, 3082-3093.

Bornstein, P. (1969). Biochemistry 8, 63-71.

Bornstein, P. (1974). Ann. Rev. Biochem. 43, 567-604.

Borysko, E. (1963). In 'Ultrastructure of Protein Fibers' (ed. R. Borasky), pp.19-37, Academic Press.

Bowden, J.K., Chapman, J.A. & Wynn, C.H. (1968). Biochim. Biophys. Acta 154, 190-195.

Bowden, J.K. & Chapman, J.A. (1972). Connective Tissue Res. 1, 109-112.

Braun-Falco, O. & Rupec, M. (1964). J. Invest. Dermatol. 42, 15-19.

Brenner, S. & Horne, R.W. (1959). Biochim. Biophys. Acta 34, 103-110.

Brodsky-Doyle, B., Haas, J., Hulmes, D.J.S., Ibel, K., Jenkin, G.T., Miller, A., Timmins, P. & White, J.W. (1975). In 'Proceedings of the 1975 Brockhaven Symposium on Neutron Scattering in Biology' (ed. B.P. Schoenborn).

Bruns, R.R. & Gross, J. (1973). Biochemistry 12, 808-815.

Bruns, R.R., Trelstad, R.L. & Gross, J. (1973). Science 181, 269-271.

Bruns, R.R. & Gross, J. (1974). Biopolymers 13, 931-941.

Burge, R.E. (1965). In 'Structure and Function of Connective and Skeletal Tissue' (ed. S. Fitton-Jackson et al.), pp.2-6, Butterworth, London.

Burge, R.E. (1963). J. Mol. Biol. 7, 213-224.

Burge, R.E. & Randall, J.T. (1955). Proc. Roy. Soc. (Lond.). A 233, 1-16.

Butler, W.T. (1970). Biochemistry 9, 44-50.

Butler, W.T. & Ponds, S.L. (1971). Biochemistry 10, 2076-2081.

Chandross, R.J. & Bear, R.S. (1973). Biophys. J. 13, 1030-1048.

Chapman, J.A. & Armitage, P.M. (1972). Connective Tissue Res. 1, 31-37.

Chapman, J.A. (1974). Connective Tissue Res. 2, 137-150.

Chapman, J.A. & Hardcastle, R.A. (1974). Connective Tissue Res. 2, 151-159.

Chothia, C.H. (1975). Nature 254, 304-308.

Clark, G.L., Parker, E.A., Schaad, J.A. & Warren, W.J. (1935). J. Amer. Chem. Soc. 57, 1509.

Clark, G.L. & Schaad, J.A. (1936). Radiology 27, 339- .

Cochran, W., Crick, F.H.C. & Vand, V. (1952). Acta Crys. 5, 581-586.

Cohen, C. (1954). Ph.D. Thesis, M.I.T., Cambridge, Mass.

Corey, R.B. & Wyckoff, R.W.G. (1936). J. Biol. Chem. 114, 407-414.

Cowan, P.M. & McGavin, S. (1955). Nature 176, 501-503.

Cowan, P.M., McGavin, S. & North, A.C.T. (1955a). Nature 176, 1062-1064.

Cowan, P.M., North, A.C.T. & Randall, J.T. (1955b). Symp. Soc. Exp. Biol. 9, 115-126.

Cox, R.W., Grant, R.A. & Horne, R.W. (1967). J. Roy. Microsc. Soc. 87, 123-142.

Crick, F.H.C. (1953). Acta Crys. 6, 689-697.

Davson, H. (1964). 'A Textbook of General Physiology', Little Brown & Co., Boston.

Dickerson, R.E. & Geis, I. (1969). 'The Structure and Action of Proteins', Harper & Row.

Dickinson, P.M. (1969). M.Sc. Thesis, Manchester University.

Dorner, R.W. (1968). Arch. Biophys. Biochem. 128, 34-39.

Doyle, B.B., Hulmes, D.J.S., Miller, A., Parry, D.A.D., Piez, K.A. & Woodhead-Galloway, J. (1974a). Proc. Roy. Soc. (Lond.) B 187, 37-46.

