A Cross-correlation of Amino Acid Sequence Data to Explain Fibril Formation in a Number of Collagen Subtypes

Thesis submitted to Cardiff University for the degree of

Doctor of Philosophy

Cristian Pinali

Structural Biophysics Research Group
School of Optometry and Vision Sciences
Cardiff University

January 2008

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Collagen is the most abundant protein in vertebrates. In particular, fibrillar collagens are found in bone, tendon, dermis, ligament, cartilage, cornea and blood vessel walls to mention but a few. Fibrillar collagens can be thought of as rigid linear structures and therefore are relatively simple to study from a theoretical point of view.

In this thesis, a study on fibrillar collagen packing is presented. Since these collagens can be represented by a linear sequence of elements, it is possible to represent the hydrophobic or the electrostatic interactions between two collagen molecules by means of a cross correlation function. In fact, in this way we represent how many amino acids with a given property face each as a function of the displacement between the two sequences they belong to. Since cross correlation can be thought of as a scoring of the interactions between two adjacent molecules we will refer to it as the scoring method. In the first instance, the scoring method was applied to a linear amino acids sequence of type I collagen to probe the hydrophobic and attractive electrostatic interactions acting on it. Then, the method was applied systematically to type II, type III type XI collagens and to mixtures of type II and type III collagens.

The appearances that these collagens would have in an electron microscope were calculated by a simulated staining method, and a comparison between real collagen fibrils and the models built in accordance to the findings of the scoring method was carried out.

The scoring method allowed us to predict the correct stagger for collagen fibrils in parallel configuration. It also allowed us to predict and explain the correct axial stagger between type I collagen fibrils oriented in an antiparallel fashion.

The scoring method was also used to explain oblique banding patterns found in reconstituted type II collagen fibrils.

When applied to type II and type III collagen fibrils oriented in an antiparallel fashion, it shed light on a possible supercoiled structure.

The validity of the scoring method was confirmed by its comparison with real collagen fibrils and, in principle, it could be extended to all proteins that can be represented as a simple linear sequence of amino acids.

Acknowledgements

I want to thank Dr. Carlo Knupp for his profound patience with me and his genuine interest in my work.

I would like to thank Professor Tim Wess for his guidance and availability during my thesis work.

I would like to thank Dr. Rob Young for his micrographs that let us confirm the validity of the scoring method and make some interesting discoveries.

I would like to thank Dr. Anne Vaughan-Thomas and Prof. Vic Duance for their samples

I would also like to thank Dr. Clark Angus Maxwell for his friendship.

Contents

Abstract 1

Acknowledgements 3

CHAPTER 1 INTRODUCTION 24

1 Introduction 25
1.1 Amino acids 25

1.2 Polypeptides 29

1.3 Molecular interactions 31

1.3.1 Electrostatic forces 31

1.3.2 Hydrophobic forces 35

1.3.3 Hydrogen bond 36

1.3.4 Van der Waal's forces 36

1.4 Conformation of polypeptide chains 39

1.4.1 Polyproline 39

1.5 Collagen general features 41

1.5.1 The Rich and Crick model 44

1.5.2 The Ramachandran model 45

1.5.3 The Okuyama model 47

1.6 Collagen in the extracellular matrix 49

1.7 Biosynthesis of collagens 50

1.8 Collagen families 51

1.8.1 Fibril-forming collagens 52

1.8.2 Fibril-associated collagens with interrupted triple helices (FACIT) 54

1.8.3 Network-forming collagens 55

1.8.4 Anchoring fibrils collagens 57

1.8.5 Transmembrane collagens 58

1.8.6 Multiplexins 59

1.8.7 Type XXVI and XXVIII collagens 59

1.9 Supramolecular structure of collagen 60

1.10 Collagen fibril formation 62

1.11 Telopeptides conformation and their role in collagen fibril formation 68

1.12 Molecular packing in fibrillar collagen as a first step toward the understanding of protein interactions 69

CHAPTER 2 MATERIALS AND METHODS 70

2 Introduction 71

2.1 Cross Correlation 72

2.2 Autocorrelation 73

2.3 Convolution 77

2.4 Fourier Transforms and Convolution Theorem 79

2.5 Transmission Electron Microscope 83

2.5.1 Illuminating unit 85

2.5.2 Specimen holder unit 85

2.5.3 Imaging unit 86

2.5.4 Image formation unit 87

2.6 Specimen preparation 87

2.7 Sectioning 90

2.8 Positive staining 91

2.8.1 Phosphotungstic Acid 92

2.8.2 Uranyl Acetate 93

2.9 Simulated staining 93

2.10 Sample preparation 97

2.11 Software applications 98

2.12 Collagen sequences. 99

CHAPTER 3 MODELLING FIBRILLAR COLLAGENS 100

3 Introduction 101

3.1 Scoring method as a means to study interactions between collagen molecules 101

3.2 Hydrophobic autocorrelation for murine type I collagen triple helix 105

3.3 Multiple parallel hydrophobic autoscoring 122

3.4 Multiple Parallel Hydrophobic interaction for type I collagen with telopeptide contribution 128

3.5 Distribution of the hydrophobic amino acids along the molecule 134

3.6 Multiple parallel electrostatic interaction for collagen sequence with folded telopeptides 142

3.7 Antiparallel hydrophobic scoring 149

3.8 Antiparallel electrostatic scoring for type I collagen 161

3.9 Discussions about the scoring method 165

3.9.1 Relationship between scoring method and collagen packing 165

3.9.2 A model representing the hydrophobic amino acids distribution 169

3.9.3 Electrostatic contribution to collagen fibril formation 171

3.9.4 Observations relative to the antiparallel aggregation 172

3.10 Concluding remarks for chapter 3 173

CHAPTER 4 VALIDATING THE FIBRILLAR COLLAGEN MODELS .. 174

4 Introduction 175
4.1 Modelling collagen microfibrils 176

4.2 Comparing experimental data and a model of an antiparallel fibril aggregation of type I collagen 180

4.3 Comparison between modelled parallel type II collagen microfibrils and experimentally obtained microfibrils 188

4.3.1 An oblique banding pattern for type II rat collagen microfibrils 194

4.4 A model for type III collagen fibrils 203

4.5 A special case: antiparallel aggregation of reconstituted fibrils of type II and type III collagen 208

4.6 Comparison between model and experimental fibrils of type XI collagen 221

4.7 Comparison of the hydrophobic scoring for type I, II, III and XI collagen 228

4.8 Discussions for chapter four 230

4.8.1 Validation of predictions 230

4.8.2 Antiparallel packing 231

4.8.3 Diagonal banding 234

4.8.4 Parallel fibril aggregation 236

4.9 Concluding remarks for chapter 4 238

CHAPTER 5 CONCLUSIONS 239

Final remarks 240
Further steps 243

APPENDIX 244

Appendix 245

Cross_correlation.cpp.....246

Interpolation.cpp.....248

Fast Fourier_Transform.cpp.....249

Fft.h.....251

Arctan.h.....253

REFERENCES 254

Articles and books 254

Websites 265

List of figures

Figure 1.1: Schematic representation of an amino acid (Figure adapted from Darnell et al., 1990).....26

Figure 1.2: Representation of the side chains of the 20 fundamental amino acids. Three letters and single letter abbreviation are shown as well. (from Creighton, 1997).....27

Figure 1.3: Schematic representation of resonance in the peptide bond.....30

Figure 1.4: Repulsive and attractive energy profiles of the Van der Waal's interactions as a function of distance between the center of the atoms. Their sum gives the Lennard-Jones potential. (from Creighton, 1997) 38

Figure 1.5 Polyproline II left handed helix (Adapted from Stapley and Creamer, 1999).....40

Figure 1.6: Idealised representation of a collagen molecule. The three α-chains are in different colours.....42

Figure 1.7: Representation of the Rich and Crick (10/3) collagen model (adapted from www.tuat.ac.jp/~x-ray/Research/R_2-models.html).....45

Figure 1.8: Representation of the Ramachandran et al. collagen model (Adapted from Ramachandran, 1967).....46

Figure 1.9 Representation of the Okuyama et al. (7/2) collagen model (adapted from www.tuat.ac.jp/~x-ray/Research/R_2-models.html).....48

Figure 1.10: Model for the binding of type IX collagen to type II collagen. Black lines represent triple helical domains of collagens II and XI. Circles indicate where the triple helical domain is interrupted. (from: Van der Rest and Mayne, 1988)....55

Figure 1.11: Model for lateral aggregations of collagen type VI as proposed by Knupp et al. 2002(a) (From Knupp et al. 2002(a)).....57

Figure 1.12: Diagrammatic representation of a collagen fibril. Molecules join together with a D-stagger (234 a.a.). The repetition of five molecules joined in such a way constitutes what we refer to as microfibril.....64

Figure 2.1: Diagrammatic representation of the Convolution Theorem. A periodic function is given by the convolution of a motif with a δ-comb. The Fourier transform of the periodic function is equal to the multiplication of the Fourier transform of the two convoluted functions 82

Figure 2.2 Comparison between transmission electron microscope (TEM) and an optical projection microscope (LM).

(Adapted from www.uiowa.edu/~cemrf/methodology/tem/index.htm).....84

Figure: 2.3 Diagrammatic representation of the simulated staining procedure for a portion of type I collagen microfibril.....96

Figure 3.1: Steps undertaken to build a type I collagen sequence for hydrophobic residues. The type I collagen triple helix is inserted in an array, the value 1 is assigned to hydrophobic residues (F, I, L, M, V, Y), the value 0 is assigned to the other residues. Residues on the same row are finally summed together.....103

Figure 3.2: Hydrophobic scoring function for two type I murine collagen triple helices. To facilitate the graph's interpretation the peak at position 0 a.a. is not shown.....106

Figure 3.3: Fourier transform of the hydrophobic scoring function for two murine type I collagen triple helices.....107

Figure 3.4: Gaussian smoothing of the normalised hydrophobic scoring function for murine type I collagen.....108

Figure 3.5: All major peaks of the hydrophobic scoring function are related by a 234 a.a. periodicity. Peaks in green can be thought of as corresponding to the interactions felt by two molecules when they slide toward each other. Peaks in red correspond to the interactions feld by molecules sliding away from each other...111

Figure 3.6: Mirror-like distribution of the peaks about position 116.5. Corresponding peaks are indicated with the same letter.....112

Figure 3.7: Periodic regions are delimited by red lines at positions 0, 233, 465, 701, and 933 a.a.. Planes of symmetry at positions 116.5, 349, 583 and 817 a.a. are in

green.....112

Figure 3.8: Comparison between the real peak distribution of the hydrophobic scoring function (red) and the "ideal" peaks distribution (green). The "ideal" distribution was normalised to make the comparison clearer.....116

Figure 3.9: Fourier analysis of the measured peak positions as shown in table3.3. The major frequency peak visible is at about 39 a.a.. A second peak is visible in position 21 a.a. No peak is visible at position 234 a.a.....117

Figure 3.10: Fourier analysis of the measured peaks as shown in table 3.3 between 0 and 233 a.a.....118

Figure 3.11: Comparison between the Fourier transform of the motif between 0 and 234 a.a. and a δ-comb with a periodicity of 234 a.a.. No peak at 234 a.a. can be seen in the Fourier transform of the scoring function because the transform of the motif has a zero in that position.....119

Figure 3.12: Comparison between the first half and the second half of region of the scoring function 1. From about 82 a.a. the two curves have a very similar peaks distribution.....120

Figure 3.13: Comparison between the Fourier transform of the motif between position 0 and 234 a.a. (blue) and the Fourier transform of half motif i.e. between. position 0 and 117 a.a. (green) The Fourier transform of the δ two δ-function separated by about 117 a.a. is shown (red).....121

Figure 3.14: Comparison between the Fourier transform of the motif of the hydrophobic autoscoring function between 0 and 234 a.a. shift (green) and the

product of the Fourier transform of the two δ-functions 117 a.a. apart with the Fourier transform of half the motif (red).....122

Figure 3.15: Diagram representing multiple parallel hydrophobic interactions....123

Figure 3.16: Parallel multiple hydrophobic scoring function of a molecule of murine type I collagen without telopeptides against a sequence of similar molecules separated by a gap of 154 amino acids. Periodic regions are delimited by red lines. Planes of symmetry are delineated with green lines.....126

Figure 3.17: Diagrammatic representation of a collagen fibril. The microfibril making up the fibril is coloured in green.....128

Figure 3.18: N-telopeptide configuration for murine type I collagen. Residues in the telopeptides are in green. Lysines responsible for cross-links are in red.....129

Figure 3.19: C-telopeptide configuration for murine type I collagen. Residues in the telopeptide are in green. Lysines responsible for cross-links are in red.....130

Figure 3.20: Multiple parallel hydrophobic scoring for a collagen molecule with straight telopeptides. Periodic regions are delimited by red lines. Planes of symmetry are delimited by green lines.....132

Figure 3.21: Multiple parallel hydrophobic scoring for a collagen molecule with folded telopeptides. Periodic regions are delimited by red lines. Planes of symmetry are delineated by green lines.....132

Figure 3.22: Comparison among the parallel hydrophobic autoscoring for the murine type I collagen molecule for different telopeptide configurations. Blue: no telopeptides; red: straight telopeptides; green: folded telopeptides.....134

Figure 3.23: Gaussian smoothing of the axial hydrophobic amino acid distribution for the configuration with folded telopeptides. Hydrophobic residues clusters are distributed in a similar way inside the periodic regions.....135

Figure 3.24: Comparison between the real hydrophobic interaction (red) and the model-generated hydrophobic interaction (green) as a. function of the axial stagger (a.a.). A constant value was added to the real hydrophobic interaction (red) to make the plot less cluttered.....137

Figure 3.25: Comparison between a portion of the multiple hydrophobic scoring function for the real amino acid distribution (red) and the model distribution (green).....139

Figure 3.26: Diagram representing different axial staggers for the modelled hydrophobic sequence.....140

Figure 3.27: Comparison between murine type I collagen residue distribution and the modelled hydrophobic residue distribution (green lines) according to the {42, 42, 42, 42, 66} distribution. F (magenta), M (yellow), Y (red).....141

Figure 3.28: Electrostatic scoring function of a positive charge sequence representing one molecule, against a negative charge sequence representing a multiple sequence of molecules.....144

Figure 3.29: Electrostatic scoring function of a negative charge sequence representing one molecule, against a positive charge sequence representing a multiple sequence of molecules.....144

Figure 3.30: Multiple attractive electrostatic scoring function for murine type I

collagen molecule. Regions of periodicity are delimited in red, mirror-planes are in green.....145

Figure 3.31: Point by point multiplication between the hydrophobic and the attractive electrostatic scoring functions for murine type I collagen. This operation represents the cooperative effect of both forces when they act together.....148

Figure 3.32:A sequence of hydrophobic residues according to the {42, 42, 42, 42, 66}*4+{42,42} distribution and oriented from the C- to the N-terminal is shifted with respect to an identical sequence oriented form the N-to the C-terminal. The cross-correlation is highest when the shift is equal to 42*2 a.a.....150

Figure 3.33: Hydrophobic scoring function for two antiparallel collagen molecules with folded telopeptides.....151

Figure 3.34: Multiple antiparallel hydrophobic scoring functions for molecules with folded telopeptides.....154

Figure 3.35: Comparison between the multiple parallel (green) and antiparallel (red) hydrophobic scoring functions.....154

Figure 3.36: Possible cross-link formation between two murine type I collagen molecules in an antiparallel configuration with -2 a.a.-stagger. a) cross-links can arise between telopeptides. b) cross-links can arise between lysines belonging to the triple helix. Residues in the telopeptides are in green, putative lysines responsible for cross-links are in red.....156

Figure 3.37: Cross-links formation between two murine type I collagen molecules in antiparallel configuration with an 80a.a. stagger. Residues in the telopeptides are

in green. Putative lysines responsible for cross-links are in red.....157

Figure 3.38: Optimised telopeptide folding configuration that maximises lysine contacts for the antiparallel interaction case. The residues in the telopeptide are in green. Lysines responsible for cross-links are in red.....158

Figure 3.39: Multiple antiparallel hydrophobic scoring. for molecules with telopeptides in the configuration shown in figure 3.38.....159

Figure 3.40: Cross-link formation between two murine type I collagen molecules in an antiparallel configuration with an 86 a.a. stagger. Residues in the telopeptides are in green. Putative lysines responsible for cross-links are in red. The telopeptides are folded according to figure 3.38.....160

Figure 3.41: Electrostatic scoring function due to one collagen molecule which is positively charged against a multiple sequence of negatively charged molecules in an antiparallel configuration.....162

Figure 3.42: Electrostatic scoring function due to one collagen molecule which is negatively charged against a multiple sequence of positively charged molecules in an antiparallel configuration.....162

Figure 3.43: Multiple antiparallel attractive electrostatic interaction as a function of the distance.....163

Figure 3.44: Point by point multiplication between the hydrophobic and the attractive electrostatic antiparallel scoring functions for murine type I collagen...164

Figure 3.45: Diagram representing the stagger for the antiparallel aggregation obtained when a single collagen molecule is scored against a multiple sequence of

molecules.....172

Figure 4.1 Scheme representing the type I and type XI collagen molecule models used in this chapter.....178

Figure 4.2: a) portion of type I collagen microfibril. Its hydrophobic residues are represented by the value 1 in b) and the lateral sum of the hydrophobic values in c).....181

Figure 4.3: Antiparallel hydrophobic scoring function for two type I murine collagen modelled microfibrils as a function of the distance in a.a.....182

Figure 4.4: Antiparallel attractive electrostatic scoring function for two type I murine collagen modelled microfibrils as a function of the distance in a.a.....182

Figure 4.5: Diagrammatic representation of an inverted molecule with an 86 a.a. axial stagger with respect to a forward sequence of molecules.....183

Figure 4.6: Comparison between the axial stagger of two antiparallel fibres. The molecule in the inverted fibril has an 86 a.a. stagger with respect to the corresponding molecule in the forward sequence. This means that the distance of the forward molecule from the first amino acid of the inverted microfibril is (140-86)=54 a.a.. Since the periodic subunit of the microfibril is 234 a.a. long, the distance between the inverted microfibril and the forward microfibril is (234-54)= 180 a.a. long. The same approach can be used to find the correspondence between the peaks at 4 and 43 a.a. positions in graph 3.39 and the peaks at 97 and 138 a.a. positions in graph 4.3.....183

Figure 4.7: Total attractive interaction (scoring function) for two type I collagen

antiparallel.....185

Figure 4.8: Portions of murine type I collagen in an antiparallel configuration with a 181 a.a. stagger. The lysines responsible for cross-link formation are in red. Lysines belonging to the parallel microfibril can form cross-links with lysines of the antiparallel microfibril and *vice versa* because they face each other. Residues in the telopeptides are in green.....186

Figure 4.9: Antiparallel aggregation for a type I rat tail collagen microfibril reconstituted in vitro and positively stained with UA. Taken from Bruns, (Bruns, 1976). The outlined portion is the one used for measurements.....186

Figure 4.10: Comparison between the density profile of a real type I collagen antiparallel microfibril (in red) and the modelled type I antiparallel collagen microfibril with a 181 a.a. staggering (in green).....187

Figure 4.11: Normalised, Gaussian-smoothed, parallel hydrophobic scoring function for two type II collagen molecules without telopeptides.....188

Figure 4.12: Normalised, Gaussian-smoothed, parallel attractive electrostatic scoring function for two type II collagen molecules without telopeptides.....189

Figure 4.13: Folding configuration used for murine type II collagen telopeptides. Lysines responsible for cross-link formations are coloured in red. Residues belonging to the telopeptides are in green.....190

Figure 4.14: Portion of the modelled murine type II collagen microfibril containing telopeptides. Residues belonging to the telopeptides are in green, lysines responsible for cross-links formation are in red.....192

Figure 4.15: Micrograph of a portion of reconstituted type II collagen fibrils from rat chondrosarcoma. The outlined portion was used to calculate the density profile. 193

Figure 4.16: Comparison between the density profile of a real type II collagen microfibril and the simulated density profile of a modelled type II collagen microfibril (a constant value is added to the experimental profile for clarity).....193

Figure 4.17: Portion of a type II collagen fibril: the region analysed is outlined in red. Some diagonal bands are highlighted by green arrows, and some orthogonal banding is highlighted by vertical red arrows.....194

Figure 4.18: Smoothed density profile of the portion outlined in figure 4.16.....195

Figure 4.19: Hydrophobic scoring function for two parallel modelled type II collagen microfibrils.....196

Figure 4.20: Electrostatic attractive scoring function for two parallel modelled type II collagen microfibrils.....197

Figure 4.21: Total attractive interaction (scoring function) for two parallel modelled type II collagen microfibrils.....198

Figure 4.22: Five type II collagen microfibril with a 39 a.a. stagger. Lysines responsible for cross-link formation are in red. Cross-links can arise between lysines belonging to different microfibrils.....199

Figure 4.23 Five type II collagen microfibril with a 86 a.a. stagger. Lysines responsible for cross-link formation are in red. Cross-links can arise between lysines belonging to different microfibrils.....200

Figure 4.24: Comparison between the density profile of type II collagen fibril (red) and the simulated fibrils made of microfibrils with a 39 a.a. stagger (green).....201

Figure 4.25: Comparison between the density profile of type II collagen fibril (red) and the simulated fibrils made of microfibrils with a 86 a.a. stagger (green).....202

Figure 4.26 Normalised Gaussian-smoothed hydrophobic parallel scoring function for murine type III collagen.....203

Figure 4.27: Normalised Gaussian-smoothed attractive electrostatic scoring function for two murine type III mouse collagen molecules.....204

Figure 4.28: Folded conformation used for the type III collagen telopeptides. Residues belonging to the telopeptides are coloured in green, while lysines responsible for cross-link formation are in red. Note that the C-telopeptides fold around an A-G or a G-G pair. G and A are the smallest amino acids so that steric hindrance is reduced to a minimum there.....205

Figure 4.29: Portion of a fibril of type III collagen with the section used for our measurements outlined in red.....206

Figure 4.30: Comparison between the simulated density profile of a modelled microfibril of type III collagen (green) and the density profile of the portion of type III collagen fibril outlined in figure 4.26 (red).....207

Figure 4.31 Diagrammatic representation of a periodic structure associated in an antiparallel manner.....208

Figure 4.32: : Portion of an antiparallel fibril made of a mixture of type II and type III collagen in 1:1 proportion in stock.....209

Figure 4.33: Fourier filtered density profile of the portion of fibril outlined in figure 4.30. A family of mirror-planes is drawn in red.....209

Figure 4.34: Comparison between the simulated density profile of a type III collagen microfibril (red) and a collagen type II microfibril (green). Best agreement is reached when type II microfibril density profile is shifted twelve a.a. forward with respect to type III microfibril.....211

Figure 4.35: Antiparallel hydrophobic scoring function for type III collagen microfibril against a type II collagen microfibril.....212

Figure 4.36: Attractive electrostatic scoring function for a type III collagen microfibril against a type II collagen microfibril.....212

Figure 4.37: Total attractive interaction (scoring function) for type II and type III collagen microfibril in an antiparallel orientation.....213

Figure 4.38: Comparison between modelled antiparallel microfibrils of type III and type II collagen in 1:1 proportion with a 115 a.a. stagger (green) and the real microfibril (red).....214

Figure 4.39 Comparison between modelled antiparallel microfibrils of type III and type II collagen in 1:1 proportion with a 200 a.a. stagger (green) and the real microfibril (red).....215

Figure 4.40: Comparison between modelled antiparallel microfibrils of type III and type II collagen in 1:1 proportion with a 178 a.a. stagger (green) and the real microfibril (red).....216

Figure 4.41: Representation of the supercoiling between type III and type II

antiparallel collagen fibrils. Type II collagen wounds around type III collagen with an 8° angle. The density profile of the experimentally obtained antiparallel fibril is given by the density profile of the type III collagen microfibril and that of the axially projected type II collagen microfibril combined together.....217

Figure 4.42: Antiparallel hydrophobic scoring function for type III and type II collagen microfibril with an 8° coiling angle.....219

Figure 4.43: Antiparallel attractive electrostatic scoring for type III and type II collagen microfibril with an 8° coiling angle.....219

Figure 4.44: total attractive interaction for an antiparallel microfibril with type two collagen wound around type III collagen with an 8° coiling angle.....220

Figure 4.45: Comparison between simulated staining for an antiparallel supercoiled fibril with an 8° angle (green) density profile and the density profile of a real antiparallel fibril (red).....221

Figure 4.46: Hydrophobic scoring function for two parallel type XI collagen molecules without telopeptides.....223

Figure 4.47: Attractive electrostatic scoring function for two type XI parallel collagen molecules without telopeptides.....224

Figure 4.48 Combined attractive interaction (scoring function) for two type XI parallel collagen molecules without telopeptides.....224

Figure 4.49: Folding configuration used for murine type XI collagen telopeptides. Lysines responsible for cross-link formations are coloured in red. Residues belonging to the telopeptides are in green.....225

Figure 4.50: Portion of the modelled murine type XI collagen microfibril containing telopeptides. Residues belonging to the telopeptides are in green, lysines responsible for cross-links formation in red.....226

Figure 4.51: Portion of a type XI reconstituted microfibril with the region used in this study outlined in red.....227

Figure 4.52: Fourier-filtered density profile of the portion outlined in red in figure 4.51.....227

Figure 4.53: Comparison between the Fourier-filtered density profile of real fibril (red) and the simulated density profile of the 234 a.a. stagger modelled microfibril (green).....228

Figure 4.54: Comparison of the hydrophobic scoring function for type I collagen (green), type II collagen (red) type III collagen (black) and type XI collagen (blue).....229

Figure 4.55: Diagram for type I collagen antiparallel aggregation.....232

Figure 4.56: Comparison between a real collagen fibril with an oblique banding and the modelled fibril made of parallel microfibril with a 39 a.a. stagger.....236

List of tables

Table 1.1: Cross-linking positions for murine type I, II, XI and for bovine type III collagens.....67

Table 3.1: Peaks positions outlined in graph 3.5 and their relation to the 234 a.a. periodicity.....109

Table 3.2: Peaks positions outlined in graph 3.5 and their relation to the 234 a.a. periodicity.....110

Table 3.3: Comparison between measured peak positions and estimated ones for the murine type I collagen scoring function, based on the mirror like distribution about the central axis of each periodic region. Peaks positions delimiting periodic regions are coloured in yellow.....116

Table 3.4: Axial positions of hydrophobic residue clusters in murine type I collagen molecules. The distance between corresponding amino acids positions in different regions is also shown.....136

Table 3.5: Axial distribution of the hydrophobic residue clusters and the relative distance between successive peaks. Regions of periodic distributions are coloured in yellow.....137

Table 3.6: Comparison between estimated peak positions and their real positions for the total electrostatic attractive scoring function. Periodic regions are in yellow.....147

Table 3.7: Comparison between estimated positions and real peak positions for the antiparallel hydrophobic scoring function.....152

CHAPTER 1

INTRODUCTION

1 Introduction

In this chapter, we introduce the background that is useful to understand the work presented in this thesis. Since we are going to present a work on fibrillar collagen aggregation, we begin by giving a brief description of amino acids properties and collagen molecules. The forces between amino acids are briefly described in order to understand protein interactions. These drive amino acid aggregation and polypeptide conformations.

A section is dedicated to fibrillar collagens and to other collagen families. A brief description of collagen fibril aggregation is given with particular attention to the role of hydroxylysines in cross-link formation and to the conformation of the telopeptides.

1.1 Amino acids

Nature uses only about 20 different amino acids to build all the proteins found in animals. Chemically, all amino acids share a common structure: a central carbon atom, called the α carbon or c_α, which is linked to an amino group (NH_2), to a carboxyl group (COOH), to a hydrogen atom (H) and to a side chain (R), which is the only part that varies from amino acid to amino acid. Among the amino acids, proline is a special one because its side chain links back to the imino group -NH-. Under physiological conditions, the carboxyl group and the amino group are ionized as COO^- and NH_3^+ (Darnell et al., 1990).

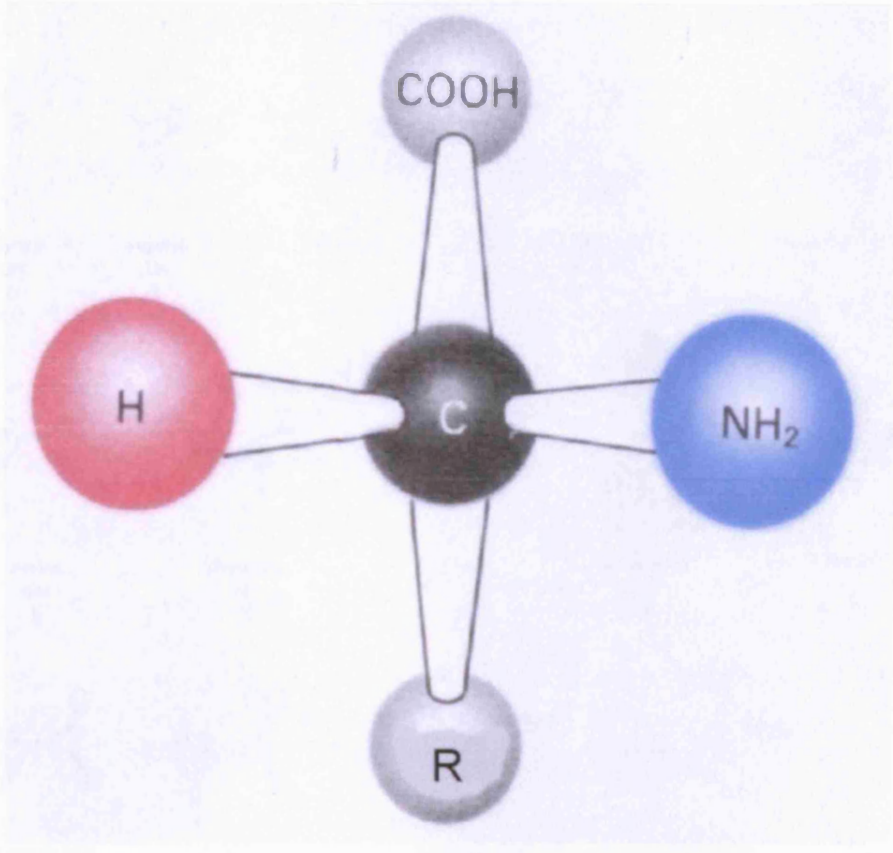


Figure 1.1: Schematic representation of an amino acid (Figure adapted from Darnell et al., 1990)

Figure 1.1 illustrates a typical amino acid. This is drawn as if the plane of the page were intersecting the α-carbon with the carboxyl group and the side chain entering the page and the amino group and the hydrogen atom coming out of the page. Conventionally, the carbon atoms that make up the side chains are designated by Greek letters beginning with the letter β. Thus, the atom nearest to the α-carbon will be designated by the letter β and the others by the letters γ, δ, ε, ζ and so on. Figure 1.2 gives a representation of the 20 amino acids normally found in proteins.

The different properties of the side chains can be used to classify the amino acids in different categories. One classification comprises the hydrophilic, the hydrophobic and the special amino acids.

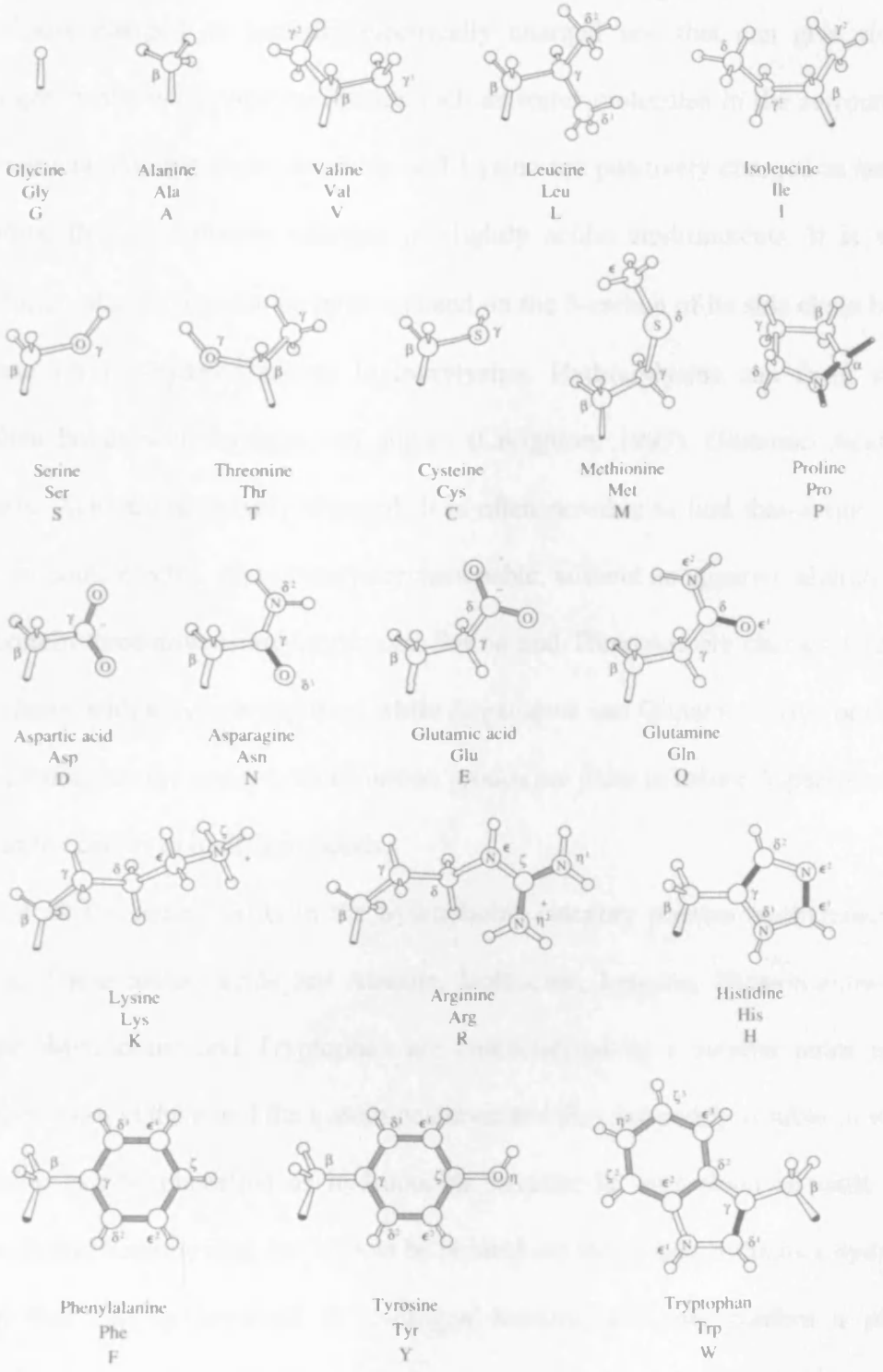


Figure 1.2: Representation of the side chains of the 20 fundamental amino acids. Three letters and single letter abbreviation are shown as well. (from Creighton, 1997)

The hydrophilic category comprises those amino acids whose side chains are electrically charged or partially electrically charged and that can give rise to hydrogen bonds with polar molecules such as water molecules in the surrounding environment. Among them, Arginine and Lysine are positively charged as well as Histidine that is positively charged in slightly acidic environments. It is worth mentioning that Lysine can be hydroxylated on the δ-carbon of its side chain by the enzyme lysyl 5-hydroxylase to hydroxylysine. Hydroxylysine can form strong covalent bonds with Lysines and sugars (Creighton, 1997). Glutamic Acid and Aspartic Acid are negatively charged. It is often possible to find that amino acids with the same electric charge are interchangeable, without an apparent alteration of the protein three-dimensional structure. Serine and Threonine are characterized by side chains with a hydroxyl group, while Asparagine and Glutamine have branches with terminal amino groups. Since amino groups are polar in nature Asparagine and Glutamine can form hydrogen bonds.

Almost all the amino acids in the hydrophobic category possess methylenic side chains. These amino acids are Alanine, Isoleucine, Leucine, Phenylalanine and Valine. Methionine and Tryptophan are characterized by a sulphur atom and a nitrogen atom at the δ and the ε atom positions and they are poorly soluble in water. Tyrosine can be classified as hydrophobic because its side chain is made of a hydrophobic benzene ring, but it must be pointed out that it also contains a hydroxyl group that can be involved in hydrogen bonding and that confers a partial hydrophilic nature to the amino acid.

The remaining amino acids, Glycine, Proline and Cysteine, have peculiar properties that go beyond the simple classification above.

Glycine is the simplest and the smallest amino acid. Its side chain consists of only one hydrogen atom. Because of its small side chain, Glycine is subjected to fewer conformational constraints if compared to the other amino acids.

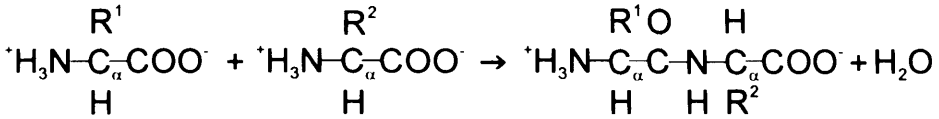
The side chain of Proline is made of three methylic groups that form a covalent bond via the δ-carbon, with the nitrogen atom in the backbone. Thus, all nitrogen bonds are saturated and this results in an imino group that can not form hydrogen bonds. The five membered ring is rigid and confers an inflexible structure to the hosting polypeptide. The pentagonal side chain of proline is puckered, with the N, C_α, C_β and C_δ atoms lying on the same plane and the C_γ atom being 0.5 Å far from it.

Cysteine side chain is characterized by the extremely reactive sulphydric group. These groups can give rise to disulphide covalent bonds between different Cysteines thus linking different regions of a polypeptide together.

1.2 Polypeptides

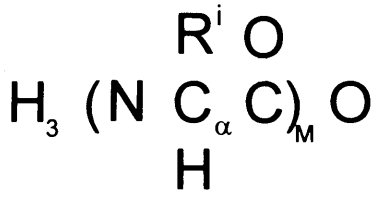
The carboxyl group of an amino acid can react with an amino group of another through a condensation reaction to form a peptide bond. This process can involve a large number of amino acids to form a peptide chain. When the number of amino acids is fewer than 30, such a chain takes the name of an oligopeptide, otherwise it is usually called a polypeptide. By convention, when describing the linear succession of amino acids an orientation is given to the polypeptide chains going from the amino group to the carboxyl group. The first amino group and the last carboxyl group are not involved in the condensation reaction and therefore maintain their ionized form in the cellular environment.

The following formula represents the condensation process between two amino acids.



1.1

The bond between the C=O group and the N-H group of the second amino acid is called the peptide bond. Generally, an entire polypeptide chain, made of M residues, can be represented by the repetition of a common unit:



1.2

where the repeated unit of every polypeptide is in brackets. The nitrogen atom, the α-carbon and the β-carbon atom of the carboxyl group are called the backbone atoms, while the entire structure in brackets is called the residue. Another way of representing a polypeptide chain is by means of the peptide units i.e. those portions of a chain that go from a c_α to the successive one. It is interesting to point out that the peptide bond has partial double-bonded character due to resonance as shown below in figure 1.3:

Figure 1.3: Schematic representation of resonance in the peptide bond.

Therefore, a rotation about the C=N bond is difficult and the atoms represented in figure 1.3 can be considered as coplanar. Peptide groups can assume a trans- or a cis-form. In the trans-form (represented in figure 1.3), the C=O and the N-H groups are pointing towards opposite directions. The trans-form is about 1000 times more stable and more frequent than the cis-form because the side chains of adjacent residues are well separated. In the cis-form, the amino and the carboxyl groups of a peptide are pointing towards the same direction. It is worth noting that the position occupied by the side chain of a proline does not change considerably from a trans- to a cis-configuration. Thus in the case of a proline, the trans-configuration is only 4 times more probable than the cis-configuration.

1.3 Molecular interactions

The main interactions that influence the protein structure are:

- Electrostatic forces
- Hydrophobic forces
- Hydrogen bonds
- Van der Waal's forces

1.3.1 Electrostatic forces

Electrostatic forces are essentially due to charge-charge interactions or to dipolar interactions.

Two charged bodies A and B, *in vacuo*, separated by a distance r_{AB} and with electric charge Z_A, and Z_B, are subjected to an electrostatic force that is described by Coulomb's law. Indicating the electrostatic force between the two bodies with F_{AB}, Coulomb's law *in vacuo* has the following scalar form:

$$F_{AB} = \frac{1}{4\pi\varepsilon_0}\frac{Z_A Z_B}{r_{AB}^2} \qquad 1.3$$

where ε_0 is the dielectric constant of *vacuum,* also known as permittivity of free space. The electrostatic force F_{AB} is associated to the electrostatic energy E_{AB} that, using the definitions previously introduced is defined as:

$$E_{AB} = \frac{1}{4\pi\varepsilon_0}\frac{Z_A Z_B}{r_{AB}} \qquad 1.4$$

E_{AB} is a function of the charges of the two bodies and it is a long-range interaction because, being dependent on $1/r_{AB}$, it is still felt at relatively long ranges. If the charges are surrounded by a medium, it is necessary to consider its presence by introducing the dielectric constant ε_r representing the effect of the medium. ε_r is also called relative static permittivity, is always greater than 1, it is specific to the medium and is incorporated in the formula for the electrostatic energy as:

$$E_{AB} = \frac{1}{4\pi\varepsilon_0\varepsilon_r}\frac{Z_A Z_B}{r_{AB}} \qquad 1.5$$

Therefore, the medium has a screening effect on the electric energy developed between the two charged bodies surrounded by it. Typical values for dielectric constant are between 2 and 110 depending on the medium (Creighton, 1997). It is thus evident how the dielectric contribution lowers the energy developed between two charged bodies.

However, Coulomb's law uses dielectric constants that are average values of the dielectric properties of the medium. In this macroscopic representation, the medium is seen as a continuum surrounding the charged body. If instead we consider the electric interactions between the charged side chains of a protein with the surrounding medium, it is necessary to apply a microscopic description of the phenomenon. In this case, the averaged dielectric constant can not be used and it is necessary to use a slightly different approach.

At the microscopic level the dimensions of the molecule belonging to the medium are comparable to those of the charged bodies immersed in the medium, thus it is possible to represent their interactions as if they were in free space.

Another model consists of calculating local values for the dielectric constants associated to small portions of a macromolecule and using it in Coulomb's law (Sharp and Honig, 1990: Nakamura et al., 1988). However, it is not clear which model is the best suited to represent the interaction of biological molecules with the medium.

A dipole is a system made of two identical but opposite electrical charges that are in close proximity. This system is mathematically represented by the dipole, a vector whose intensity is the product of the charge times the separation distance of the two charges:

$$\vec{d} = q\vec{s} \qquad 1.6$$

with $\vec{d}$ being the dipole moment, q the charge and $\vec{s}$ the separation distance between the two charges. The energy involved in dipolar interactions is inversely proportional to the second or third power of the separation distance if the interacting dipoles are fixed in space; otherwise, it is inversely proportional to the sixth power. Dipoles can interact with dipoles, quadrupoles, and point charges.

If the dipole is immersed in a medium, the energy is also inversely proportional to the dielectric constant of the medium if the macroscopic representation is used. Otherwise, in the microscopic description, it is necessary to consider the interactions between charges and dipoles as if they were in free space, thus an approach similar to the one mentioned above can be used (Sharp and Honig, 1990).

Water is the typical medium proteins are immersed in. Electrostatic interactions in water among proteins are thus less intense than *in vacuo* because of its dielectric constant. If in addition, we consider the presence of small ions in water such as Na^+ and Cl^- the apparent dielectric constant of water increases because ions tend to concentrate around the protein creating an ionic atmosphere around it as is described in the Debye-Hückel theory. The charge of the ionic atmosphere is opposite to the one of the protein. Therefore, a screening between proteins happens. The effect of the Debye-Hückel screening is described by an effective dielectric constant D_{eff} that is proportional to the dielectric constant of water D_W by the exponential of the distance d between the charged bodies as shown below:

$$D_{eff} = D_w e^{(kd)} \qquad 1.7$$

where k is proportional to the square of the ionic strength.

1/k is called Debye screening distance and is an approximate measure of the thickness of the ionic atmosphere around an ion. In other words, it indicates the distance over which an ion extends its electrostatic field. (Moore, 1972). A typical value for the Debye distance at physiological ionic strength is about 8 Å (Creighton, 1997).

The Debye screening decreases the electrostatic free energy of a protein increasing its solubility. Therefore adding salts to a water-protein solution increases the solubility of the protein independently of the type of salt used at lower concentrations. However at higher salt concentrations the solubility tend to decrease presumably because of an electrostatic repulsion occurring between the exceeding salts ant the less polar interior of the protein (Creighton, 1997; Zhang and Cremer, 2006)

1.3.2 Hydrophobic forces

Hydrophobic forces arise when apolar molecules are immersed in a polar fluid, for example water. In this case, water molecules attract each other forcing the apolar molecules to group together. This has a great effect on protein folding since proteins exist in an aqueous environment. In fact, apolar amino acids tend to be buried inside the proteins and polar and charged amino acids tend to be on the surface. "The conflict between the hydrophobic side chains and the polar nature of the water guides these groups to collapse and be shielded from water by forming a tightly packed core that contains more than 80% of the non polar side chains of a typical protein" (Levy and Onuchic, 2006).

1.3.3 Hydrogen bond

The hydrogen bond is formed between two electronegative atoms that are held together by means of a shared hydrogen atom. The hydrogen bond is characterized by almost a collinear junction among the atoms involved. Sometimes, there can be more than two atoms involved. Usually the hydrogen atom is closer to one of the atoms, (the donor atom), making a covalent bond with it. The other atom is conventionally called the acceptor. Such a bond is usually indicated as:

$$D—H\cdots A \qquad 1.8$$

where the letters D, H and A indicate the donor atom, the hydrogen and acceptor atom respectively.

In collagen, the most important hydrogen bond because of its stabilising effects, is that involving the donor amide group, N-H, and the acceptor carboxyl group, C=O, of a polypeptide backbone. Water too can easily form hydrogen bonds working either as an acceptor or as a donor. More specifically, the oxygen atom, in a water molecule, possesses two lone electron pairs that can act as the acceptors of the hydrogen atom belonging to another water molecule. Consequently, hydrogen bonding is the characterising interaction among water molecules.

1.3.4 Van der Waal's forces

Van der Waal's forces are attractive, weak and close range. Their energy of interaction is inversely proportional to the sixth power of the distance between the molecules involved. They are a result of the mutual attraction between two

permanent molecular dipoles, between a permanent and an induced dipole, or between two mutually inducing dipoles. In this latter case, Van der Waal's forces are also called dispersion forces. They are produced by shifting clouds of electrons around the atom's nucleus that acquires a transient dipolar nature and induces dipoles in the nearby atoms.

Van der Waals forces can also be repulsive. These arise when two atoms or molecules are in close proximity. In this situation, their electronic clouds should be penetrating each other, but this is quantum-mechanically forbidden by Pauli's exclusion principle. The energy associated with these forces is inversely proportional to the twelfth power of the distance between the interested atoms. The increase in repulsive energy of two atoms in close proximity is so high that it is natural to associate impenetrable volumes to the atoms involved. These volumes are described by the Van der Waal's radius. The rigid sphere models that have been employed by Ramachandran and Sasisekharan in creating the Ramachandran plot (Ramachandran and Sasisekharan, 1968), make use of the Van der Waal's radii. Repulsive and attractive Van der Waals forces are often combined in what is called the Lennard-Jones potential, represented in figure 1.4 and described by equation 1.9 in which d is the distance between the interacting molecules and C_{12} and C_6 are constants:

$$E(d) = \frac{C_{12}}{d^{12}} - \frac{C_6}{d^6} \qquad 1.9$$

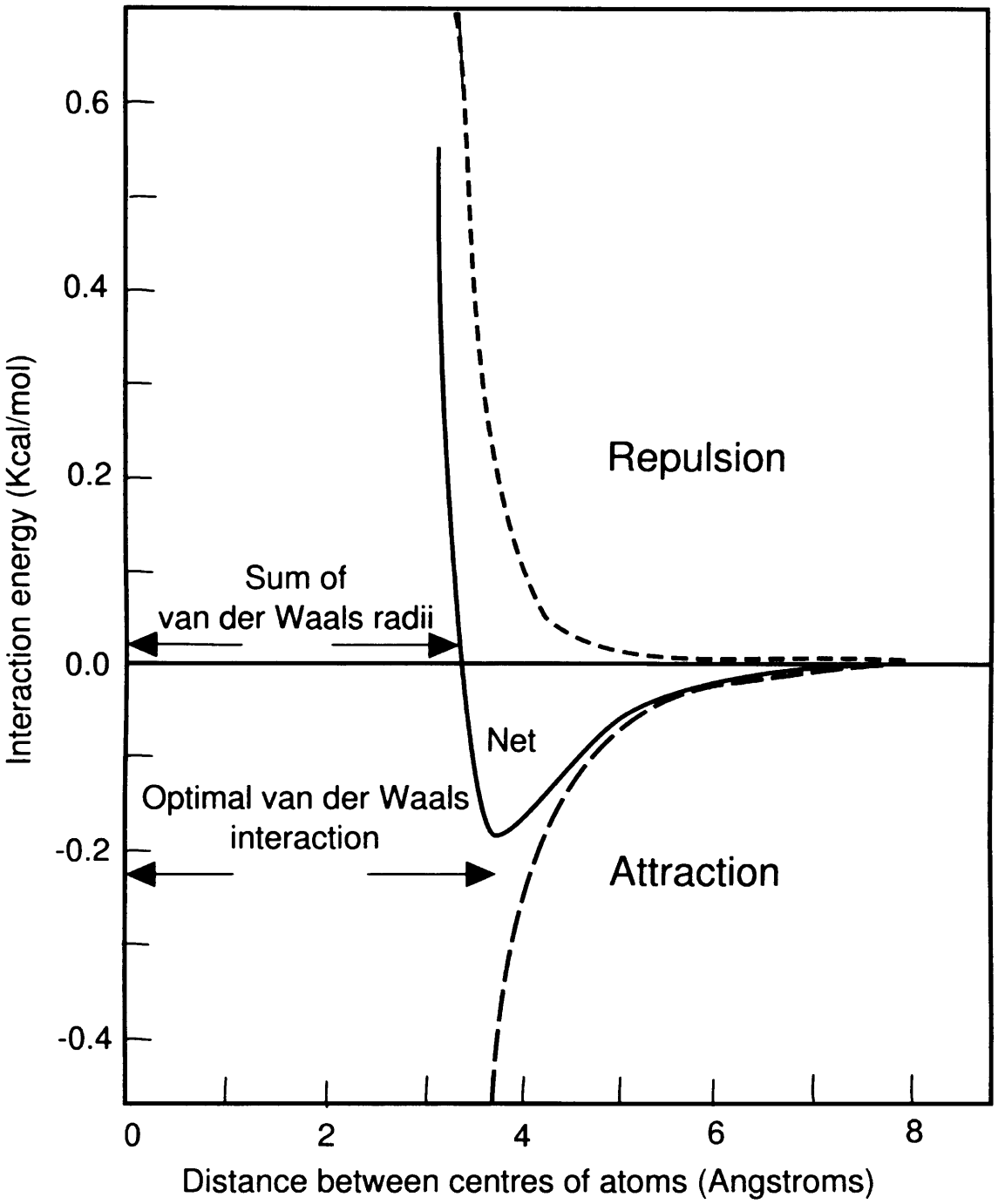


Figure 1.4: Repulsive and attractive energy profiles of the Van der Waal's interactions as a function of distance between the centres of the atoms. Their sum gives the Lennard-Jones potential. (Creighton, 1997).

1.4 Conformation of polypeptide chains

Steric constraints and the interactions described above cooperate to give proteins their final three-dimensional structures. There is a number of folding motifs that are found in most proteins. The most important of these are the α-helix, the β-pleated sheet and the collagen triple helix.

Since fibrillar collagen interactions are the main focus of this thesis, we describe some of the structures proposed over the years for the collagen triple helix, starting from their precursor, the polyproline molecule.

1.4.1 Polyproline

Polyproline is a secondary structure of particular interest to us because of its chemical similarity with the collagen molecule. It is a helical polypeptide structure modelled by stringent geometrical constraints due to the ringed proline side chains, and by the fact that no hydrogen bonds can form between the carboxyl group and the imino group of the prolines. It can occur in two configurations, one is named polyproline I the other polyproline II. The former is a right-handed helix with 3.3 prolines per turn in a cis-conformation and its energetic minimum is obtained when the φ and ψ torsion angles have values of -75° and +160° respectively (φ and ψ are the torsional angles around the N-C_α and the C_α-C=O bonds along a polypeptide chain). The latter is a left-handed helix with three prolines per turn in a trans-conformation, with a minimum of energy obtained when the values for φ and ψ are -75° and +145° respectively (Adzhubei and Sternberg, 1993). The displacement per residue along its axis is 3.12 Å, so that the pitch is 9.36 Å, resulting in a much more extended chain compared to the α-helix. In normal

aqueous buffer, polyproline II is the most stable form (Cantor and Schimmel, 1980). Figure 1.5 represents a polyproline structure.

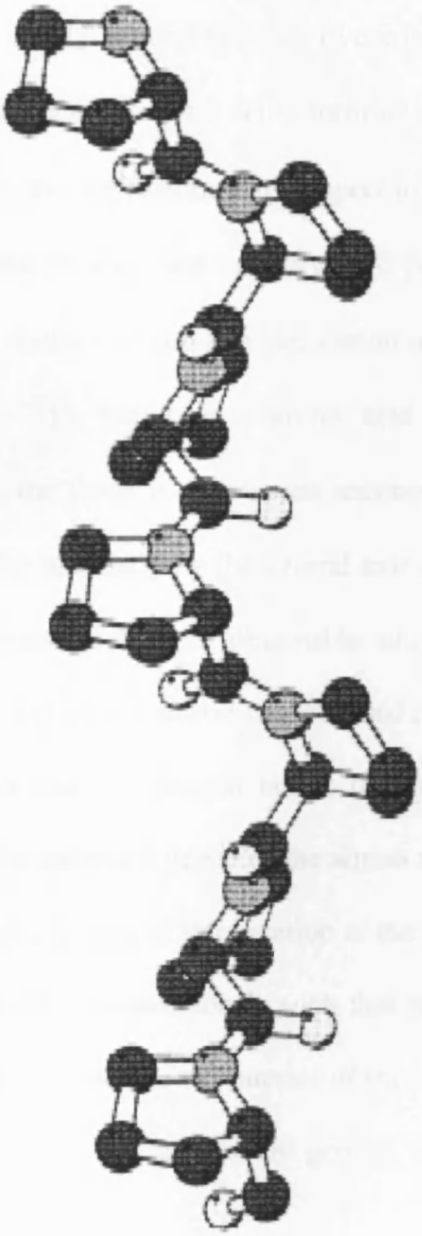


Figure 1.5 Polyproline II left handed helix (Adapted from Stapley and Creamer, 1999).

1.5 Collagen general features

The common structural feature of all collagens is that they possess one or more triple helical domains. These domains are made of three polypeptide chains wound around a common axis in a right-handed fashion. Every individual chain (called an α-chain) is made of a sequence of amino acids forming a left-handed coil. Every α-chain is staggered axially by one residue with respect to its neighbour chains, and its sequence is characterized by a glycine in every third position. Thus, the general amino acid sequence of a single α-chain is a succession of motifs of the kind Gly-Xxx-Yyy with Xxx and Yyy being any amino acid in positions X and Y respectively. In this way, the three α-chains can accommodate the glycines (the smallest of the amino acids) internally by the central axis of the collagen molecule. This results in a tight packing that is obtainable only because of the small dimensions of the glycine. Figure 1.6 shows the idealised representation of a typical collagen triple helix. Inter-chain hydrogen bonds that arise between the amino group of the glycine and the carbonyl group of the amino acid in the X position are the primary, but not the only, source of stabilization of the entire collagen molecule. In addition, the triple helical conformation is such that the peptide bonds linking adjacent amino acids are buried within the interior of the molecule. Thus, the triple helical region is highly resistant to attacks by general proteases such as pepsin (Lisenmayer, 1991).

The residues in X and Y positions are located in such a way that their side chains face the external environment. These amino acids make up the surface of the collagen and mediate all the interactions with cells and other molecules. In fact, for example, residues with charged or partially charged side chains are frequent and they form salt bridges and polar interactions with other molecules.

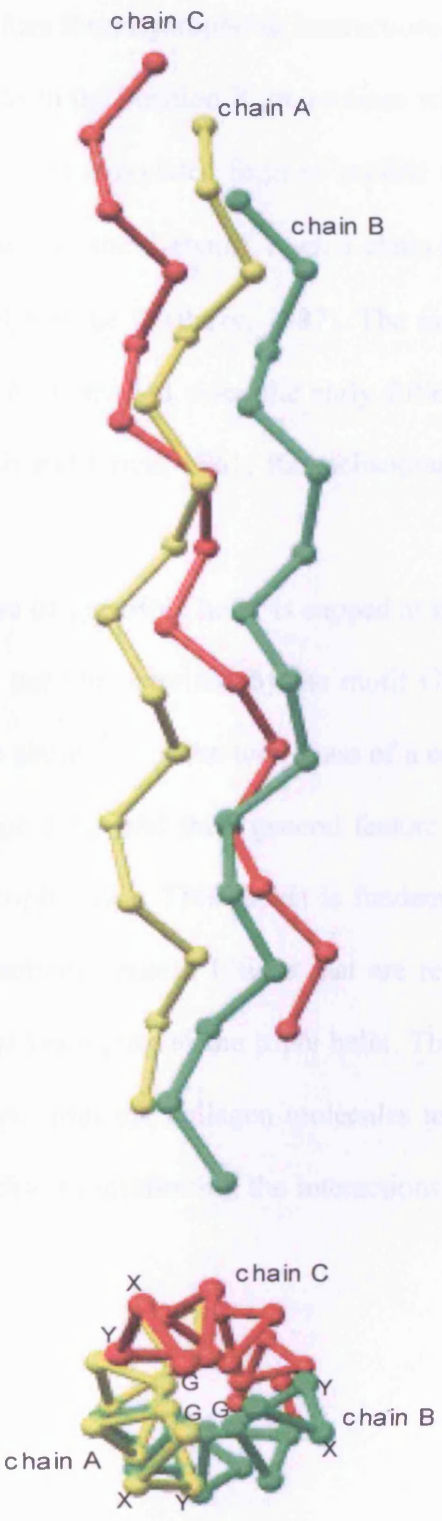


Figure 1.6: Idealised representation of a collagen triple helix. The three α-chains are in different colours.

Similarly, non-polar residues form hydrophobic interactions with other molecules. 10-12% of the amino acids in the position X are prolines while 6-10% of the amino acids in position Y are the hydroxylated form of proline (Rich and Crick, 1961). This confers a rigid structure to the α-chains. Each α-chain has a structure similar to that described for the polyproline II (Parry, 1987). The molecular structure of the collagen triple helix has been studied since the early fifties and numerous models have been proposed (Rich and Crick, 1961; Ramachandran, 1967; Okuyama et al, 1981).

The well ordered structure of the triple helix is capped at the extremities by amino acids whose sequence is not characterised by the motif Gly-Xxx-Yyy. In fibrillar collagens, they constitute about 2% of the total mass of a collagen molecule. These structures are called telopeptides and their general feature is more disordered and flexible that that of the triple helix. This aspect is fundamental for collagen fibril formation. In fact, telopeptides contain lysines that are responsible for cross-link formation with the lysines belonging to the triple helix. Thus after the electrostatic and the hydrophobic forces join the collagen molecules together, the telopeptides operate as cross-link mediators reinforcing the interactions and in last analysis the fibres.

1.5.1 The Rich and Crick model

The model of Rich and Crick (Rich and Crick, 1961) is of particular interest. They used chemical and X-ray diffraction data to delineate a model of the collagen molecule formed by three parallel left-handed polypeptide chains as depicted in figure 1.7. They assumed that every chain was a sequence of (Gly-Xxx-Yyy) triplets and that the triplets in a chain were staggered by one residue position with respect to those in the next chain. In addition, they postulated that prolines could occur only in positions X or Y, and that all peptides bonds were in the trans-conformation. By model building, they found that there was one intra-chain hydrogen bond per triplet, between the amino group of a glycine and the carboxyl group of the residue in position X of the neighbouring chain. These bonds would confer the necessary stability to the collagen molecule. Their model was supported by quantitative observations, obtained from X-ray diffraction photographs of rat-tail tendon collagen, which was consistent with the collagen molecule having a helical screw axis with a residual translation of 2.86 Å and a rotation of approximately 108°. Moreover, density considerations led them to think that there were three residues every 2.86 Å along the molecular axis (Rich and Crick, 1961). In their model, every α-chain was similar to polyproline II and each α-chain was staggered by one position with respect to the others. The overall helical symmetry was 10/3.

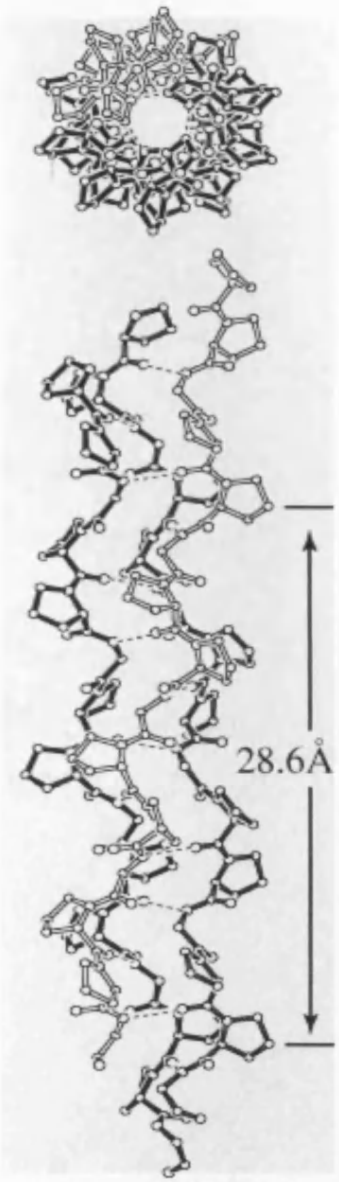


Figure 1.7: Representation of the Rich and Crick (10/3) collagen model (adapted from www.tuat.ac.jp/~x-ray/Research/R_2-models.html)

1.5.2 The Ramachandran model

The model proposed by Ramachandran and colleagues (figure 1.8), consisted of a triple helix very similar to that of Rich and Crick but, in their case, the axial displacement was of 2.91 Å per residue and each collagen molecule possessed a 36/11 supramolecular symmetry (Ramachandran, 1967).

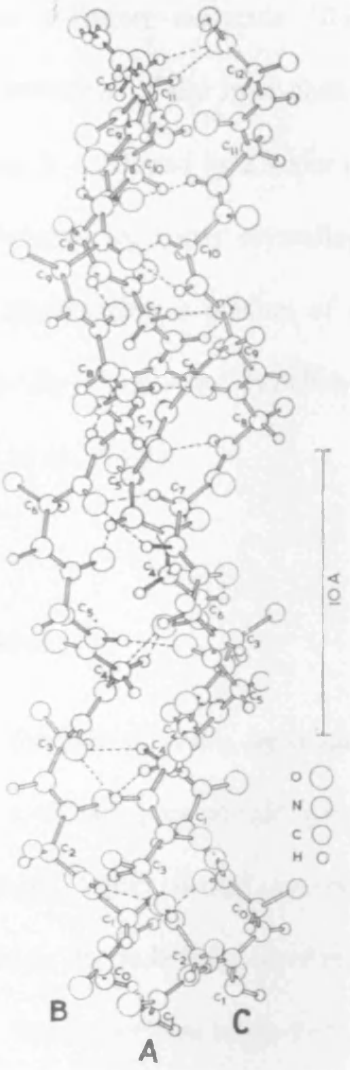


Figure 1.8: Representation of the Ramachandran et al. collagen model (Adapted from Ramachandran, 1967).

Their model predicted the existence of more than one hydrogen bond per triplet contributing to collagen stability. To explain the coiling of the α-chains, they hypothesized the presence of a hydrogen bond between the carboxyl group of a glycine and the amino group of a residue in position X in a neighbouring chain that was mediated by a water molecule. In addition, they suggested that this other

molecule could form a hydrogen bond between the C=O group of the glycine and the O-H group of next hydroxyproline belonging to the same chain, thus enhancing the overall stability of the collagen molecule (Ramachandran et al., 1973; Ramachandran, 1988). The presence of the inter-chain bond between the glycine and the amino acid in position X, mediated by a water molecule, was proven by the findings of Kramer and colleagues by X-ray crystallography of single crystals of synthetic collagen peptides representing a portion of type III collagen (Pro-Hyp-Gly)$_3$-(Ile-Thr-Gly)-(Ala-Arg-Gly)-(Leu-Ala-Gly)-(Pro-Hyp-Gly)-(Pro-Hyp-Gly)$_3$ (Kramer et al., 1999; Kramer et al., 2001).

1.5.3 The Okuyama model

The group of Okuyama was the first to study, by means of X-ray crystallography, the molecular structure of a synthetic polypeptide with the collagen-like sequence (Pro-Pro-Gly)$_{10}$ (Okuyama et al., 1981). Based on their findings, they proposed a model of a triple helix in which the molecular axial displacement was 2.87 Å per residue (very similar to what Rich and Crick suggested), and a 7/2 helical symmetry for the collagen molecule (figure 1.9). The stability of the molecule was ensured by direct hydrogen bonds equivalent to those proposed by Rich and Crick, and, in addition, they suggested a hydrogen bond mediated by a water molecule between the carboxyl group of a glycine and that of a proline in the Y position of the same α-chain. They also suggested the presence of a second water molecule bonded to the carboxyl group of the proline in the Y position and a third water molecule that connects the first two (Okuyama et al., 1981).

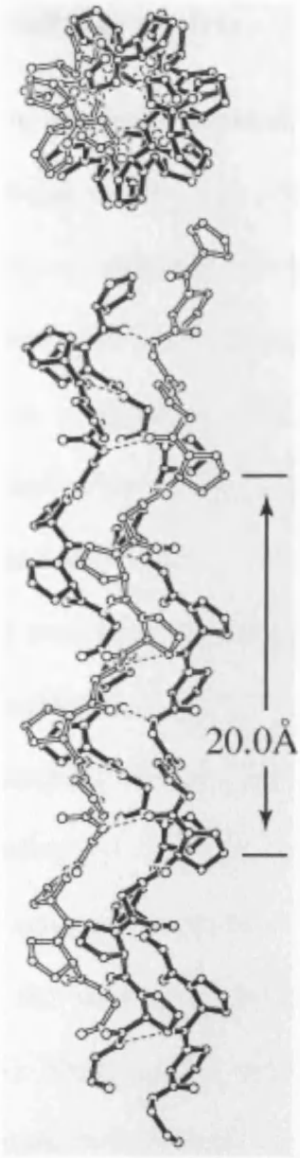


Figure 1.9 Representation of the Okuyama et al. (7/2) collagen model (adapted from www.tuat.ac.jp/~x-ray/Research/R_2-models.html).

The models summarised above, describe the structure of a single collagen molecule. In nature, collagen molecules are rarely found as single monomers, and often they tend to interact with each other to form not only homotypic but also heterotypic structures (see for example collagen I and collagen III in fibrils). It is therefore important to illustrate the different types of collagen secreted in the extracellular matrix for a complete understanding of the possible mechanisms driving the interactions among collagen molecules.

1.6 Collagen in the extracellular matrix

Animal cells are surrounded by a complex network of proteins and carbohydrates that constitutes their extracellular matrix. The extracellular matrix is a highly specialized framework carrying out different functions. For example, it can play structural roles in tendons, function as filters in the kidneys, or help cell adhesion via the collagen-rich basement membranes. The role of the matrix is also a developmental one. In fact, it can vehiculate hormones and growth factors that are essential during the evolution and differentiation of the cells.

Collagen is the most abundant protein of the extracellular matrix. The function of collagen has always been considered simply a structural one, but the collagen family is made of several genetically distinct types that carry out many different functions. For example, collagens are involved in cell attachment and differentiation, as chemotactic agents, as antigens in immunopathological processes, and are also recognised as the defective components in certain pathological conditions (Lisenmayer, 1991). Studying the molecular structure, assembly, and turnover of collagen is important to understand that part of cell physiology that involves the extracellular matrix, some embryonic and foetal developmental processes, and those pathologies of the connective tissue that are linked to many diseases. The study of the expression and the function of the different collagen types should contribute to a better understanding of diseases that are due to genetic defects of the collagen molecule such as chondrodysplasias, osteogenesis imperfecta, Alport syndrome, Ehlers Danlos syndrome, and epidermolysis bullosa (Gelse et al., 2003).

1.7 Biosynthesis of collagens

The process of formation of fibrillar collagens is determined by intracellular and extracellular events and it is thought to be similar for all collagens (Prockop and Kivirikko 1995; Gelse et al., 2003). Inside the rough endoplasmic reticulum, after the removal of the signal peptide by a signal peptidase, the single α-chains making up the collagen molecule undergo a post-translational modification that hydroxylates the prolines and lysines in the Y position changing them into 4-hydroxyprolines and hydroxylysines respectively. The presence of hydroxyprolines is fundamental since it endows the collagen molecule with water-mediated intramolecular hydrogen bonds that enhance the thermal stability of the triple helical domain (Bella et al., 1995; Brodsky and Ramshaw, 1997). Some other modifications such as the hydroxylation of some prolines in the X position or the addition of sugars to hydroxylysines do also occur. The extent of these modifications depends on the collagen type and on its tissue distribution (Eyre et al., 1984) subsequently, three α-chains join together at the carboxy-terminals (C-terminals). Specific enzymes like PPI (peptidyl-prolyl cis-trans-isomerase) PDI (protein disulphide isomerase) or collagen specific chaperons like HSP47 drive the correct folding of the α-chains into a so-called procollagen molecule (Gelse et al., 2003).

Once the procollagen molecule is created, it is secreted into the extracellular matrix where the N- and C-terminals of the α-chains are cleaved by specific proteinases. A second modification of some lysines and hydroxylysines takes place that turns them into aldehyde derivatives that have a fundamental role in cross-link formation during fibrillogenesis (Prockop and Kivirikko 1995; Gelse et al., 2003).

1.8 Collagen families

The roles and tissue distribution of collagen are varied. At least twenty-eight genetically distinct types of collagen exist (Veit et al., 2005) and they can be subdivided into the following families (Van der Rest and Garrone, 1991; Prockop and Kivirikko, 1995; Gelse et al., 2003):

- Fibril-forming collagens
- Fibril-Associated Collagens with Interrupted Triple helices (FACIT)
- Network-forming collagens
- Anchoring fibril collagens
- Transmembrane collagens
- Multiplexins

Due primarily to technological limitations, the first type of collagen to be studied (type I collagen) was also the most abundant and it is found mainly in skin, bone and tendon. During the years, it became apparent that there were many different types of collagen. To each one of them a Roman numeral was associated (I, II, III, etc.) that was chosen according to the chronological order in which the collagen type was discovered. If a collagen molecule is made of three identical α-chains it is called homotrimer otherwise it is a heterotrimer. Since some collagen types can be made up of different α-chains, it is conventional to describe each collagen molecule by the Roman numeral of its type with the identifier of the α-chains making up the particular collagen studied. For example, the heterotrimeric form of type I collagen

made of two α_1(I) chains and one α_2(I) chain is indicated by $[\alpha_1(I)]_2[\alpha_2(I)]_1$ while the rarer homotrimeric form is indicated by $[\alpha_1(I)]_3$.

1.8.1 Fibril-forming collagens

Collagen types I, II, III, V, and XI belong to what is classically referred to as the family of fibril forming collagens. These collagens are similar to each other in size and they possess a single triple helical domain containing about 340 triplets. Their most striking feature is the staggered manner in which they assemble. In fact, they give rise to fibrils with electron dense bands with a 67 nm axial periodicity.

Type I collagen is the most frequent of all collagens and it is found usually in heterotrimeric form ($[\alpha_1(I)]_2[\alpha_2(I)]_1$) throughout the body in bones, dermis, tendons ligaments and cornea. It is also synthesized in response to injury or formed as a consequence to some fibrotic diseases (Gelse et al., 2003). A rarer homotrimeric form ($[\alpha_1(I)]_3$) does exist but it is for the most part limited to wound healing and to embryonic development (Phillips et al., 2002).

Type II collagen is a homotrimer ($[\alpha_1(II)]_3$), it is found in cartilage, developing cornea and vitreous humour (Gelse et al., 2003). In cartilage, it is associated with type XI and Type IX collagens that are supposed to limit the lateral growth of type II collagen fibrils (Eyre, 2004; Eyre et al 2006).

Type III collagen is also a homotrimer ($[\alpha_1(III)]_3$), it is found in the walls of arteries and other hollow organs and it usually forms heterotypic fibrils with type I collagen in skin and reticular fibres in lung, liver and spleen (Gelse et al., 2003). Heterofibrils made of type II and type III collagens are found in cartilage (Young et al., 2000).

Type V collagen is the one with the most heterogeneous structure because its three α-chains belong to at least four distinct genetic types: α_1(V), α_2(V), α_3(V) and α_4(V). Moreover, it can contain a type XI collagen α-chain (Ricard-Blum and Ruggiero, 2005). It associates with type I (Birk, 2001) and type III collagen in bone matrix, the corneal stroma and interstitial matrix of muscles, liver, lungs and placenta (Gelse et al., 2003).

Type XI collagen is a heterotrimer whose α_3(XI)-chain is a post-translational modification of α_1(II) chains, even though it can also be substituted by α_1(V) (Ricard-Blum and Reggiero, 2005). It is found together with type II collagen in articular cartilage (Wu and Eyre, 1995).

A common feature of both type V and XI collagen is that they retain part of their terminal domains after cleavage (Birk, 2001; Wu and Eyre 1995). Since they are thought to be the central core of hetero-fibrils, it is likely that they retain some terminal domain to limit the lateral accretion of the fibrils that they form (Wess, 2005; Kadler et al., 1996).

The recently sequenced type XXIV and XXVII collagens have been added to the family of the fibril forming collagens. They are characterised by a shorter triple helical domain (about 330 triplets long) and by some imperfections in the repetitions of the triplet motif (Gly-Xxx-Yyy). Type XXIV collagen forms hetero-fibrils with Type V and Type I collagen in bones and cornea while Type XXVII is usually associated to Type II and XI collagens in cartilage (Ricard-Blum and Ruggiero, 2005)

1.8.2 Fibril-associated collagens with interrupted triple helices (FACIT)

Collagens type IX, XII, XIV, XVI, XIX, XX, XXI and XXII belong to the FACIT group. These non-fibrillar collagens are made of triple helical domains that are interrupted with non-collagenous domains. Normally, they are linked to the surface of a host fibrillar collagen via the collagenous domains at one extremity of the molecule while the rest is projected outwards towards the external environment. A non-collagenous domain acts as a hinge for this purpose. This conformation is thought to facilitate the interactions with other proteins of the extracellular matrix or with other cells.

Type IX collagen is the best characterized of this family and, as shown in figure 1.10, consists of two collagenous domains attached to a host type II collagen and a further domain pointing towards the matrix. Type IX collagen is found in hyaline cartilage and the vitreous body of the eye (Van der Rest and Garrone, 1991).

Types XII and XIV collagens are similar in structure and associate or colocalize with type I collagen in skin, perichondrium, periosteum, tendons, lung, liver, placenta and vessels walls.

Type XVI collagen is found in cartilage heterofibrils and it is suspected to have a role in fibroblasts stabilisation in the extracellular matrix.

Type XIX collagen is expressed in the epithelial basement membranes. It is supposed to be involved in the early stages of the skeletal muscle cells differentiation.

Collagen type XX and XXI are similar to collagen XII and XIV respectively. Type XX collagen is localised to corneal epithelium whereas type XXI collagen is expressed at foetal stages and it is mostly found in vascular walls.

Type XXII collagen is considered a marker of tissue junction and it is thought to interact with microfibrils rather than with collagen fibrils (Ricard-Blum and Ruggiero, 2005; Gelse et al., 2003).

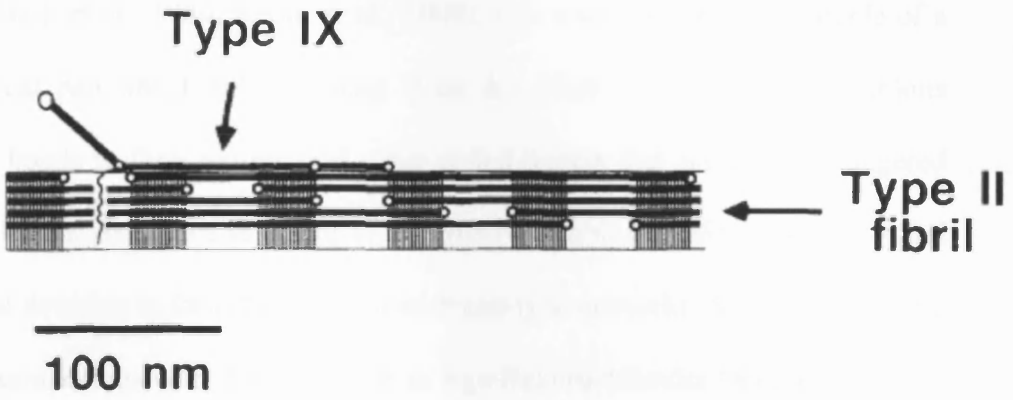


Figure 1.10: Model for the binding of type IX collagen to type II collagen. Black lines represent triple helical domains of collagens II and XI. Circles indicate where the triple helical domain is interrupted. (from: Van der Rest and Mayne, 1988)

1.8.3 Network-forming collagens

Collagen types IV, VI, VIII and X belong to the network-forming collagen family (Knupp and Squire, 2003).

Type IV collagen is primarily found in basement membranes. Its α-chains are about 400 nm long and are characterized by a voluminous C-terminal domain and by triple helical interruptions that give rise to 26 irregularly spaced sites. These sites give the whole molecule great flexibility (Yurchenco and Schntty, 1990), which is probably necessary to form networks. Collagen IV assembles in a network-like

structure via C-terminal with C-terminal interactions and N-terminal with N-terminal interactions. These interactions give rise to dimers and tetramers that assemble laterally to complete network formation.

Collagen type VI is widespread throughout the human body and is found in skin, cartilage, placenta, intervertebral disc and cornea (Eyre, 2002; Poole et al., 1992; Von der Mark et al., 1984; Reale et al., 2000). It is a rod-like molecule made of a triple helical part about 105 nm long (Von der Mark et al., 1984). It exploits disulfides bonds to form antiparallel super coiled dimers that are axially staggered by about 30 nm. As shown in figure 1.11 dimers assemble laterally via their C- and N-terminal domains to form basement membrane-type networks. Such networks are found in associations with diseases such as Age-Related Macular Degeneration and Sorsby's Fundus Dystrophy. They are also found in human cornea (Knupp et al. 2002(a); Knupp et al., 2002(b); Knupp et al 2006).

Type VIII collagen is a non-fibrillar short chain collagen about 130 nm long. It is a major constituent of Descemet's membrane and is also found in vascular subendothelial matrices, heart, liver, kidney and lung as well as in malignant tumours (Kvansakul et al., 2003; Shuttleworth, 1997). In Descemet's membrane, it is known to assemble in stacks of hexagonal lattices. These lattices are made via the large globular domains.

Type X collagen is among the most specialized collagens and is synthesized by hypertrophic chondrocytes in the deep calcified zone of the cartilage. The assembled form of type X collagen resembles the hexagonal lattice that type VIII forms in Descemet's membrane (Prockop and Kivirikko, 1995).

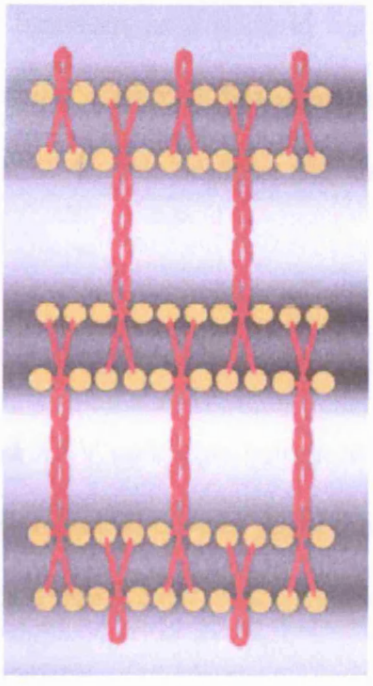


Figure 1.11: Model for lateral aggregations of collagen type VI as proposed by Knupp et al. 2002(a) (From Knupp et al. 2002(a))

1.8.4 Anchoring fibrils collagens

Type VII collagen is the only known member of the fibril anchoring collagen family. It is present in skin, dermal-epidermal junctions, oral mucosa and in the cervix (Gelse et al., 2003). It is also found within the basement membrane beneath stratified squamous epithelium. Type VII collagen is a major component of anchoring fibrils, which are attachment structures within the basement membrane of the dermis of the human skin (Chen et al., 2001). It is made of a very large discontinuous triple helical region about 420 nm long. At the C-terminal, the α-chains unravel to give rise to three arms about 50 nm long. (Van der Rest and Garrone, 1991). In the extracellular matrix, collagen VII aggregates to form antiparallel dimers by overlapping their N-terminal domains by 60 nm. The dimers associate laterally to form a complex structure connecting the lamina densa to the anchoring plaques of the basement membrane.

The network thus formed functions as a scaffold for the interstitial collagens by stabilizing the adhesion of the epithelial basement membrane to the underlying stromal matrix (Lisenmayer, 1991; Van der Rest and Garrone, 1991).

1.8.5 Transmembrane collagens

Type XIII, XVII, XXIII and XXV collagens belong to the transmembrane collagen family. They all contain a domain that crosses the cell membrane and for this reason are thought not to be secreted completely in the extracellular matrix (Prockop and Kivirikko, 1995).

Type XIII collagen is found in many tissues including skin, bone and muscle (Nykvist et al., 2000). Its primary structure is made of three collagenous domains that are interrupted by non-collagenous domains. The short cytosolic domain and the transmembrane domain encompass about half of the first non-collagenous domain, while the rest of the molecule forms the ectodomain, which is a rod-like structure about 150 nm long with two flexible hinges corresponding to the intermediate non-collagenous domains (Latvanlehto et al., 2003).

Collagen type XVII is found mainly in the hemidesmosomes of the skin (Prockop and Kivirikko, 1995) and appears as an asymmetric molecule with a 466-residue long globular head which forms the intracellular domain, a short central transmembrane rod of only 26 amino acids and a flexible 1008-residues long tail that form the extracellular domain. Its functions are not clearly understood but it is possible that it creates a link between intracellular and extracellular structural elements and that it anchors the epithelium to its basement membrane (Schacke et al., 1998).

Collagen type XXIII is expressed in normal human heart, in the retina and in metastatic prostate cancer cells. Similarly to the other collagens in this family, it is made of an amino terminal cytoplasmic domain, a brief transmembrane region, and three extracellular collagenous domains followed by short non-collagenous domains (Banyard et al., 2003).

Type XXV collagen is believed to play an important role in the formation and function of adherens junctions in neurons, which have been identified as puncta or synaptic adherens junctions (Hashimoto et al., 2002).

1.8.6 Multiplexins

Collagens type XV and XVIII belong to the multiplexin collagen family. These two homologous collagens are formed by polypeptide chains with 9 or 10 distinct triple helical domains (Oh et al., 1994).

Type XV collagen is found in basement membranes surrounding skeletal, cardiac and small intestine muscles, and in vascular systems.

Type XVIII collagen is distributed in sinusoidal areas, in the liver, in Bowman's layer, in the dermal basement membrane, in arterial walls and in the capillary basement membranes (Tomono et al, 2002).

1.8.7 Type XXVI and XXVIII collagens

Type XXVI collagen is difficult to classify because it does not share similarities with collagens of other families. It is mainly found, in testis and ovaries during the embryo development (Sato et al., 2002)

Type XXVIII collagen is characterised by some similarities to type VI collagen but can not be classified as a network forming collagen. It is a component of the basement membrane in the peripheral nerve of the nervous system (Veit et al., 2005).

The different forms and functions of collagens are defined by their amino acids sequence. Fibrillar collagens are, seemingly, the most simple to study. Establishing a relationship between amino acids sequence and the aggregation properties of collagen is the first step toward the understanding of protein interactions.

1.9 Supramolecular structure of collagen

Over the years, different methods were tried to establish a relationship between the amino acid sequence of collagen and its supramolecular structure. These methods can be broadly grouped in two main classes, one based on a statistical approach and the other based on modelling.

Structural studies (Salem and Traub 1974; Traub and Fietzk 1976; Traub 1978) found a relationship between the structure of collagen and the amino acid regularities in its sequence. At first, it was noticed that some triplets occur much more frequently than others do and their presence can be related to molecular stability. Salem and Traub (Salem and Traub, 1974) for example, studied the preference of particular amino acids for certain positions within the triplets. For this purpose they used the collagen type I α_1(I) chain, that was the only sequence available at the time of the study. They found, for example, that the Gly-Glu-Lys triplet was encountered 2.3% of the times and that it was more common then the

Gly-Lys-Glu triplet that was almost never found. With the aid of a hard sphere model, they explained the amino acid preference for particular positions in terms of steric interactions (Salem and Traub, 1974).

Hulmes and colleagues (Hulmes et al., 1973) looked at the interaction energy between amino acids to explain the D periodicity typical of type I collagen. They used two α_1(I) collagen sequences and studied the hydrophobic and charge interactions arising between amino acids when one chain was shifted past the other. They found that the energetically favourable interactions were maximised periodically every 234 amino acids, corresponding to a D period.

More recently, Ramshaw and colleagues (Ramshaw et al., 1998) used synthetic peptide models to explain position propensities of some amino acids within a collagen triplet. In their experiment, they capped the amino terminal and the carboxyl terminal of the triplets they were interested in, Gly-Xxx-Yyy, with a few triplets of Gly-Pro-Hyp in order to add stability to the entire structure. Studying the melting temperature of their models, they confirmed that the presence of hydroxyproline in the Y position increases collagen stability (Gustavsson, 1955; Ramachandran, 1973). But, surprisingly, they also found that triplets with very different frequency of occurrence (for example Gly-Glu-Lys about 2.5% and Gly-Lys-Glu less than 0.75%) are characterized by very similar melting temperature, suggesting that the strong preference of some amino acids for the X or Y position is probably motivated by effects arising at higher level of association rather than by simple molecular stability.

The same technique was applied by Kramer and colleagues (Kramer et al., 2000; Kramer et al., 2001) in an attempt to reproduce and study a more realistic collagen-like peptide sequence. They used a core sequence representing a portion of type III collagen ((Ile-Thr-Gly)(Ala-Arg-Gly)(Leu-Ala-Gly)(Pro-Hyp-Gly)) capped at the extremities by three triplets of (Pro-Hyp-Gly). By X-ray crystallography, they were able to demonstrate that the symmetry of the triple helix depends on the type of amino acids making up the α-chain. Thus, for example, the type III collagen sequence was well explained by the Rich and Crick model while the ending caps were well explained by the Okuyama model.

Another very important result arising from their study was that they were able to determine the distribution of water molecules around the collagen-like peptide. In fact, they found that the water molecules act as hydrogen bond mediators between different groups of the collagen α-chains, adding stability to the overall molecule as already suggested by Ramachandran. However, what is perhaps more striking, is the fact that water molecules are regularly distributed around the collagen molecule and form regular hydration shells (Bella et al., 1995). This may mediate higher levels of organization, such as the formation of fibrils.

1.10 Collagen fibril formation

Type I, II, III, V and XI collagen molecules are rigid, rod-like proteins that have charged and hydrophobic amino acids distributed along their axis. They are subject to geometrical constraints. The rod-like structure favours a kind of parallel aggregation while the electrostatic interactions drive molecules to pack together. Subsequently, hydrophobic interactions have a major role in stabilising the packing

that is finalised by the cross-links occurring between the telopeptides and the triple helices. Thus, lateral accretion takes place forming microfibrils and eventually the fibrils.

Negatively stained fibrils show darker and lighter bands alternating along their axis. When viewed in an electron microscope, this basic repeating pattern is about 67 nm long and it is often referred to as D period. Since the length of a collagen molecule is about 300 nm or about 4.47 D, it was suggest by Hodge and Petruska in 1963, that the collagen fibrils arose from a lateral association of the collagen molecules with a stagger of 1D with respect to one another. This process is presumably driven by electrostatic forces and then stabilised by hydrophobic interactions between amino acids and results in the formation of microfibrils where succeeding collagen molecules are 0.53 D apart and side by side molecules are shifted by 1D with respect to each other (as visible in the diagram of figure 1.12). The basic structural element of a collagen fibre is made of five adjacent molecules staggered by D. We call this elemental structure a microfibril coloured in green in figure 1.12. We can thus think of a fibril as the repetition of adjacent microfibrils. The alternating darker and lighter bands seen in the negatively stained electron micrographs simply represent the projection of stains that have filled the gap between the collagen molecules or the overlap regions that do not uptake staining.

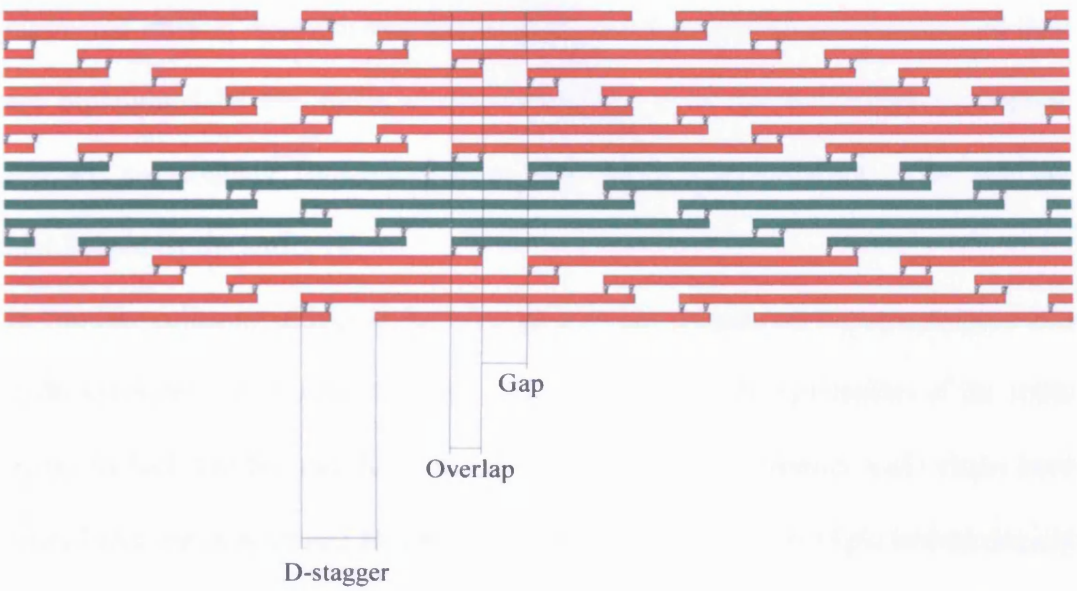


Figure 1.12: Diagrammatic representation of a collagen fibril (red). Molecules join together with a D-stagger (234 amino acids). The repetition of five molecules joined in such a way constitutes what we refer to as microfibril (green).

The analysis made by Hulmes and colleagues pointed out that this particular staggering (i.e. 234 a.a.) is explained by hydrophobic and electrostatic interactions along the collagen molecules. However, the interactions along the triple helix are necessary but not sufficient for the formation of stable collagen fibrils. Once the molecules are aligned, they are linked covalently.

In fibrils, this is achieved primarily through cross-links between the telopeptides at the N- and C-terminals and the triple helix of neighbouring molecules, see for example, collagen types I, II and III. Even though collagen types I, II and III are not identical, they all have equivalent sites of cross-linking on the triple helix. They all have two cross-linking sites on the telopeptides and two other corresponding sites on the triple helical domain.

In our analysis, we use the convention that the amino acids belonging to the α-chains forming the triple helix are numbered starting from the first glycine at the amino-terminal side of the chain. The amino acids of the C-terminal telopeptide are

numbered starting from the first amino acid after the triple helical domain and they are highlighted by the suffix C. The amino acids of the N-terminal telopeptide instead, are counted backwards from the one before the triple helix and are highlighted by the suffix N.

In fibrillar collagen, aldehyde-forming lysines are present on the telopeptides and hydroxylysines are located, in a symmetrical way, near the extremities of the triple helix. In fact, the N- and the C-terminal telopeptides of murine α_1(I) chain have cross-links site in positions 8N and 16C respectively. While its triple helical domain has possible cross linking site at positions 87 and 930. Murine α_2(I) has a short C-telopeptide that does not contain any derivatives of lysines, however its N-telopeptide has an aldehyde forming lysine in position 7N. Triple helical cross-linking lysines are at positions 87 and possibly at position 933.

Murine type II and type III collagens are homotrimer with triple helical cross-linking sites in position 87 and 930 for the former and 96 and possibly 942 for the latter. The telopeptide cross-linking sites are at positions 11N and 17C for murine type II collagen and at positions 16C for murine type III. According to Toman and de Combrugghe (Toman and de Combrugghe, 1994) the N-terminal telopeptide for murine type III collagen is very short and has a lysine at position 2N. For this reason, in this case, it does not have the necessary mobility to face the hydroxylysine in position 942 and to form a cross-link.

However, according to Henkel (Henkel, 1996) bovine type III collagen is characterised by cross-links between lysine 5N and hydroxylysine 939 that are homologous to lysine 2N and hydroxylysine 942 respectively. Thus, we indicate this kind of cross-link as possible even though it is not acceptable if the sequence for type III mouse collagen is used.

These hydroxylysines and aldehyde forming lysines mediate head to tail cross-links formation between adjacent molecules. For example in type II collagen, the lysines of the C-telopeptides (17C) are brought in contact with those at position 87 and those of the N-telopeptides (11N) are in contact with those in position 930. In this way, every molecule is cross-linked with a 4D staggering to the adjacent one (Eyre et al., 1984; Bailey, 2001).