Doyle, B.B., Hulmes, D.J.S., Miller, A., Parry, D.A.D., Piez, K.A. & Woodhead-Galloway, J. (1974b). Proc. Roy. Soc. (Lond.). B 186, 67-74.

Doyle, B.B., Hukins, D.W.L., Hulmes, D.J.S., Miller, A. & Woodhead-Galloway, J. (1975). J. Mol. Biol. 91, 79-99.

Drake, M.P., Davison, P.F., Bump, S. & Schmitt, F.O. (1966). Biochemistry 5, 301-312.

Ellis, D.O. & McGavin, S. (1970). J. Ultrastructure Res. 32, 191-211.

Epstein, C.J. (1970). In 'Aspects of Protein Synthesis, Part A'. (ed. C.B. Anfinsen), pp.367-432, Academic Press.

Ericson, L.G. & Tomlin, S.G. (1959). Proc. Roy. Soc. (Lond.). A 252, 197-216.

Fietzek, P.P., Kell, I. & Kuhn, K. (1972a). FEBS Lett. 26, 66-68.

Fietzek, P.P., Wendt, P., Kell, I. & Kuhn, K. (1972b). FEBS Lett. 26, 74-76.

Fietzek, P.P., Rexrodt, F.W., Wendt, P., Stark, M. & Kuhn, K. (1972c). Eur. J. Biochem. 30, 163-168.

Fietzek, P.P., Breitkrentz, D. & Kuhn, K. (1974). Biochim. Biophys. Acta 365, 305-310.

Fietzek, P.P. & Kuhn, K. (1975). Int. Rev. Connective Tissue Res. 7, in press.

Fitton-Jackson, S. (1953). In 'Nature & Structure of Collagen'. (ed. J.T. Randall), pp.140-151, Butterworth, London.

Fitton-Jackson, S. & Randall, J.T. (1953). In 'Nature & Structure of Collagen' (ed. J.T. Randall), pp.181-191, Butterworth, London.

Fitton-Jackson, S. (1956). Proc. Roy. Soc. (Lond.). B 144, 556-572

Fitton-Jackson, S. (1968). In 'Treatise on Collagen. Vol. 2B' (ed. B.S. Gould), pp.1-66, Academic Press, New York.

Frankel-Conrat, H. & Williams, R.C. (1955). Proc. Nat. Acad. Sci. (Wash.) 41, 690-698.

Frankel-Conrat, H. & Singer, B. (1959). Biochim. Biophys. Acta 33, 359-370.

Fraser, R.D.B., McRae, T.P. & Miller, A. (1965). J. Mol. Biol. 14, 432-442.

Fraser, R.D.B. & McRae, T.P. (1973). 'Conformation in Fibrous Proteins and Related Synthetic Polypeptides', Academic Press, New York.

Fraser, R.D.B., Miller, A. & Parry, D.A.D. (1974). J. Mol. Biol. 83, 281-283.

French, G.S. (1975). D.Phil Thesis, Oxford University.

Gallop, P.M., Blumenfeld, O.O. & Seifter, S. (1972). Ann. Rev. Biochem. 41, 617-672.

Gelman, R.A. & Blackwell, J. (1973). Connective Tissue Res. 2, 31-35.

Gelman, R.A. & Blackwell, J. (1974). Biochim. Biophys. Acta 342, 254-261.

Gerber, B.R., Routledge, L.M. & Takashima, S. (1972). J. Mol. Biol. 71, 317-337.

Gieseking, R. (1962). Z. Zellforsch. 58, 160-169.

Glimcher, M.J. & Krane, S.M. (1968). In 'Treatise on Collagen, Vol. 2B' (ed. B.S. Gould), pp.67-251, Academic Press, New York.

Goldberger, R.F., Epstein, C.J. & Anfinsen, C.B. (1963). J. Biol. Chem. 238, 628-635.

Goldsack, D.E. & Chalifoux, R.C. (1973). J. Theor. Biol. 39, 645-651.

Gross, J. & Schmitt, F.O. (1948). J. Exp. Med. 88, 555-567.

Gross, J., Highberger, J.H. & Schmitt, F.O. (1954). Proc. Nat. Acad. Sci. (Wash.), 40, 679-688.