The cross-links sites of type XI collagen are less well defined. According to Wu and Eyre (Wu and Eyre, 1995), who employed HPLC (High Performance Liquid Chromatography), they are chain specific and occurring only between telopeptides of an α-helix and the triple helical domain of a defined α-helix. The hydroxylysines situated within the triple helix that form cross-links are in positions 84 and 924 for the α_1(XI) chain while they are in positions 924 and 930 for the α_2(XI) and α_3(XI) chains of the triple helix respectively. According to Wu and Eyre (Wu and Eyre, 1995) type XI collagen retains a large portion of the N-propeptide which is used to make cross-links in the extra cellular matrix of the cartilage. Thus, α_1(XI) and α_2(XI) have both an aldehyde forming hydroxylysine in their N-terminal telopeptides at position 24N. They do not appear to have cross-linking sites in their C-telopeptides. Since the α_3(XI) chain is a post-translational modification of the α_1(II) chain, its telopeptides posses cross-linking lysines at positions 11N and 17C in common with type II collagen. The authors (Wu and Eyre, 1995) suggest that cross-linking can occur only between the C-telopeptide of the α_3(XI) chain and the hydroxylysine at position 84 of the α_1(XI) chain, and between the N-telopeptides of the α_1(XI) α_2(XI) and α_3(XI) chains and the hydroxylysines of the α_3(XI), α_1(XI) and α_2(XI) chains respectively.

However using the sequence available from the database UniProt it is possible to see how the α_1(XI) and α_2(XI) chains have lysines capable of forming cross-links in positions 84 and 924 while α_3(XI) have hydroxylysines in positions 87 and 930.

α_1(XI) and α_2(XI) chains have hydroxylysines at positions 11C in their C-telopeptides while they do not possess hydroxylysines in the N-telopeptide. The α3(XI) chain possesses hydroxylysines at position 11N and 17C. It is possible that these lysines and their derivatives are forming cross-links in cartilage (Wu and Eyre, 1995). In table 1.1, the cross-linking positions for murine type I, II and XI and for bovine type III collagens are given.

Fibrillar collagens from other species do have cross-linking sites in similar positions.

Cross links sites positions for murine collagen				
	N-telopeptide	Triple helix head	Triple helix tail	C-telopeptide
α1(I)	8	87	930	16
α2(I)	7	87	933?	_____
α1(II)	11	87	930	17
α1(III) (bovine)	5	96	939	16
α1(XI)	_____	84	924	11
α2(XI)	_____	84	924	11
α3(XI)	11	87	930	17

Table 1.1: Cross-linking positions for murine type I, II, XI and for bovine type III collagens.

1.11 Telopeptides conformation and their role in collagen fibril formation

The three-dimensional conformation of the telopeptides is still not well understood. However, it is generally accepted that they fold back to bring the lysines into contact. X-ray diffraction study of type I collagen was carried out by Orgel and colleagues (Orgel et al., 2000) in which they modelled the electron density profile of its C-telopeptides. They were able to determine that the folding occur at about amino acids 13C and 14C so that lysines involved in cross-links face each other during fibril formation. On the other hand, Malone and collaborators (Malone et al., 2004) used the docking technique to study the interaction of the N-telopeptides with the triple helix finding that the ideal folding point for the α_1(I)-telopeptides is about the lysine in position 9N. They also suggested that the particular folding conformation adopted by the telopeptides is sterically conditioned by the surrounding triple helices in the fibril.

The loss of the telopeptides has major effects on fibril formation. The complete loss of the N-telopeptide is associated with the formation of a fibril with a D-periodic antiparallel staggering, while the partial removal of the C-telopeptides is associated with the formation of D-periodic tactoids (Kadler et al., 1996).

The complete removal of both telopeptides inhibits the formation of D-periodic fibres allowing nonetheless the formation of fibrous aggregates. These observations seem to imply that N-telopeptides are involved in fibre polarisation and C-telopeptides are involved in the formation of lateral aggregates and in the lateral accretion of collagen fibres (Kadler et al., 1996).

1.12 Molecular packing in fibrillar collagen as a first step toward the understanding of protein interactions

Fibrillar collagen molecules are made of a rigid linear triple helical domain flanked by two telopeptides in folded configurations. When they interact to form fibrils, they are in first instance guided by the electrostatic forces and then subsequently compacted by the hydrophobic interactions. Finally, the head to tail cross-link formations concur in locking the molecules together. The fibrillar packing is also favoured by the geometrical constraints imposed by the rod-like structure of collagen proteins.

Hulmes and colleagues (Hulmes et al., 1973) were among the first to study the relationship between residues distribution and the linear structure of collagen molecules on one hand, and collagen fibril formation on the other. The homology search algorithm they used for type I collagen came as a natural choice when they had to analyse proteins that could be represented as linear sequences of elements. They found a simple and effective way to link the primary structure of collagen to its quaternary structure.

Similarity algorithms show how interactions develop among collagen molecules in function of their reciprocal displacement. It is thus possible to find a relationship between the pattern of the interactions and the residues distribution responsible for them. This approach commonly used with collagen molecules in parallel configuration can be extended to molecules in antiparallel configurations predicting their state of aggregation. The role of the regular distribution of hydrophobic amino acids is fundamental in explaining this unusual packing, even though it is not often found in nature (Mallinger et al., 1992).

We describe these algorithms in the next chapter.

CHAPTER 2

MATERIALS AND METHODS

2 Introduction

To study the interactions among fibrillar collagens we make some simplifications that allow us to represent collagen molecules as linear sequences of a.a. and the interactions between them as cross-correlation functions. It is therefore necessary to introduce the mathematical tools to represent these interactions. In addition, the definition of convolution is given since it is extensively used during image analysis. The Fourier transform is also briefly introduced since it is used to study the amino acids sequences. The classical book written by Bracewell (Bracewell, 2000) is the main reference used in writing this chapter.

Our predictions from the modelled interactions among collagen fibrils must be tested against real experimental data to prove their validity. This will be done mainly by using electron micrographs. We thus describe the general layout of an electron microscope and how images are formed. Specimen preparation is also briefly illustrated since it can heavily influence the quality of the electron micrographs. Finally, positive specimen staining is analysed because it defines the appearance of an electron micrograph. We can also simulate the effect of staining on an amino acid sequence so that we can compare theoretical fibril aggregation models with real collagen microfibrils. The basics of Transmission electron microscopy, sample preparations and staining were taken from Hayat (Hayat, 1989) and Bozzola and Russel (Bozzola and Russel, 1992).

The chapter ends with a description of the materials and the software applications actually used in the thesis

2.1 Cross Correlation

The cross correlation C(x) of two continuous real functions f(x) and g(x) is defined as:

$$C(x) = \int_{-\infty}^{\infty} f(u-x)g(u)du \qquad 2.1$$

and for a given displacement x it describes how much f(x) and g(x) correlate to each other. The higher the value of C(x) the more similar f(x) and g(x) are. Equation 2.1 can easily be extended to the case of numerical sequences. Thus if f{i} and g{i} represent two sequences with N elements each, their cross correlation C{j} is a function of the displacement j of one sequence with respect to the other and is defined by equation 2.2.

$$C\{j\} = \sum_{i=0}^{N-1} f\{i-j\}g\{i\} \qquad 2.2$$

Cross correlations are a useful mathematical tool to quantify the similarities between two different functions or numerical sequences of elements. More specifically, they can be used to quantify similarities between two polypeptides sequences once numerical values representing, for example, the hydrophobicity or the electric charge are assigned to the sequence. A normalised version of equation 2.2 is usually employed in practice and it is simply referred to as Correlation (r):

$$r(j)=\frac{\sum_{i=0}^{N-1}\left[(f\{i-j\}-\bar{f})(g\{i\}-\bar{g})\right]}{\sqrt{\sum_{i=0}^{N-1}(f\{i-j\}-\bar{f})^2}\sqrt{\sum_{i=0}^{N-1}(g\{i\}-\bar{g})^2}} \qquad 2.3$$

where $\bar{f}$ and $\bar{g}$ are the average values of the sequences f{i} and g{i} respectively. The value of r(j) for two sequences f{i} and g{i}, can vary between 1 and -1 indicating a strong correlation for r=1 and a strong anticorrelation for r=-1. If r(j) is close to 0 then the two sequences have no correlation at all.

2.2 Autocorrelation

The cross correlation C{j} of a numerical sequence with itself takes the name of autocorrelation (equation 2.4)

$$C\{j\}=\sum_{i=0}^{N-1} f\{i-j\}f\{i\} \qquad 2.4$$

An autocorrelation is characterised by two important properties, it is symmetrical with respect to the origin of the displacement j (i.e. C{j}=C{-j}) and it has an absolute maximum in it (i.e. C{0}≥C{j} ∀j). The first statement is simply demonstrated using a dummy variable d=i-j. In this case, we write:

$$C\{j\}=\sum f\{i-j\}f\{i\}=\sum f\{d\}f\{d+j\} \qquad 2.5$$

since i-j=d and d+j=i. But

$$\sum f\{d\}f\{d+j\} = \sum f\{d+j\}f\{d\} \qquad 2.6$$

because of the commutative property of multiplication. Finally, by applying the definition of autocorrelation (equation 2.4) we obtain:

$$\sum f\{d+j\}f\{d\} = C\{-j\} \qquad 2.7$$

Combining 2.5 and 2.7 we obtain:

$$C\{j\}=C\{-j\} \qquad 2.8$$

The second statement can be easily demonstrated for a sequence with an infinite number of elements.

An autocorrelation C{j}, of an infinite sequence f{i}, has a maximum for j=0 if:

$$C\{0\} = \sum_{i=-\infty}^{\infty}[f\{i\}]^2 \geq \sum_{i=-\infty}^{\infty} f\{i\}f\{i+j\} \qquad \forall j \in Z \qquad 2.9$$

If we introduce an $\varepsilon \in \mathbb{R}$, the following equation is always true being a sum squared of numbers:

$$\sum_{i=-\infty}^{\infty}[f\{i\} + \varepsilon \cdot f\{i+j\}]^2 \geq 0 \qquad 2.10$$

We can now expand equation 2.10 to obtain

$$\sum_{i=-\infty}^{\infty}[f\{i\}]^2 + 2\varepsilon \cdot \sum_{i=-\infty}^{\infty}[f\{i\}f\{i+j\}] + \varepsilon^2 \cdot \sum_{i=-\infty}^{\infty}[f\{i+j\}]^2 \geq 0 \qquad 2.11$$

Now, the first and the third terms in 2.11 have the same value since f{i+j} is the same as f{i} (it is simply shifted by an amount j, but the sum is calculated over all elements) therefore

$$\sum_{i=-\infty}^{\infty}[f\{i\}]^2 = \sum_{i=-\infty}^{\infty}[f\{i+j\}]^2 \qquad 2.12$$

For ease of representation, we introduce the following notation

$$a = \sum_{i=-\infty}^{\infty}[f\{i\}]^2 \qquad 2.13a$$

$$b = \sum_{i=-\infty}^{\infty}[f\{i\}f\{i+j\}] \qquad 2.13b$$

Thus equation 2.11 can be written as

$$a \cdot \varepsilon^2 + 2b \cdot \varepsilon + a \geq 0 \qquad 2.14$$

For equation 2.14 to be satisfied, it must be that the discriminant b^2-a^2 of the quadratic polynomial

$$a \cdot \varepsilon^2 + 2b \cdot \varepsilon + a = 0 \qquad 2.15$$

is equal or less than zero i.e.:

$$b^2 - a^2 \leq 0 \qquad 2.16$$

this is true when

$$a \geq b \geq -a \qquad 2.17$$

Substituting definitions 2.13a and 2.13b in inequality 2.17 we obtain

$$\sum_{i=-\infty}^{\infty}[f\{i\}]^2 \geq \sum_{i=-\infty}^{\infty}[f\{i\}f\{i+j\}] \geq -\sum_{i-\infty}^{\infty}[f\{i\}]^2 \qquad 2.18$$

Or considering only the first part of equation 2.18 and applying the definition 2.4 and 2.8 we have that:

$$C\{0\} \geq C\{j\} \qquad 2.19$$

for all values of j.

This result is valid for finite sequences too, since a finite sequence can be thought of as an infinite sequence whose additional elements are always zero.

In summary, the value of an autocorrelation tends to decrease as soon as the displacement j between the sequences increases. An auto correlation can be used to study the interaction energies between two identical polypeptide sequences. Since cross correlations and autocorrelations are used to quantify or "score", the

interaction energies due to hydrophobic or electrostatic forces between two sequences of peptides as a function of their mutual displacement, they will be referred here also as scoring functions.

2.3 Convolution

The convolution h(x) between two real functions f(x) and g(x) is defined as:

$$h(x) = \int_{-\infty}^{\infty} f(u)g(x-u)du \qquad 2.20$$

It is similar to a cross correlation but it requires the reversal of the function g(x) with respect to the origin before applying a displacement x. Equation 2.20 can be written for numerical sequences as follows

$$h\{j\} = f\{i\} * g\{i\} = \sum_{i=0}^{N-1} f\{i\}g\{j-i\} \qquad 2.21$$

The convolution operation, between two sequences f{i} and g{i}, is denoted by the symbol *. Convolution is commutative that is f{i}*g{i}=f{i}*g{i}. To see this we apply first the definition 2.20 to calculate the convolution of f{i}*g{i}. It is convenient here to make f{i} and g{i} infinite sequences by adding an infinite number of zero elements. Equation 2.20 becomes then

$$f\{i\} * g\{i\} = \sum_{i=-\infty}^{\infty} f\{i\}g\{j-i\} \qquad 2.22a$$

By substituting j-i with the dummy variable k, so that i=j-k and j-i=k we obtain:

$$\sum_{i=-\infty}^{\infty} f\{i\}g\{j-i\} = \sum_{k=\infty}^{-\infty} f\{j-k\}g\{k\} \qquad 2.22b$$

The sum is calculated for i=-∞→∞ or equivalently using the dummy variable k, for k=-∞→∞. Therefore

$$\sum_{k=\infty}^{-\infty} f\{j-k\}g\{k\} = \sum_{k=-\infty}^{\infty} g\{k\}f\{j-k\} = g\{k\} * f\{k\} \qquad 2.23$$

The convolution operation is also associative and distributive with respect to addition

$$f\{i\} * (g\{i\} * h\{i\}) = (f\{i\} * g\{i\}) * h\{i\} \qquad 2.24$$

$$f\{i\} * (g\{i\} + h\{i\}) = f\{i\} * g\{i\} + f\{i\} * h\{i\} \qquad 2.25$$

and it can be shown very easily by applying the definition of convolution 2.20 and using the associative and the distributive properties of addition. Below we prove briefly the distributive property:

$$f\{i\} * (g\{i\} + h\{i\}) = \sum_{i=0}^{N-1} f\{i\}(g\{i-j\} + h\{i-j\}) = \sum_{i=0}^{N-1} (f\{i\}g\{i-j\} + f\{i\}h\{i-j\}) =$$

$$= \sum_{i=0}^{N-1} f\{i\}g\{i-j\} + \sum_{i=0}^{N-1} f\{i\}h\{i-j\} = f\{i\} * g\{i\} + f\{i\} * h\{i\} \qquad 2.26$$

The associative property can be proved using the substitution i+k=l:

$$f\{i\} * (g\{i\} * h\{i\}) = f\{i\} * (h\{i\} * g\{i\}) = \sum_{j=0}^{N-1} f\{j\} \sum_{i=0}^{N-1} h\{i\} g\{i + k - j\} =$$

$$= \sum_{j=0}^{N-1} f\{j\} \sum_{l=0}^{N-1} h\{k - l\} g\{l - j\} = \sum_{l=0}^{N-1} (\sum_{j=0}^{N-1} f\{j\} h\{l - j\}) g\{k - l\} = (f\{i\} * h\{i\}) * g\{i\}$$

2.27

Below we will use electron micrographs taken with a transmission electron microscope. The convolution operation can help us interpret these micrographs because from a physical point of view the action of any instrument (including sample preparation) over a measurable quantity f(x) can be interpreted as a convolution g(x) * f(x) where g(x) represents a distortion due to the preparation and the measurement process. Thus, the final measure h(x) of a quantity f(x) is limited by the measurement process through g(x).

2.4 Fourier Transforms and Convolution Theorem

The Fourier Transform F(s) of a real function f(x) is defined as:

$$F(s) = \int_{-\infty}^{\infty} f(x) e^{-i 2\pi x s} dx \qquad \text{with } i = \sqrt{-1} \qquad 2.28$$

The inverse Fourier transform of F(s) is defined as:

$$F^{-1}(x) = \int_{-\infty}^{\infty} F(s) e^{i 2\pi x s} ds \qquad 2.29$$

The Fourier transform and the inverse Fourier transform for a real sequence f{j} are defined as:

$$F(k) = \sum_{j=0}^{N-1} f(j) e^{-2i\pi(k/N)j} \qquad 2.30$$

and

$$F^{-1}(j) = \frac{1}{N} \sum_{k=0}^{N-1} F(k) e^{2\pi i(j/N)k} \qquad 2.31$$

The Fourier transform highlights the spectral components of a function, or a sequence, so that they are the best instruments to analyse frequency distributions. The Fourier transform can be used in combination with the Convolution Theorem to study the periodic structure of a function.

The convolution theorem states that if F(s) and G(s) are the Fourier transforms of f(x) and g(x) respectively, then f(x)*g(x) has Fourier transform F(s)·G(s), that is:

the Fourier transform of the convolution of two functions is the multiplication of the Fourier transform of the functions:

$$F(f(x) * g(x)) = \int_{-\infty}^{\infty} \left[\int_{\infty}^{\infty} f(y) g(x-y) dy \right] e^{-2i\pi xs} dx = \int_{-\infty}^{\infty} f(y) \left[\int_{-\infty}^{\infty} g(x-y) e^{-2i\pi xs} dx \right] dy =$$

$$\int_{-\infty}^{\infty} f(y) e^{-2i\pi ys} G(s) dy = F(s)G(s) \qquad 2.32$$

Let us suppose that we are examining a periodic structure P(x) (figure 2.1). This means that we have a motif that is equally spaced several times. If we represent the motif as f(x) then P(x) is given by the convolution of f(x) with a δ-comb III(x), that is:

$$P(x)=f(x)*III(x) \qquad 2.33$$

We can apply the convolution theorem to P(x) to calculate its Fourier transform:

$$F(P(x))=F(f(x)*(III(x)))=F(f(x))F(III(x)) \qquad 2.34$$

As a result we have that the Fourier transform of the periodic structure is given by the Fourier transform of the motif times the Fourier transform of the δ-comb. We will use this theorem in the next chapter.

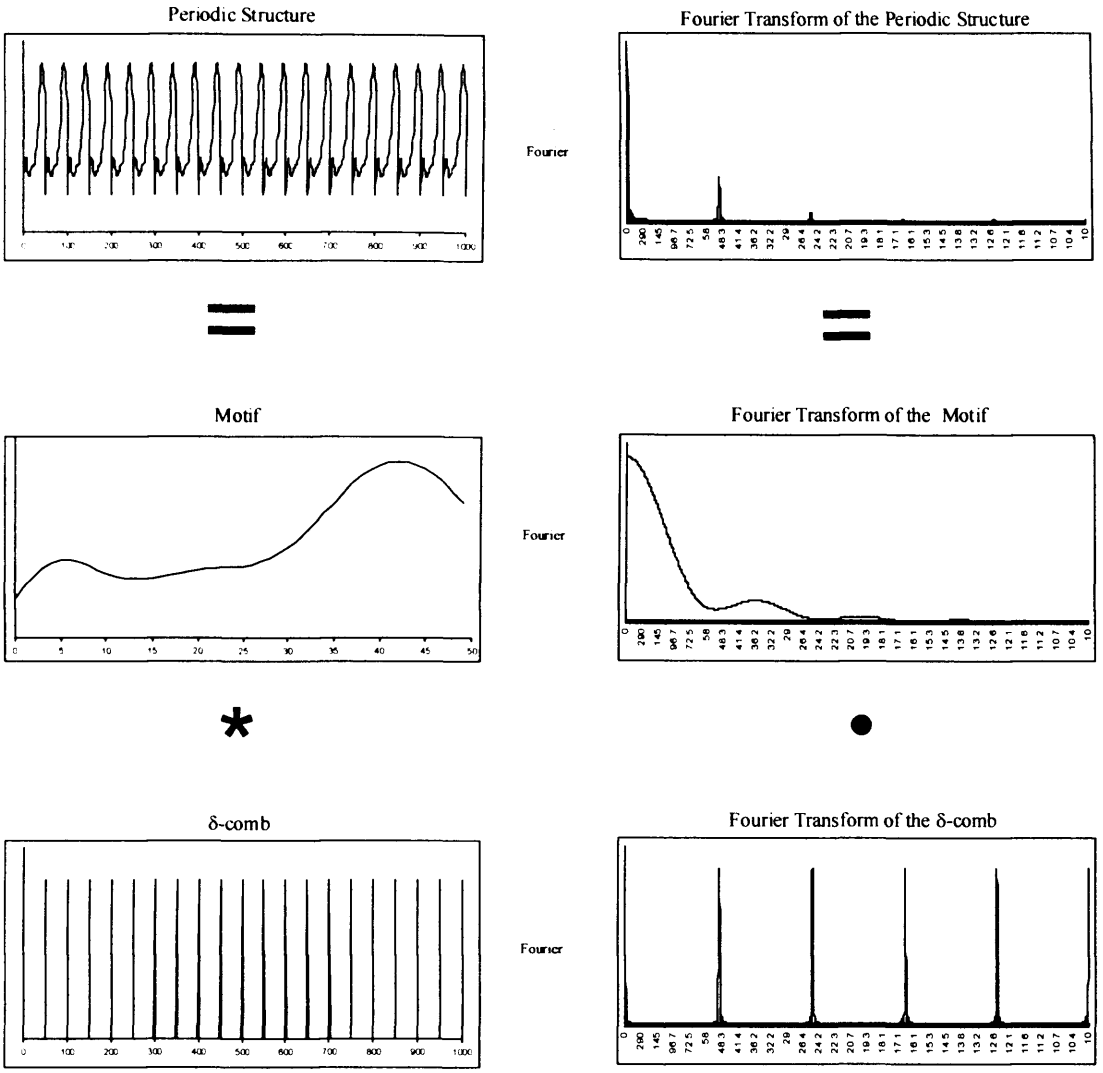


Figure 2.1: Diagrammatic representation of the Convolution Theorem. A periodic function is given by the convolution of a motif with a δ-comb. The Fourier transform of the periodic function is equal to the multiplication of the Fourier transform of the two convoluted functions.

2.5 Transmission Electron Microscope

From a conceptual point of view, a transmission electron microscope is very similar to an optical projection microscope. In fact, as shown in figure 2.2, both kinds of microscope are made of 1) a radiation source and a set of condenser lenses, that are used to select the amount of radiation reaching the specimen, 2) an objective lens that gathers and magnifies the radiation that has passed through the sample, 3) a series of projector lenses that magnify the image and 4) a system where the final image is formed and recorded. In the older designs, images were usually recorded on films treated with silver salts. More recently, the development of accurate CCD cameras permits the storage of images directly on a computer in a digital format.

The biggest difference between the two types of microscope is the type of radiation used: visible light for an optical microscope and focussed electrons in the case of the Transmission Electron Microscope (TEM). The wavelength of the radiation used determines the theoretical resolution of a microscope because it can not be greater than the size of the object observed. The resolution limit for a light microscope is therefore about 200 nm, while a very good resolution for an electron transmission microscope is about 0.1 nm. However, values around 2 nm for the latter case are more realistic.

As electrons have mass, it is extremely important that they travel through a *vacuum* chamber to avoid any scattering due to air molecules. For this reason, the air pressure in the column of a TEM is maintained at very low levels.

More specifically, the transmission electron microscope is essentially made of four units:

- An illuminating unit
- A specimen holder unit
- An imaging unit
- An imagine formation unit.

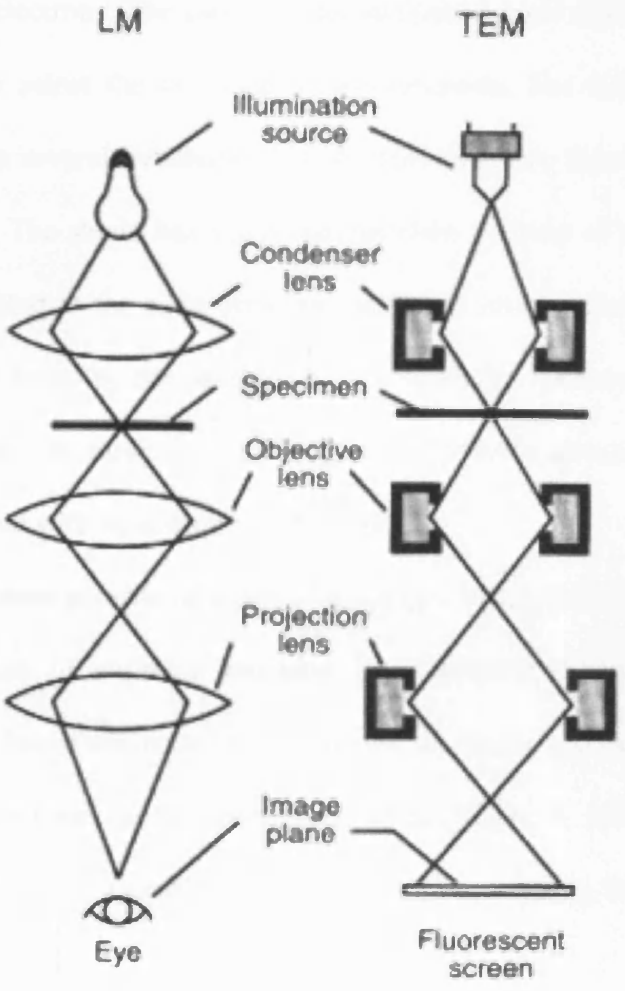


Figure 2.2 Comparison between transmission electron microscope (TEM) and an optical projection microscope (LM).
(Adapted from www.uiowa.edu/~cemrf/methodology/tem/index.htm).

2.5.1 Illuminating unit

The illuminating unit is made of the electron gun and a set of condenser lenses. The electron gun is a device that produces electrons via thermo-ionic emission from heated metal. Usually a negative voltage is imposed on a thin (about 0.1 mm in diameter) tungsten filament and then it is heated to force the emission of electrons. Tungsten is used because of the high melting point (around 3650 K) and because of the high yield of electrons. The emission due to thermo-ionic effect is isotropic. It is thus necessary to select the direction of the electrons. For this reason, a shield whose potential is several hundreds of volts more negative then the filament (100 kV) surrounds it. The shield has a 2-3 mm aperture in front of the filament. Only the electrons emitted in the right direction and with enough energy to breach the negative field produced by the shield can go trough the aperture. As soon as this happens they are forced away from the negative shield and attracted by a grounded anode, a metal plate with an aperture in the centre.

The condenser system is made of a pair of magnetic lenses that have the purpose of control the amount of radiation reaching the specimen. Usually the first lens demagnifies the size of the cross section of the electron beam coming out of the anode by about ten times, while the second one magnifies it. This configuration is necessary to control the amount of electron radiation reaching the sample to limit damage to the specimen.

2.5.2 Specimen holder unit

The specimen manipulating system is made of a specimen holder and a specimen stage. The specimen holder is a bar with a molybdenum strip at the end where it is

possible to insert the copper grid containing the specimen. The rod is inserted in an air lock system and pushed in the middle of the electron microscope column. Once the rod is in position, the stage manipulator allows to move the specimen in the plane orthogonal to the electron beam. Movements can be as small as 10 nm. It is also possible to tilt the specimen along the rod axis in order to collect images suitable for three-dimensional reconstructions.

2.5.3 Imaging unit

The imaging unit gathers the radiation that has passed through the specimen and focuses it on the focal plane where the screen (or the camera) is. It is composed of an objective lens, an intermediate lens and a series of projector lenses. The objective lens has a very short focal length of about 1 mm and it is used mainly to focus the image, while the intermediate and the projector lenses magnify it. When the electrons pass through the specimen, they undergo two main processes that give rise to a contrasted image. They can be backscattered by the specimen, or they can interact with it without any appreciable deviation in their trajectory but undergoing a phase shift. They can also be scattered without leaving the electron beam thus blurring the final image and lowering the contrast. To avoid this effect, a selectable aperture is placed below the objective lens and it is used to eliminate scattered electrons from the beam. The intermediate lens is below the objective lens and magnifies the image produced by the objective. The illuminating system terminates with the projector lenses that further magnify the image coming from the intermediate lens. They are used to increase the depth of focus since the viewing

screen and the camera system do not lie at the same distance from the intermediate lenses.

2.5.4 Image formation unit

The image is formed on a phosphorescent screen where it can be examined with the aid of a stereomicroscope. The viewing screen has marks to help positioning the image directly upon the film contained in a chamber underneath the screen. Once the image has been focused and centred, the chamber is open, the screen is removed, and the image taken on film.

2.6 Specimen preparation

The image formed in the transmission electron microscope is due to the electrons that pass through the specimen. This means that the specimen undergoes mechanical and thermal stresses and shrinkage caused by the impinging electrons. Moreover, sectioning itself can damage the sample. It is also necessary to preserve the specimen for a long period minimising as much as possible any alteration. Therefore, specimen must be processed before being examined in the electron microscope. Ideally, specimen preparation consists in fixing the constitutive elements of the specimen such as cells and proteins in the extra cellular matrix, and then in stabilising the architecture of the constituents by means of resins before sectioning. In a more detailed way, the specimen preparation protocol can be subdivided in the following fundamental steps:

- Primary fixation
- Washing
- Secondary fixation
- Dehydration
- Embedding.

Fixation and embedding being the most important elements.

Fixation is needed to preserve the structure of the specimen. The fixatives employed are different, but glutaraldehyde and osmium tetroxide are the most widely used for primary and secondary fixation respectively. Glutaraldehyde is composed of two aldehyde groups at the extremity of three methylenic groups as shown below for the monomer form:

$$CHO\text{-}CH_2\text{-}CH_2\text{-}CH_2\text{-}CHO \qquad 2.35$$

During the fixation procedure, the aldehyde groups react with the α-amino group of the lysines belonging to the proteins of the specimen. Other reactions involving amino acids other than lysine can also occur. Independently of the reactions occurring between glutaraldehyde and the proteins amino acids, cross-link formation is wide spread across the sample. This results in the formation of a solid meshwork interconnecting the proteins of the specimen. Elements that are not directly bound to glutaraldehyde are thought to be entrapped in the tight network. It is as if a rigid scaffolding interconnecting the soft tissues of the specimen were built.

Osmium tetroxide reacts with the unsaturated bonds of the fatty acids oxidising them. During this process, osmium tetroxide is reduced to black metallic osmium that has good properties of density and staining enhancing the contrast of the image of the specimen obtained in the electron microscope. Both fixatives are diluted in buffer solution to maintain the pH of the specimen near the biological values. As glutaraldehyde does not bind to fatty acids, they can be extracted during the dehydration procedure. To avoid this, glutaraldehyde fixation is followed by osmium tetroxide fixation. Specimen should be prepared in very thin slices (≈1mm) because the rate of penetration of the fixatives is very slow and this could cause inhomogeneous fixation of the specimen. Unreacted glutaraldehyde can be oxidised by osmium tetroxide forming undesired spots into the specimen. It is therefore necessary to wash the treated sample to eliminate free glutaraldehyde molecules.

Living tissues are naturally surrounded by water. Therefore, water can be thought of a natural embedding medium supporting specimen constituents. Sectioning is a very stressful event for the specimen that undergoes shears pressures that can alter its structure in a permanent way. To avoid this it is necessary to remove the water molecules in the specimen and to substitute them with a more resistant medium this is done with dehydration and embedding.

Dehydration consists in extracting the water from the cellular and extra cellular constituents with solvents. Common solvents are ethanol and acetone. The specimen is immersed in different solutions of the solvent with increasing concentrations. Depending upon the different protocols used, a transitional solvent is used before embedding. The dehydrating agent is substituted with the transitional agent (usually propylene oxide) that has the property to mix with the embedding resin facilitating its penetration into the sample.

To embed the tissue, the transitional solvent is mixed with a plastic resin, usually epoxy. The concentration of the resin in the solvent is slowly increased in steps to substitute the solvent with the resin. To improve permeation the vials containing the resin are delicately agitated. After a good permeation has been achieved, the resin block is polymerised putting it into an oven or illuminating it with an UV lamp. Eventually, the resin block is removed from the vial and trimmed with a razor or a more sophisticated instrument in order to obtain a pyramid shaped block whose apex contains the specimen.

2.7 Sectioning

The transmission electron microscope uses an electron beam to probe the specimen. Typical values of the accelerating potential are in the range between 50 and 100 kV. This implies that electrons do not have enough energy to penetrate excessively thick specimens. For this reason, it is important to cut the specimen in extremely thin sections. Once the plastic block has been roughly trimmed, in order to expose the specimen, it is mounted in a microtome. The microtome is an instrument that is constituted of a swinging arm, a knife and a collection trough. At the end of the arm, the specimen block is clutched and it is then slowly moved toward the knife. This is achieved by thermal dilatation or by mechanical advancement of the arm. The vertical movement that produces the actual sectioning is obtained by letting the arm fall delicately via an internal mechanism. The knives are usually made of glass or diamond depending upon the quality of the sectioning needed, diamond being used for the thinnest sections. Once the arm touches the edge of the knife an ultrathin slice is deposited into the collection trough, a very small tank filled with

distilled water. If the procedure is correctly performed, specimen slices form a ribbon that can be collected using a specimen grid. After the water has been absorbed with filter paper and the grid has dried, the specimen is ready for the staining procedure.

2.8 Positive staining

The most abundant atoms in biological specimens are carbon, nitrogen, oxygen and hydrogen. They are light atoms that are almost transparent to the electron beam passing through the specimen. In addition, the embedding medium is made of the same type of atoms thus the contrast in the microscope is very poor. As a result, specimens that have been only fixed and embedded produce electron micrographs that have very low contrast and detail. To increase contrast it is necessary to increase the differential backscattering ratio between the specimen and the medium. This is achieved by staining the specimen. It is commonly performed after embedding and it consists in attaching heavy metal atoms to membranes, proteins or macromolecules in the specimen to increase their density with respect to the surrounding environment. The enhanced contrast is a direct consequence of the differential density among different portion of the specimen because stains are electron opaque and backscatter electrons in an effective way. In addition, as the stain attaches to the surface of the molecules examined, it shields them from the electron bombardment enhancing the duration of the specimen. Ideally, a good staining should also be site specific for the molecules studied. Unfortunately, the staining mechanism is not fully understood, thus it is not possible to establish a definite relationship with all specimens. However, an empirical procedure could

consist in varying the duration of staining or its concentration in the solution in order to attach it only to those sites that react more quickly to its action. It is nonetheless possible to establish a relationships between charged amino acids and polar stains as will be shown later. Positive staining also presents some minor drawbacks that must be considered. As we have seen, staining consists in heavy metal atoms, or more precisely, groups of heavy metal atoms adhering to the specimen. This implies that the electron microscope resolution is heavily limited by the size of the metal aggregates. The lateral dimension of these aggregates can vary widely from 1 to 8 nm. according to the type used. Moreover, as the metal clusters deposit on the molecules surfaces, they can increase their apparent volume.

The major part of the stains forms ions that bind to the charged sites in the proteins. Among them phosphotungstic acid and uranyl acetate are those used to treat collagen proteins.

2.8.1 Phosphotungstic Acid

Phosphotungstic acid (PTA) belongs to the family of polar or ionic stains. Tungsten is a metal whose atomic number is 74. It binds to phosphate and oxygen to form a variety of anionic compounds whose ultimate structure depends upon the pH of the solution. For lower pH, $PW_{12}O_{40}^{3-}$ is the form most widely diffused, whereas the form $PW_{11}O_{39}^{7-}$ is more common for pH values between 5.5 and 7. The mode of binding with the substrate is not precisely known. However, it is reasonable to consider it as essentially electrostatic. Thus, it can be used as a selective stain to highlight positively charged sites in proteins. In particular, it binds to basic residues of collagen such as arginine and lysine. In this way, we can establish a direct

relationship between stains and the distribution of positive amino acids along fibrillar collagen molecules.

2.8.2 Uranyl Acetate

Uranium is a metal with atomic number 92. It binds to oxygen to give uranyl ions UO_2^{2+}, that form a family of staining agents among which uranyl acetate (UA) is the most used. Uranyl acetate ($UO_2(C_2H_3O_2)_2$) dissolves in solution with acetic acid forming a saturated solution at pH values at about 3.5-4. In this pH range, different types of positive and negative uranyl ions are found. Thus, there is no polar site specificity as compared to PTA. In other words, it is not possible to associate UA only with the distribution of negative residues along a protein. In solutions with higher pH values, the uranyl ions form complexes with acetate ions to give once again a series of cations and anions. Nonetheless, in protein staining, UA is supposed to bind primarily with the side chains of aspartic acid and glutamic acid. Thus, it can be used to highlight the distribution of negative, even though not exclusively, residues along collagen molecules.

2.9 Simulated staining

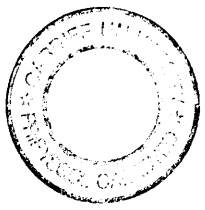
Phosphotungstic acid and uranyl acetate are polar staining agents commonly used for collagen. Their combined effect is to mark selectively the distribution of charged residues along the collagen molecule or other proteins. Therefore, it is possible to identify charged groups in a protein. If the amino acids sequence of a protein is known, it is possible to model the effects that staining would have by

assigning a numeric value to its charged amino acids that mimics their staining uptake as we are going to show later on. This principle is particularly straightforward with a "simple" molecule such as collagen of the fibril forming families. In fact, the rod-like triple helical domain of fibrillar collagen can be essentially thought of as a rigid linear structure where it is easy to pinpoint the staining pattern from the amino acids sequence. This idea was first proposed by Meek and colleagues (Meek et al 1979; Chapman et al. 1981), with modern computer its application is straightforward, and was applied even recently by Ortolani and colleagues (Ortolani et al 2000). One uncertainty in this method is that the extent of the stain uptake of the amino acids is not known accurately and it is difficult to assign a correct stain value to the charged residues.

Here, we have assigned a value of 1 to arginine, lysine and hydroxylysine, glutamic acid and aspartic acid and 0 to all other amino acids.

In practice, the protocol is quite simple, and for type I collagen for example, we assume that the triple helical domain is made of three rigid linear chains. We assign a stagger of one amino acid position i.e. 2.86 Å, to one chain with respect to the others according to the distribution α_2-α_1-α_1 (α_2 being the reference chain and the two α_1 chains being shifted by 2.86 and 2*2.86 Å respectively along the molecule axis). The particular order chosen for the chain distribution must be specified only for heterotrimeric fibrillar collagens. However, as we explain later we think that, at this stage of the analysis, it does not affect the representation of the collagen fibril. In an Excel spreadsheet, we write down the α-chains amino acids sequence as array columns that are parallel to each other. We substitute each amino acid with its corresponding numerical value and calculate the lateral sum of the arrays. In this way, we obtain a one-dimensional sequence of values that represents the stain

pattern along the collagen molecule. It is important to observe that with this construction, each element of an α-chain sequence is one residue position away from the adjacent one. Since the simulated staining highlights the position of the single charged amino acid, the simulated staining sequence would have a resolution of one amino acid position by construction. However, this is not what gets measured in practice in an electron microscope. In fact, the resolution obtained in an electron microscope is greater than 2.86 Å. To compare the simulated staining to the experimental one it is necessary to process the simulated staining pattern further. First, it is necessary to reduce its resolution to one comparable with what it is seen in the electron microscope micrographs. This is achieved by convoluting the simulated staining pattern with a normalised Gaussian function with σ=3 amino acids. The full width at half maximum (FWHM) of this sequence is about 7 residues wide and so the convoluted sequence has a resolution of about 2 nm that is about the resolution for biological specimen of an electron microscope. The outcome is a "blurred" stain distribution that fits more neatly the experimental data. Therefore, also for the consequences due to this procedure, we consider that the order chosen for the collagen chains does not influence in an appreciable way the simulated staining pattern. A further refinement consists in scaling the simulated staining pattern to compare it better to the experimental one. It is obtained by multiplying the sequence values by a scalar factor; this procedure does not affect the relative values of a simulated stain. The same procedure is applied to a collagen microfibril that is the basic unit in a collagen fibril. In this way, an immediate comparison between the model and the actual collagen fibrils is obtained. Figure 2.3 represents a diagram of the simulated staining technique for a portion of a type I collagen microfibril.



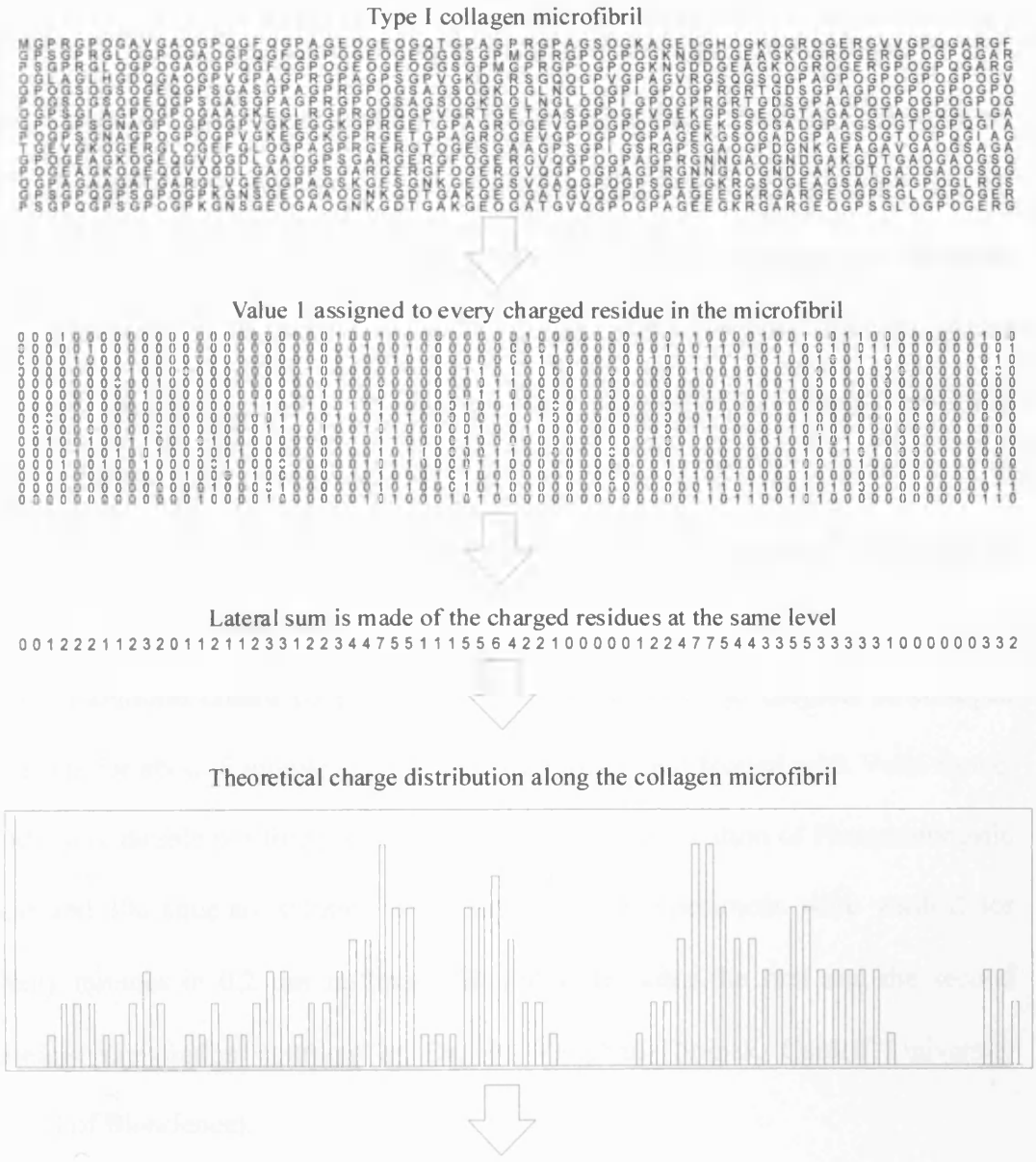


Figure: 2.3 Diagrammatic representation of the simulated staining procedure for a portion of type I collagen microfibril.

2.10 Sample preparation

Collagen was acid extracted from tissue. Individual collagen types were isolated by salt precipitation in different NaCl solutions concentrations. Purified solid collagen was obtained from freeze drying. Solid collagen extracts were solubilised in diluted acetic acid concentrations at 4°C. The mixtures of collagen in acetic acid solutions were transferred in dialysis tube and dialysed against neutral PBS (Phosphate Buffered Solution) at 4°C overnight. Fibril assembly was finally obtained transferring the tubes in warm bath at 37°C.

Type II and type XI collagen samples were from rat chondrosarcoma, while type III collagen was from rabbit skin.

G 300 Pioloform coated copper grids were floated on 25 μl droplets of collagen solutions for about 5 minutes. Then they were dried and blotted with Velin tissue. Grids were double positively stained with 1% aqueous solution of Phosphotungstic Acid and 1% aqueous solution of Uranyl Acetate. Specimens were washed for twenty minutes in 0.2 μm millipore filtered water after the first and the second staining (preparation courtesy of Dr. A. Vaughan-Thomas, Cardiff University School of Bioscience).

Micrographs were taken by Dr Rob Young with a Philips EM 208 electron microscope.

We digitised the micrographs using an Epson Perfection 4990 Photo scanner with the following settings: professional mode, film with area guide, B&W negative film, 16-bit grey-scale, 2400 dpi resolution, Jpeg format file.

2.11 Software applications

We used some software applications in the public domain to analyse the electron micrographs, to build models of collagen molecules and to visualise them:

- Image processing application: ImageJ 1.37 was used to analyse the electron micrographs. More specifically, images were rotated and cropped to bring them to the desired size suitable for image analysis. Subsequently a Gaussian blur filter with 8 pixels radius was applied to reduce the noise levels. To finish, the regions of interest were selected and their density profiles were calculated and saved in txt format.
- Model building application: Deep View/ Swiss-PdbViewer 3.7 was used to build the sequences representing a collagen molecule
- Model visualisation application: UCSF Chimera beta version 1 build 2065 2004/12/15 was used to represent models of collagen molecule.

Microsoft office Excel 2003 was used to analyse the data, while Corel Draw 11 was used to prepare some images.

I wrote two C++ programs to calculate the scoring function of the collagen sequences (Scoring_function.cpp) and to perform linear interpolation of the functions (Interpolation.cpp).

I used a C++ program written by Dr. Carlo Knupp to calculate the Fourier Transform of the functions analysed (Fast_Fourier_Transform.ccp, arctan.h, fft.h).

In the Appendix, the human readable format of the programs mentioned above is presented.

2.12 Collagen sequences.

α-chains collagen sequences published in UniProt (http://www.ebi.uniprot.org/index.shtml) were used. The different types of collagens studied are associated to the following primary access numbers:

Type I collagen:

- α_1(I): P11087
- α_2(I): Q01149

Type II collagen:

- α_1(II): P28481

Type III collagen:

- α_1(III): P08121

Type XI collagen:

- α_1(XI): Q61245
- α_2(XI): Q64739
- α_3(XI)=α_1(II): P28481

CHAPTER 3

MODELLING FIBRILLAR COLLAGENS

3 Introduction

In this chapter, we present a study of the interactions between murine type I collagen molecules which is based on a scoring technique first pioneered by Hulmes and colleagues (Hulmes et al., 1973). First, we concentrate our attention on the triple helical region of two type I collagen molecules. This is to facilitate the initial stages of the analysis. In fact, the molecular structure of the triple helical region is well known, so that it can be built in our models and only the intermolecular interactions need to be determined. Once this is done, it is easier to deal with the telopeptide regions, whose three-dimensional structure is not determined in collagen fibres. Moreover, the distribution of the hydrophobic amino acids in the triple helical region is such that the scoring function between two molecules is made of well-defined peaks. This makes it easier to find regular patterns that can be exploited to build preliminary models that are used to probe the interactions between molecules and sequences of molecules.

Finally, the interactions between a molecule and a sequence of molecules oriented in an antiparallel fashion are also considered to understand how collagen packs in an antiparallel fashion.

3.1 Scoring method as a means to study interactions between collagen molecules

One of the most important features of the fibrillar collagens is their rigid rod-like structure. If in addition, we consider that the helical rise per residue is constant, it is possible to represent a collagen triple helix as an array of values where the columns

represent α-chains and the rows represent residues at the same axial level. Figure 3.1 represents an idealised diagram where it is possible to see the first 21 residues of the type I collagen triple helix. To study the interactions between two collagen molecules, we assign a value to every residue that is responsible for the particular interaction studied. If, for example, we want to study hydrophobic interactions we assign the value 1 to all the major hydrophobic amino acids residues in the array (F, I, L, M, V, Y), while we assign the value 0 to all the remaining residues. This simple assumption is necessary because hydrophobic values presented in literature are not consistent with each other (Kyte and Doolittle, 1982; Wimley and White, 1996; Hessa et al., 2005). Moreover, they are obtained from macroscopic measurements; it is thus difficult to assign them a value that represents the interactions between two amino acids. A similar approach is used in dealing with charged amino acids because the values of the electrostatic interactions at the microscopic level are not known due to the difficulties in calculating the dielectric constants for proteins (Nakamura et al., 1988). With our representation, we are thus describing if there is interaction between two facing amino acids (when both have the value 1) or if there is no interaction at all (when at least one has the value 0). The array thus obtained is then collapsed as a linear sequence by means of a lateral sum of all the hydrophobic values at the same axial position.

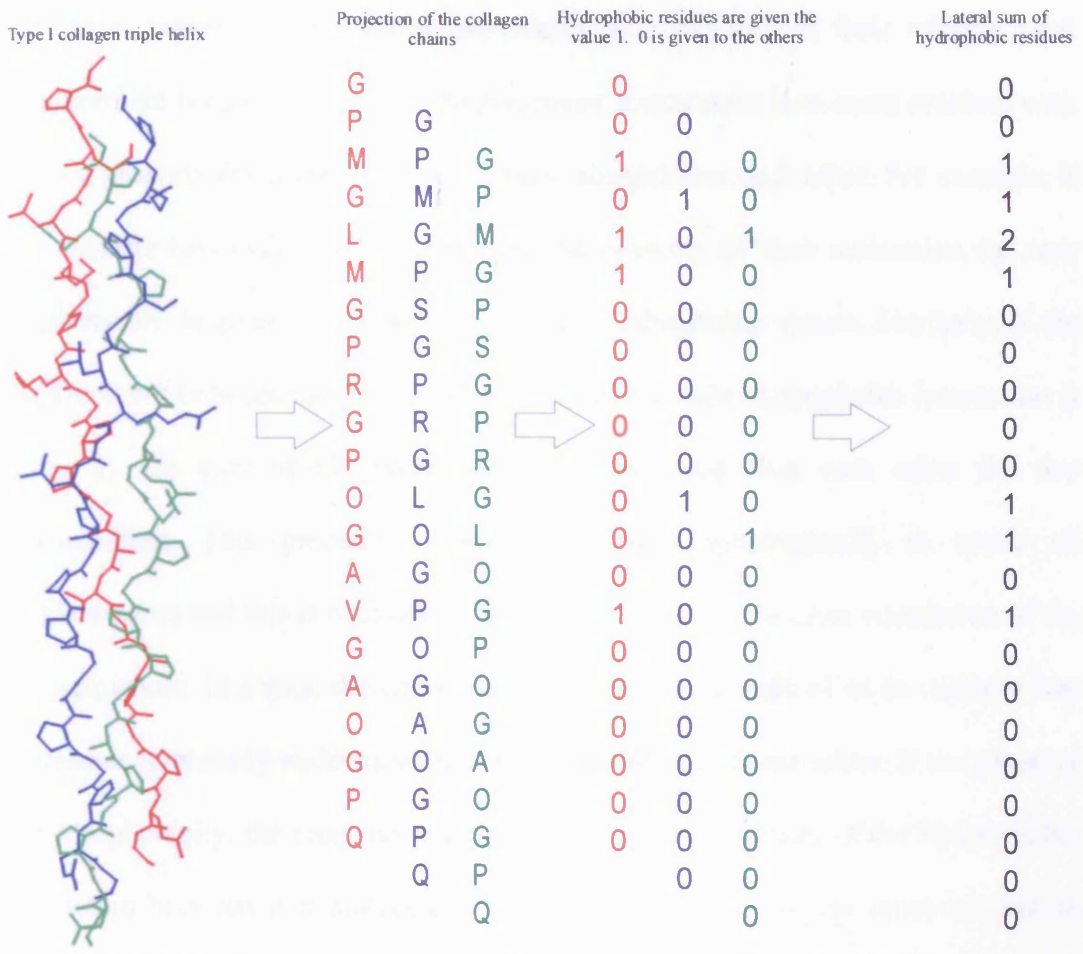


Figure 3.1: Steps undertaken to build a type I collagen sequence for hydrophobic residues. The type I collagen triple helix is inserted in an array, the value 1 is assigned to hydrophobic residues (F, I, L, M, V, Y), the value 0 is assigned to the other residues. Residues on the same row are finally summed together.

This procedure is applied with minor and straightforward alterations to build arrays that represent the positively charged amino acid residues or the negatively charged amino acids residues in a collagen molecule.

With this new representation, the numerical sequence obtained and the collagen triple helix can be thought of as being equivalent. Once a collagen triple helix is represented as a one-dimensional sequence of numerical values, it is possible to calculate its cross-correlation (equation, 2.2) to probe the interactions between two collagen triple helices. In this way, we are assuming that the two collagen molecules are in contact one with the other and that the interaction occurs between facing amino acids. The cross-correlation is a good representation of the intensity of

the forces between two adjacent molecules as a function of their relative axial displacement because for a given displacement it expresses how many residues with a given property (hydrophobicity or electric charge) face each other. For example, if we consider two hydrophobic sequences the intensity of their interaction for zero displacement is given by the sum of all their hydrophobic values. Similarly, if the displacement between the two molecules increases, their hydrophobic interaction is given by the sum of all hydrophobic residues that face each other for that displacement. This procedure can be repeated systematically to cover all displacements and this is equivalent to the calculation of the cross correlation of the two sequences. In a way, the cross correlation can be thought of as an operator that represents how many residues with a given property face each other. If the property is hydrophobicity, the cross correlation represents the intensity of the hydrophobic interaction between two molecules. The higher the peaks in the cross-correlation graph, the higher is the intensity of the hydrophobic interaction at the axial displacement defined by the peaks. The contribution of the cross-link forming hydroxylysines is not taken into account by this treatment and must be considered in a separated treatment that is used to devise an optimal folding conformation for the telopeptides (see below). However, this simple model does not take into account the electrostatic forces between amino acids as a function of their distance, because it is applied only to facing amino acids, or the local dielectric properties of the collagen molecule because it is not possible to calculate it in a unique way. Similarly, the extension of surface contact occurring between hydrophobic amino acids is not considered, therefore our representation must be considered as a simplified model and not a physical description of the interactions actually occurring among collagen molecules. If we consider two identical triple helices,

their hydrophobic sequences are identical. Thus, their cross-correlation is in fact an autocorrelation (see paragraph 2.2). The autocorrelation has the property of being symmetrical with respect to the origin (equation 2.8) and of having a decreasing intensity as the axial displacement increases (equation 2.19). If we think of two identical triple helices facing each other with zero displacement and then we increase the axial displacement as one helix slides past the other, it is apparent how increasingly smaller regions of the helices are in contact. This results in smaller and smaller intensities for their cross-correlation simply because fewer amino acids can interact with each other.

Since the cross correlation and autocorrelation can be thought of as a scoring of the interaction of two adjacent molecules, we will refer often to these methods as scoring or autoscoring methods respectively.

3.2 Hydrophobic autocorrelation for murine type I collagen triple helix

Initially, we used all the possible combination of the α-chains (α_1-α_1-α_2, α_1-α_2-α_1, α_2-α_1-α_1) to represent the murine type I collagen triple helix. Subsequently we built their hydrophobic sequence and calculated the autoscoring as described above. However, the α_2-α_1-α_1 combination was the one that gave the neatest profile, in the sense that the main peaks of the autocorrelation had greatest intensity and a better defined shape. In addition, densitometric comparisons between micrographs and models made by Bender and colleagues (Bender at al., 1982) suggested that this was the most probable configuration. For these reasons, we used this combination

to model the collagen molecule even if the results obtained using the other combinations are, in essence, comparable.

At this point of the analysis, we focussed our attention only on the hydrophobic interactions because in an aqueous environment hydrophobic forces are easier to understand than electrostatic ones (see below for a discussion). Figure 3.2 presents the hydrophobic autoscoring for the murine type I collagen triple helix as a function of the displacement between the two molecules expressed in amino acids (a.a.). As explained above, the autocorrelation presents an absolute maximum at the origin and it decreases with increasing intermolecular displacement.

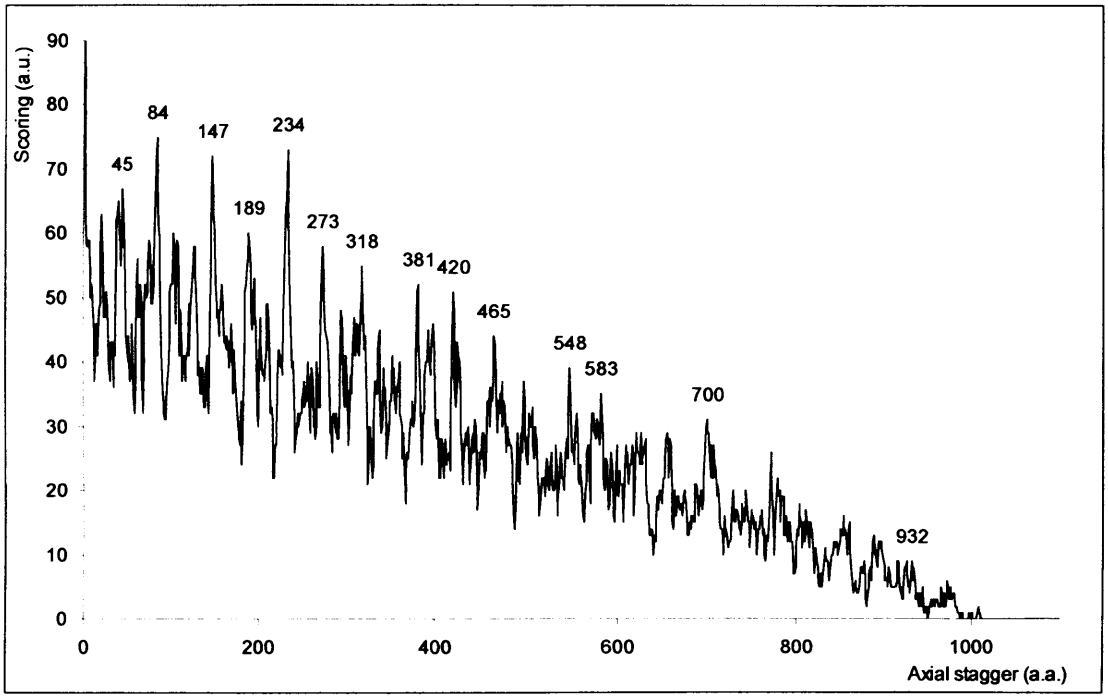


Figure 3.2: Hydrophobic scoring function for two type I murine collagen triple helices. To facilitate the graph's interpretation the peak at position 0 a.a. is not shown.

As pointed out by Hulmes et al. (1973), it is possible to notice intense peaks in position 234, 465, 700 and 932 that are separated by a distance of about 233-234 a.a.. If these peaks were the most intense, they would tell us that the intensity of the hydrophobic interactions is greatest when the molecules are displaced axially by

234, 465, 700 and 932 a.a.. These distances correspond to integer multiples of 67 nm (234*2.86Å) and would explain the characteristic banding seen in collagen fibres. However, they do not appear to be the peaks with the highest intensity in the graph (peaks at positions 84 or 147 and 234 are the most intense). Moreover, there are other peaks at positions 45, 189, 273, 318 etc. that are very intense and are not linked to each other by a periodicity of 234 a.a.. It is thus not possible to conclude definitely, by using only this graph, that the 234 a.a. periodicity is the defining one. This is also confirmed by the Fourier analysis of the peak distribution of figure 3.2 that we present in figure 3.3.

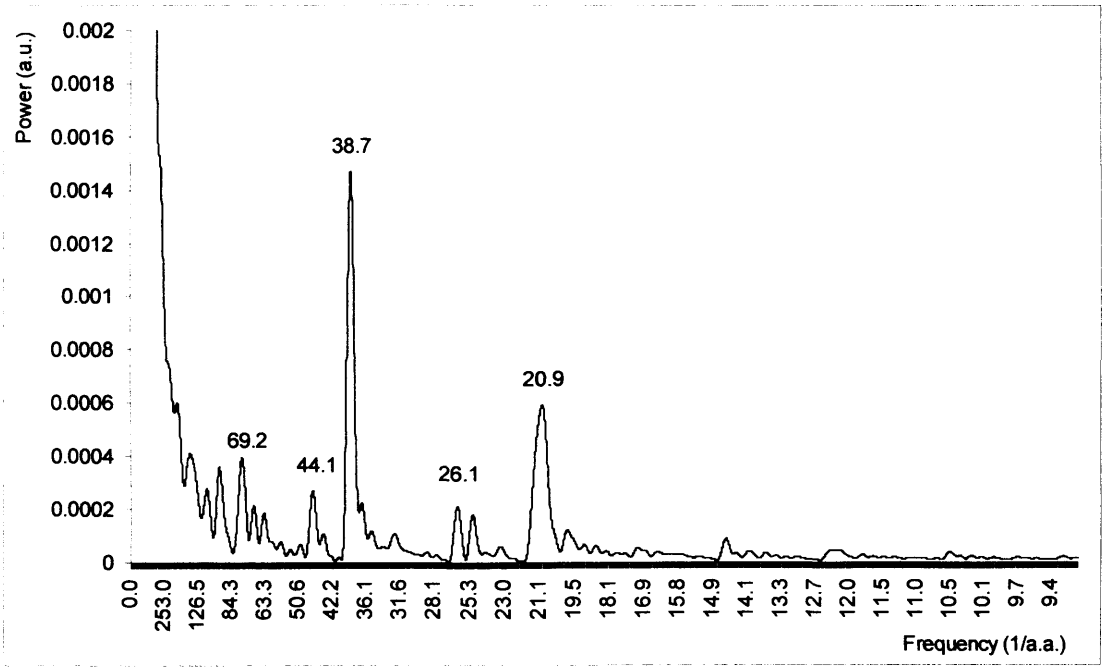


Figure 3.3: Fourier transform of the hydrophobic scoring function for two murine type I collagen triple helices.

From figure 3.3 it is apparent that there are two main periodicities in the peak distribution, namely at about 39 and 21 a.a.. This seems in contrast with an underlying periodicity of about 234 a.a. since no obvious peak with this periodicity is apparent. The analysis of graph 3.2 is also made more difficult by the high

frequency noise and by the fact that the intensity of the autoscoring function is decreasing with increasing displacements. To overcome these problems it was necessary to operate on the resolution and on the slope of graph 3.2.

The first goal, that of eliminating high frequency noise, was achieved by convoluting the hydrophobic autoscoring function with a normalised Gaussian function with σ=3 whose FWHM (Full Width at Half Maximum) is about seven a.a.. This is a reasonable value because it is comparable to the distance covered by the fully stretched side chains of two voluminous amino acids. This means that two amino acids that are seven amino acid positions apart along the collagen molecule can interact with each other. In this way, the peaks that are due to short range interactions, related to small displacements between two molecules, are smoothed away so that the most prominent peaks due to long range interactions, related to big displacements between the molecules, get enhanced.

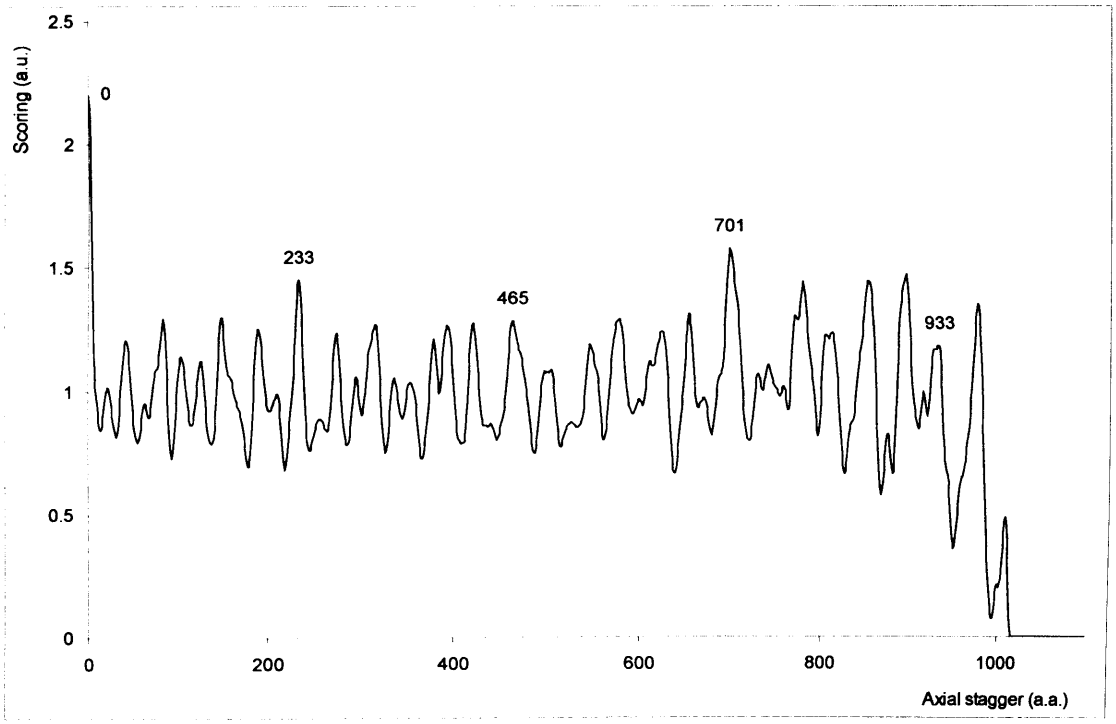


Figure 3.4: Gaussian smoothing of the normalised hydrophobic scoring function for murine type I collagen.

The second goal, that of normalising the peak intensities as the axial displacement increases, was achieved dividing point by point the graph of figure 3.2 with its linear fit. The result of this operation is presented in graph 3.4 where the distribution of the peaks appears better defined at all displacements.

The graph in figure 3.4, is characterised by some strong peaks at positions 233, 465, 701 and 933 a.a. that are separated by an average distance of about 233-234 a.a.. However, a more attentive scrutiny shows that, this separation is typical of all the peaks in figure 3.4.

In fact, as shown in figure 3.5, which reproduces the graph in figure 3.4, the local maximum in position 976 a.a. is preceded by one in position 743 a.a. (≈ 976 a.a.-234 a.a.) and this is preceded by one in position 507 a.a. (≈ 743 a.a.-234 a.a.). This relationship can be extended to find the peaks in position 275 and 42 a.a. all related by the same shift of about 233-234 a.a..

In table 3.1 we present the peaks found in this way along with their 234 a.a. relationship.

Peaks positions	Relation to periodicity (234 a.a.)
42	42
275	≈42+234
507	≈42+234*2
743	≈42+234*3
976	≈42+234*4

Table 3.1: Peaks positions outlined in graph 3.5 and their relation to the 234 a.a. periodicity.

This relationship can be further applied to find the peak that should precede the one in position 42. If such a peak existed, it should be in position -191. We saw above (equation 2.8), that the autocorrelation function is symmetrical with respect to the origin. This means that if there is a peak in position -191, a mirror peak is expected in position 191. Figure 3.5 shows a peak in position 189 in accordance with the prediction. Now, the relationship can be applied again to find the peaks in positions 423, 656 and 896 that are shown in table 3.2 together with their 234 a.a. relationship.

Peaks positions	Relation to periodicity (234 a.a.)
189	189
423	≈189+234
656	≈189+234*2
896	≈189+234*3

Table 3.2: Peaks positions outlined in graph 3.5 and their relation to the 234 a.a. periodicity.

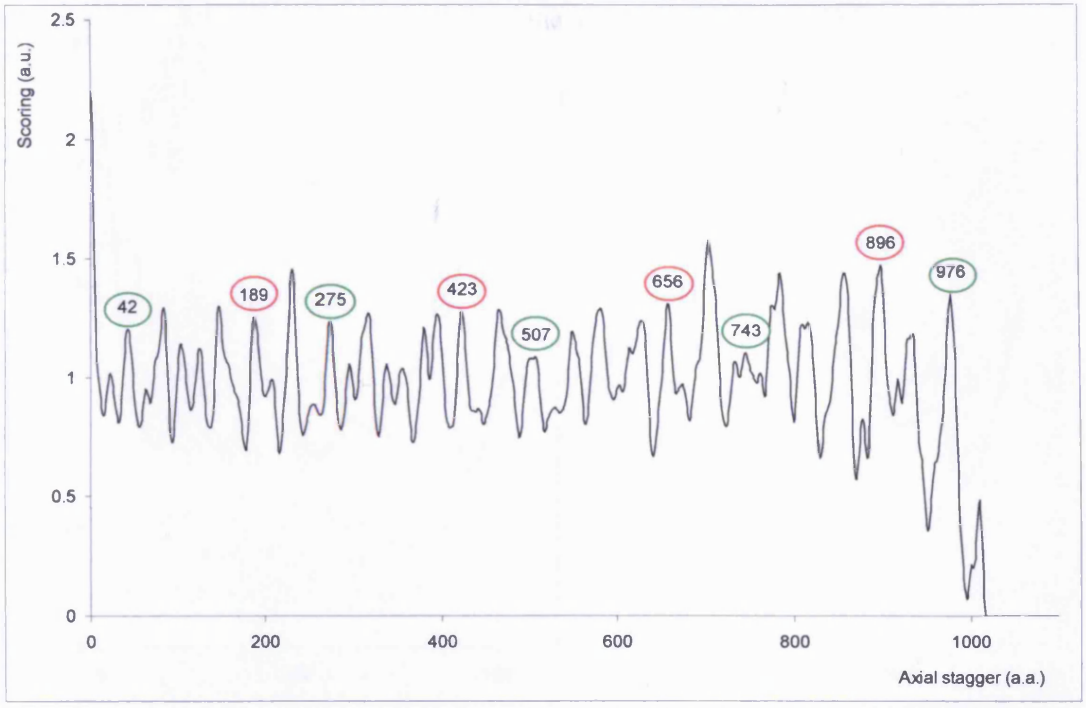


Figure 3.5: All major peaks of the hydrophobic scoring function are related by a 234 a.a. periodicity. Peaks in green can be thought of as corresponding to the interactions felt by two molecules when they slide toward each other. Peaks in red correspond to the interactions feld by molecules sliding away from each other.

From a mathematical point of view, the peaks in figure 3.5 are symmetrical with respect to the origin, and each major peak is related to another one situated 234 a.a. apart. The intensity of the peaks describes the hydrophobic interaction felt by one triple helix when it slides past the other. Thus from a physical point of view, the peaks in positions 976, 743, 507, 275 and 42 a.a. correspond to an axial shift between two molecules becoming smaller. Similarly, the peaks in positions 189, 423, 656 and 896 a.a. correspond to an axial shift becoming bigger.

It is also possible to observe the emergence of a pattern where there are symmetrical peaks that are bounded by the peaks in position 0 and 233 (region 1) as highlighted in figure 3.6 which reproduces figures 3.4 and 3.5 up to an axial stagger of about 240 a.a..

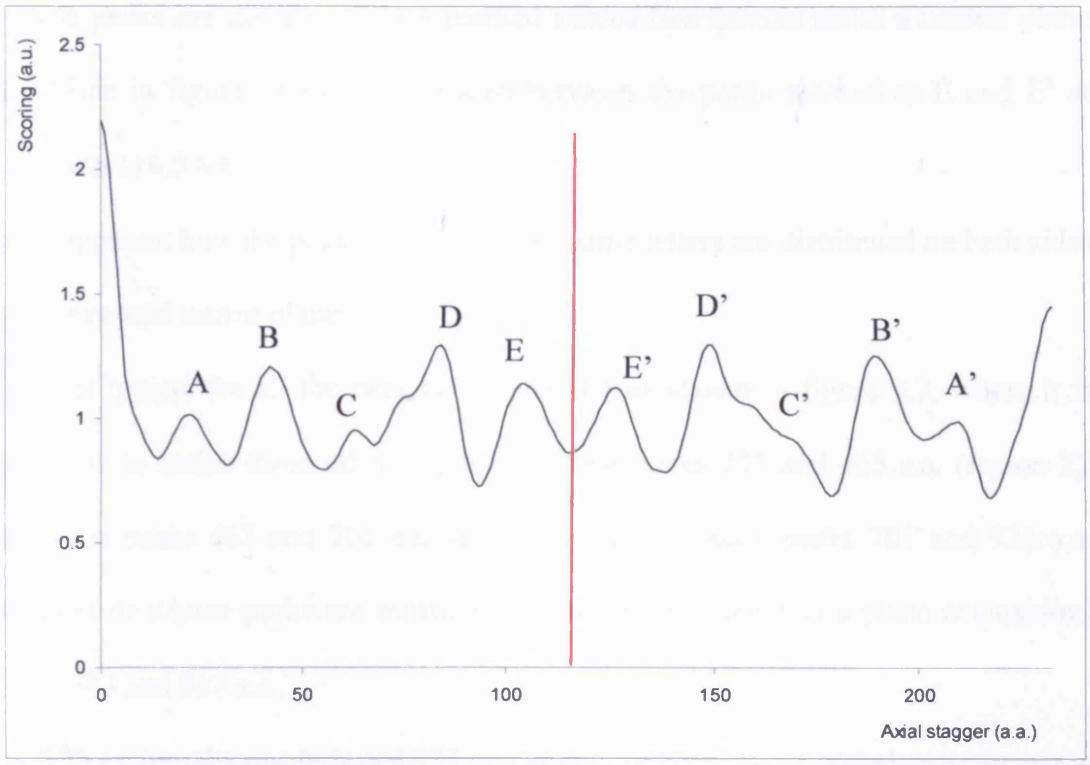


Figure 3.6: Mirror-like distribution of the peaks about position 116.5. Corresponding peaks are indicated with the same letter.

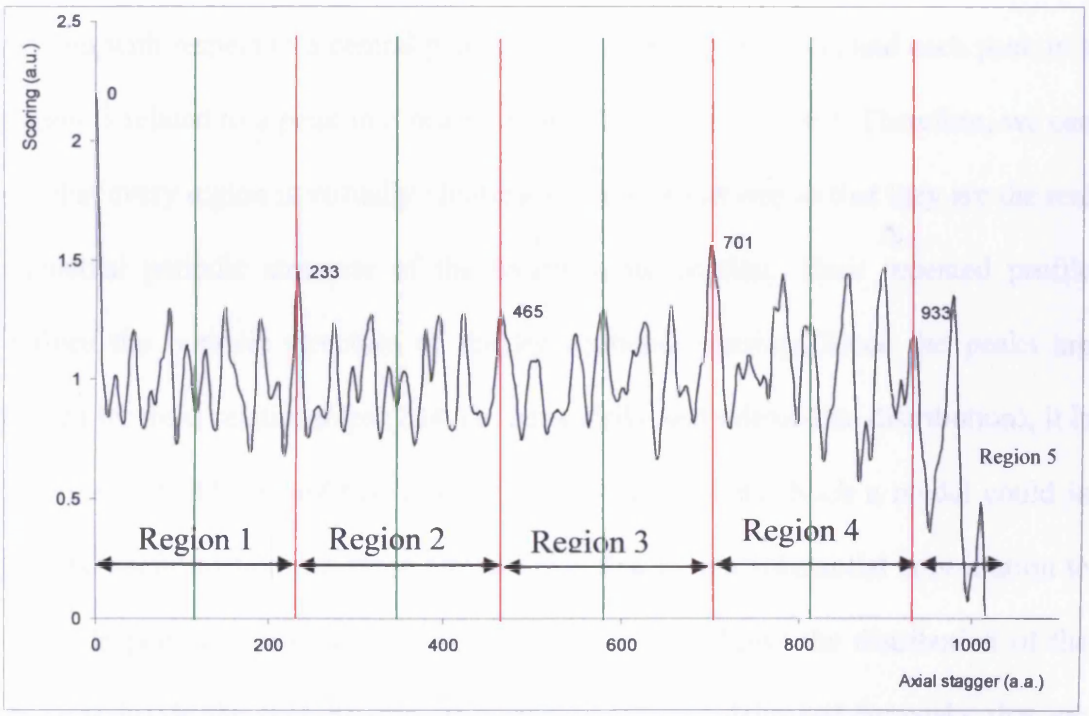


Figure 3.7: Periodic regions are delimited by red lines at positions 0, 233, 465, 701, and 933 a.a.. Planes of symmetry at positions 116.5, 349, 583 and 817 a.a. are in green.

These peaks are distributed in a marked mirror-like fashion about a central plane (red line in figure 3.6) that is located between the peaks marked as E and E' at position 116.5 a.a..

It is apparent how the peaks marked by the same letters are distributed on both sides of the central mirror plane.

This is typical for all the peaks of figure 3.4 as shown in figure 3.7, where it is possible to notice three other regions between peaks 233 and 465 a.a. (region 2), between peaks 465 and 701 a.a. (region 3) and between peaks 701 and 933 a.a. (region 4) whose peaks are mirror symmetric with respect to a plane at positions 349, 583 and 817 a.a..

A fifth region, the one between 933 and about 1000 a.a. is not complete as shown in figure 3.7 so that this behaviour can not be found in it.

All these regions have two common features: their peaks are located in a mirror-like fashion with respect to a central plane (green lines in figure 3.7) and each peak in a region is related to a peak in a nearby region by a 234 a.a. shift. Therefore, we can see that every region is virtually identical to an adjacent one so that they are the real elemental periodic structure of the hydrophobic scoring. Their repeated profile defines the periodic structure of the hydrophobic scoring. Since the peaks are related by these relationships (234 a.a. periodicity and mirror-like distribution), it is possible to build a simplified model that represents them. Such a model could in fact, be useful to find out those amino acids that give a substantial contribution to the hydrophobic interactions (see below). Table 3.3 shows the distribution of the maxima inside the regions (region 1, region 2, etc.) delimited by peaks that are situated at 0 plus multiple integer of 234. The regions are made of peaks with a similar distribution. Since in each region, the peaks have the same mirror-

symmetric distribution, we measure their relative distance form the peaks delimiting the regions (red lines in figure 3.7) to build a model of mirror-distributed peaks. Thus, for example, if we consider figure 3.6, we calculated the distance of peak A from the peak at the origin and the distance of the mirror-peak A' from the peak at position 233. We did the same for the peaks B and B', C and C', D and D' and E and E'. We repeated the same procedure for all the peaks belonging to the other regions. In addition, we calculated the weighted average of the distances according to the overlap between the two collagen sequences. The result is an "ideal" scoring function distribution where the peaks are located in a symmetric way with respect of the central plane of the periodic regions.

The result is shown in table 3.3 and in figure 3.8 where the comparison between the real and the estimated peak positions is presented. The distance between estimated peak positions is also shown to make the periodic and the symmetric feature of the modelled sequence apparent. The average distance between corresponding peaks is about 233.6±3 a.a., this value is approximated to 234 a.a. to simplify our calculations. Because of noise, the accordance between real and estimated peak positions becomes worse with increasing distance from the origin. The average discrepancy between the real peak positions and the estimated ones is about 2.3 a.a.. However, if we consider only the peaks belonging to the first two regions, where the overlapping between the two molecules during scoring calculations is greatest, the average discrepancy is about 1.7 a.a. since the signal is more intense.

By looking at the column of the distances between the "ideal" peak positions in table 3.3, it is apparent how the underlying exact periodicity in the hydrophobic scoring is not the one that links a peak to the adjacent one at about 21 a.a. shift (in fact very few peaks are separated by 21 a.a.), but it is the one that relates each

region to the adjacent one 234 a.a. away, since adjacent regions can be superimposed exactly.

Measured positions	Estimated positions	Discrepancy	Distance between modelled peak positions in amino acids
0	**0**	**0**	**23.5**
A 22	23.5	1.5	19
B 42	42.5	0.5	20
C 63	62.5	0.5	12.5
77	75	2	10
D 84	85	1	22
E 104	107	3	20
E' 126	127	1	22
D' 149	149	0	10
159	159	0	12.5
C' 170	171.5	1.5	20
B' 189	191.5	2.5	19
A' 210	210.5	0.5	23.5
233	**234**	**1**	**23.5**
257	257.5	0.5	19
275	276.5	1.5	20
296	296.5	0.5	12.5
310	309	1	10
317	319	2	22
337	341	4	20
355	361	6	22
380	383	3	10
395	393	2	12.5
missing	405.5		20
423	425.5	2.5	19
441	444.5	3.5	23.5
465	**468**	**3**	**23.5**
missing	491.5		19
507	510.5	3.5	20
529	530.5	1.5	12.5
missing	543		10
549	553	4	22
missing	575		20
missing	595		22
613	617	4	10
627	627	0	12.5
missing	639.5		20
656	659.5	3.5	19
672	678.5	6.5	23.5
701	**702**	**1**	**23.5**
missing	725.5		19
743	744.5	1.5	20
760	764.5	4.5	12.5
773	777	4	10

Measured positions	Estimated positions	Discrepancy	Distance between modelled peak positions in amino acids
782	787	5	22
807	809	2	20
missing	829		22
845	851	6	10
855	861	6	12.5
missing	873.5		20
896	893.5	2.5	19
916	912.5	3.5	23.5
933	**936**	**3**	**23.5**
959	959.5	0.5	19
976	978.5	2.5	20
1000	998.5	1.5	12.5
1009	1011	2	

Table 3.3: Comparison between measured peak positions and estimated ones for the murine type I collagen scoring function, based on the mirror-like distribution about the central axis of each periodic region. Peaks positions delimiting periodic regions are coloured in yellow.

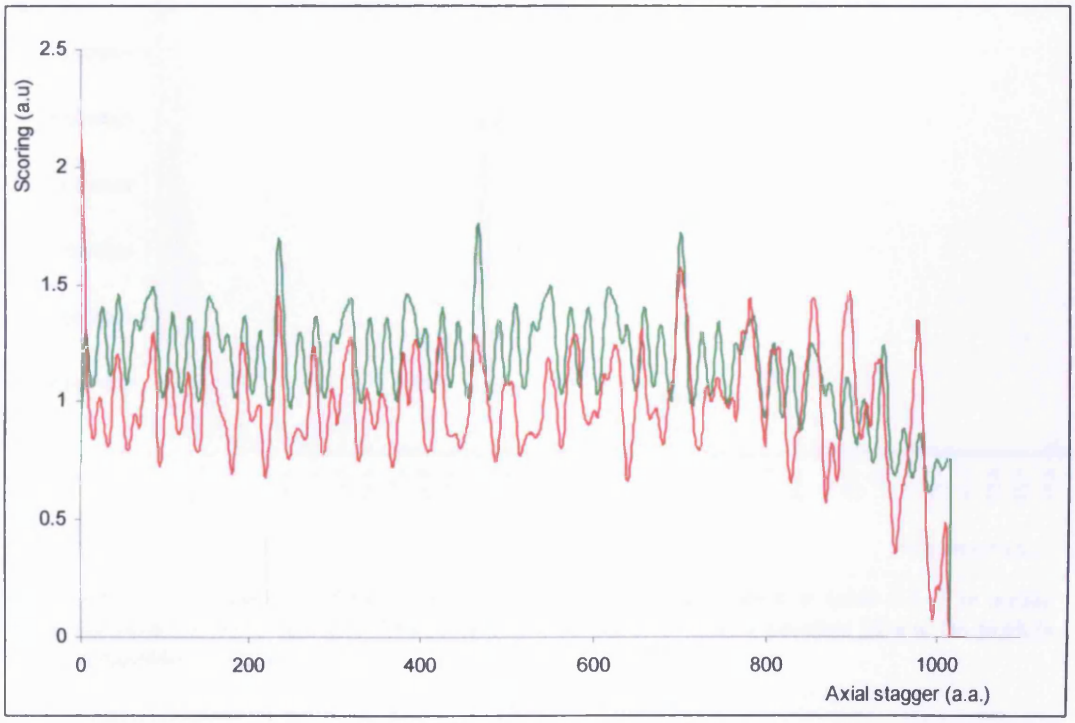


Figure 3.8: Comparison between the real peaks distribution of the hydrophobic scoring function (red) and the "ideal" peaks distribution (green). The "ideal" distribution was normalised to make the comparison clearer

However, as seen above, Fourier analysis of the real peak positions in table 3.3 seems to contradict such an observation. In figure 3.9, we present the Fourier transform of the real peak positions of table 3.3. It is possible to notice two major peaks at shift of about 39 and 21 a.a. whereas no apparent peak at a 234 a.a. shift is visible. This can be explained if we consider that the Fourier transform of the positions of the real maxima is given by the convolution of the motif (the profile between 0 and 234 a.a.) with a series of delta function at every 234 a.a. (paragraph 2.4). In figure 3.10, we present the Fourier transform of the motif. We can see two maxima at shift of about 37 and 21 a.a. in positions close to the ones previously found.

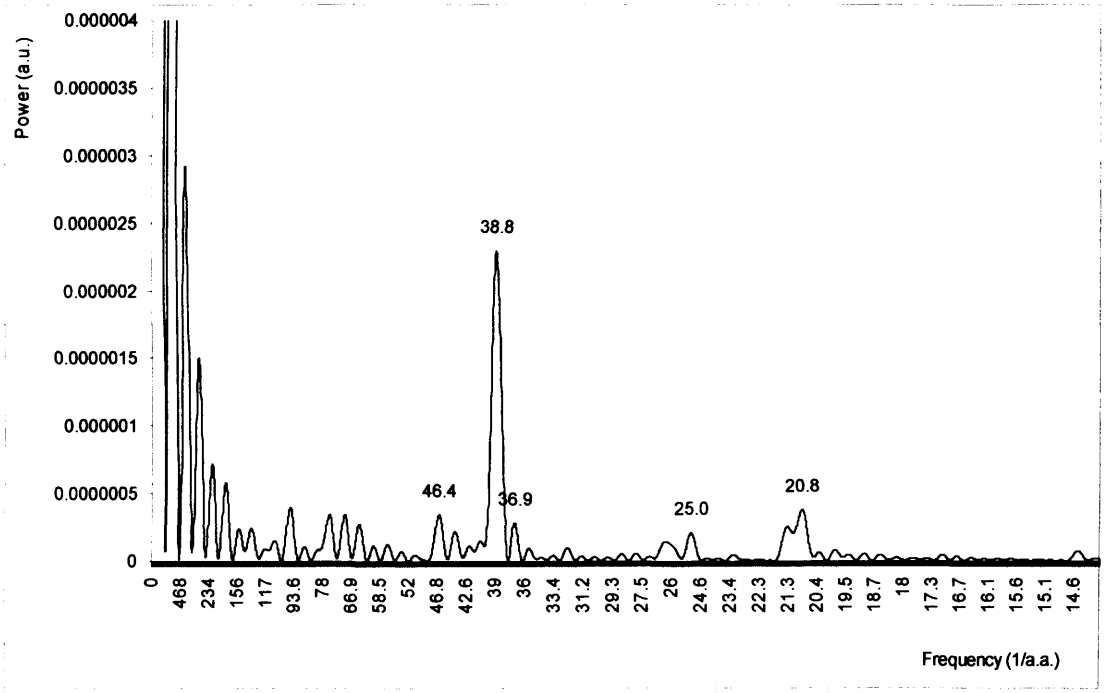


Figure 3.9: Fourier analysis of the measured peak positions as shown in table 3.3. The major frequency peak visible is at about 39 a.a.. A second peak is visible in position 21 a.a. No peak is visible at position 234 a.a.

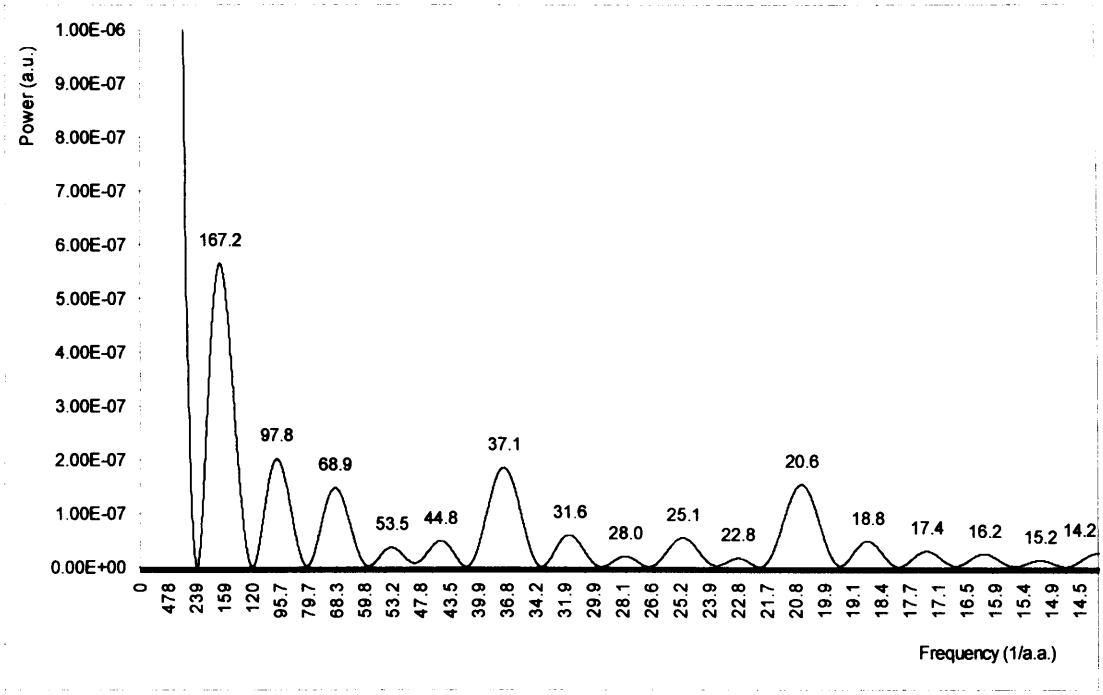


Figure 3.10: Fourier analysis of the measured peaks as shown in table 3.3 between 0 and 233 a.a.

In figure 3.11 we superimpose the Fourier transform of the motif with the Fourier transform of a delta comb with a periodicity of 234 a.a.. It can be readily seen how the maxima of the motif compare with the maxima of the delta comb. In particular, the first order maximum of the delta comb, the one in position 234, corresponds to a minimum in the Fourier transform of the motif. Thus, when they are multiplied together to give the Fourier transform of the scoring function, they do not give rise to any appreciable peak at 234 a.a. shift.

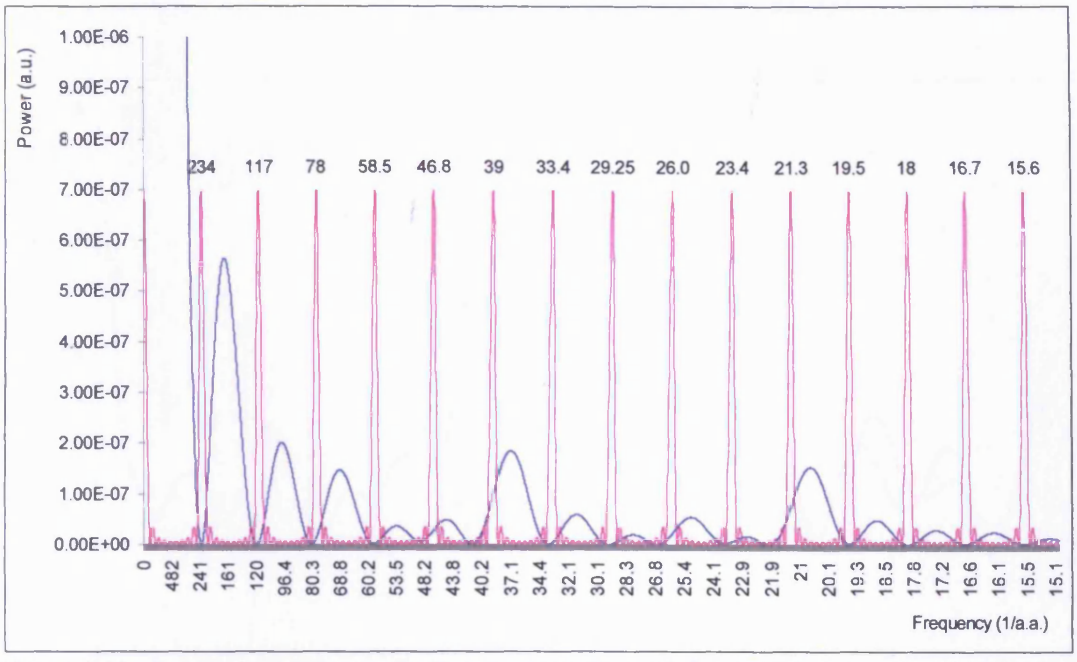


Figure 3.11: Comparison between the Fourier transform of the motif between 0 and 234 a.a. and a δ-comb with a periodicity of 234 a.a.. No peak at 234 a.a. can be seen in the Fourier transform of the scoring function because the transform of the motif has a zero in that position.

The fact that the Fourier transform of the motif has a minimum at a 234 a.a. shift is due to the fact that the profile can be subdivided into two similar regions.

This is confirmed visually by graph 3.12 where we show region 1 compared to itself after an 82 a.a. shift.

The peaks of the two curves are distributed in a similar way. This is also confirmed in table 3.3 where the peak distribution is repeated after about an 84 a.a. shift. Thus, we can approximate region 1 as the convolution of the first half of the region with two δ-functions displaced at the centres of the two half-regions that is, at positions 58 and 175. We can apply once again the convolution theorem (paragraph 2.4) to find out that the Fourier transform of the peak distribution of the motif (region 1) is given by the product of the Fourier transform of half the motif with the Fourier transform of the two δ-functions.