Gross, J. (1956). J. Biophys. Biochem. Cytol. 2, (suppl.), 261-273.

Hall, C.E., Jakus, M.A. & Schmitt, F.O. (1942). J. Amer. Chem. Soc. 64, 1234.

Hall, C.E. & Doty, P. (1958). J. Amer. Chem. Soc. 80, 1269-1274.

Hannig, K. & Nordwig, A. (1967). In 'Treatise on Collagen, Vol. 1' (ed. G.N. Ramachandran), pp.73-102, Academic Press, New York.

Harkness, R.D. (1961). Biol. Rev. 36, 399-463.

Hascall, V.C. & Sadjera, S.W. (1969). J. Biol. Chem. 244, 2384-2396.

Highberger, J.H., Gross, J. & Schmitt, F.O. (1950). J. Amer. Chem. Soc. 72, 3321-3322.

Highberger, J.H., Gross, J. & Schmitt, F.O. (1951). Proc. Natl. Acad. Sci. (Wash.), 37, 286-291.

Highberger, J.H., Kang, A.H. & Gross, J. (1971). Biochemistry, 10, 610-615.

Hodge, A.J. & Schmitt, F.O. (1960). Proc. Nat. Acad. Sci. (Wash.), 46, 186-197.

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

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

Hodge, A.J., Petruska, J.A. & Bailey, A.J. (1965). In 'Structure and Function of Connective and Skeletal Tissue' (ed. S. Fitton-Jackson et al.), pp.31-41, Butterworth, London.

Hodge, A.J. (1967), In 'Treatise on Collagen, Vol. 1' (ed. G.N. Ramachandran), pp.185-205, Academic Press, New York.

Hormann, H. & von Wilm, M. (1964). Naturwiss. 51, 464-465.

Hosemann, R., Dreissig, W. & Nemetschek, Th. (1974). J. Mol. Biol. 83, 275-280.

Huxley, H. (1963). J. Mol. Biol. 7, 281-308.

Hulmes, D.J.S., Miller, A., Parry, D.A.D., Piez, K.A. & Woodhead-Galloway, J. (1973). J. Mol. Biol. 79, 137-148.

Hvidt, A. & Nielsen, S.O. (1966). Adv. Prot. Chem. 21, 288-386.

Jackson, D.S. & Bentley, J.P. (1968). In 'Treatise on Collagen, Vol. 2A' (ed. B.S. Gould), pp 189-214, Academic Press, New York.

Jardetsky, O., Thielman, H., Arata, Y., Markley, J.L. & Williams, M.N. (1971). Cold Spring Harbour Symp. Quant. Biol. 36, 257-261.

Kaesberg, P. & Shurmann, M.M. (1953). Biochim. Biophys. Acta 11, 1-6.

Kajikawa, K. (1961). J. Electron Microscopy 10, 1-6.

Kang, A.H., Bornstein, P. & Piez, K.A. (1967). Biochemistry, 7, 788-795.

Kauzmam, W. (1959). Adv. Prot. Chem. 14, 1-64.

Keech, M.K. (1961). J. Biophys. Biochem. Cytol. 9, 193-209.

Kuhn, K., Grassman, W. & Hoffmann, U. (1960). Naturmiss 47, 258-259.

Kuhn, K. & Zimmer, E. (1961a). Naturwiss. 48, 219-220.

Kuhn, K. & Zimmer, E. (1961b). Naturwiss. 48, 220.

Kuhn, K. & Zimmer, E. (1961c). Z. Naturforsch 16b, 648-658.

Kuhn, K., Kuhn, J. & Schuppler, G. (1964). Naturwiss 51, 337.

Kuhn, K., Schuppler, G. & Kuhn, J. (1965). In 'Structure and Function of Connective and Skeletal Tissue'. (ed. S. Fitton-Jackson et al.), pp.64-69. Butterworth, London.

Kuhn, K., Fietzek, P.P. & Kuhn, J. (1966). Biochem. Z. 344, 418-434.

Kuhn, K. (1967). Naturwiss. 54, 101-109.

Kuhn, K. (1969). Essays in Biochemistry 5, 59-87.