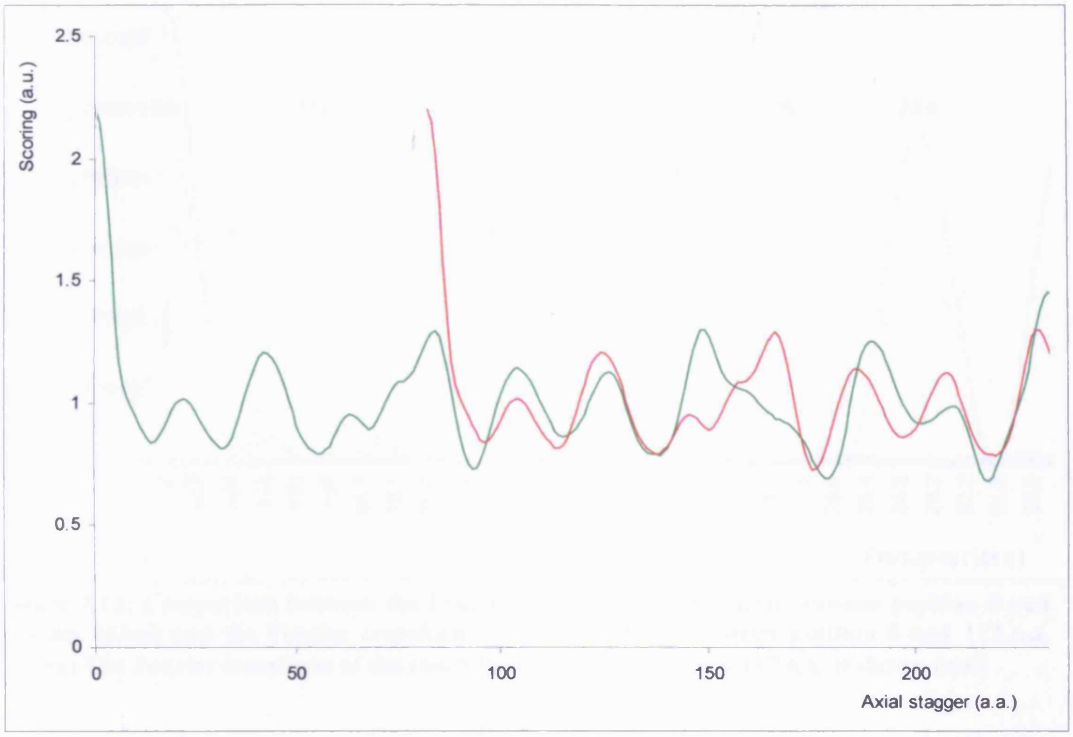


Figure 3.12: Comparison between the first half and the second half of region 1 of the scoring function. From about 82 a.a. the two curves have a very similar peaks distribution.

In fact, figure 3.13 shows the Fourier transform of the motif (blue), of half motif (green) and of two δ-functions (red) that are separated by about 117 a.a.. It is immediately seen that the transform of the two δ-functions is zero for a 234 a.a. displacement. If we multiply this curve for the transform of the motif, we obtain a transform which is zero at 234 a.a. stagger explaining graph 3.11.

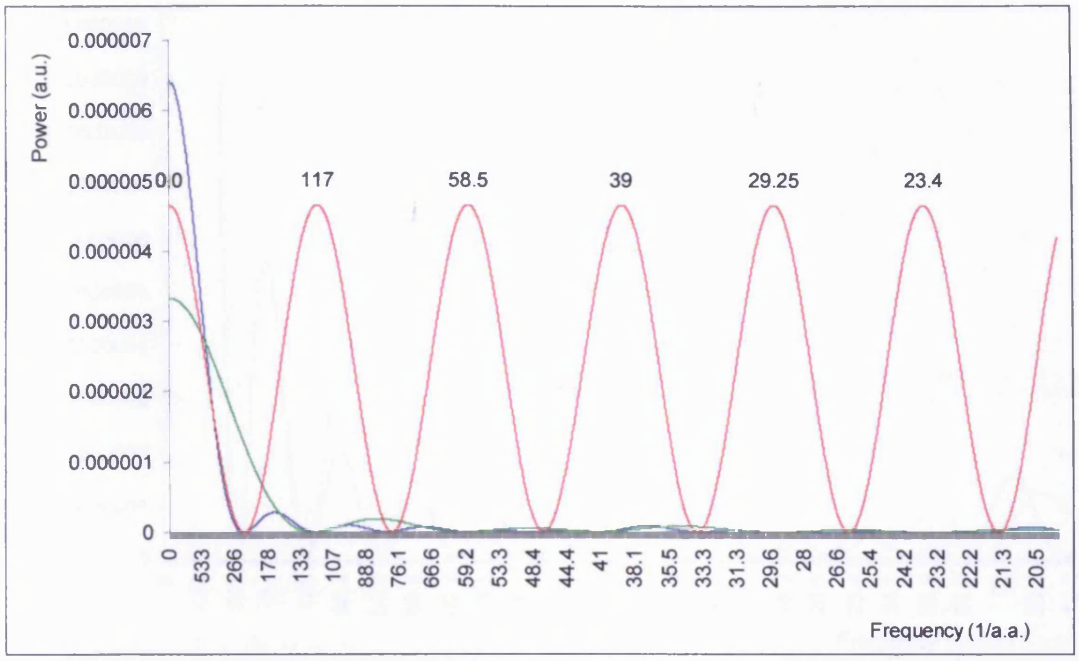


Figure 3.13: Comparison between the Fourier transform of the motif between position 0 and 234 a.a. (blue) and the Fourier transform of half motif i.e. between position 0 and 117 a.a. (green) The Fourier transform of the two δ-functions separated by 117 a.a. is shown (red)

In figure 3.14, we present the graph of the Fourier transform of the motif (green) and the product of the Fourier transform of the two δ-functions with the Fourier transform of half the motif (red). The close resemblance of the two curves is apparent showing that the convolution theorem can be used to explain why the Fourier transform of the motif is zero for 234 a.a. stagger.

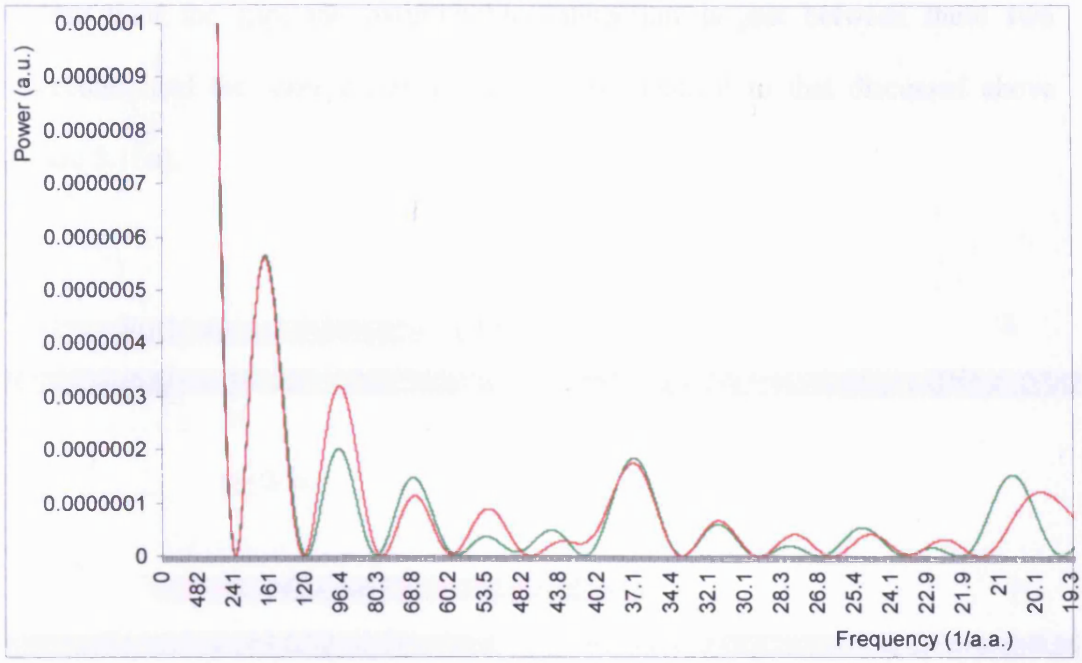


Figure 3.14: Comparison between the Fourier transform of the motif of the hydrophobic autoscoring function between 0 and 234 a.a. shift (green) and the product of the Fourier transform of the two δ-functions 117 a.a. apart with the Fourier transform of half the motif (red).

3.3 Multiple parallel hydrophobic autoscoring

The main structural role of type I collagen is to form fibrils. To study the hydrophobic interactions during fibrillogenesis we need to evaluate how many molecules can interact together. We build therefore a model made of a linear sequence of identical molecules, with the same orientation and direction and separated by a gap G. We refer to the sequence obtained in this way as a multiple sequence.

We use a moving molecule, with the same orientation of the molecules in the multiple sequence, to calculate the hydrophobic scoring function when it slides past the other molecules as shown in figure 3.15. When the distance (Δx) between the moving molecule (molecule P) and the first molecule of the sequence (molecule 1)

is less than the gap, the hydrophobic interaction is just between these two molecules, and the scoring function profile is identical to that discussed above (figure 3.15a).

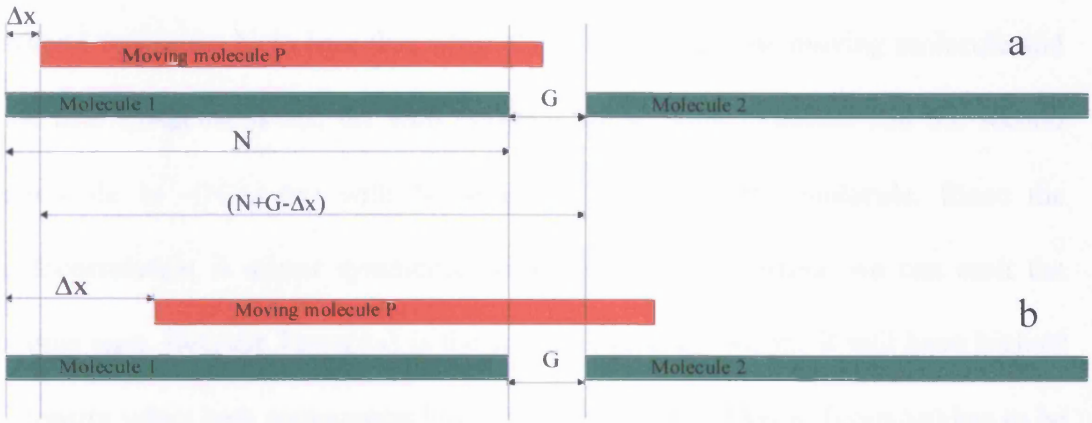


Figure 3.15: Diagram representing multiple parallel hydrophobic interactions.

However, when the distance Δx becomes greater than G the force felt by the moving molecule is due to the combined effect of two molecules in the multiple sequence, (molecules 1 and 2, figure 3.15b).

As before, we use the hydrophobic autocorrelation function (equation 2.4) to describe the hydrophobic interactions between two collagen molecules. Thus, the intensity of the total interaction can be thought of as the sum of two autocorrelation functions, one that describes the interaction of the moving molecule P with molecule 1 and the other, the interaction of the moving molecule P with molecule 2 (equation 3.1):

$$Tcorr(\Delta x) = corr_{P\text{-}1}(\Delta x) + corr_{P\text{-}2}(N\text{-}\Delta x+G) \qquad 3.1$$

Where Tcorr(Δx) is the total hydrophobic interaction felt by the moving molecule (P) as a function of its distance (Δx) from the first amino acid of the multiple sequence. $corr_{P-1}(\Delta x)$ is the interaction between the moving molecule and the first molecule, while $corr_{P-2}(N-\Delta x+G)$ is the interaction between the moving and the second molecule. Note here that while the shift between the moving molecule and the first molecule is Δx, the shift between the moving molecule and the second molecule is –(N+G-Δx) with N being the length of the molecule. Since the autocorrelation is mirror symmetric with respect to the origin, we can omit the minus sign. Because Tcorr(Δx) is the sum of two components, it will have highest intensity when both components have highest intensity. That is Tcorr(Δx) has to be a maximum when $corr_{p-1}(\Delta x)$ and $corr_{p-2}(N-\Delta x+G)$ have simultaneous maxima at positions Δx and (N-Δx+G) respectively. From figure 3.4, it is apparent that the hydrophobic scoring has the highest values at positions that are integer multiples of the periodicity T=234 a.a.. Therefore, Tcorr(Δx) will have maximum intensity if both Δx and (N-Δx+G) are integer multiple of T as shown in equations 3.2 and 3.3.

$$\Delta x = nT \qquad n \in \mathbb{N} \qquad 3.2$$

$$N-\Delta x+G = mT \qquad m \in \mathbb{N} \qquad 3.3$$

We can now solve the system of equations 3.2 and 3.3 to find the following

$$N-nT+G = mT \qquad n, m \in \mathbb{N} \qquad 3.4$$

Equation 3.4 can be solved for G to give

$$G = (n+m)T-N = kT-N \qquad \text{with } k=(n+m)\in\mathbb{N} \qquad 3.5$$

From figure 3.15 it is apparent how kT can not be less than N otherwise molecule 1 and molecule 2 would clash. On the other hand, kT must be as small as possible to maximise the number of hydrophobic contacts between the moving molecule and both molecules of the multiple sequence. If we consider these constraints, that the length of projected triple helical domain is N=1016 a.a. and that the period T=234 a.a. we can write equation 3.5 as:

$$G = 5T-N = (5*234-1016)\text{ a.a.} = (1170-1016)\text{ a.a.} = 154\text{ a.a.} \qquad 3.6$$

with k=5 being the smallest allowed value.

We can now use this value for G to build the multiple sequence of molecules previously described.

We use the autocorrelation function (equation 2.4) to study the hydrophobic interaction between a moving molecule without telopeptides and a sequence of identical molecules. We call the scoring of this kind of configuration (i.e. one molecule against many) the multiple scoring. In this case, the orientation of the probe and the sequence is parallel. In figure 3.16 we can see the hydrophobic interaction (multiple scoring) in arbitrary units as a function of the shift in a.a..

The autoscoring function has been smoothed with the same Gaussian sequence described previously (paragraph 3.2). Since, the whole moving molecule is interacting with the multiple sequence, it is not necessary to normalise the autoscoring function.

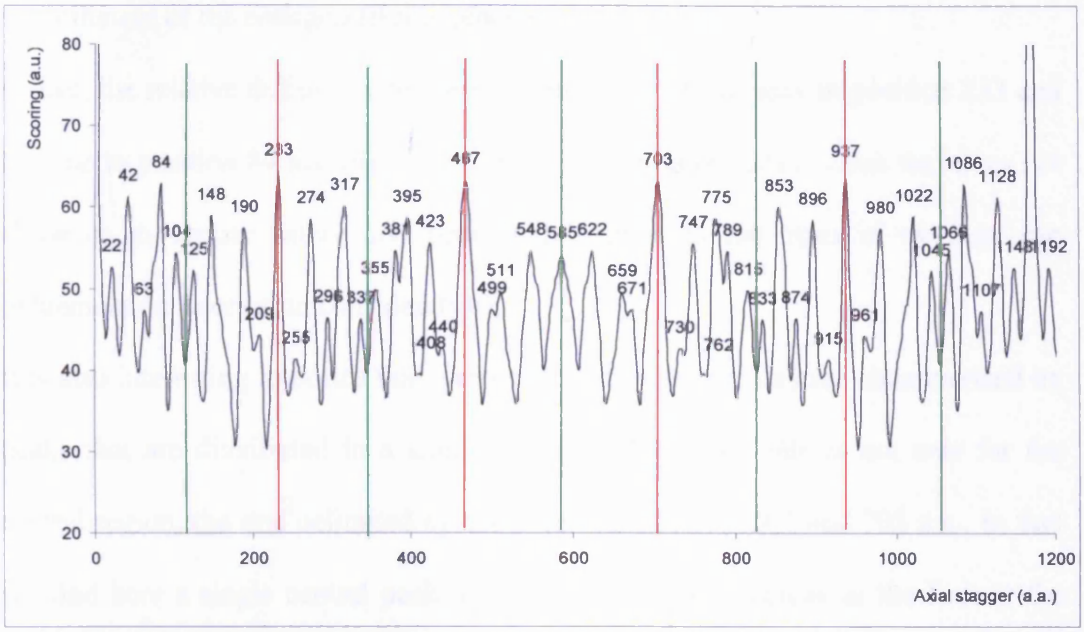


Figure 3.16: Parallel multiple hydrophobic scoring function of a molecule of murine type I collagen without telopeptides against a sequence of similar molecules separated by a gap of 154 amino acids. Periodic regions are delimited by red lines. Planes of symmetry are delineated with green lines.

The positions of the peaks have varied only slightly if compared to those obtained for the scoring of one molecule against another (figure 3.4). Their positions are symmetrical with respect to the peaks in positions 0, 233, 467, 703 and 937 (red lines in figure 3.16) forming regions of symmetry similar to those found in the case of the scoring of one triple helix against another. Their distribution is also mirror-symmetric with respect to the central plane of such regions (green lines in figure 3.16). As stated by equation 2.19 we have an absolute maximum when the shift is zero or equal to integer multiple of 1170 because in such instances the molecules are exactly in register. The most intense peaks after these ones, are those in position 233 a.a. and its mirror-symmetric in position 937 a.a. and that in position 467 a.a. and its mirror-symmetric in position 703 a.a..

Nonetheless, their intensity does not look strong enough to make them stand out as the initiators of the collagen fibrillogenesis.

In fact, the relative difference between the intensity of the peak in position 233 and the one in position 84 a.a. (the second tallest) is of about 2% a value that does not allow us to choose safely one peak with respect to the other if we base our judgement on interaction considerations.

It is also interesting to notice that almost all symmetry regions are characterised by peaks that are distributed in a similar manner. However, this is not true for the central region, the one delimited by the peaks in position 467 and 703 a.a.. In fact we find here a single central peak in position 585 a.a., whereas in the first or the second region we have a pair of peaks in the corresponding positions (104-125)a.a. and (337-355) a.a..

However, the results obtained until now are strongly biased by the amino acids sequence used that did not include telopeptides. To try to understand what really happens during fibrillogenesis, it is important to use a more realistic model for the collagen molecule that includes telopeptides. In what follows, we will be studying the interactions of two or more multiple sequences, since we have defined all the parameters that describe them.

3.4 Multiple Parallel Hydrophobic interaction for type I collagen with telopeptide contribution

In our model, we used two different configurations for type I collagen. A simple one with unfolded telopeptides and a more realistic one with folded telopeptides. It is in fact believed (Orgel et al., 2000; Malone et al., 2004; Ortolani et al., 2000; Otter et al., 1988) that the telopeptides are usually folded back towards the centre of the molecule. The folding configuration is determined by the telopeptide function inside the microfibril which is to make cross-links (Malone et al., 2004; Otter et al., 1988). To have a realistic representation of the telopeptide folding, we need therefore to study their behaviour in the fibril context. We use the results found for the hydrophobic autoscoring of multiple sequences of the triple helices to build a "fibril" made of molecules with a 234 a.a. staggering. Inside the fibril, it is now possible to highlight a "microfibril" made by lateral aggregations of five parallel linear sequences of molecules that are staggered by 234 a.a. with respect to each other and represents the basic element of the fibril (in green in figure 3.17).

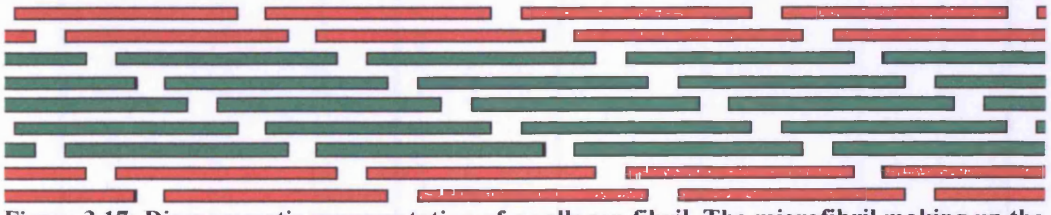


Figure 3.17: Diagrammatic representation of a collagen fibril. The microfibril making up the fibril is coloured in green.

The telopeptides of the single collagen molecule were folded in a way that had the lysines, responsible of cross-link formation, facing each other in order to minimise the interaction energy: we folded the α_2(I) telopeptide so that its 7N hydroxylysine

faces exactly lysine 933 of the α_2(I)-chain of an adjacent molecule in the microfibril. Similarly, we folded the N-telopeptides of the second and the third helix to make their 8N lysines face exactly hydroxylysine 930 of the second and the third helix of an adjacent molecule in the microfibril, respectively. In this way, the telopeptide folding occurs between residues five and six for the first helix and between residues seven and eight for the second and the third helix

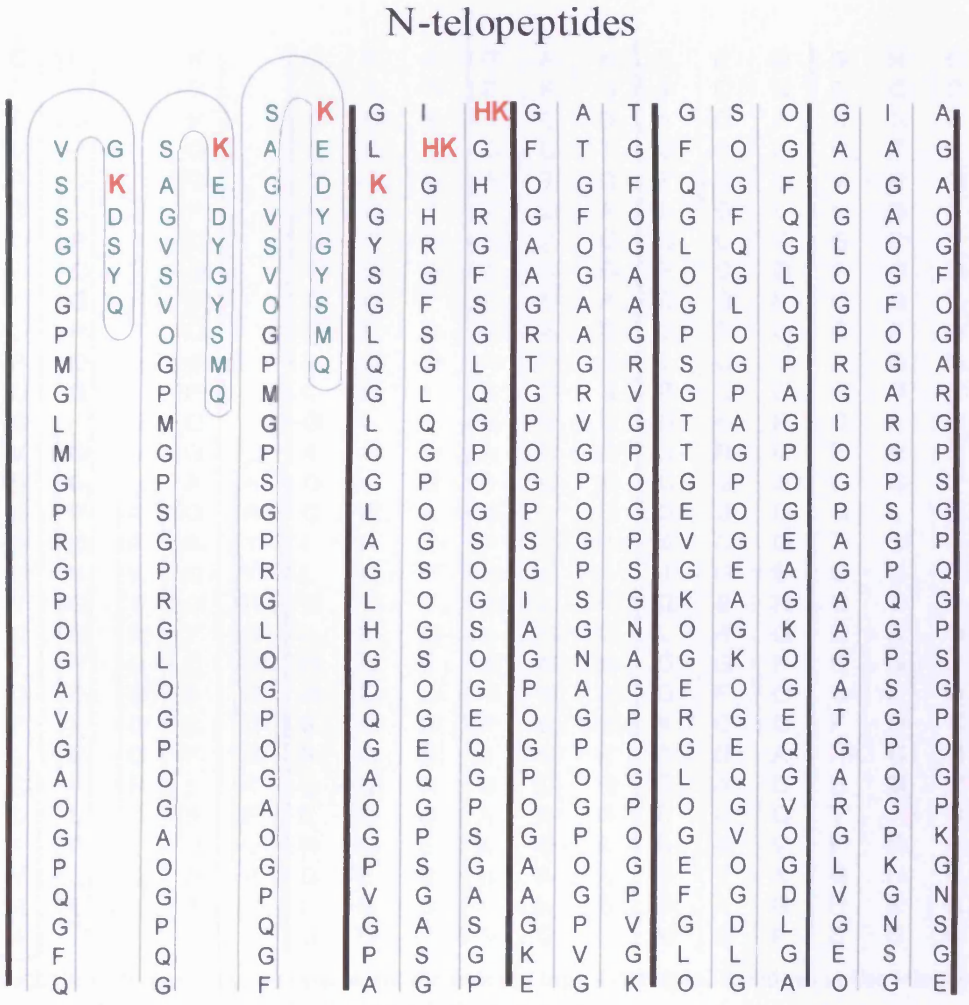


Figure 3.18: N-telopeptide configuration for murine type I collagen. Residues in the telopeptides are in green. Lysines responsible for cross-links are in red

The same approach was used in folding the C-telopeptides of the second and third α-chains only, because the α_2-chain does not have any cross-link forming lysine.

The α_1-telopeptides were folded to have lysine 16C facing hydroxylysine 87 of the triple helix. Thus, both telopeptides were folded between residues 12C and 13C.

In figure 3.18 and 3.19, we show a diagram of the configuration used for the N-telopeptides and the C-telopeptides where the lysines responsible for cross-link formation are highlighted in red.

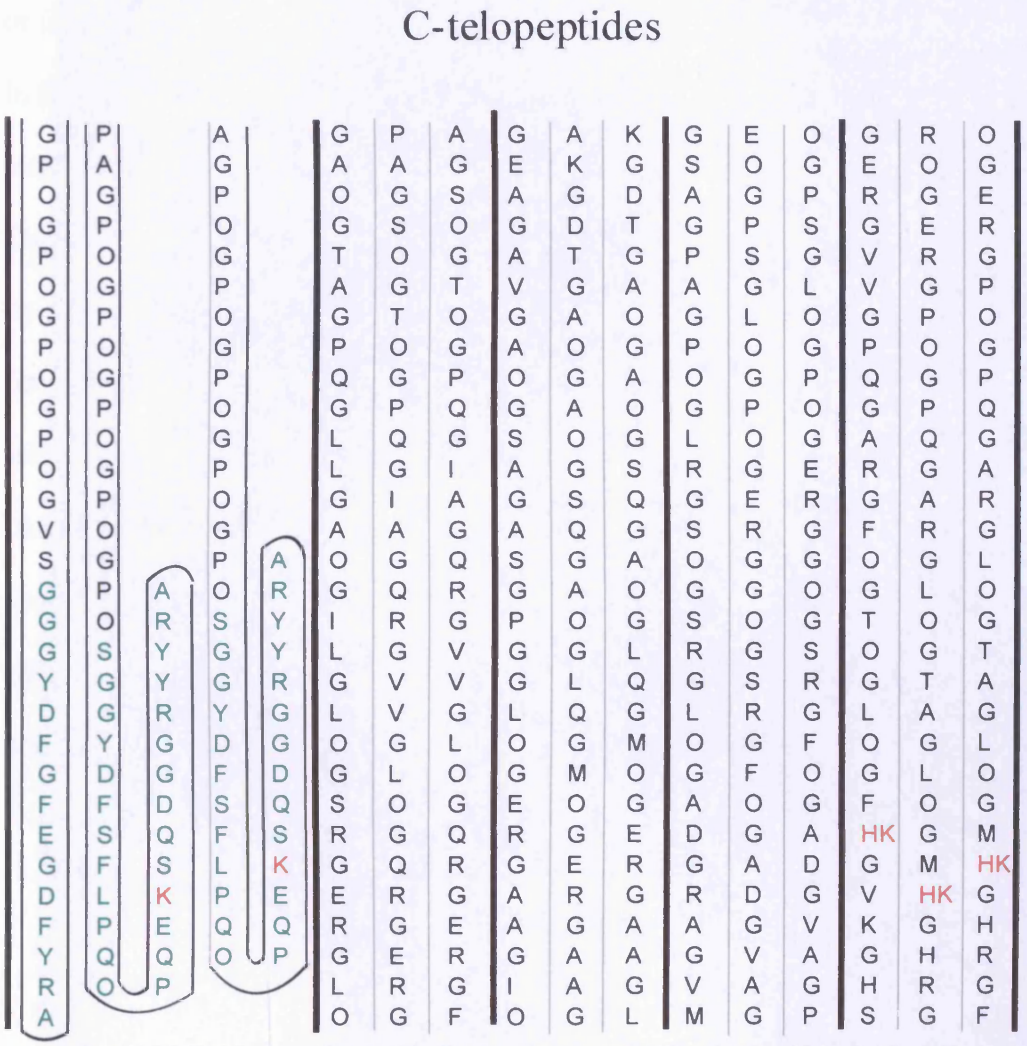


Figure 3.19: C-telopeptide configuration for murine type I collagen. Residues in the telopeptide are in green. Lysines responsible for cross-links are in red

As described above, we then assigned the value 1 to the major hydrophobic amino acids of the collagen molecule (triple helix and telopeptides) and we summed the values laterally to obtain a sequence of values that represent the distribution of the

hydrophobic amino acids along the collagen axis that are comprehensive of the telopeptide contribution.

We used equation 3.6 to calculate the correct gap G and to build multiple sequences of similarly oriented collagen molecules with unfolded (G=113) and folded telopeptides (G=135). We employed a moving collagen molecule identical to those used to build the multiple sequences, to study the multiple hydrophobic interaction of fibrillar collagen in a parallel configuration.

In figure 3.20 and 3.21, we represent the multiple parallel hydrophobic scoring for murine type I collagen with unfolded and folded telopeptides respectively. Qualitatively, it is possible to notice the formation of symmetric regions delimited by the peaks situated at position that are integer multiples of about 234 a.a. (red lines). The peaks in these regions are also distributed in a mirror-symmetrical way with respect to the central plane of the regions (green lines). In addition, we can notice that the central region is now characterised by a couple of peaks about its central axis (585 a.a. shift) whereas we found a single peak with the autoscoring calculated without using telopeptides (figure 3.16). This is immediately visible comparing figures 3.20 and 3.21 with figure 3.16.

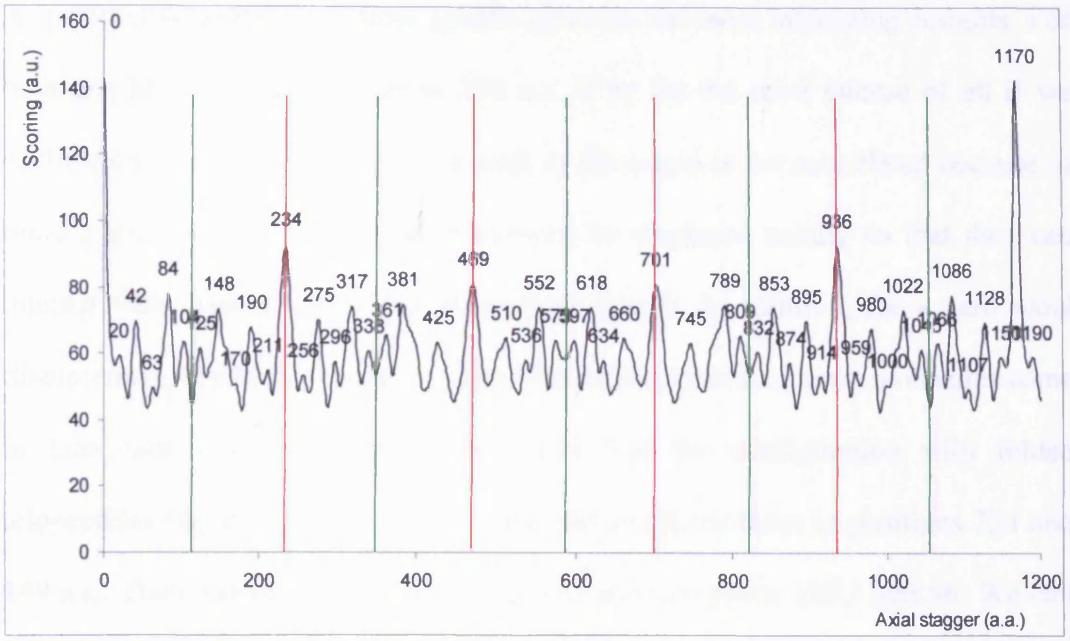


Figure 3.20: Multiple parallel hydrophobic scoring for a collagen molecule with straight telopeptides. Periodic regions are delimited by red lines. Planes of symmetry are delineated by green lines.

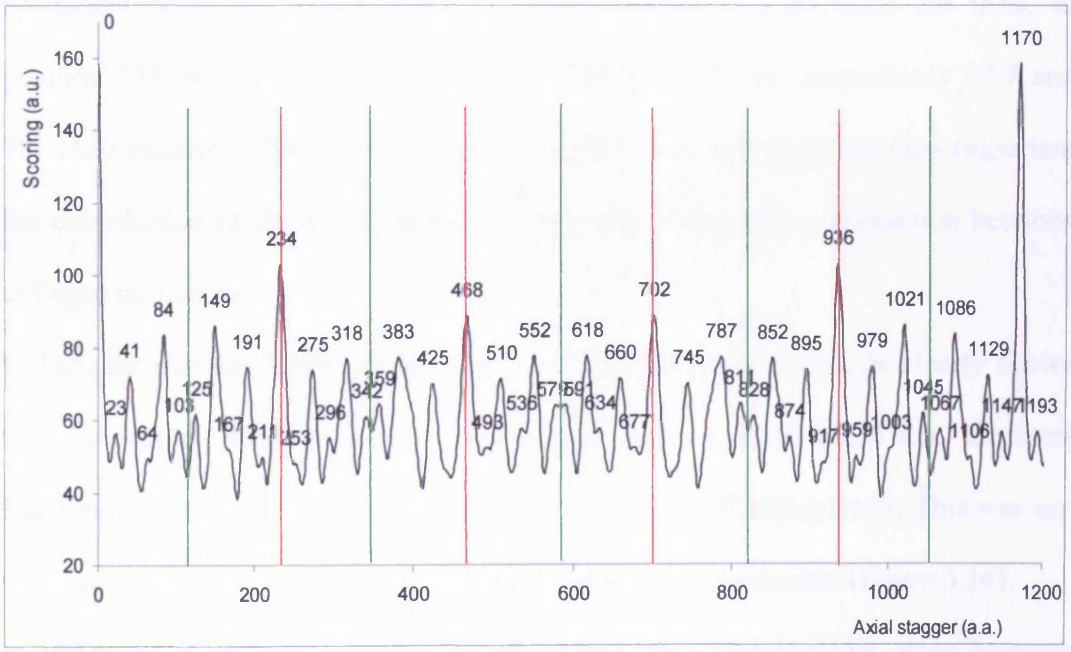


Figure 3.21: Multiple parallel hydrophobic scoring for a collagen molecule with folded telopeptides. Periodic regions are delimited by red lines. Planes of symmetry are delineated by green lines.

A quantitative analysis of these graphs gives us the most interesting insights. For both graphs, the peak in position 234 a.a. is by far the most intense of all if we exclude the peak at the origin. The peak at the origin is not considered because to build a solid structure, the molecules must be displaced axially so that they can interact with several molecules along their length. In addition, for a zero axial displacement, similarly charged residues belonging to two adjacent molecules come to face each other and repel each other. For the configuration with folded telopeptides (figure 3.21) the most prominent peaks are those in positions 234 and 149 a.a.. Their values, expressed in arbitrary units are about 103.3 and 86. We call the difference in amplitude of the minor peak with respect to the major, relative difference. The relative difference between the two peaks just mentioned is thus about 20%. This difference is much higher than the one found when the telopeptides are not considered (2%; figure 3.16). In a similar way, the two most prominent peaks for the configuration with unfolded telopeptides are those at position 234 and 84 a.a. whose values in arbitrary units are respectively 92.7 and 78. Their relative difference is thus about 18.8%. It is thus apparent how important the contribution of the telopeptides is in the total hydrophobic interaction between collagen molecules.

In fact, the multiple hydrophobic autoscoring for collagen molecule clearly shows that the peak in position 234 a.a. is the most intense, making 234 a.a., the most likely relative shift between collagen molecules during fibrillogenesis. This was not the case when the telopeptides were omitted from the calculations (figure 3.16).

In addition, it is clear that the configuration with folded telopeptides gives the most prominent peak at position 234 (figure 3.21). The relative difference with the corresponding peaks in figure 3.16 and 3.20 is about 61 and 12 % respectively.

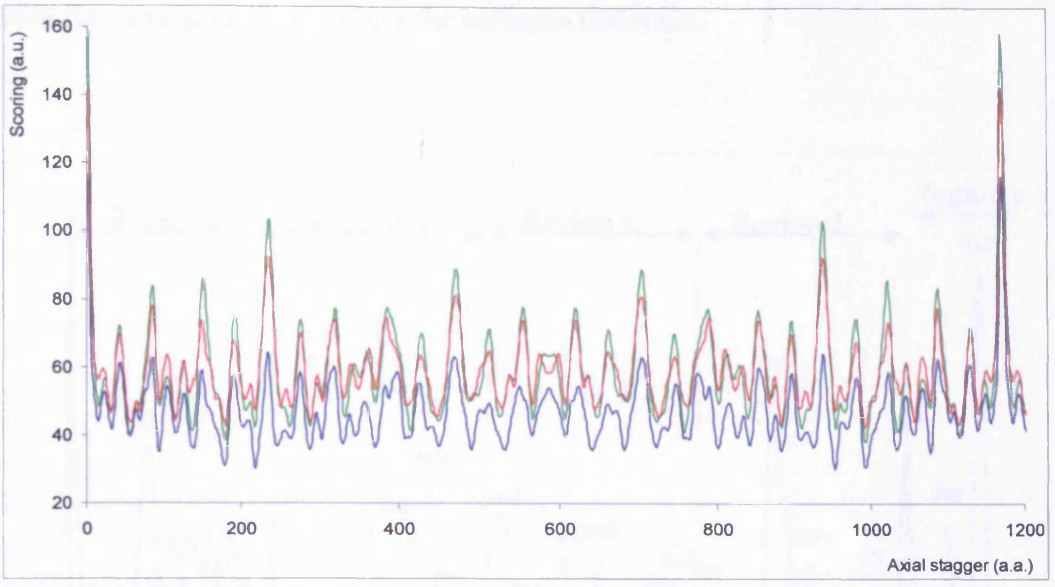


Figure 3.22: Comparison among the parallel hydrophobic autoscoring for the murine type I collagen molecule for different telopeptide configurations. Blue: no telopeptides; red: straight telopeptides; green: folded telopeptides.

In figure 3.22, we compare the hydrophobic scoring functions of the triple helix and of the collagen molecules with straight and folded telopeptides, the differences are apparent. The distribution of maxima is however very similar for all configurations except in the central region of the autoscoring function for the collagen without telopeptides were there is a single peak instead of the double one mentioned before.

3.5 Distribution of the hydrophobic amino acids along the molecule

The 234 a.a. periodic regions are the fundamental feature of the hydrophobic scoring function for type I collagen. These regions can be used to identify those amino acids that contribute the most to the hydrophobic interaction. Since the scoring function is a sequence of similar regions that repeat every 234 a.a. it is

reasonable to think that they are generated by a sequence of amino acids that repeat with the same periodicity along the collagen molecule.

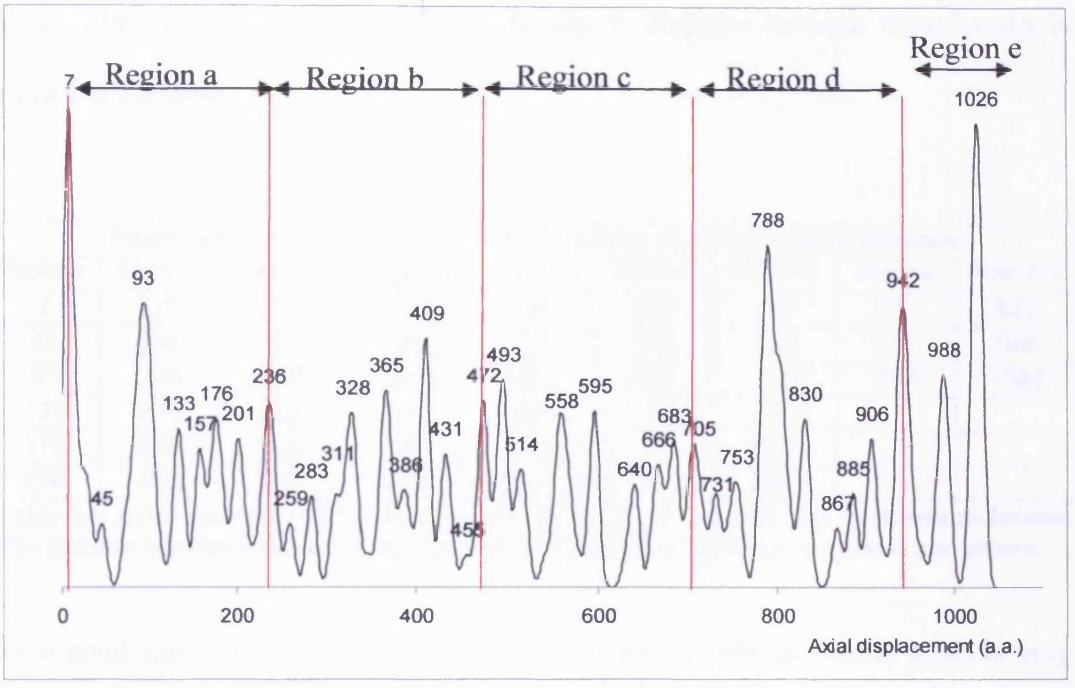


Figure 3.23: Gaussian smoothing of the axial hydrophobic amino acid distribution for the configuration with folded telopeptides. Hydrophobic residues clusters are distributed in a similar way inside the periodic regions.

To verify this hypothesis we have considered the sequence of hydrophobic amino acids with folded telopeptides. We have then applied a Gaussian smoothing to it to highlight those regions where the amino acids cluster together to form higher peaks. The result is presented in figure 3.23 where we can notice four regions between a.a. 7 and 236 (region a), between a.a. 236 and 472 (region b), between a.a. 472 and 705 (region c) and between a.a. 705 and 942 (region d) that are characterised by a similar distribution of the hydrophobic residues clusters. A fifth incomplete region between a.a. 942 and 1026 is also seen (region e). It is important to point out here that the positions of these peaks do not correspond to real positions of the amino

acids along the collagen molecule but they represent the positions of the centre of mass of a.a. clusters.

In table 3.4 we show the positions of the peaks. These are linked by a periodicity of about 234 a.a. (233.3±3.5) a.a.. The measured distance between these peaks is shown in the table.

Peaks positions belonging to different regions and their relative distance								
Region a	Distance	Region b	Distance	Region c	Distance	Region d	Distance	Region e
7	229	236	236	472	233	705	237	942
45	238	283	231	514	239	753	235	988
93	235	328	230	558	230	788	238	1026
133	232	365	230	595	235	830		
176	233	409	231	640	227	867		
236	236	472	233	705	237	942		

Table 3.4: Axial positions of hydrophobic residue clusters in murine type I collagen molecules. The distance between corresponding amino acids positions in different regions is also shown.

To a good approximation, the distribution of the hydrophobic amino acids is very similar inside the regions as shown in table 3.5 where the relative distance between a residue and the successive one is shown. It is also possible to notice that the hydrophobic residue clusters are separated by two main distances corresponding to about 42 a.a. or to about 66 a.a.. We use the values for the distances among the hydrophobic clusters of table 3.5 to build a simplified model that represents the hydrophobic amino acids distribution along the collagen molecule. This model is thus represented by the repetition of four complete regions of hydrophobic amino acids according to the distribution {42, 42, 42, 42, 66} in which five hydrophobic clusters are separated by 42 a.a. and are followed by a gap 66 a.a. long, plus a fifth incomplete region where the amino acids are positioned according to the distribution {42, 42} (table 3.5). We calculated then the multiple parallel hydrophobic autoscoring for this simplified model and compared it to the real one (figure 3.24).

Axial distribution of hydrophobic residues clusters	
position	distance
7	38
45	48
93	40
133	43
176	60
236	47
283	45
328	37
365	44
409	63
472	42
514	44
558	37
595	45
640	65
705	48
753	35
788	42
830	37
867	75
942	46
988	38
1026	

Table 3.5: Axial distribution of the hydrophobic residue clusters and the relative distance between successive peaks. Regions of periodic distributions are coloured in yellow.

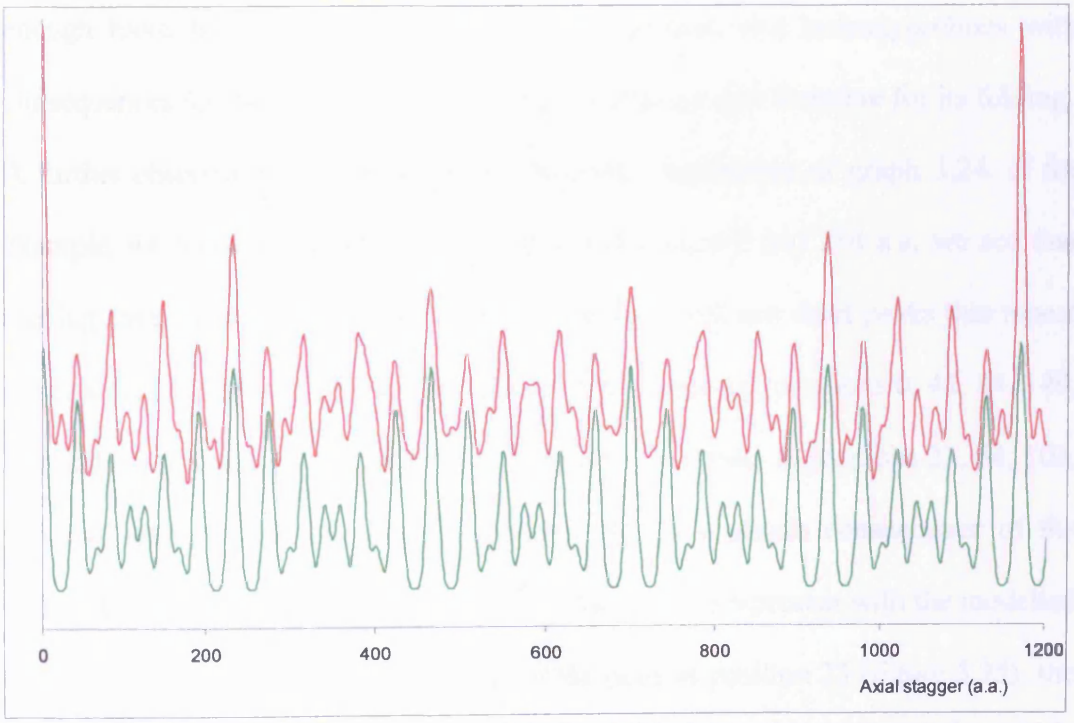


Figure 3.24: Comparison between the real hydrophobic interaction (red) and the model-generated hydrophobic interaction (green) as a function of the axial stagger (a.a.). A constant value was added to the real hydrophobic interaction (red) to make the plot less cluttered.

Qualitatively the autoscoring functions of the simplified model and of the real amino acid distribution are very similar. They are characterised by the same regions of symmetry were the maxima have the same distribution. We used equation 2.3 (with j=0) to calculate the correlation coefficient r between the two sequences. In this case, r=0.84, a very good value if we consider the high number of elements (1170) in the two sequences. It seems thus reasonable to use the simplified model to represent the hydrophobic distribution that gives rise to the hydrophobic interaction during the fibrillogenesis process.

It seems that the 42 a.a. distance between hydrophobic residues is a very favourable distance to drive the hydrophobic interaction. A possible explanation could be that if the hydrophobic residues were more widely spaced probably the hydrophobic interaction would be too weak. Conversely, if they were more closely spaced the charged amino acids would be too few making their interaction ineffective and ruining the coming together of the molecules. In addition, there would not be enough room for other amino acids such as prolines and hydroxyprolines with consequences for the stiffness of the collagen molecule and therefore for its folding.

A further observation can be made on the peak distribution of graph 3.24. If for example, we focus our attention to the region between 0 and 234 a.a. we see that starting from the peak at position zero a.a. we have tall and short peaks that repeat in sequence. Figure 3.25 shows this region, where peaks at positions 0, 41, 84, 149, 191 and 234 a.a. are the most prominent whereas the peaks at positions 23, 64, 103, 125, 167 and 211 a.a. are less prominent. This is a simple consequence of the distribution of the hydrophobic amino acids that we can represent with the modelled distribution. In fact, with the exception of the peak at position 23 (figure 3.25), the autoscoring function of the modelled distribution reproduces very well the real one.

The modelled amino acid distribution can explain this behaviour. The diagram of figure 3.26 represents the amino acids distribution according to the {42, 42, 42, 42, 66}*4+{42, 42} model distribution. It is easy to see that for a shift of about 21 a.a. there is no scoring (no hydrophobic clusters on the two molecules face each other). While for a shift of about 42 a.a. 18 clusters face each other. Again, for a stagger of 66 a.a. only four clusters face each other and for a shift of 84, 13 amino acids clusters face each other. This can be repeated to reproduce all the tall and short peaks in graph 3.25.

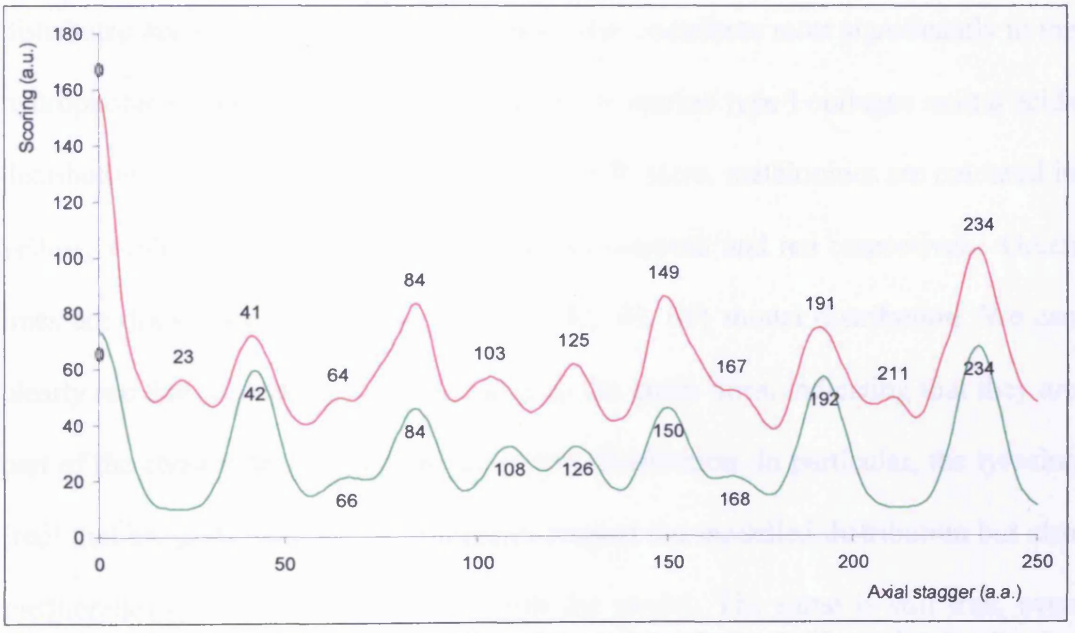


Figure 3.25: Comparison between a portion of the multiple hydrophobic scoring function for the real amino acid distribution (red) and the model distribution (green).

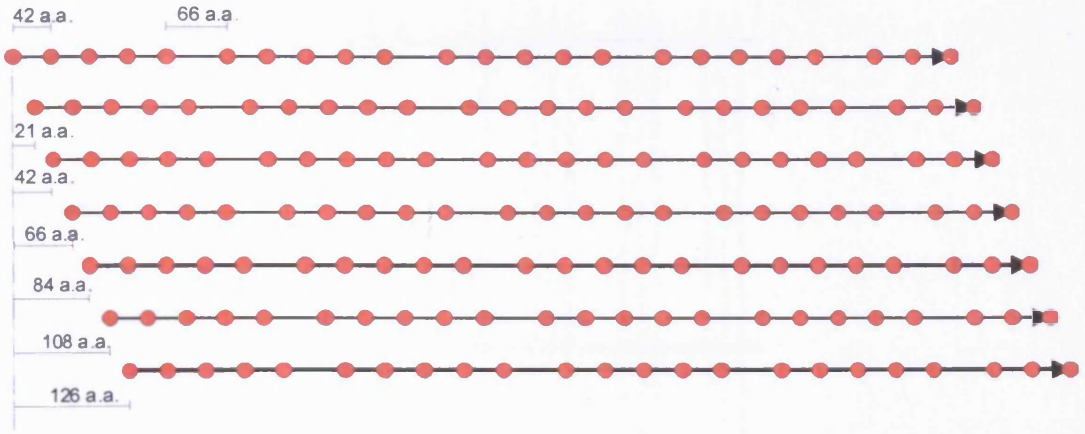


Figure 3.26: Diagram representing different axial staggers for the modelled hydrophobic sequence.

Since the scoring for the model distribution represents very well the scoring of the real distribution, we can assume that the amino acids in the clusters that are distributed according to our model are those that contribute most significantly to the hydrophobic interaction. Figure 3.27 shows the murine type I collagen amino acids distribution aligned with a 234 amino acids shift. Here, methionines are coloured in yellow, while phenylalanines and tyrosine in magenta and red respectively. Green lines are drawn according to the {42, 42, 42, 42, 66} model distribution. We can clearly see that M, F and T lie very close to the green lines, indicating that they are part of the clusters that suggested our model distribution. In particular, the tyrosine (red) that are gathered at the telopeptides respect the modelled distribution but also methionines (yellow) do agree well with the model. The same is still true, even though less markedly, for the phenylalanines (magenta) that are distributed on the first, the third and the fifth green line to a good approximation.

This probably means that the most voluminous hydrophobic amino acids, those with the biggest and the most mobile side chains (M, F, T), are those that contribute the most to hydrophobic collagen packing. The same could hold true for protein aggregation in general.

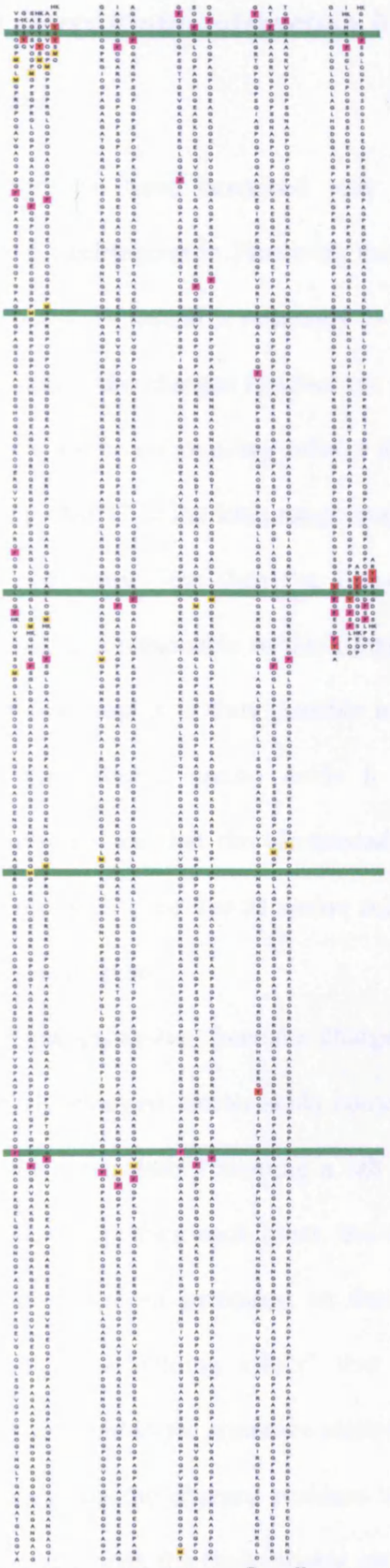


Figure 3.27: Comparison between murine type I collagen residue distribution and the modelled hydrophobic residue distribution (green lines) according to the {42, 42, 42, 42, 66} distribution. F (magenta), M (yellow), Y (red)

3.6 Multiple parallel electrostatic interaction for collagen sequence with folded telopeptides

In the previous paragraphs, we have described only the contribution of the hydrophobic residues during fibrillogenesis. However, they are not the only amino acids acting during this process. In fact, it is necessary to consider the contribution of the charged residues as well. The charged residues are made of voluminous and mobile side chains immersed in the surrounding solvent that can isolate them from each other with a "shielding effect" in the long range interactions. Thus, only side chains that are in close proximity, less than the Debye distance, can interact significantly. For this reason, it is reasonable to think that the charged interactions are essentially short range ones and it is thus possible to apply the same scoring technique used for the hydrophobic amino acids to study their interaction properties. In this way, we postulate that the electrostatic forces arising between adjacent parallel collagen molecules are due to amino acids with opposite charges that come into contact with each other.

Ideally, this limitation is not excessive, because charged residues have mobile chains. When two oppositely charged amino acids come sufficiently close, their side chains attract each other presumably forming a salt bond. Conversely, when two residues with the same charge face each other, the mobile chains repel each other and the distance between them increases, so that much solvent flows in between them. This causes a "shielding effect" that makes the electrostatic repulsion much smaller if compared to the attractive electrostatic force.

We built a sequence representing the charged residues in the collagen molecule, using the same philosophy used with the hydrophobic amino acids (see paragraph 3.1). The collagen molecule used here is that with folded telopeptides found during

the hydrophobic scoring analysis (paragraph 3.4). The value 1 was assigned to the positively charged amino acids (K, HK and R) and the values from the three α-chains were summed. The same was done for the negatively charged amino acids (D and E). Finally we built a sequence made of multiple molecules using the gap G=135 a.a. found during of the hydrophobic scoring analysis. Since both sequences of positive and negative residues are represented by ones, when they are scored together, the correlation function always has a positive value. This allows us to interpret the new scoring function in a way similar to that of the hydrophobic interactions.

We then scored a single moving charged molecule against a multiple sequence of charged molecules in a parallel configuration. To study the attractive electrostatic force we chose to break it down into two components. The one obtained sliding a positive molecule against a multiple sequence of negative molecules and the opposite one, with a negative molecule sliding against a positive multiple sequence. The total attractive force is then obtained by summing both contributions. In this case, it is not possible to use the autocorrelation function (equation 2.4) but it is necessary to use the more general cross-correlation function (equation 2.2) because the two interacting sequences are not identical. Cross-correlations do not necessarily have an absolute maximum in the origin (equation 2.19) and it is not symmetrical with respect to the origin (equation 2.8). In figure 3.28 and 3.29, we present the electrostatic scoring function due to a positive molecule sliding past a multiple sequence of negative molecules and that of a negative molecule sliding past a multiple sequence of positive molecules. Both graphs have a periodicity of 1170 a.a. by construction, in addition they are mirror-symmetric with respect to the origin.

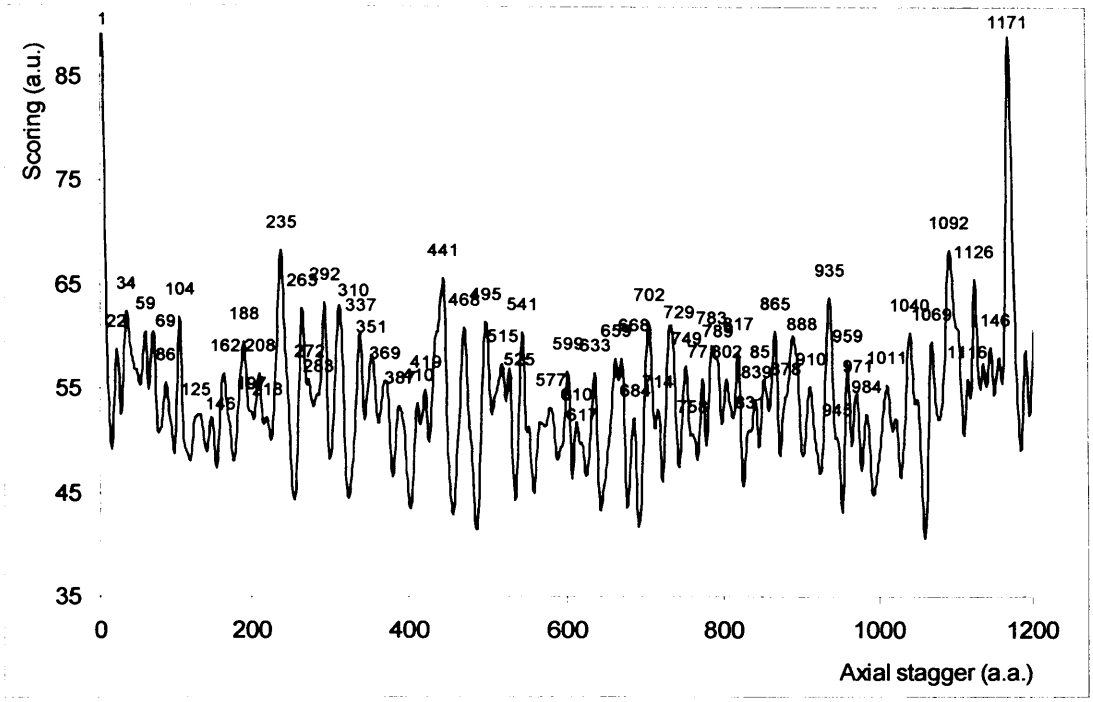


Figure 3.28: Electrostatic scoring function of a positive charge sequence representing one molecule, against a negative charge sequence representing a multiple sequence of molecules.

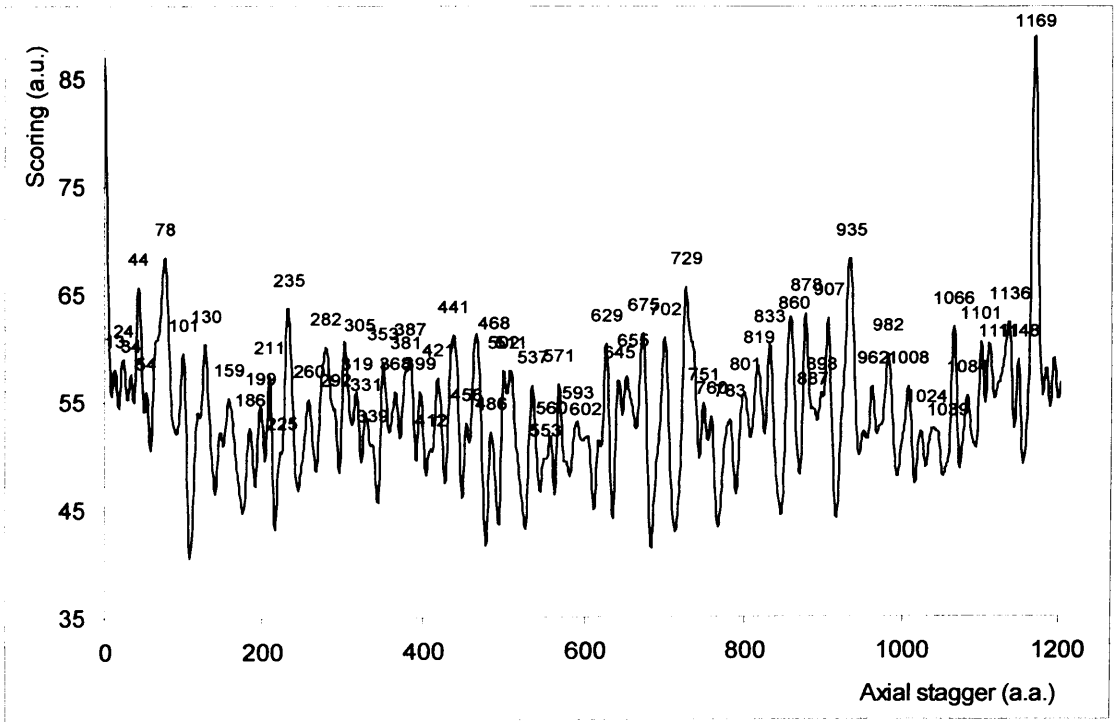


Figure 3.29: Electrostatic scoring function of a negative charge sequence representing one molecule, against a positive charge sequence representing a multiple sequence of molecules.

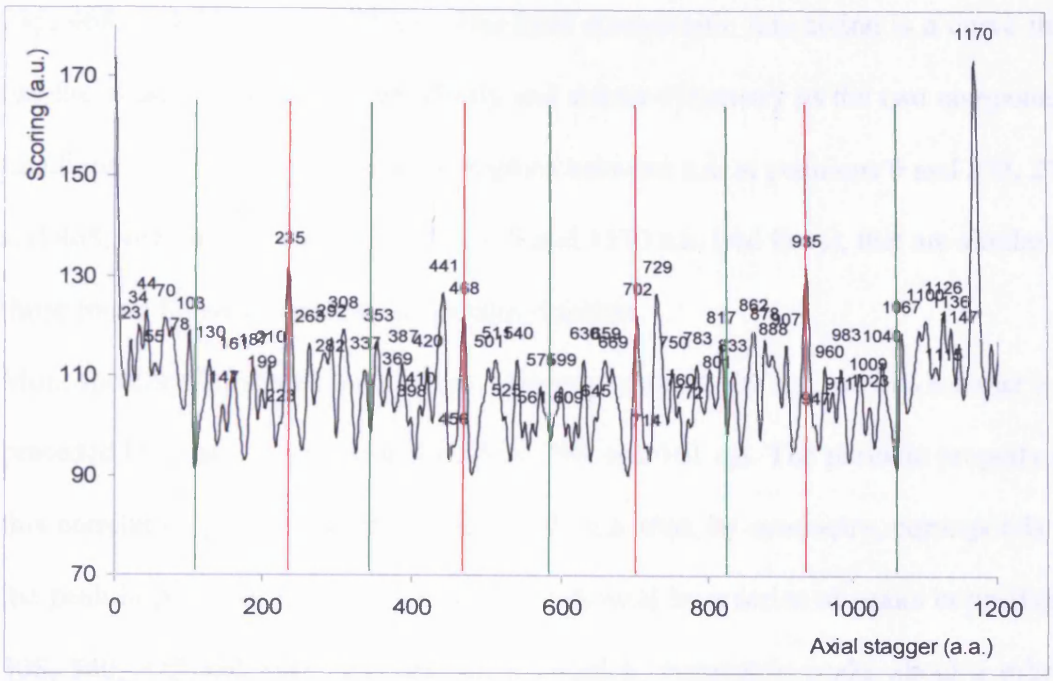


Figure 3.30: Multiple attractive electrostatic scoring function for murine type I collagen molecule. Regions of periodicity are delimited in red, mirror-planes are in green.

In graph 3.30, we show the total attractive electrostatic interaction obtained by summing both contributions. This graph has a 1170 a.a. periodicity and it is symmetric with respect to the origin. It is also symmetric with respect to the plane at 585 a.a.. This is because we summed two antisymmetrical functions. If we now consider the graph in figure 3.28 representing the scoring of a positive molecule against a multiple sequence of negative ones, we can notice that some of its peaks are characterised by a periodicity of about 234 a.a. typical of fibrillar collagens. For example, those at positions 235, 468, 702, 935 and 1171 a.a. are apparent and correspond to peaks at the same positions found for the hydrophobic interactions (figure 3.21). Other less prominent peaks as those in positions 188, 419, 659, 888 and 1126 a.a., have the same periodicity.

The graph in figure 3.29 has exactly the same peaks as those in figure 3.28 but they are mirror-symmetric with respect to the origin. These peaks are those in positions

235, 468, 701 935 and 1169 a.a.. The total electrostatic interaction is a curve that has the same characteristic periodicity and mirror-symmetry as the two component functions. Figure 3.30 has periodic regions between a.a. at positions 0 and 235, 235 and 468, 468 and 702, 702 and 935, 935 and 1170 a.a. (red lines), that are similar to those found for the hydrophobic scoring function.

More specifically, if we consider the peak at position 1100 a.a., we can see that it is preceded by peaks at positions 862, 630, 398 and 161 a.a. The periodic property of this correlation graph predicts a peak at -73 a.a. that, by symmetry, corresponds to the peak in position 70 a.a.. This peak is followed by a series of peaks in positions 308, 540, 772 and 1009 a.a. that correspond to symmetric peaks about a mirror plane in position 585 a.a..

In this way, we can highlight the formation of five periodic regions where four pairs of peaks are mirror-symmetric with respect to a central plane (green lines). Since the peak positions are predictable, we can build a model that describes their positions, in a way similar to that used for the hydrophobic peaks in paragraph 3.2. In table 3.6, we show the positions of the peaks in the correlation function and those that are estimated. The average discrepancy between the real and the estimated peaks is about 1.5 a.a.. This confirms the well ordered distribution of the peaks.

Peaks Positions	Estimated Peaks Positions	Discrepancy
0	0	0
23	27	4
34	36	2
44	47	3
70	72	2
103	103	0
130	131	1
161	162	1
187	187	0
199	198	1
210	207	3
235	234	1

Peaks Positions	Estimated Peaks Positions	Discrepancy
263 (≈23+234)	261	2
missing	270	
282 (≈44+234)	281	1
308 (≈70+234)	306	2
337 (≈103+234)	337	0
369 (≈130+234)	365	4
398 (=161+234)	396	2
420 (≈187+234)	421	1
missing	432	
441 (≈210+234)	441	0
468 (≈235+234)	468	0
missing	495	
501 (≈34+234*2)	504	3
511 (≈44+234*2)	515	4
540 (≈70+234*2)	540	0
571 (≈103+234*2)	571	0
599 (≈130+234*2)	599	0
630 (≈161+234*2)	630	0
659 (≈187+234*2)	655	4
669 (≈199+234*2)	666	3
missing	675	
702 (≈235+234*2)	702	0
729 (≈23+234*3)	729	0
missing	738	
750 (≈44+234*3)	749	1
772 (≈70+234*3)	774	2
801 (≈103+234*3)	805	4
833 (≈130+234*3)	833	0
862 (≈161+234*3)	864	2
888 (≈187+234*3)	889	1
missing	900	
907 (≈210+234*3)	909	2
935 (≈235+234*3)	936	1
960 (≈23+234*4)	963	3
971 (≈34+234*4)	972	1
983 (≈44+234*4)	983	0
1009 (≈70+234*4)	1008	1
1040 (≈103+234*4)	1039	1
1067 (≈130+234*4)	1067	0
1100 (≈161+234*4)	1098	2
1126 (≈187+234*4)	1123	3
1136 (≈199+234*4)	1134	2
1147 (≈210+234*4)	1143	4
1170 (≈235+234*5)	1170	0

Table 3.6: Comparison between estimated peak positions and their real positions for the total electrostatic attractive scoring function. Periodic regions are in yellow.

The peak distribution of the electrostatic scoring is such that it reinforces the peak distribution of the hydrophobic scoring function. This is particularly apparent if we compare the multiple attractive electrostatic scoring in figure 3.30 with the multiple hydrophobic scoring in figure 3.21, where peaks at integer multiple of 234 a.a. are the most prominent for both graphs.

In order to represent the combined effect of both kinds of interactions, when they act together, we multiply point by point the two graphs. In this way, this multiplication is used as an operator similar to the logical operator "AND" that produces a graph that has peaks where both multiplied graphs have peaks. A sum would not work as well because the scoring function tells us nothing about the absolute values of the interactions. Their sum would thus be a misleading quantity.

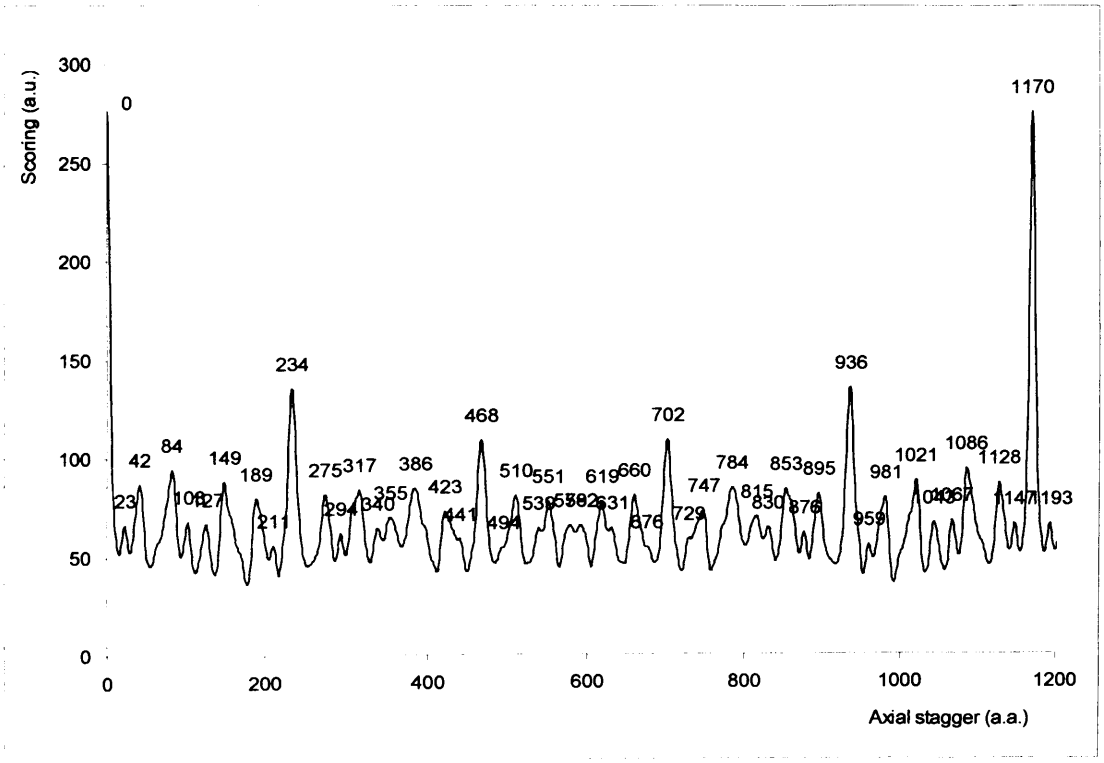


Figure 3.31: Point by point multiplication between the hydrophobic and the attractive electrostatic scoring functions for murine type I collagen. This operation represents the cooperative effect of both forces when they act together.

Figure 3.31 shows the graph of the two multiplied interactions. Since the hydrophobic scoring is characterised by peaks whose amplitude is much bigger than the one of the electrostatic scoring, the graph in figure 3.31 is very similar to that in figure 3.21. It is as if the hydrophobic scoring had modulated the electrostatic scoring. However since both interactions have dominant peaks that are integer multiple of 234 a.a., it is immediate to see that they combine to give the highest peak in position 234 a.a. (if we exclude the one at the origin that does not have any physical sense) confirming the results of the two separated analysis.

3.7 Antiparallel hydrophobic scoring

Until now, we have considered the hydrophobic and the electrostatic interactions arising between collagen molecules in a parallel configuration. Now we can apply the same method to study the case of the hydrophobic interaction when the molecules are in an antiparallel configuration. This approach can help us to understand this particular kind of aggregation that is sometimes found in collagen molecules reconstituted *in vitro* (Bruns, 1976; Williams et al., 1978) and *in vivo* (Mallinger et al., 1992). In addition, if our predictions are verified, we have a confirmation of the validity of the method.

We use the amino acid sequence previously built to study the parallel hydrophobic scoring with folded telopeptides. However, we invert one sequence with respect to the other. In this case, since the two sequences are not identical, we must use the cross correlation function (equation 2.2).

Before proceeding, we recall that the hydrophobic a.a. sequence is made of periodic regions that repeat every 234 a.a.. We saw that the real distribution of hydrophobic

amino acids along a type I collagen molecule can be represented by a repeat according to the {42, 42, 42, 42, 66}*4+{42, 42} a.a. model distribution (equation 3.7), that is:

$$H_{NC}=\{42,42,42,42,66\}*4+\{42,42\} \quad 3.7$$

where, H_{NC} is the hydrophobic distribution oriented from the N-telopeptide to the C-telopeptide. Similarly, we can represent the reversed sequence, that oriented from the C-telopeptide to the N-telopeptide as:

$$H_{CN}=\{42,42\}+\{66,42,42,42,42\}*4 \quad 3.8$$

From a comparison between equations 3.7 and 3.8, it is apparent that if the inverted sequence is shifted along the straight sequence by about 42*2=84 a.a., the two oppositely oriented sequences are again in register (figure 3.32). It is thus reasonable to expect a maximum for the hydrophobic interaction around this shift value between the two sequences.

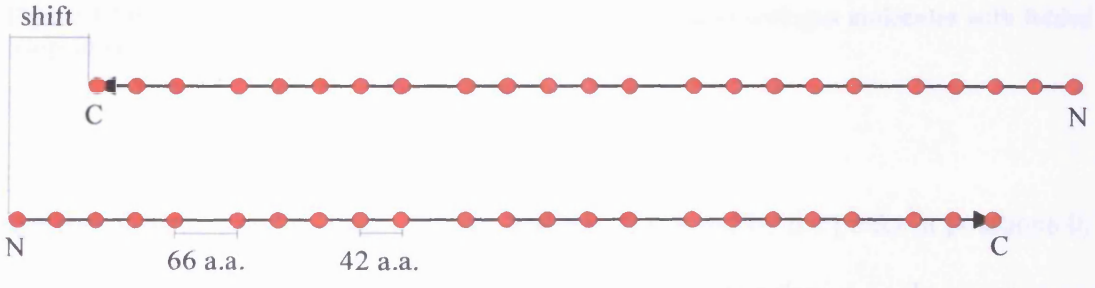


Figure 3.32:A sequence of hydrophobic residues according to the {42, 42, 42, 42, 66}*4+{42,42} distribution and oriented from the C- to the N-terminal is shifted with respect to an identical sequence oriented form the N-to the C-terminal. The cross-correlation is highest when the shift is equal to 42*2 a.a.

In figure 3.33, we present the antiparallel hydrophobic scoring for two molecules with folded telopeptides after having applied a Gaussian smoothing. The region between a displacement of 0 and 228-236 a.a. has peaks that are very similar to those found for the parallel hydrophobic autoscoring shown in figure 3.4. We can also see a few mirror-symmetric peaks, with respect to plane drawn at 117 a.a. positions, (red line in figure 3.33) such as those in positions 38 and 191, 57 and 176, 80 and 163, 102 and 127 a.a.. Interestingly one of the most prominent peaks is at position 80 a value that is very close to that predicted above.

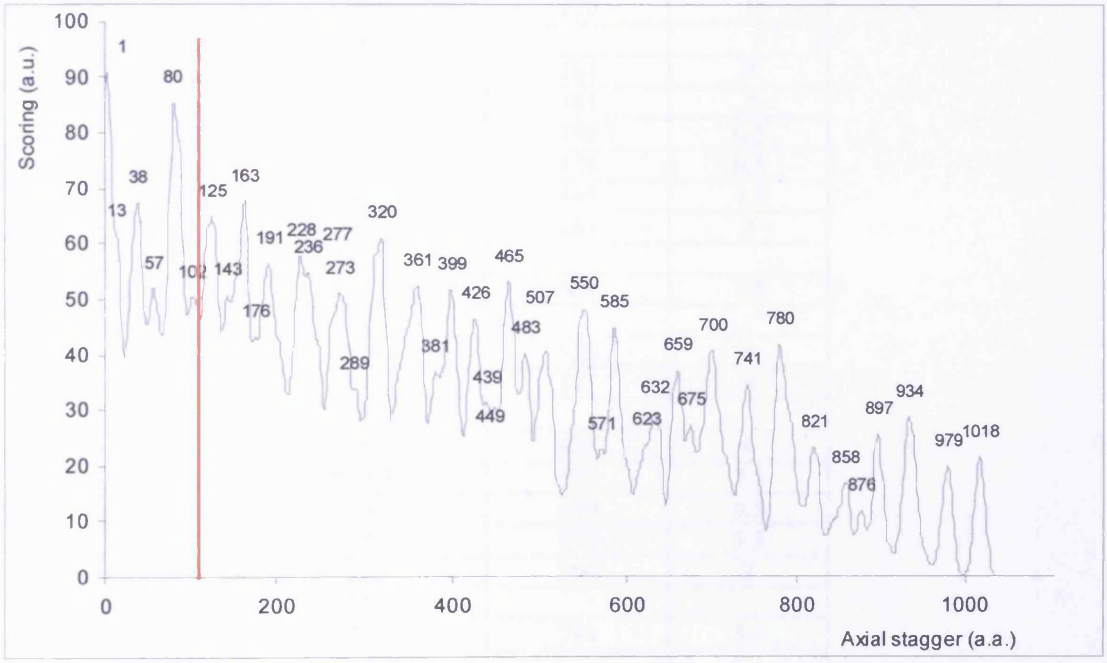


Figure 3.33: Hydrophobic scoring function for two antiparallel collagen molecules with folded telopeptides.

We can observe that there are periodic regions delimited by the peaks at positions 0, 228-336, 465, 700 and 934 a.a. where the peaks are distributed in the same way. The average distance between homologous peaks belonging to different regions is about 234 a.a. (234.25±4.7 a.a.). These peaks have a similar distribution to those

found by hydrophobic autoscoring. We can compare them with the same distribution obtained for the parallel configuration model as shown in table 3.7.

Estimated positions	Real Positions	Discrepancy
0	1	1
23.5		
42.5	38	4.5
62.5	57	5.5
85	80	5
107	102	5
127	125	2
149	143	6
171.5	176	4.5
191.5	191	0.5
210.5		
234	236	2
257.5		
276.5	277	0.5
296.5		
319	320	1
341		
361	361	0
383	381	2
405.5	399	6.5
425.5	426	0.5
444.5	449	4.5
468	465	3
491.5		
510.5	507	3.5
530.5		
553	550	3
575	571	4
595		
617	623	6
639.5	632	7.5
659.5	659	0.5
678.5	675	3.5
702	700	2
725.5		
744.5	741	3.5
764.5		
787		
809		
829		
851	858	7
873.5	876	2.5
893.5	897	3.5
912.5		
936	934	2
959.5		
978.5	979	0.5
998.5		
1021	1018	3

Table 3.7: Comparison between estimated positions and real peak positions for the antiparallel hydrophobic scoring function.

From this comparison, we can see how the two different configurations generate a very similar pattern (the average discrepancy between the two data sets is about 3.3±2.2 a.a.). This is a direct consequence of the similar distribution of the hydrophobic amino acids for the parallel and the antiparallel sequences.

However, the multiple antiparallel configuration is more interesting because it represents a more realistic interaction between a molecule and a fibril during fibrillogenesis. To represent it, we use the molecule with folded telopeptides previously described (paragraph 3.4). Since this configuration is 1035 a.a. long and the period T is about 234 a.a., by applying equation 3.6 we obtain that the gap G=135 a.a.. We use this value to build a multiple sequence of molecules that are separated by a gap G. We then calculate the hydrophobic scoring of this multiple sequence with an identical single inverted molecule. In figure 3.34, we report the result obtained.

Figure 3.35 is a comparison of the parallel hydrophobic autoscoring (green) with the antiparallel hydrophobic scoring (red). It is immediately possible to see that the two different scoring functions are very similar. This is because of the similar distribution of the hydrophobic residues for the parallel and antiparallel sequence. The cross correlation coefficient r, for the two functions is about 0.77 when calculated on 1170 elements.

Figure 3.34 shows the antiparallel hydrophobic scoring. Two main families of major peaks are visible. One starting from position -2 (=1168-1170) with peaks at positions -2, 230-237 (≈-2+234), 467 (≈-2+234*2), 700 (≈-2+234*3), 932 (≈-2+234*4) and 1168 (=-2) a.a.; the other starting with the peak in position 80 with following peaks at positions 319 (≈80+234), 551 (≈80+234*2), 782 (≈80+234*3) and1017 (≈80+234*4) a.a..

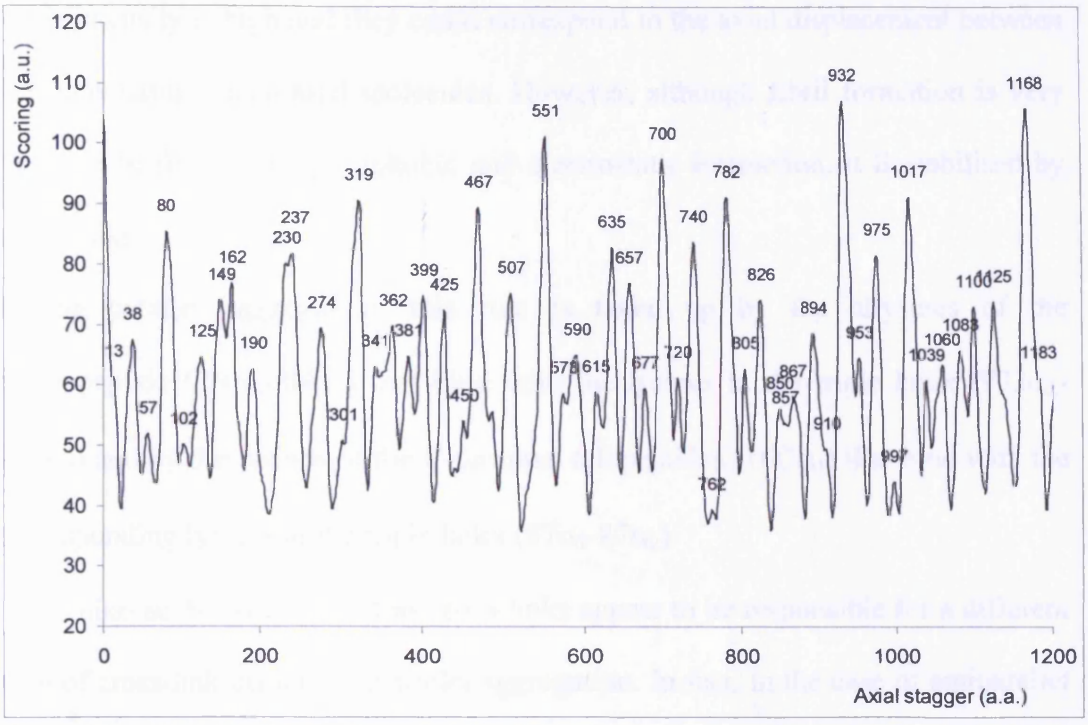


Figure 3.34: Multiple antiparallel hydrophobic scoring functions for molecules with folded telopeptides.

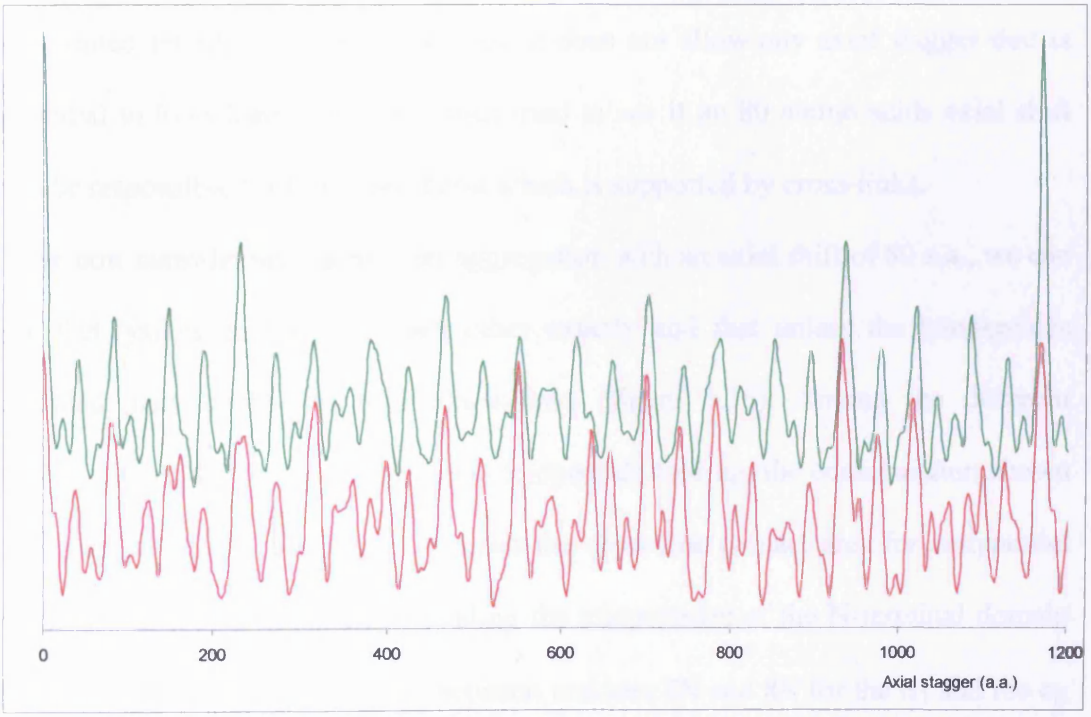


Figure 3.35: Comparison between the multiple parallel (green) and antiparallel (red) hydrophobic scoring functions.

Their intensity is high and they could correspond to the axial displacement between two interacting antiparallel molecules. However, although fibril formation is very likely to be driven by hydrophobic and electrostatic interaction, it is stabilised by cross-links.

During parallel aggregation, this role is taken up by the allysines of the N-telopeptide ($7N\alpha_2$-$8N\alpha_1$) that bind with the lysines in the triple helix ($933\alpha_2$-$930\alpha_1$) and by the lysines of the C-terminal telopeptides ($16C\alpha_1$) that bind with the corresponding lysines in the triple helix ($87\alpha_1$-$87\alpha_2$).

The amino acids involved in this cross-links appear to be responsible for a different type of cross-link during antiparallel aggregation. In fact, in the case of antiparallel aggregation with -2 a.a. axial shift, or an axial shift of -2 a.a. plus integer multiples of 234 (figure 3.36), the lysines of the telopeptides and the triple helix face each other and could form cross-links. However, we think that this configuration is not best suited for fibril formation because it does not allow any axial stagger that is essential to form long fibrils. We thus tried to see if an 80 amino acids axial shift can be responsible for fibril formation which is supported by cross-links.

If we now consider the antiparallel aggregation with an axial shift of 80 a.a., we can see that lysines do not face each other exactly and that unless the telopeptides refolded, they could not form cross-links (figure 3.36). Among the different configurations that we tried for a new telopeptide folding, the conformation shown in figure 3.37 is the one that maximises the cross-link contact area for antiparallel aggregation. It was obtained by folding the telopeptides of the N-terminal domain between residues 6N and 7N and between residues 7N and 8N for the α_1 and the α_2 chains respectively. The telopeptides of the C-terminal domain were instead folded at residue positions 10C and 11C and at positions 8C and 9C for the α_1 and the α_2

chains respectively. In figure 3.39, we present the antiparallel hydrophobic scoring for this specific configuration. We can see peaks similar to those of graph 3.34 for the previous antiparallel folding configuration. The new peak in position 86 a.a. corresponds to the one in position 80 a.a. in the previous configuration.

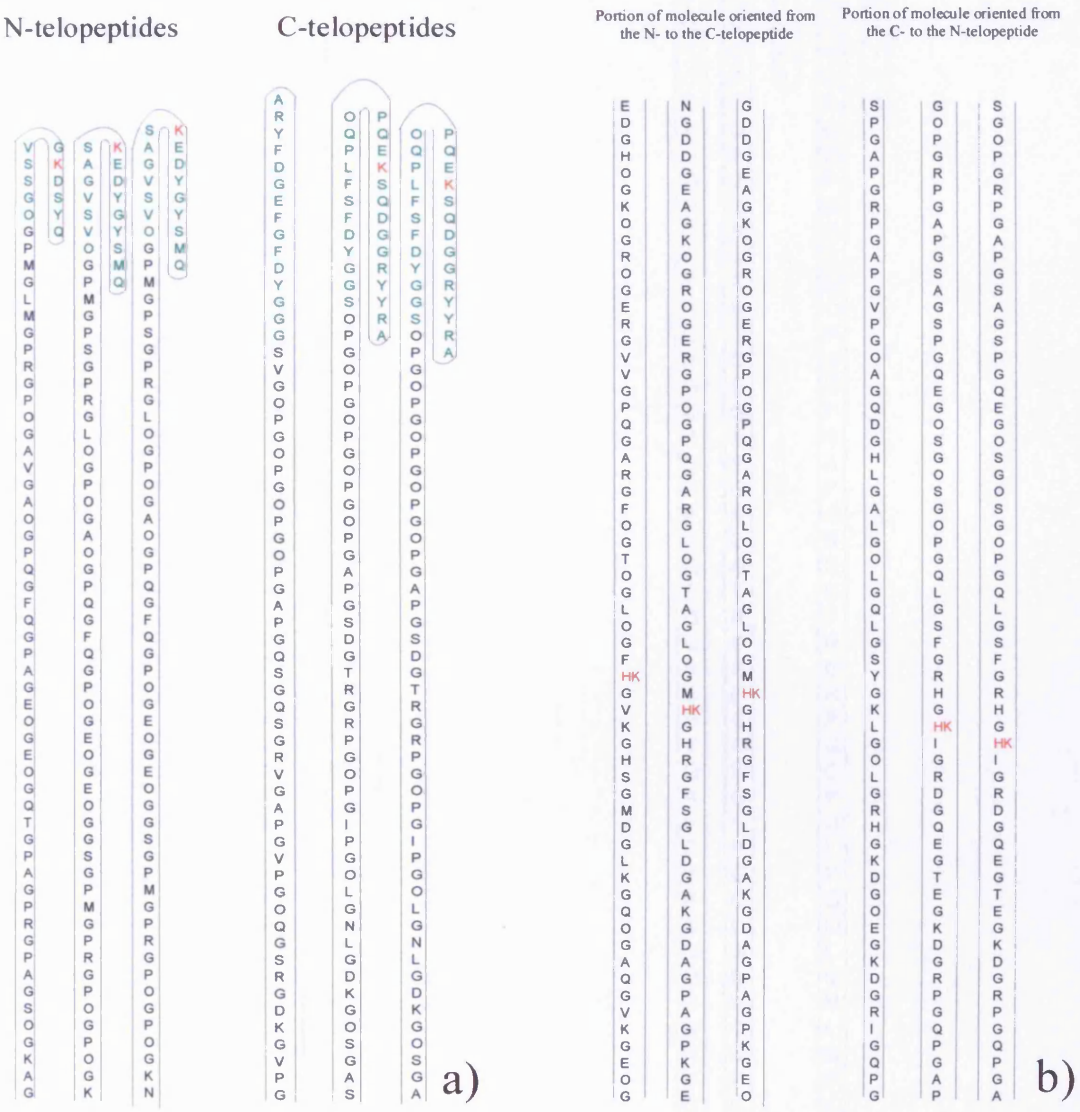


Figure 3.36: Possible cross-link formation between two murine type I collagen molecules in an antiparallel configuration with -2 a.a.-stagger. a) cross-links can arise between telopeptides. b) cross-links can arise between lysines belonging to the triple helix. Residues in the telopeptides are in green, putative lysines responsible for cross-links are in red.