Kushner, D.J. (1969). Bact. Rev. 33, 302-345.

Kratky, O. & Sekora, A. (1943). J. Makromol. Chem. 1, 113-121.

Lapière, C.M., Lenaers, A. & Kohn, L.D. (1971). Proc. Nat. Acad. Sci. (Wash.), 68, 3054-3058.

Lauffer, M.A. (1975). 'Entropy Driven Processes in Biology', Springer-Verlag.

Leibovich, S.J. & Weiss, J.B. (1970). Biochim. Biophys. Acta 214, 445-454.

Luse, S.A., Zopf, D. & Cox, J.W. (1963). Anat. Rec. 145, 254-255.

McFarlane, E.F. (1971). Search 2, 171-172.

McLachlan, A.D. & Stewart, M. (1975). Nature 257, 331-333.

Marks, M.H., Bear, R.S. & Blake, C.H. (1949). J. Exp. Zool. 111, 55-78.

Mathews, M.B. & Lozaityte, I. (1958). Arch. Biochem. Biophys. 74, 158-174.

Mathews, M.B. (1965). Biochem. J. 96, 710-716.

Mathews, M.B. & Decker, L. (1968). Biochem. J. 109, 517-526.

Mathews, M.B. (1971). Biochem. J. 125, 37-46.

Mathews, M.B. (1975). 'Connective Tissue', Springer-Verlag.

Maurice, D.M. (1969). In 'The Eye' (ed. H. Davson), 2nd ed. Vol. 1, pp.489-600, Academic Press, New York.

Meyer, K., Davidson, E., Linker, A. & Hoffman, P. (1956). Biochim. Biophys. Acta 21, 506-518.

Miller, A. & Wray, J.S. (1971). Nature 230, 437-439.

Miller, A. & Parry, D.A.D. (1971). J. Mol. Biol. 75, 441-447.

Miller, A., Parry, D.A.D. & Wray, J.S. (1975), in preparation.

Muir, H. (1964). Int. Rev. Connective Tissue Res. 2, 101-154.

Nakao, K. & Bashey, R.I. (1972). Exp. Mol. Pathol. 17, 6-13.

Nemetschek, Th. (1965). Naturwiss. 52, 438-439.

Nemetschek, Th. & Hosemann, R. (1973). Kolloid-Z.u.Z. Polymere 251, 1044-1056.

North, A.C.T., Cowan, P.M. & Randall, J.T. (1954). Nature 174, 1142-1143.

North, A.C.T. (1955). Ph.D. Thesis, University of London.

Nutting, G.C. & Borasky, R. (1948). J. Amer. Leather Chem. Assoc. 43, 96- .

Obrink, B. (1973). Eur. J. Biochem. 33, 387-400.

Offer, G. (1974). In 'Companion to Biochemistry'. (ed. A.T. Bull et al.), pp.623-672. Longman.

Olsen, B.R. (1963). Z. Zellforsch. Mikrosk. Anat. 59, 184-213.

Parry, D.A.D. (1974). Biochem. Biophys. Res. Comm. 57, 216-224.

Pauling, L. (1960). 'The Nature of the Chemical Bond', 3rd. edition, Cornell University Press.

Pauling, L. & Corey, R.B. (1951). Proc. Nat. Acad. Sci. (Wash.), 37, 235-272.

Pérez-Tamayo, R. (1972). Connective Tissue Res. 1, 55-60.

Picken, L. (1960). 'The Organisation of Cells', Clarendon Press, Oxford.

Porter, K.R. & Pappas, G.D. (1959). J. Biophys. Biochem. Cytol. 5, 153-166.

Ramachandran, G.N. & Kartha, G. (1955). Nature 176, 593-595.

Ramachandran, G.N. & Sasisekharan, V. (1956). Arch. Biochem. Biophys. 63, 255-257.

Ramachandran, G.N. (1964). In 'Advanced Methods in Crystallography' (ed. G.N. Ramachandran), p.53, Academic Press, New York.

Ramachandran, G.N. (1967). In 'Treatise on Collagen, Vol. 1', (ed. G.N. Ramachandran), pp.103-184, Academic Press, New York.