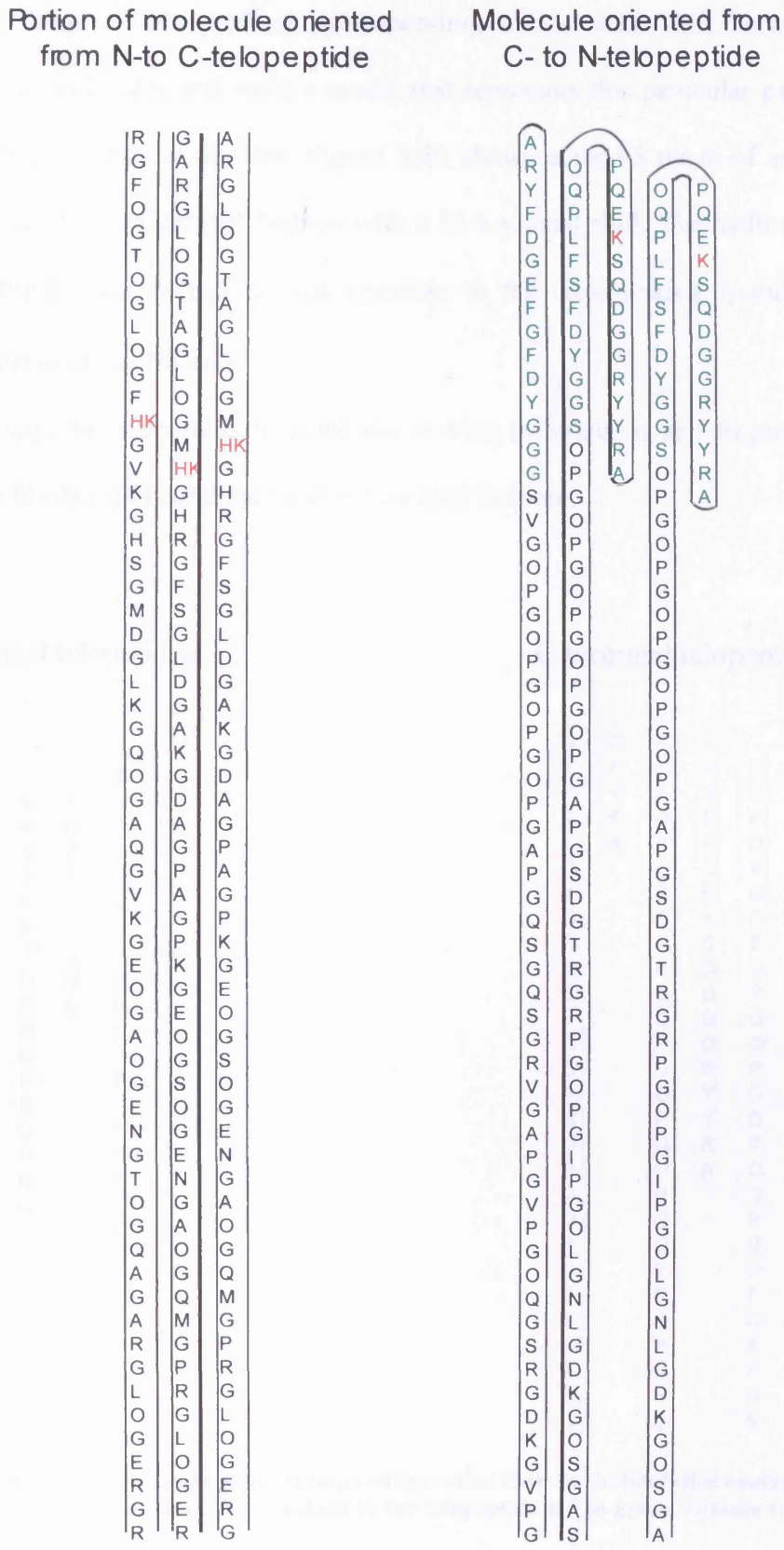


Figure 3.37: Cross-links formation between two murine type I collagen molecules in antiparallel configuration with an 80 a.a. stagger. Residues in the telopeptides are in green. Putative lysines responsible for cross-links are in red.

We can think of this peak as corresponding to the axial shift between two antiparallel molecules and build a model that represents this particular packing of antiparallel collagen molecules. Figure 3.40 shows a model made of molecules aggregated in an antiparallel fashion with a 86 a.a. axial shift. The hydroxylysines of the triple helix ($87\alpha_1$) is now opposite to the cross-linking lysines of the C-terminal domain ($16C\alpha_1$).

In summary, the antiparallel hydrophobic scoring technique is able to predict how collagen fibrils could combine in an antiparallel fashion.

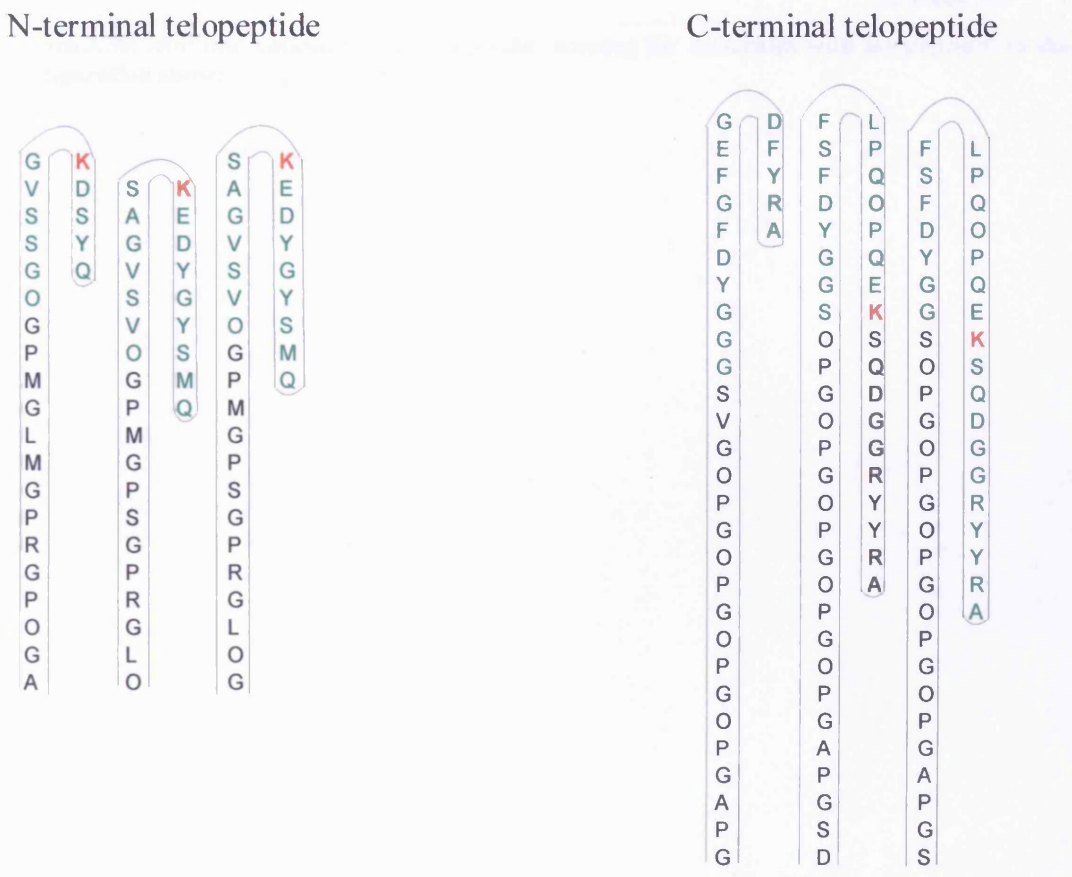


Figure 3.38: Optimised telopeptide folding configuration that maximises lysine contacts for the antiparallel interaction case. The residues in the telopeptide are in green. Lysines responsible for cross-links are in red

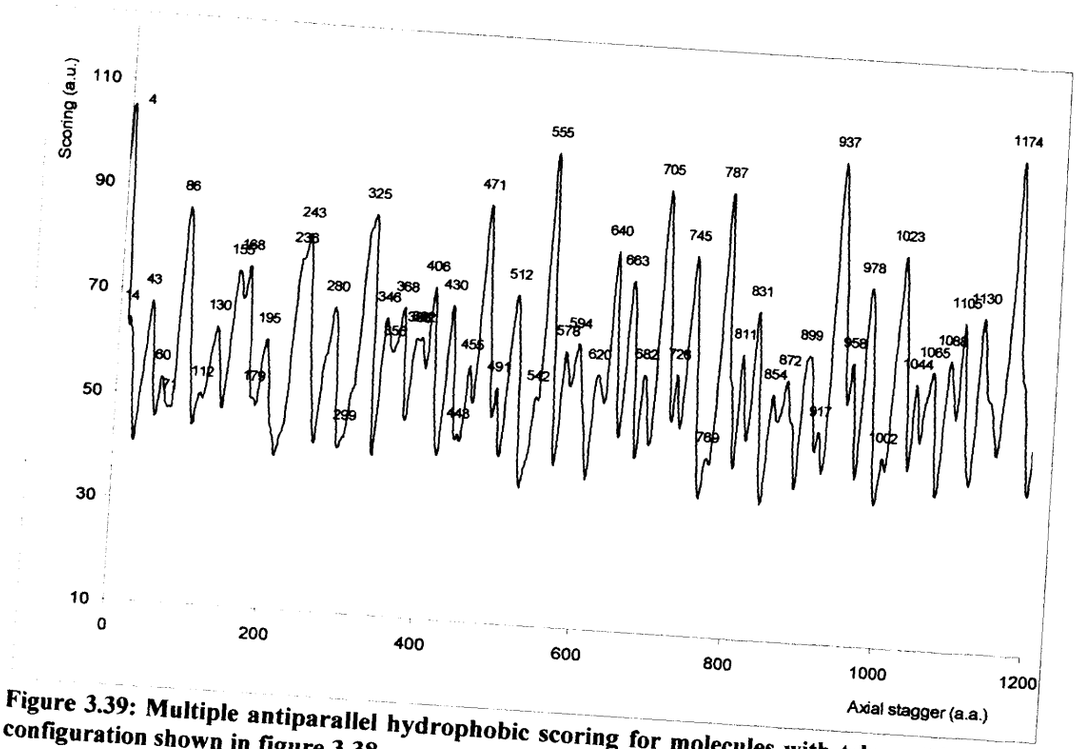


Figure 3.39: Multiple antiparallel hydrophobic scoring for molecules with telopeptides in the configuration shown in figure 3.38.

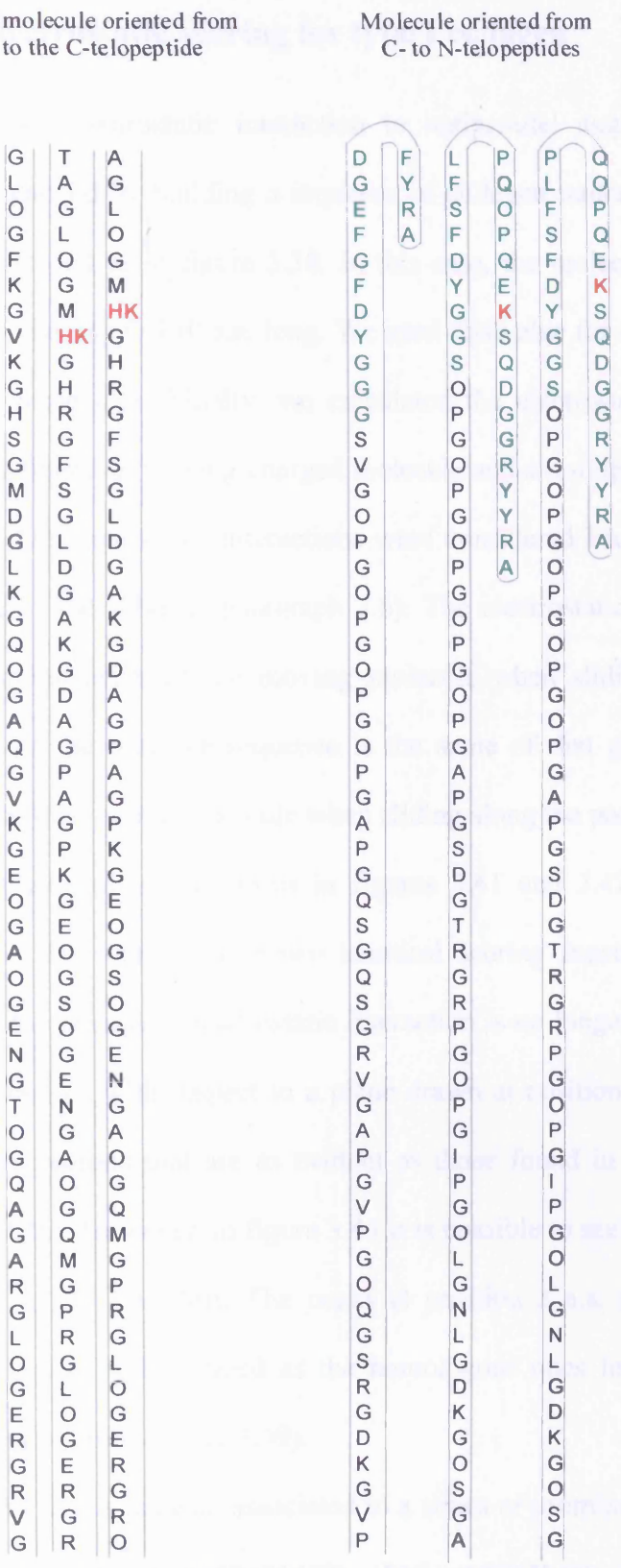


Figure 3.40: Cross-link formation between two murine type I collagen molecules in an antiparallel configuration with an 86 a.a. stagger. Residues in the telopeptides are in green. Putative lysines responsible for cross-links are in red. The telopeptides are folded according to figure 3.38.

3.8 Antiparallel electrostatic scoring for type I collagen

The contribution of the electrostatic interaction to antiparallel aggregation of fibrillar collagen was gauged by building a sequence of collagen molecules where the telopeptides were folded as in figure 3.38. In this case, the molecule is 1030 elements long, making the gap G 140 a.a. long. We used this value for G to build a multiple sequence of molecules. Finally, we calculated the electrostatic scoring function between an antiparallel moving charged molecule and a multiple sequence of parallel molecules. Only attractive interactions were considered because of the "shielding" properties of the solvent (paragraph 3.6). The electrostatic interaction due to the positive amino acids of the moving molecule when sliding past the negative amino acids of the multiple sequence is the same of that given by the negative amino acids of the moving molecule when sliding along the positive amino acids of the multiple sequences; as shown in figures 3.41 and 3.42. The total attractive interaction is the sum of these two identical scoring functions (figure 3.43). As we can see, the attractive electrostatic interaction is no longer symmetric with respect to the origin or with respect to a plane drawn at position 585 a.a.. It does not have periodic regions that are as evident as those found in the parallel configuration (figure 3.26). However, in figure 3.43 it is possible to see some peaks that are separated by a 234 a.a. shift. The peaks at position 3 a.a. plus integer multiples of 234 are not as well defined as the homologous ones found for the antiparallel hydrophobic scoring (figure 3.39).

The peak in position 88 a.a. is instead, associated to a series of prominent peaks in positions 320 (≈88+234), 559 (≈88+234*2), 790 (≈88+234*3) and 1024 (≈88+234*4) a.a. that correspond to peaks found for the antiparallel hydrophobic scoring function (figure 3.39).

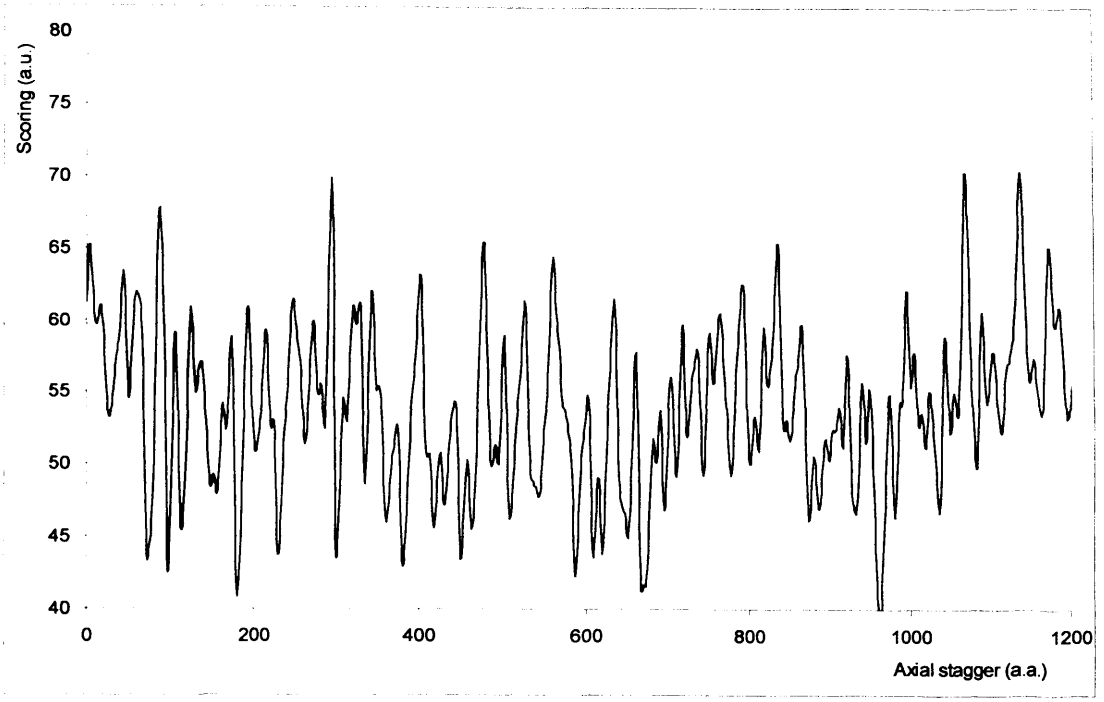


Figure 3.41: Electrostatic scoring function due to one collagen molecule which is positively charged against a multiple sequence of negatively charged molecules in an antiparallel configuration.

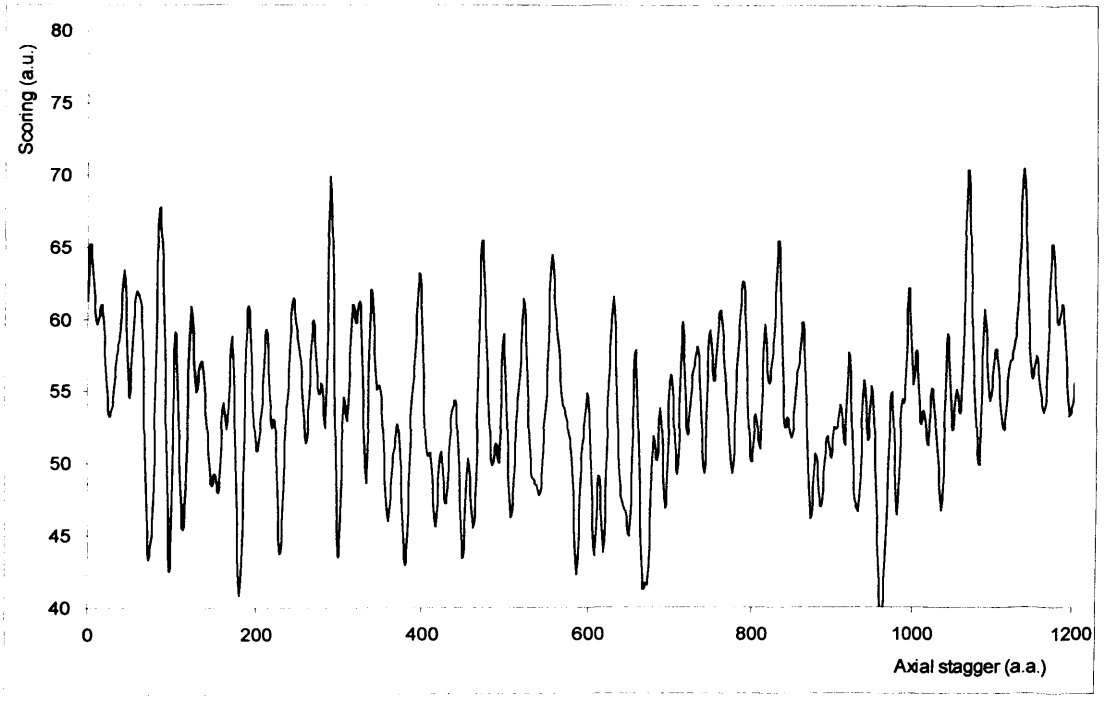


Figure 3.42: Electrostatic scoring function due to one collagen molecule which is negatively charged against a multiple sequence of positively charged molecules in an antiparallel configuration.

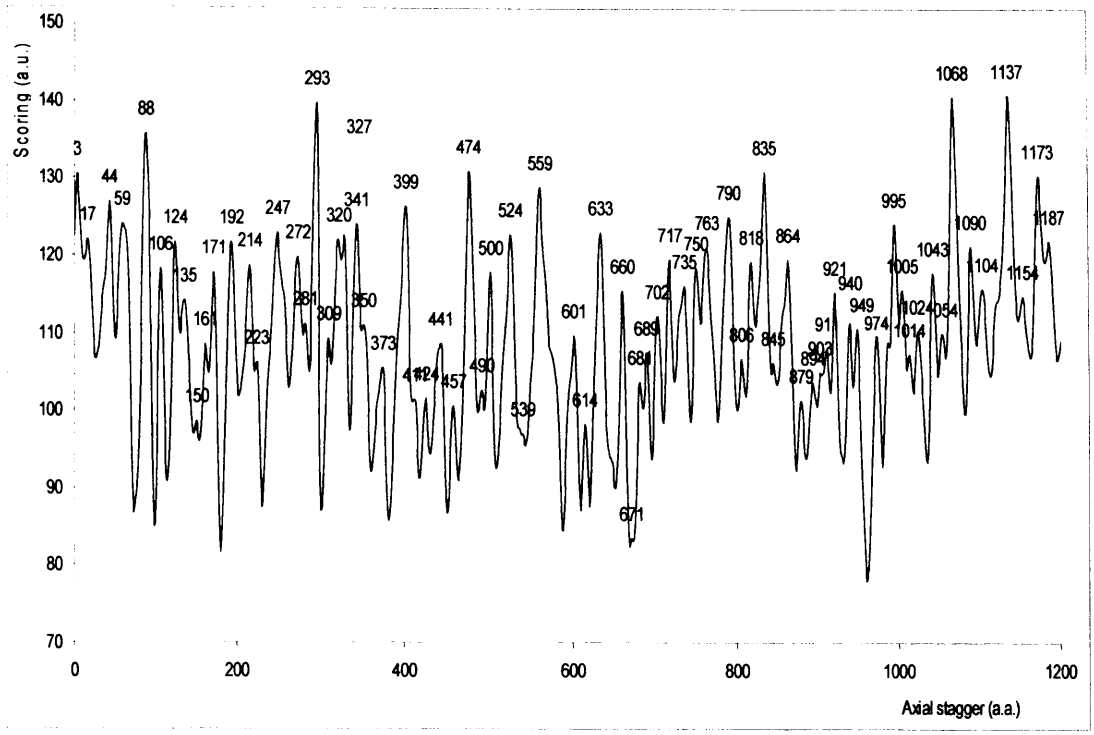


Figure 3.43: Multiple antiparallel attractive electrostatic interaction as a function of the distance.

To find out the axial stagger that is related to the combined action of the hydrophobic and the attractive electrostatic scoring, we perform a point by point multiplication between graph 3.39 and 3.43 and report the result in figure 3.44. Now both peaks at 4 and at 87 a.a. positions are among the most prominent and seem to be corresponding to an acceptable antiparallel stagger between molecules. However, as already pointed out above we reckon that the peak at position 87 is the correct one because of cross-link formation (figure 3.40) and because it allows the formation of a long fibril.

Until now, we have only proposed theoretical models. However, a model must be verified experimentally to confirm its fit. This is the main intent of the next chapter.

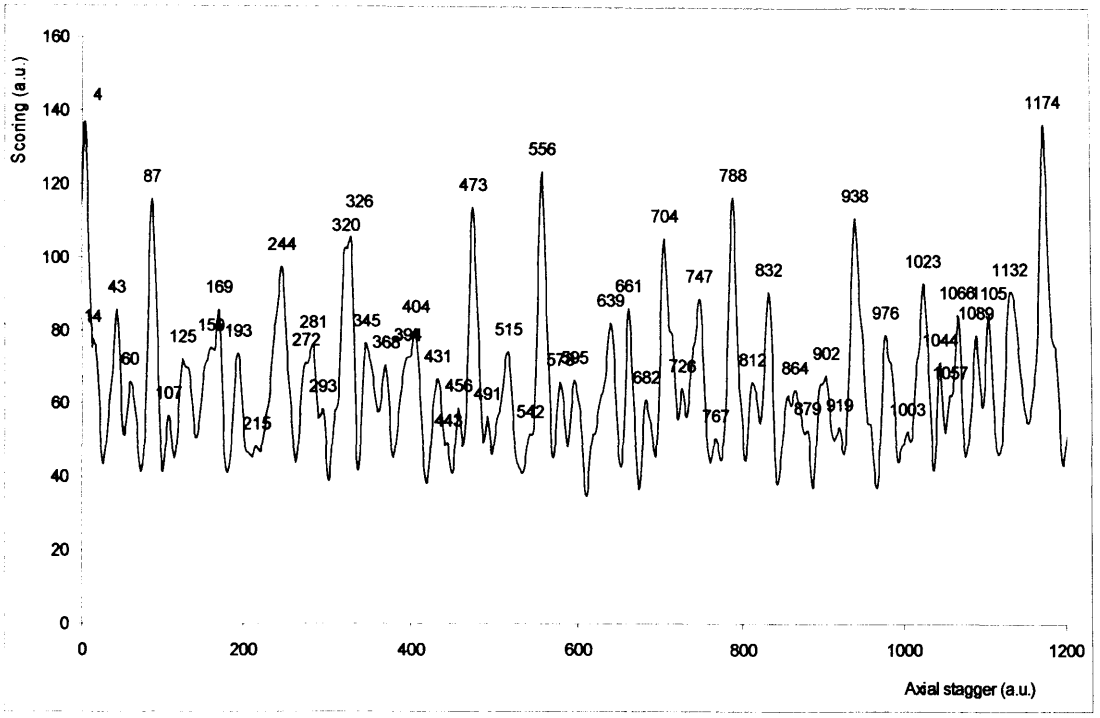


Figure 3.44: Point by point multiplication between the hydrophobic and the attractive electrostatic antiparallel scoring functions for murine type I collagen.

3.9 Discussions about the scoring method

So far, the thesis consisted of the development of a robust methodology necessary to study the interactions between fibrillar collagen molecules. Once the distribution of the hydrophobic and of the charged amino acids along the collagen molecule is known, it is possible to use a cross-correlation function to describe and predict the axial shift encountered in parallel collagen fibrils. Moreover, it can be further extended to predict the most likely axial displacement of collagen molecules in antiparallel aggregations.

3.9.1 Relationship between scoring method and collagen packing

To study the interaction between two collagen molecules we applied the scoring method first proposed by Hulmes and colleagues (Hulmes et al, 1973). In a similar way to what was done by Hulmes, we considered at first only the contribution of the triple helical domain of type I collagen. For Hulmes this was a necessity for lack of a complete set of data that was not available at time. In addition he could use in his work only one α_1(I)-chain, made from sequence fragments of different animal species. Today a complete sequence for murine type I collagen is available so that we could use in our work the information from the three α-chains to study their combined effect during fibril formation.

A first surprising result, obtained when we were using a combination of the three α-chains, consisted of the fact that the most prominent peak in the hydrophobic scoring function of two type I collagen triple helices was not at position 234 a.a., as should be expected, but at position 84 a.a.. This cast some doubts upon the

effectiveness of the scoring function to locate the interaction energy peaks that can explain collagen axial packing.

A second surprising result consisted of the fact that according to Fourier frequency analysis, the underlying periodicity in the hydrophobic scoring function was about 21 or 39 a.a. (figure 3.3) in contrast with what seen in the scoring function in real fibres that had a dominant periodicity of 234 a.a. with major peaks at 0, 234, 465, 700 and 932 a.a.. In fact, the peaks in the scoring function at positions 0, 234, 465, 700 and 932 a.a. are prominent but they are not isolated, so that they can not be dominant in the Fourier transform (figure 3.2).

It was thus necessary to apply some refinements to the scoring method. Gaussian smoothing (σ=3) and the normalisation of the scoring function was a first step. This resulted in an enhancement of the peaks that revealed a more marked 234 a.a. periodic structure It was thus possible to see that all peaks in the scoring function were located in a mirror-symmetric way with respect to planes situated at positions 117, 349, 583 and 817 a.a.. In addition, each mirror-symmetric peak was related to another mirror-symmetric peak 234 a.a. away situated in an adjacent region. In this way, we found that the hydrophobic scoring function was made of very similar regions where the peaks were distributed in a similar manner. For this reason, it was possible to think of the hydrophobic scoring function as made of several repeats. These observations suggested the use of the convolution theorem to explain the apparent inconsistency between Fourier analysis of the scoring function and the real periodicity linking one peak to another. In fact, a periodic structure such that of the hydrophobic scoring function of the type I collagen, can be thought of as the convolution of the unit repeat with a δ-comb. The convolution theorem states that the Fourier transform of two convoluted functions is equal to the product of the

Fourier transform of the two functions. This theorem allowed us to explain why there is no peak at 234 a.a. in the frequency domain of the hydrophobic scoring function (figure 3.11). In fact, as we saw in paragraph 3.2, the Fourier transform of the unit repeat between position 0 and 234 a.a., is such to have a minimum at 234 a.a. that cancels out every other peak to be found in such position. Fourier analysis combined with the convolution theorem can be used also to explain why the frequency spectrum of the unit repeat has a minimum at 234 a.a. shift. In fact (paragraph 3.2), the unit repeat can be thought of as the convolution of half of the unit repeat with two δ-functions 117 a.a. apart. However, since the Fourier transform of the two δ-functions is equal to $2\cos(234\pi\nu)$ we have that the Fourier transform of the motif is zero when $\frac{1}{\nu}$ is $\frac{2 \bullet 234}{(1+2n)}$ with n∈ℕ. This helps to explain why it is the unit period between 0 and 234 a.a. that constitutes the real periodicity of the hydrophobic scoring and why its spectral analysis does not present any peak at 234 a.a.

The hydrophobic scoring between two molecules is a good starting point to understand collagen fibril formation.

It is however necessary to take another step toward a more realistic representation of what happens in real fibrils. This is done by building a multiple sequence of molecules each separated by a gap G = (5*234 –N) a.a., where N is the number of amino acids in one collagen molecule. The study of the hydrophobic scoring of a single collagen molecule with a multiple sequence should be thought of as the interaction forces felt by a molecule during "fibrillogenesis" because it takes into account the contribution of several molecules at one time. Unsurprisingly the hydrophobic scoring obtained in this way is very similar to the normalised scoring

function between two molecules and it is made of five regions with a common repeat. A minor alteration of the common repeat is found in the central region between peaks at positions 467 and 703 a.a. (figure 3.16). In fact, in this case there is no double central peak that is clearly seen in the other regions so that the periodic nature of the scoring function is broken there.

However, this difference is taken care of by a more realistic representation of the collagen molecule. In fact, our model is partially limited by the fact that we have been considering only the triple helical domain of type I collagen. However, it is necessary to consider the contribution from the telopeptides. About the 26% of the residues making up the telopeptides is made of voluminous hydrophobic amino acids such as methionine, phenylalanine, tyrosine, and valine. This is particularly remarkable if compared with the percentage of hydrophobic residues in the triple helix, only about 8%.

We have thus considered two situations one with unfolded telopeptides and one with telopeptides folded in a way that facilitated hydroxylysines cross-links. Incidentally, as seen in figure 3.18 and 3.19 this configuration is such that the telopeptides hydrophobic residues face each other to maximise the hydrophobic interaction inside the telopeptide domains.

Qualitatively both conformations give rise to a scoring function that is made of five similar regions. In this case, the central region shows a pair of peaks that are very similar to those found in neighbouring regions. The quantitative result is however the most important. In fact, if we consider figure 3.16 for the hydrophobic scoring function of the triple helix, the peak in position 233 a.a. does not appear to be the most prominent, being only 2% higher than the peak at position 84 a.a. The situation is improved when the molecules include unfolded telopeptides where the

peak in position 234 a.a. is 19% higher than the one at 84 a.a. (figure 3.20). Even better (figure 3.21), the scoring with folded telopeptides shows a more prominent peak at position 234 a.a., 20% higher than the second most prominent peak, which is now in position 149 a.a..

To conclude, the scoring of molecules with folded telopeptides has a peak in position 234 a.a. that is about 12% and 60 % higher than the corresponding peaks in the scoring function with unfolded and simple triple helical configurations respectively. This shows how important the contribution of the telopeptides is in collagen packing and that without telopeptides, the 234 a.a. shift would not probably be there, since the interactions energies would not be greatest at that relative displacement.

It is therefore reasonable to affirm that the contribution of the telopeptides is fundamental for the hydrophobic attraction between molecules because it forms periodic regions in the scoring function that are more neatly defined, and because it increases the intensity of the peak at position 234 a.a. making it the ideal axial shift between molecules. In particular, the folded configuration of the telopeptides is the preferred one.

3.9.2 A model representing the hydrophobic amino acids distribution

The repeat of five identical regions is the most prominent feature of the hydrophobic scoring. Since the scoring function is a correlation function, it is natural to think that those amino acids that contribute the most to it are distributed with the same periodicity (234 a.a.). This means that in the type I collagen molecule

there should be hydrophobic residues 234 a.a. apart. Even though with small discrepancies it is possible to see that the hydrophobic amino acids are generally clustered according to the {42, 42, 42, 42, 66}*4+{42, 42} distribution.

Even though this distribution does not cover all hydrophobic residues in the molecule, it is a good representation because its calculated hydrophobic scoring function is very similar to the function calculated from the real distribution as shown in figure 3.24 where the correlation coefficient for the two curves is 0.84. This means that the model distribution incorporates all the essential amino acids found in the real distribution.

Probably, the 42 a.a. distance between hydrophobic residues is such that it allows a hydrophobic interaction that is the most suited to keep the molecules together. A larger distance could give rise to an overall hydrophobic interaction that is too weak for collagen packing. Conversely, a shorter distance would not allow a sufficient number of charged amino acids in between the hydrophobic ones with a decrease in the intensity of the attractive forces between molecules. In addition, in this case, prolines and hydroxyprolines could be too few to make the collagen molecule rigid enough.

Using the model distribution, it is possible to highlight those hydrophobic residues that are the most likely responsible for collagen aggregation during the fibrillogenesis. These are methionine, phenylalanine and tyrosine (figure 3.27). This is not surprising because these are highly hydrophobic amino acids with big and mobile side chains, it is also interesting to notice that tyrosine is concentrated in the telopeptides where it appears to have a fundamental role for telopeptide folding. It also makes big hydrophobic clusters distributed according to the model distribution.

The modelled distribution can be read from left to right or from right to left still giving the same residue profile. In other words if we consider a molecule oriented form the N- to the C-terminal or from the C-to the N-terminal we find the same hydrophobic amino acids distribution bar an offset (i.e. {42,42} in figure 3.32). This is consistent with the hydrophobic residues in type I collagen being distributed in a way that was meant to form parallel and also antiparallel aggregates, even though antiparallel aggregates are found very rarely *in vivo* (Mallinger et al., 1992). Probably type I collagen has evolved towards parallel molecular associations.

In addition, a possible explanation of the periodic structure of the hydrophobic residue distribution could be obtained if we hypothesise that, during the evolution, type I collagen was assembled by gene duplication. The fact that the hydrophobic amino acids do not repeat exactly every 234 a.a., could be due to differentiation that happened with time.

3.9.3 Electrostatic contribution to collagen fibril formation

In this thesis, we assumed that the attractive electrostatic interaction is much greater than the repulsive one because of the "shielding effect" of the medium. Thus, we have considered only an attractive electrostatic scoring function. We found that it is characterised by periodic regions, where peaks are mirror-symmetric with respect to a central plane, in a way similar, even though not identical, to that found for the hydrophobic interactions (figure 3.30). The most important feature is that the peaks which are integer multiples of 234 a.a. are among the most prominent which reinforce the effect of the hydrophobic amino acids. It seems thus reasonable to think that both kind of interactions work together during molecular packing. If we

combine the effect of both interactions as a point by point multiplication of the scoring functions as shown in figure 3.31 the peak at position 234 a.a. is greatly enhanced, showing that both types of interactions act to favour a 234 a.a. stagger aggregation.

3.9.4 Observations relative to the antiparallel aggregation

As already seen above, collagen molecules can interact in an antiparallel fashion. This is seen *in vivo* (Mallinger et al, 1992) and *in vitro* (Bruns, 1976; Williams et al, 1978). This infrequent packing configuration can be predicted with our scoring method applied to two antiparallel molecules. The results obtained for hydrophobic (figure 3.39) and attractive electrostatic scoring (figure 3.43) highlighted that a shift of about 87 a.a. is, the ideal axial displacement between molecules in this kind of packing (figure 3.44). This displacement allows the formation of cross-links between the hydroxylysines of the C-terminal domain of a molecule and the hydroxylysines at position 87 on the triple helix of another molecule (figure 3.38). In figure 3.41 we show an idealised diagram of how the molecules interact together as described in paragraph 3.7.

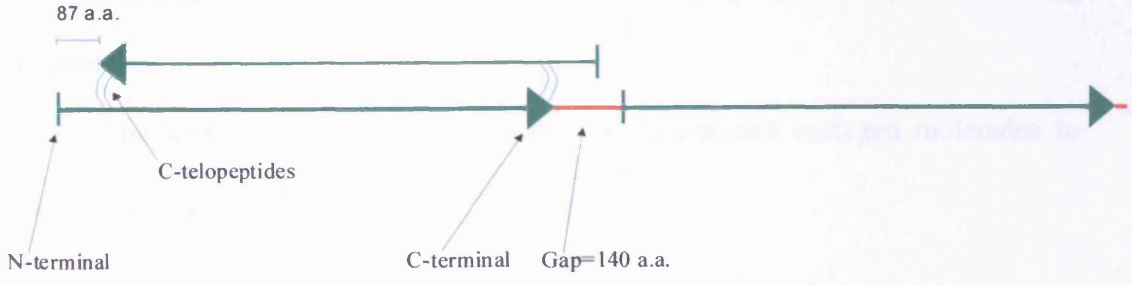


Figure 3.45: Diagram representing the stagger for the antiparallel aggregation obtained when a single collagen molecule is scored against a multiple sequence of molecules.

3.10 Concluding remarks for chapter 3

Chapter 3 was substantially theoretical. It was based on the study of the type I collagen amino acids distribution and on their interactions. It also included the creation of a method to explain amino acids interactions and to predict collagen packing. The major results can be summarized as follows.

1. The hydrophobic scoring function is made of identical regions with a 234 a.a. periodicity
2. The contribution of the folded telopeptides is fundamental to enhance the peak at position 234 a.a. in the scoring function. 234 a.a. is therefore the prefixed axial shift between two molecules
3. Hydrophobic residue clusters are equally spaced (≈42 a.a.). This imply that antiparallel packing can arise
4. Type I collagen could be the results of genetic duplication
5. The role of tyrosine, methionine and phenylalanine is fundamental for collagen fibril formation
6. The attractive electrostatic scoring function is characterised by periodic mirror-like regions whose peaks correspond to those in the hydrophobic scoring function
7. The scoring method can be applied to antiparallel collagen molecules to predict their packing.

However, predictions must be validated. This is the main intent in the next chapter.

CHAPTER 4

VALIDATING THE FIBRILLAR COLLAGEN MODELS

4 Introduction

Our scoring method allowed us to study the modes of interaction between two or more type I collagen molecules during fibre formation. It could explain for example, why a particular packing configuration was adopted during the formation of fibrils where all molecules were parallel with respect to each other. This scoring method was also employed to predict how type I collagen molecules interacted in an antiparallel configuration.

Until now, no systematic studies were carried out on antiparallel fibrillar packing to understand the forces driving it. In this chapter, we use the scoring method to make predictions on parallel and antiparallel fibril formation. To test the validity of our approach and of our deductions, we carried out comparisons between electron micrograph of collagen microfibrils recorded experimentally and our predictions.

Collagen fibrils were created *in vitro* from mixtures of purified collagen molecules (paragraph 2.10). The type II and type XI collagen preparations used were obtained from rat chondrosarcoma, while type III collagen was obtained from rabbit skin. All fibrils were prepared adjusting temperature and pH by Dr A. Vaughan-Thomas in the laboratory of Professor V. Duance at Cardiff University. They were double stained using uranyl acetate (UA) and phosphotungstic acid (PTA) and imaged using a Philips EM 208 electron microscope by Dr Rob Young.

We used ImageJ 1.37v (Rasband, 1997-2007) to measure the density profiles of the microfibrils in the micrographs. Most density profiles were subsequently filtered in Fourier space to reduce noise and enhance all the periodic components. Once the density profiles for the microfibrils were obtained, they were compared with a simulated density profile from a model of the microfibril.

Simulated density profiles were obtained from a profile of the charged a.a. of the models built according to the results obtained with our scoring method. The "theoretical" staining of our models was obtained using a method first pioneered by Keith Meek and colleagues (Meek et al. 1979, Chapman et al 1981) and subsequently improved by other authors (Ortolani et al., 2000). It is based on the fact that UA and PTA heavy metal atoms form clusters around charged amino acids in proteins. Therefore, it is possible to assign a value to every charged residue in a collagen molecule and to represent the molecule itself as a linear sequence of these values. Their linear plot can be thought of as the projection onto the molecule axis of UA and PTA stain since, effectively, only charged amino acids are stained. The linear plot can be compared with the density profile of the microfibrils to verify directly our models (paragraph 2.9).

4.1 Modelling collagen microfibrils

We used the values for the relative axial stagger found in chapter 3 by applying the scoring method to build collagen microfibrils. These microfibrils were compared to real fibrils. A single fibril forming collagen molecule can be thought of as three linear parallel α-chains that are staggered by one amino acid with respect to each other. Fibrillar collagens can be homotrimeric or heterotrimeric. In this chapter, we analyse type I, II, III and XI collagens. Type II and III collagens are homotrimers so that, the relative displacement assigned to the α-chains while building a linear representation of the collagen molecule is not important. However, type I and type XI collagen are heterotrimers and the relative displacements used for their α-chains must be specified. Type I collagen was dealt with in chapter 3, while for type XI

collagen we used the configuration α_3-α_2-α_1 as shown in figure 4.1, where we represent the N-terminal portion of the type XI collagen triple helical domain. As for type I collagen, the other combinations give results that are similar, however this combination has a scoring function that is neater with well defined peaks with a 234 a.a. periodicity.

Since complete sequences for type XI rat collagen and for type III rabbit collagen were not available in the Protein Data Bank (PDB), we used in our analysis mouse collagen sequences to build all our collagen molecule models. Most mutations are conservative in these collagens so we predict that this approach introduces only a small error.

Type I murine collagen α_1(I) and α_2(I) chain sequences were obtained from PDB (primary access number: P11087 and Q01149 respectively). The length of the triple helical domain is 1014 amino acids for both chains. The length of the N-telopeptides and the C-telopeptides is 17 and 26 amino acids for the α_1(I) chain while it is 11 and 15 amino acids for the α_2(I) chain (Dalgleish, 1997).

The amino acid sequence data used for type II murine collagen α_1(II)-chain sequence was obtained from PDB (primary access number: P28481).

The triple helical domain is 1014 residues long, while the N-telopeptide and the C-telopeptide are 19 and 27 residues long respectively (Metsäranta et al., 1991).

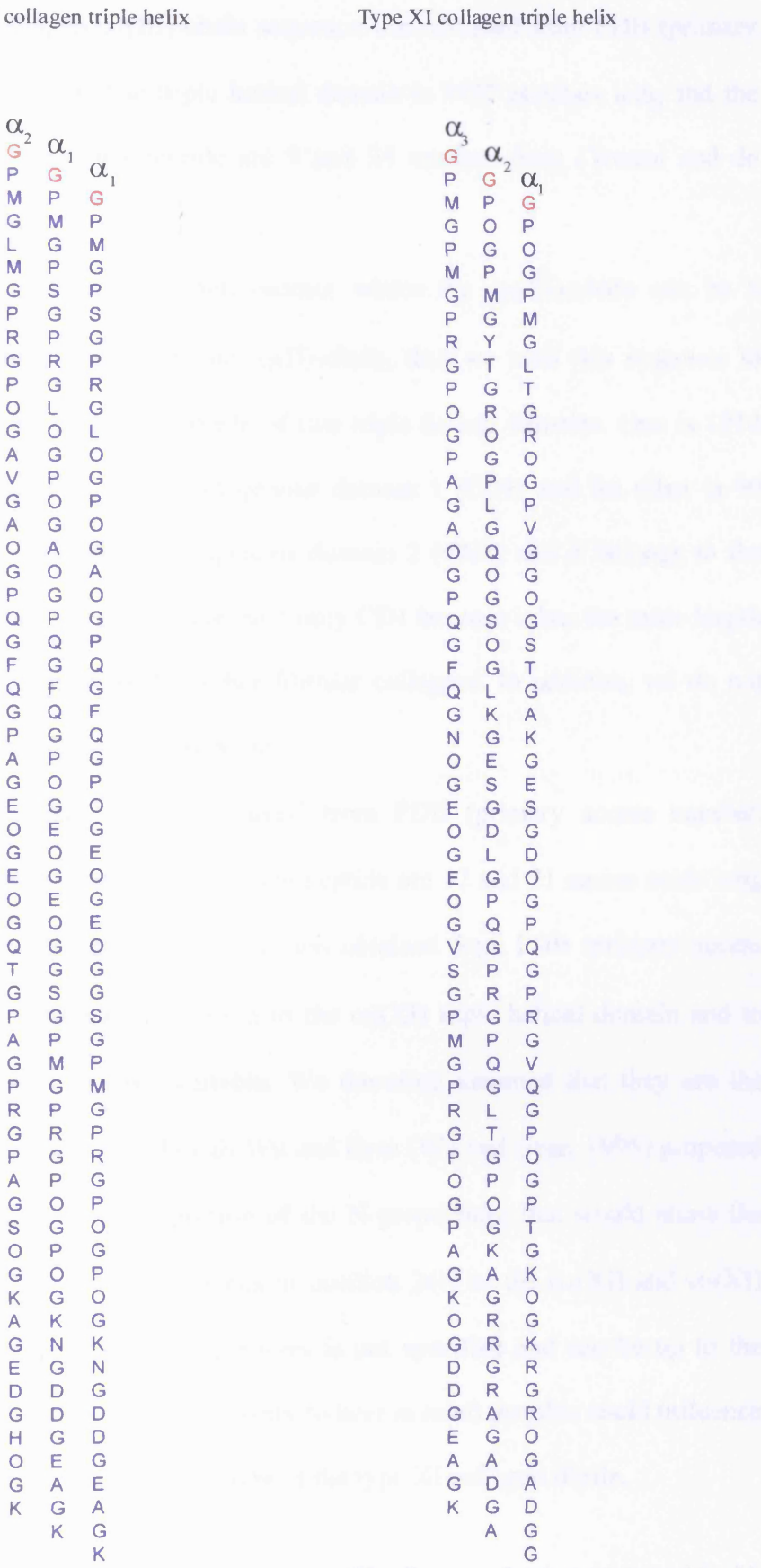


Figure 4.1 Scheme representing the type I and type XI collagen molecule models used in this chapter.

Type III murine collagen α_1(III)-chain sequence was obtained from PDB (primary access number: P08121). The triple helical domain is 1032 residues long and the N-telopeptide and the C-telopeptide are 9 and 24 residues long (Toman and de Combrugghe, 1994).

Type XI murine collagen is a heterotrimer where the α_3(XI)-chain can be a post-translational modification of the α_1(II)-chain, thus we used this sequence to represent it. Type XI collagen is made of two triple helical domains. One is 1014 amino acids long, and is called collagenous domain 1 (CD1) and the other is 90 amino acids long, it is called collagenous domain 2 (CD2) and it belongs to the N-terminal propeptide. Here we have used only CD1 because it has the same length of the triple helix of most of the other fibrillar collagens. In addition, we do not know how CD2 can fold on the molecule.

The α_1(XI)-chain sequence was obtained from PDB (primary access number: Q61245). The N-telopeptide and the C-telopeptide are 17 and 21 amino acids long respectively. The α_2(XI)-chain sequence was obtained from PDB (primary access number: Q64739). Information relating to the α_2(XI) triple helical domain and to the telopeptides length is not available. We therefore assumed that they are the same as those of α_1(XI), even though Wu and Eyre (Wu and Eyre, 1995) proposed that type XI collagen retains a portion of the N-propeptides that would allow the formation of cross-links via the lysines in position 24N of the α_1(XI) and α_2(XI) chains. The dimension of such propeptides is not specified and can be up to the 25% of the total chain mass. It is necessary to bear in mind that this could influence in a considerable way the staining feature of the type XI collagen fibrils.