Rattew, C.J. (1974). D.Phil. Thesis, Oxford University.

Rauterberg, J., Fietzek, P.P., Rexrodt, F., Becker, U., Stark, M. & Kuhn, K. (1972). FEBS Lett. 21, 75-79.

Rich, A. & Crick, F.H.C. (1955). Nature 176, 915-916.

Rich, A. & Crick, F.H.C. (1961). J.Mol. Biol. 3, 483-506.

Rodbard, S. (1970). Persp. Biol. Med. 14, 507-527.

Ross, R. & Benditt, E.P. (1961). J. Biophys. Biochem. Cytol. 11, 677-700.

Rossmann, M.G. (1960). Acta Crys. 13, 221-226.

Rougvie, M.A. & Bear, R.S. (1953). J. Amer. Leather Chem. Assoc. 48, 735-751.

Rubin, A.L., Drake, M.P., Davison, P.F., Pfahl, D., Speakman, P.T. & Schmitt, F.O. (1965). Biochemistry 4, 181-190.

Salem, G. & Traub, W. (1975). in press.

Sasisekharan, V. & Ramachandran, G.N. (1957). Proc. Indian Acad. Sci. A45, 363-376.

Scheraga, H.A. (1963). In 'The Proteins, Vol. 1' (ed. H. Neurath), pp.478-594, Academic Press.

Schmitt, F.O., Hall, C.E. & Jakus, M.A. (1942). J. Cell. Comp. Physiol. 20, 11-33.

Schmitt, F.O., Gross, J. & Highberger, J.H. (1953). Proc. Nat. Acad. Sci. (Wash.), 39, 459-470.

Schmitt, F.O., Gross, J. & Highberger, J.H. (1955). Exp. Cell. Res. Suppl. 3, 326-334.

Schmitt, F.O., Levine, L., Drake, M.P., Rubin, A.L., Pfahl, D. & Davison, P.F. (1964). Proc. Nat. Acad. Sci. (Wash.), 51, 493-497.

Schoenborn, B.P. (1975) (ed.) 'Proceedings of the 1975 Brockhaven Symposium on Neutron Scattering in Biology'.

Segrest, J.P. & Cunningham, L.W. (1971). Nature New Biol. 234, 26-28.

Segrest, J.P. & Cunningham, L.W. (1973). Biopolymers 12, 825-834.

Sela, M., White, F.H. & Anfinsen, C.B. (1957), Science 125, 691-692.

Serafini-Fracassini, A. & Smith, J.W. (1966). Proc. Roy. Soc. (Lond.) B 165, 440-449.

Smith, J.W. (1965). Nature 205, 356-358.

Smith, J.W. (1968). Nature 219, 157-158.

Spiro, R.G. (1969). J. Biol. Chem. 244, 602-612.

Spiro, R.G. (1972). In 'Glycoproteins, 2nd Ed.' (ed. A. Gottschalk) pp.964-999, Elsevier, Amsterdam.

Stark, M., Miller, E.J. & Kuhn, K. (1972). Eur. J. Biochem. 27, 192-196.

Stokes, A.R. (1955). Acta Crys. 8, 27-29.

Stone, D., Sodek, J., Johnson, P. & Smillie, L.B. (1974). Proc. 9th FEBS Mtg. Budapest, in press.

Tanzer, M.L. (1973). Science 180, 561-566.

Tanzer, M.L., Church, R.L., Yaeger, J.A., Wampler, D.E. & Park, E.-D. (1974) Proc. Nat. Acad. Sci. (Wash.) 71, 3009-3013.

Tomlin, S.G. (1955). Proc. Int. Wool Textile Res. Conf. Australia, Part B, p.187.

Tomlin, S.G. & Worthington, C.R. (1956). Proc. Roy. Soc. (Lond.), A 235, 189-201.

Tomlin, S.G. & Ericson, L.G. (1960). Acta Crys. 13, 395-402.

Toole, B.P. & Lowther, D.A. (1968). Biochem. J. 109, 857-866.

Torp, S., Baer, E. & Friedman, B. (1974). Proceedings of the 26th. Colston Symposium, Bristol University. 'Structure of Fibrous Biopolymers', Butterworth, London.

Traub, W. & Piez, K.A. (1971). Adv. Prot. Chem. 25, 243-252.

Tromans, W.J., Horne, R.W., Gresham, G.A. & Bailey, A.J. (1963). Z. Zellforsch. Mikrosk. Anat. 58, 798-802.

Veis, A., Anesey, J. & Mussell, S. (1967). Nature 215, 931-934.

Veis, A., Bhatnagar, R.S., Shuttleworth, C.A. & Mussell, S. (1970) Biochim. Biophys. Acta 200, 97-112.

Veis, A. & Yuan, L. (1975). Biopolymers 14, 895-900.

Venetianer, P. & Straub, F.B. (1963). Biochim. Biophys. Acta 67, 166-168.

von der Mark, K., Wendt, P., Rexrodt, F. & Kuhn, K. (1970). FEBS Lett. 11, 105-108.

von Hippel, P.H. (1974). In 'Treatise on Collagen, Vol. 1' (ed. G.N. Ramachandran), pp.253-338, Academic Press, New York.

Wang, S.K. & Worthington, C.R. (1975). Biophys. J. 15, 322a.

Weiss, P.A. (1956). Proc. Nat. Acad. Sci. (Wash.), 42, 819-830.

Weiss, P.A. (1959). Rev. Mod. Phys. 31, 11-20.

Wendt, P., von der Mark, K., Rexrodt, F. & Kuhn, K. (1972). Eur. J. Biochem. 30, 169-183.

White, F.H. (1961). J. Biol. Chem. 236, 1353-1358.

Wolpers, C. (1943). Klin. Wochshr. 32, 624- .

Wood, W.B. & Edgar, R.S. (1967). Scientific American, July.

Woodhead-Galloway, J., Hukins, D.W.L. & Wray, J.S. (1975). Biochem. Biophys. Res. Comm. 64, 1237-1244.

Woolfson, M.M. (1970). 'X-ray Crystallography', Cambridge University Press.

Wray, J.S. (1972). D.Phil. Thesis, Oxford University.

Wyckoff, R.W.G., Corey, R.B. & Biscoe, J. (1935). Science, 82, 175-176.

Yee, R.Y., Englander, S.W. & von Hippel, P.H. (1974). J. Mol. Biol. 83, 1-16.

LIST OF PUBLICATIONS

1. 'Analysis of the Primary Structure of Collagen for the Origins of Molecular Packing'.

   D.J.S. Hulmes, A. Miller, D.A.D. Parry, K.A. Piez & J. Woodhead-Galloway.

   J.Mol.Biol. (1973) 79, 137-148.

2. 'A D-periodic narrow filament in Collagen'

   B.B. Doyle, D.J.S. Hulmes, A. Miller, D.A.D. Parry, K.A. Piez & J. Woodhead-Galloway.

   Proc. Roy. Soc. (Lond.) (1974) B 186, 67-74.

3. 'Axially projected Collagen Structures'

   B.B. Doyle, D.J.S. Hulmes, A. Miller, D.A.D. Parry, K.A. Piez & J. Woodhead-Galloway.

   Proc. Roy. Soc. (Lond.) (1974) B187, 37-46.

4. 'Origins and Implications of the D stagger in Collagen'

   B.B. Doyle, D.W.L. Hukins, D.J.S. Hulmes, A. Miller, C.J. Rattew & J. Woodhead-Galloway.

   Biochem.Biophys. Res.Comm. (1974) 60, 858-864.

5. 'Collagen Polymorphism: Its Origins in the Amino Acid Sequence'.

   B.B. Doyle, D.W.L. Hukins, D.J.S. Hulmes, A. Miller & J. Woodhead-Galloway.

   J.Mol.Biol. (1975) 91, 79-99.

6. 'Neutron Scattering by Collagen'

   B.B. Doyle, J. Haas, D.J.S. Hulmes, K. Ibel, G.T. Jenkin, A. Miller, P. Timmins & J.W. White

   submitted to the Proceedings of the 1975 Brookhaven Symposium entitled 'Neutron Scattering for the Analysis of Biological Structures'.