4.2 Comparing experimental data and a model of an antiparallel fibril aggregation of type I collagen

We saw in chapter three how two type I collagen molecules could interact during parallel aggregation. We found that some axial displacements appear to be favoured because of high values for the scoring function and because of cross-links formation. We can use the same approach to study two interacting microfibrils in an antiparallel configuration. For example, to study hydrophobic interactions between two microfibrils, we created first the linear sequences of numerical values that represent the collagen α-chains (figure 4.2), then we assigned the value 1 at the positions of every major hydrophobic amino acid (F, I, L, M, V, Y) in the microfibril while we assigned the value 0 to the other residues. A linear sequence whose every element is made of the axial sum of the non zero representing the hydrophobic amino acids at the same level from the three α-chains is finally built. This linear sequence is used to calculate the scoring of the two antiparallel microfibrils. The attractive electrostatic scoring function is calculated in the same way but in this case, only charged amino acids are considered.

The results from the previous chapter were used to build a microfibril made of five collagen molecules that are staggered by 234 amino acids with respect to each other (paragraph 3.4). The telopeptides were folded in such a way that they would be able to cross-link according to the findings of paragraph 3.7. The hydrophobic and electrostatic scorings were calculated for two such microfibrils in antiparallel configuration.

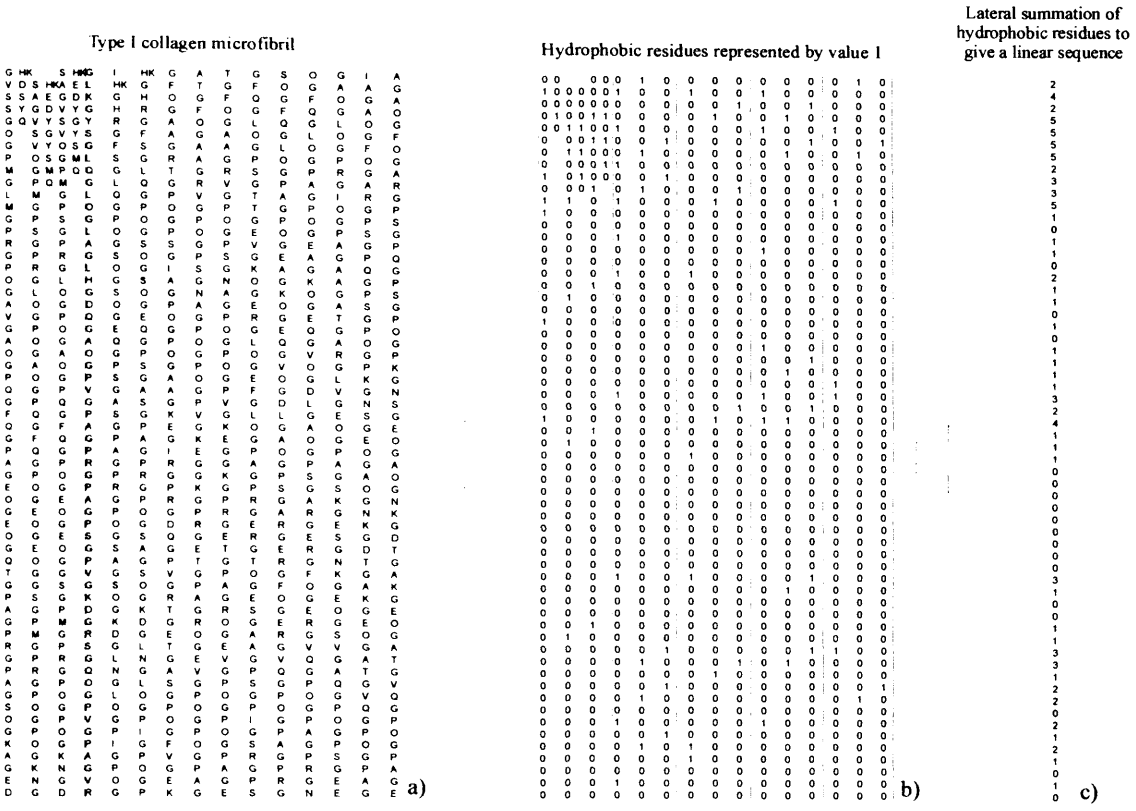


Figure 4.2: a) portion of type I collagen microfibril. Its hydrophobic residues are represented by the value 1 in b) and the lateral sum of the hydrophobic values in c).

Figures 4.3 and 4.4 represent the graphs of the hydrophobic and attractive electrostatic scorings for two antiparallel microfibrils of murine type I collagen as a function of their mutual displacement. Since microfibrils are made of molecules with 234 a.a. staggering periodicity, both graphs have this periodicity built in.

Graph 4.3 for the hydrophobic scoring function of antiparallel microfibrils presents three peaks at positions 97, 138 and 180 a.a. while the same graph calculated for the molecules in antiparallel configuration in chapter 3 (figure 3.39) has the same peaks at positions 4, 43 and 86 a.a.. This apparently incongruence is due to the way we represented the microfibrils that, when inverted, include the gap.

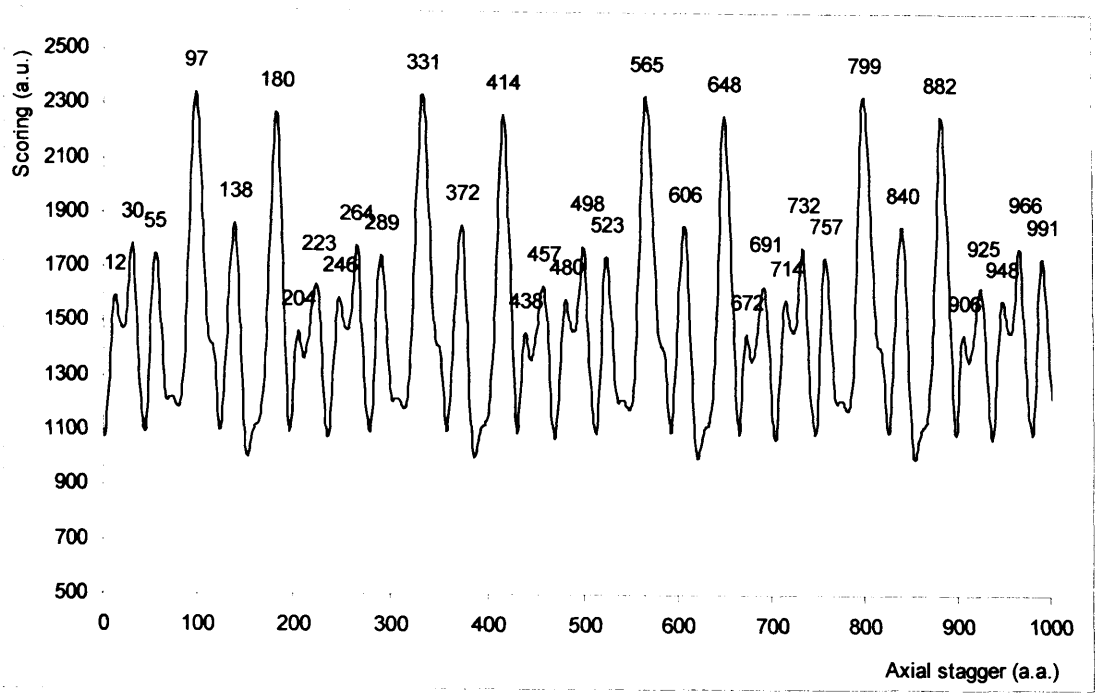


Figure 4.3: Antiparallel hydrophobic scoring function for two type I murine collagen modelled microfibrils as a function of the distance in a.a.

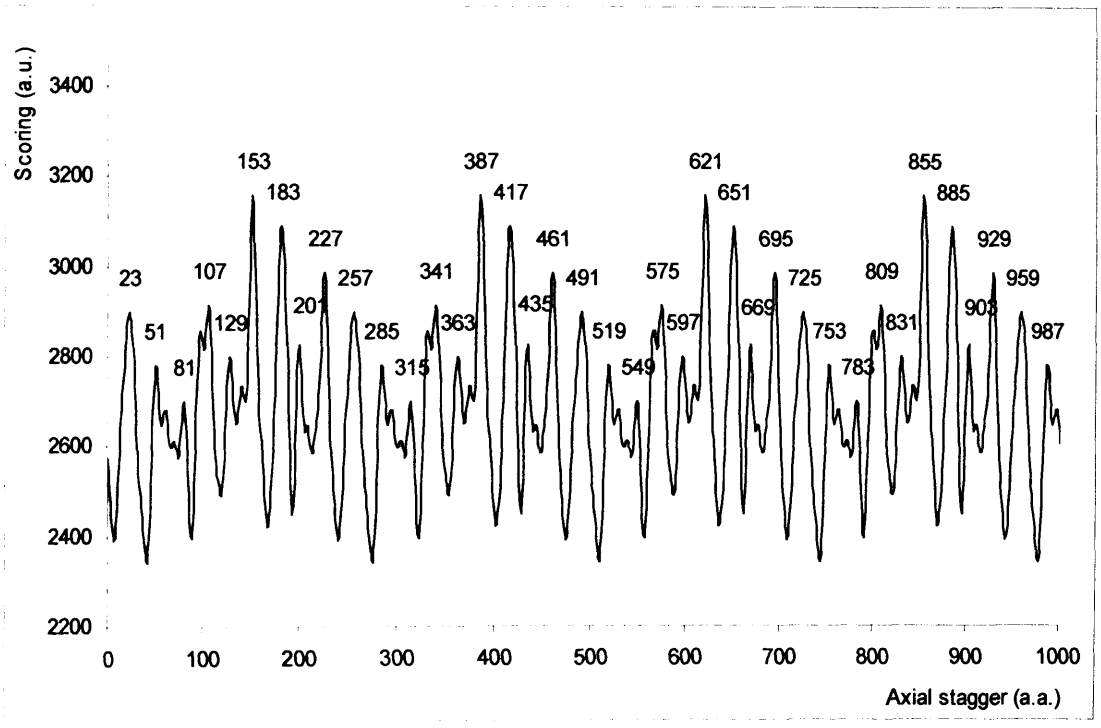


Figure 4.4: Antiparallel attractive electrostatic scoring function for two type I murine collagen modelled microfibrils as a function of the distance in a.a.

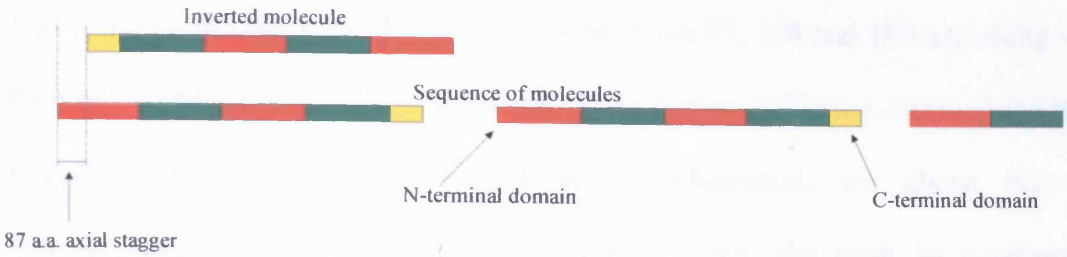


Figure 4.5: Diagrammatic representation of an inverted molecule with an 87 a.a. axial stagger with respect to a forward sequence of molecules

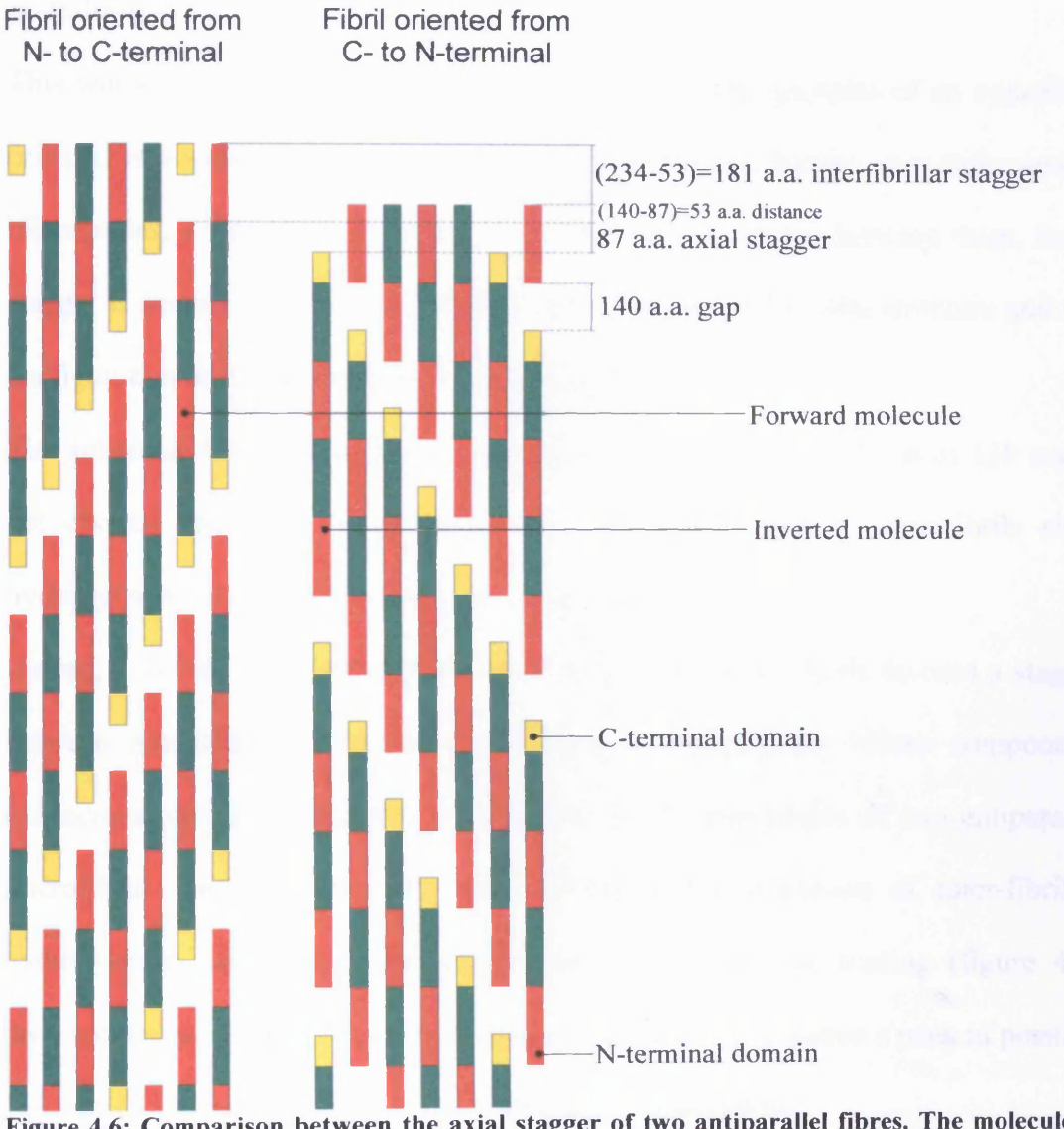


Figure 4.6: Comparison between the axial stagger of two antiparallel fibres. The molecule in the inverted fibril has an 87 a.a. stagger with respect to the corresponding molecule in the forward sequence, to compare the staggering displacement between molecules as found in paragraph 3.8 and the staggering between fibrils. This means that the distance of the forward molecule from the first amino acid of the inverted microfibril is (140-87)=53 a.a.. Since the periodic subunit of the microfibril is 234 a.a. long, the distance between the inverted microfibril and the forward microfibril is (234-53)= 181 a.a. long. The same approach can be used to find the correspondence between the peaks at 4 and 43 a.a. positions in graph 3.39 and the peaks at 97 and 138 a.a. positions in graph 4.3.

Figure 4.3 shows three dominant peaks in positions 97, 138 and 180 a.a. along with their related peaks that are shifted by integer multiples of 234 a.a.. Among the three possible candidates for the interfibrillar displacement, we chose the one corresponding to the peak in position 180. In fact, the peak in position 97 corresponds to a displacement in which the molecules in a microfibril are positioned exactly in register with the molecules of the oppositely oriented microfibril.

This would mean that the N-telopeptides face the C-telopeptides of an oppositely oriented molecule. Even though this configuration can favour cross-links among telopeptides, it would force the molecules to have no stagger between them. Some stagger is needed between the building molecules to form a solid structure and this configuration does not satisfy this requirement.

The antiparallel configuration corresponding to the peak in position at 138 a.a. is not chosen since cross-links could not be formed between microfibrils since hydroxylysine amino acids would be too far apart.

Instead, a 180 a.a. relative shift between antiparallel microfibrils favours a stagger between molecules that is conducive to a solid structure whose components (molecules) are not in register. In addition, the C-telopeptides of two antiparallel microfibrils are close to each other favouring the formation of inter-fibrillar hydroxylysine mediated cross-links. Attractive electrostatic scoring (figure 4.4) does not present prominent peaks however, it is possible to notice a peak in position 183 that reinforces the effect of the hydrophobic interaction.

To highlight the cooperative effect of both types of forces we perform a point by point multiplication between the hydrophobic and the attractive electrostatic scoring

functions. In this way, we intend to represent with an operation similar to the logical operator "AND" the combined effect of the two interactions (figure 4.7).

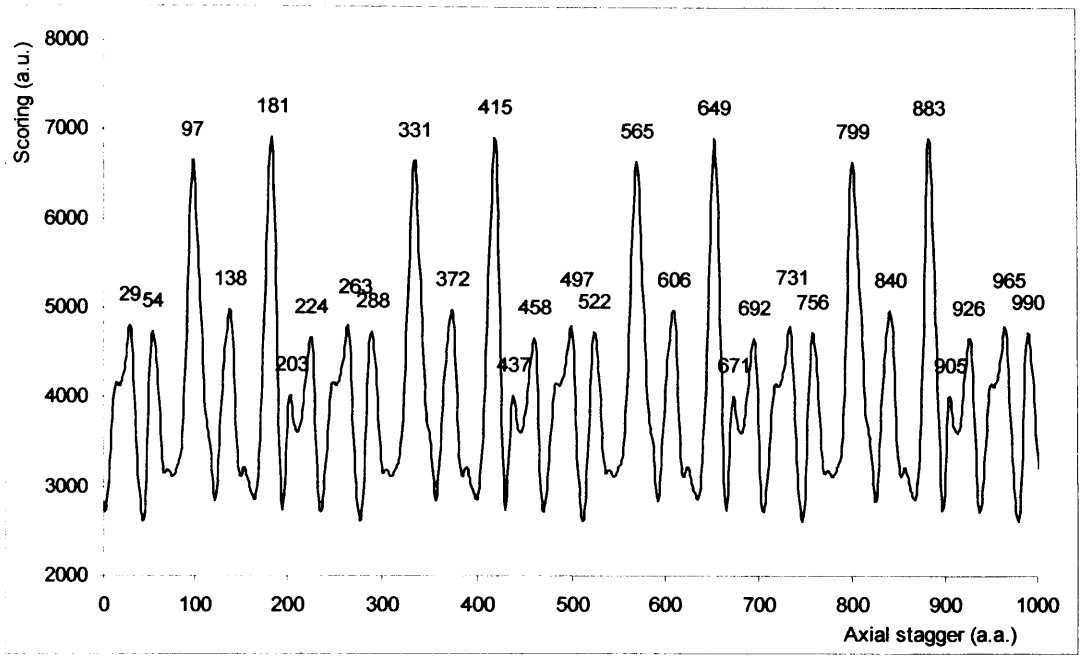


Figure 4.7: Total attractive interaction (scoring function) for two type I collagen antiparallel.

The effect of combining the scoring functions of hydrophobic and electrostatic interactions is that the most prominent peaks of the hydrophobic scoring are enhanced. The peaks in positions 97 and 181 are now the most prominent with the latter being the highest even though for a very small amount (0.4%). The conclusions drawn above are therefore confirmed and we can safely conclude that the combined action of hydrophobic and electrostatic interactions would favour an antiparallel aggregation of microfibrils with about 180 a.a. relative displacement.

In figure 4.8, we show a portion of two antiparallel microfibrils with a 181 a.a. relative stagger where the lysines responsible for possible cross-links are highlighted in red.

Parallel fibril Antiparallel fibril

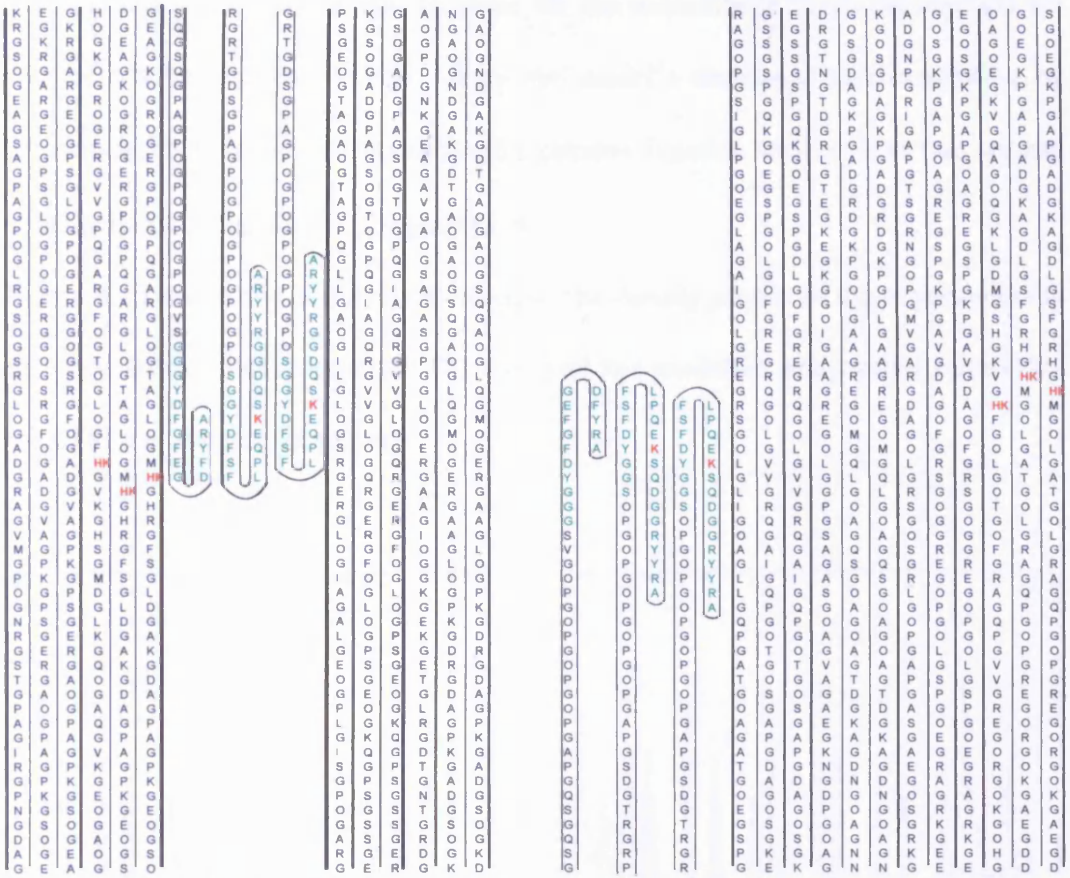


Figure 4.8: Portions of murine type I collagen in an antiparallel configuration with a 181 a.a. stagger. The lysines responsible for cross-link formation are in red. Lysines belonging to the parallel microfibril can form cross-links with lysines of the antiparallel microfibril and *vice versa* because they face each other. Residues in the telopeptides are in green.

To verify the validity of this model for the antiparallel collagen packing, it was compared with the micrograph of a murine type I collagen microfibril oriented in an antiparallel fashion published by Bruns in 1976 (Bruns, 1976) and presented in figure 4.9.

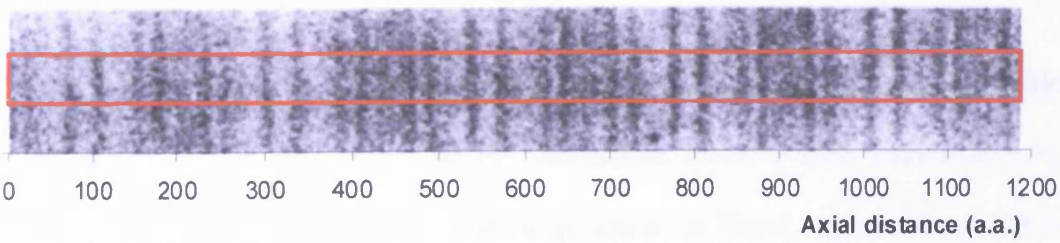


Figure 4.9: Antiparallel aggregation for a type I rat tail collagen microfibril reconstituted in vitro and positively stained with UA. Taken from Bruns, (Bruns, 1976). The outlined portion is the one used for measurements.

Since the simulated density profile allows us to separate every single charged residue, and since this can not be done for the microfibrils in the micrograph for lack of resolution, we need to reduce the model's simulated stain resolution by convoluting it twice with a normalised Gaussian function with $\sigma=3$ so that we can make a direct comparison (paragraph 2.9).

Figure 4.10 shows the comparison between the density profile of the experimentally obtained antiparallel microfibril (in red) and the modelled antiparallel microfibril with 181 a.a. stagger (in green).

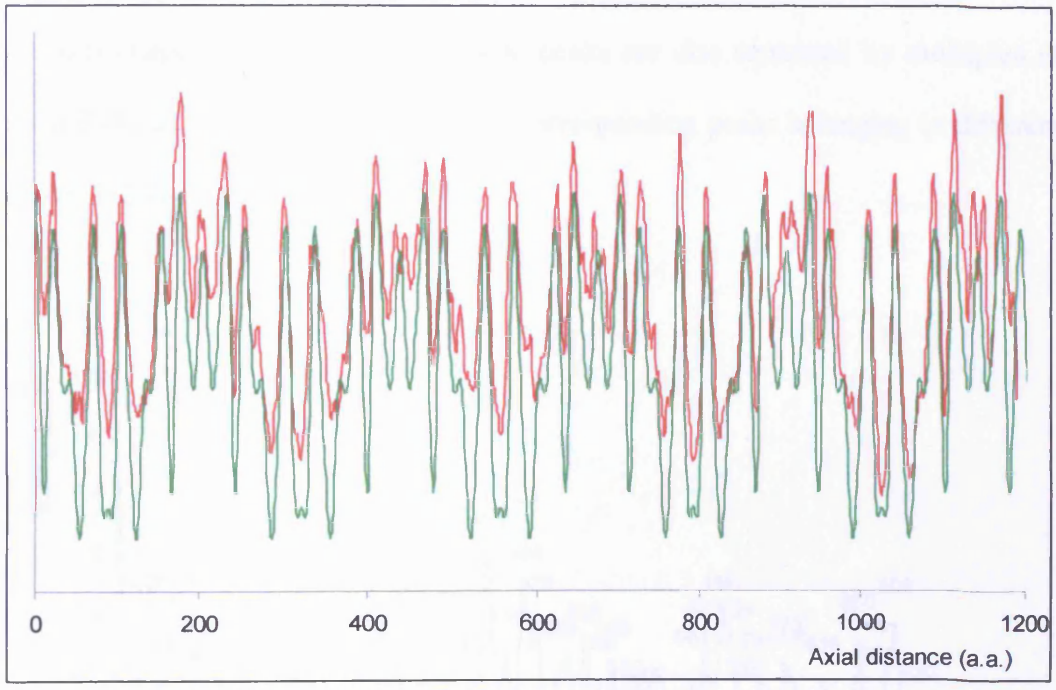


Figure 4.10: Comparison between the density profile of a real type I collagen antiparallel microfibril (in red) and the modelled type I antiparallel collagen microfibril with a 181 a.a. staggering (in green)

Visual inspection of the two graphs immediately reveals how closely they resemble each other. This is confirmed by the correlation factor $r=0.86$ (equation 2.3), obtained applying the cross correlation function in Excel, evaluated on 1370 elements, that confirms the validity of our model.

4.3 Comparison between modelled parallel type II collagen microfibrils and experimentally obtained microfibrils

The scoring method was used to study hydrophobic and electrostatic interactions between two parallel type II collagen molecules without telopeptides. This was done to predict the stagger adopted by type II collagen during molecular packing.

In figure 4.11, the hydrophobic scoring for murine type II collagen is presented. As previously seen for type I collagen, the dominant peaks at positions 0, 234, 464, 698-706 and 932 a.a. are apparent. Even in this case, they represent the boundaries of regions whose peaks respect the mirror-symmetric pattern outlined in the previous chapter (paragraph 3.2). These peaks are also separated by multiples of about 234 a.a.. The average distance of corresponding peaks belonging to different regions is 233.6±2.8 a.a..

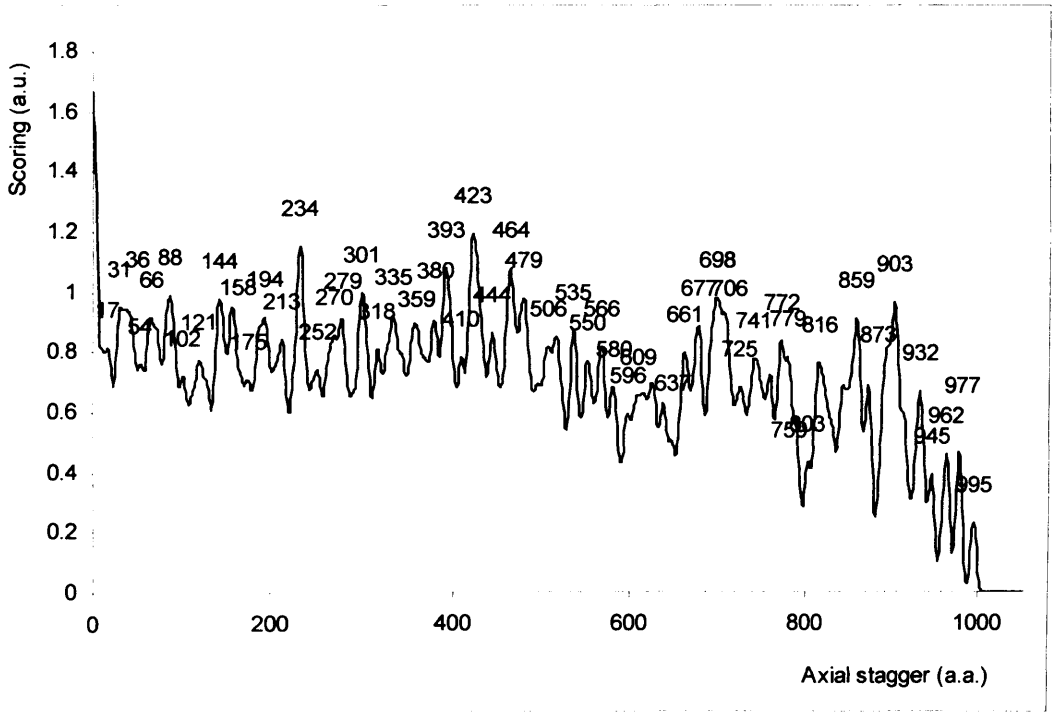


Figure 4.11: Normalised, Gaussian-smoothed, parallel hydrophobic scoring function for two type II collagen molecules without telopeptides

Similarly, the distribution of the most prominent peaks in the electrostatic attractive scoring function (figure 4.12) is very consistent with the hydrophobic peaks shown in figure 4.11.

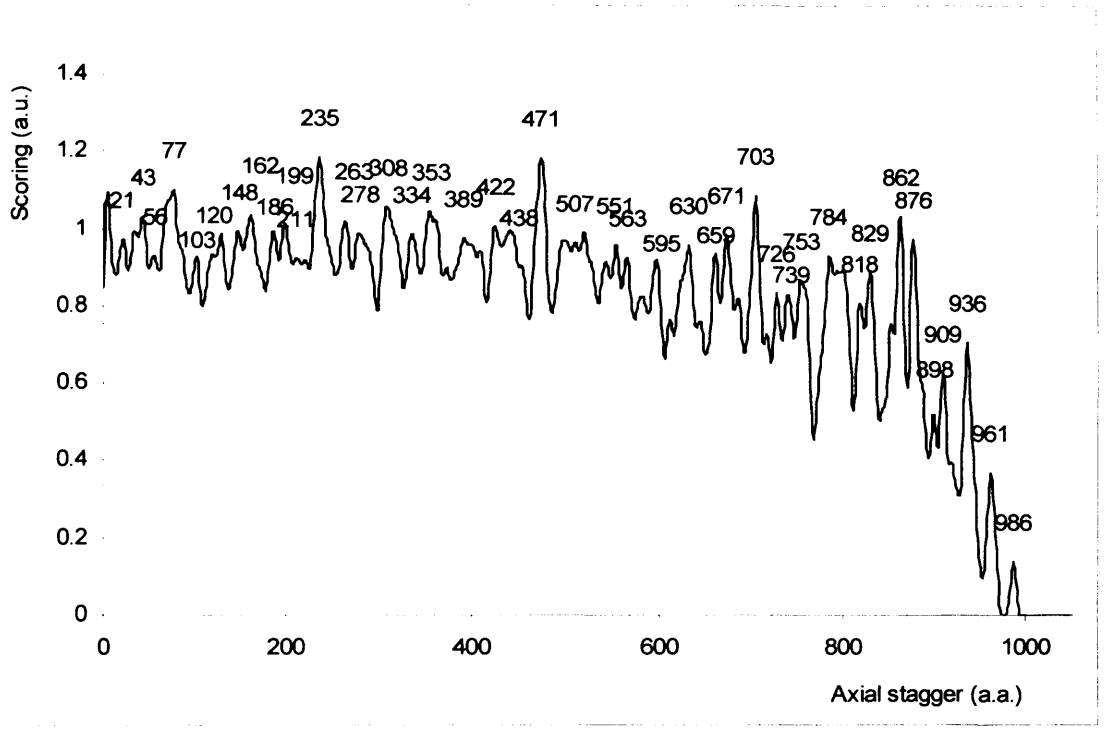


Figure 4.12: Normalised, Gaussian-smoothed, parallel attractive electrostatic scoring function for two type II collagen molecules without telopeptides

The average distance between corresponding peaks is approximated by 234 a.a., and this value is used as the shift distance among five parallel type II collagen molecules that are used to form a microfibril.

Once the shift distance was established, we folded the telopeptides to maximise the number of possible lysine-hydroxylysines cross-links between the telopeptides and the collagen triple helix. In doing this, we made use of the telopeptide glycines and alanines as hinges around which the telopeptides could fold. This is reasonable since they occupy the smallest volume and they are the ideal vertex candidates for hairpin loops.

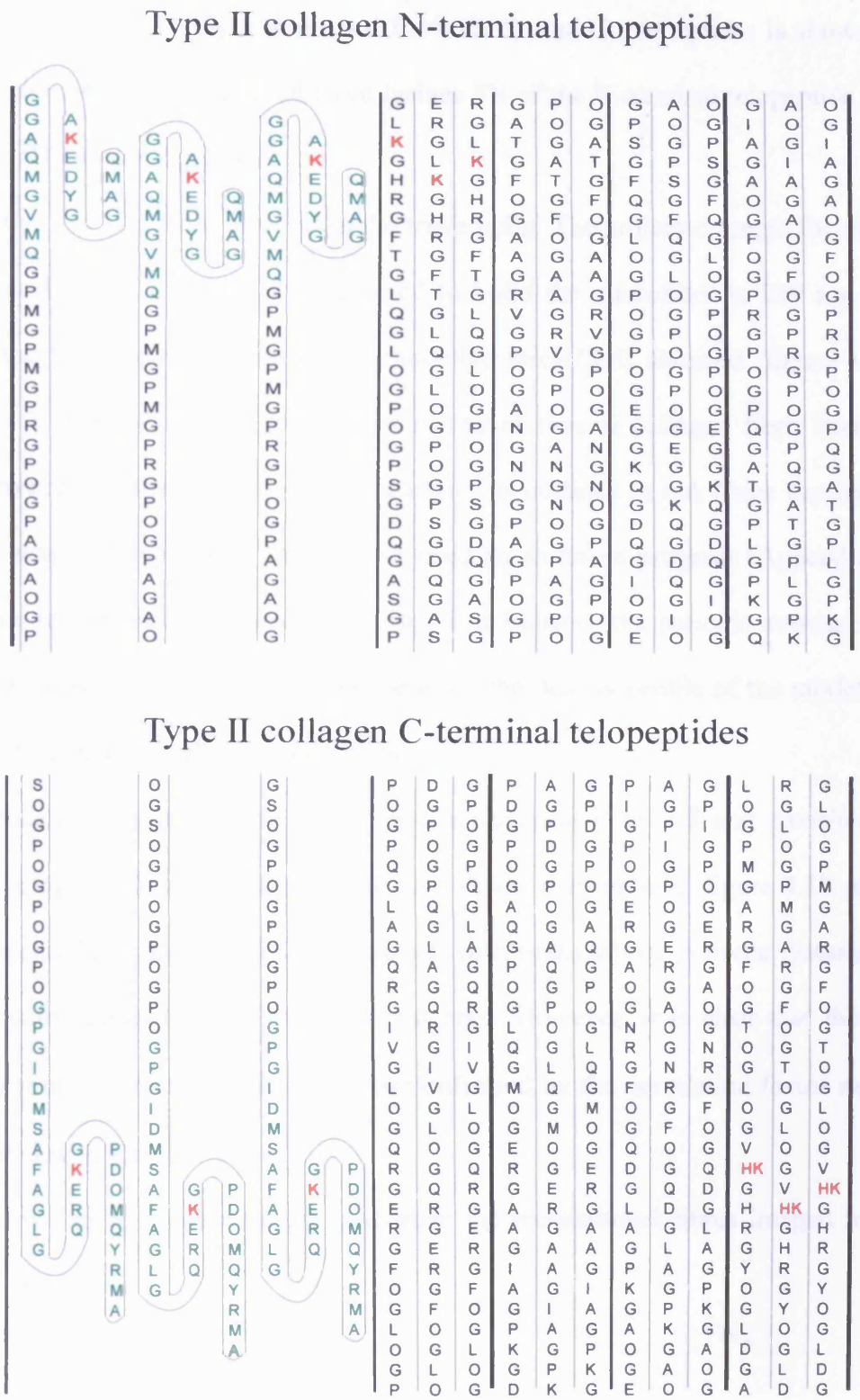


Figure 4.13: Folding configuration used for murine type II collagen telopeptides. Lysines responsible for cross-link formations are coloured in red. Residues belonging to the telopeptides are in green.

In figure 4.13, the configuration used for both N- and C-telopeptides is shown. In this particular configuration, all three lysines 5N of the N-terminal telopeptide face a lysine 930 of the triple helix.

Similarly, Lys 17C faces Hlys 87 in the triple helix. The projected length for such a molecule is 1041 a.a. thus, the gap used to build the microfibril is 129 a.a.. We simulated the staining pattern of the modelled microfibril obtained (figure 4.14). Figures 4.15 shows a portion of a reconstituted type II collagen fibril from rat chondrosarcoma, the portion actually analysed is outlined in red. Once the density profile was calculated with ImageJ, we used an in house program (Appendix) to calculate its Fourier transform and to filter it to enhance the periodic structures. In figure 4.16, we present a comparison between the density profile of the model and that of the experimentally obtained fibril.

The modelled density profile was scaled multiplying it by 1.8 and stretched by 1.218, using Excel, to ease the comparison. Visual inspection of figure 4.17 shows that the simulated density profile is sharper, with peaks whose periodic features are more neatly drawn than the experimental ones. However, it is clear that the two density profiles resemble each other, as confirmed by the correlation factor $r \approx 0.82$ calculated with Excel

This allows us to use our model to explain the reconstituted fibres imaged in the microscope.

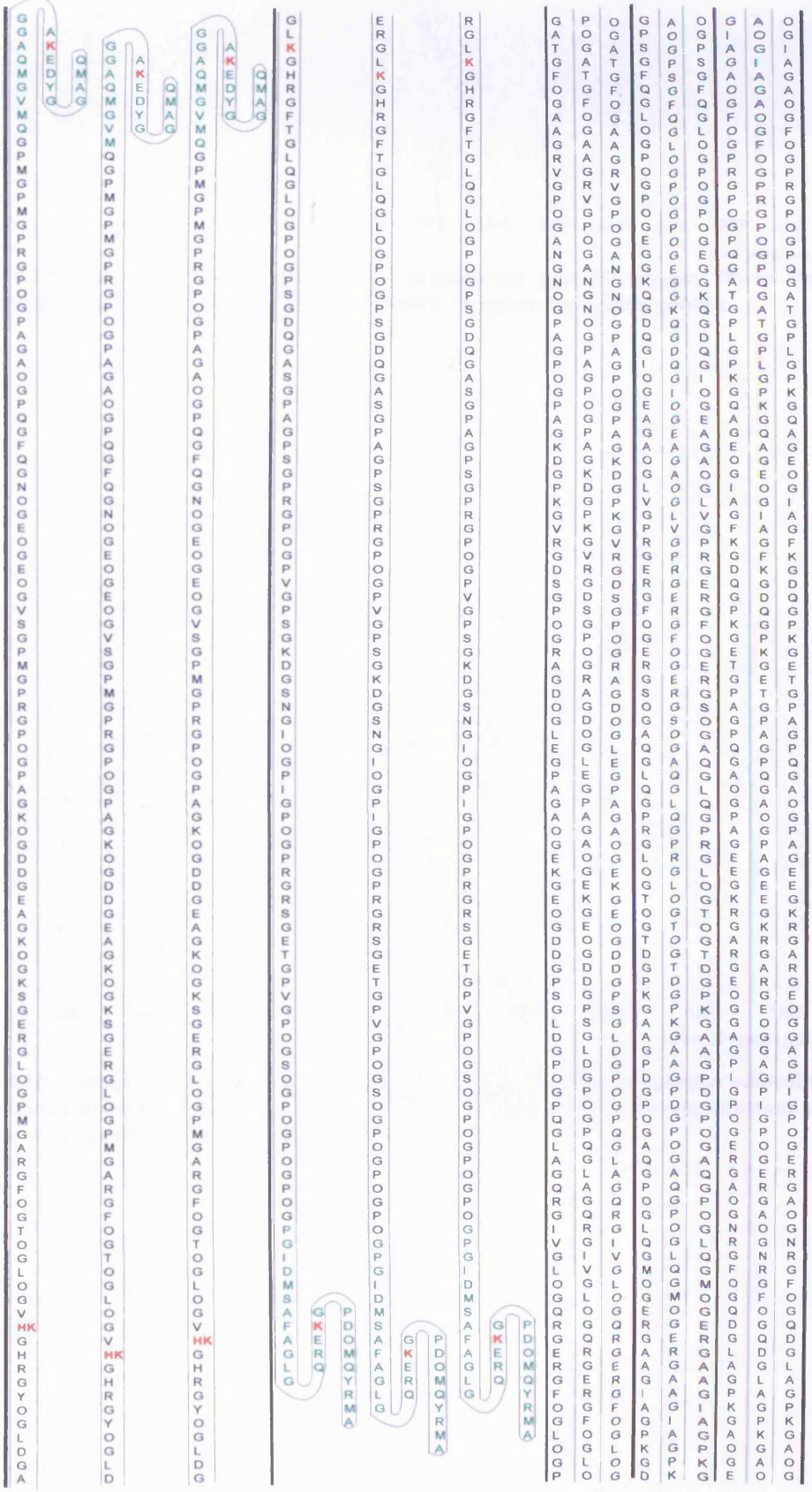


Figure 4.14: Portion of the modelled murine type II collagen microfibril containing telopeptides. Residues belonging to the telopeptides are in green, lysines responsible for cross-links formation are in red.

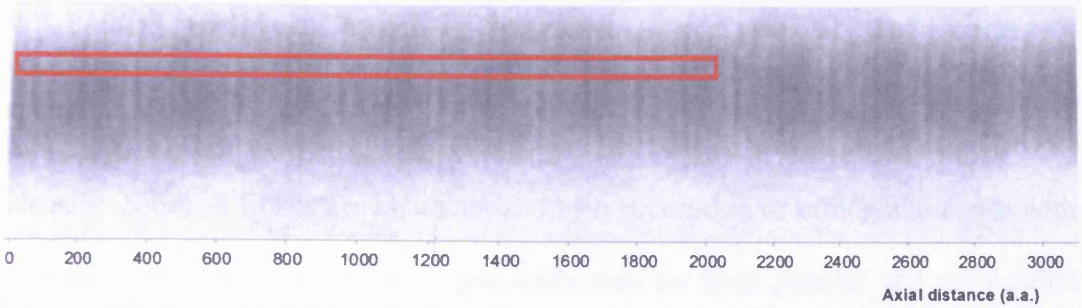


Figure 4.15: Micrograph of a portion of reconstituted type II collagen fibrils from rat chondrosarcoma. The outlined portion was used to calculate the density profile.

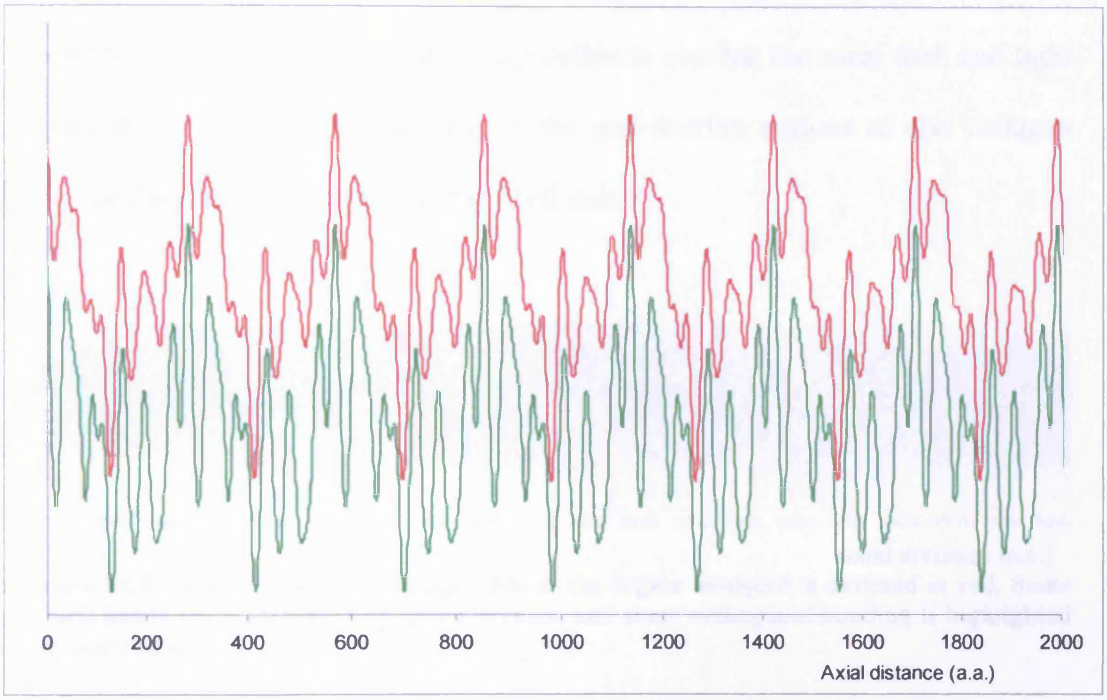


Figure 4.16: Comparison between the density profile of a real type II collagen microfibril and the simulated density profile of a modelled type II collagen microfibril (a constant value is added to the experimental profile for clarity).

4.3.1 An oblique banding pattern for type II rat collagen microfibrils

Usually, collagen fibrils are characterised by a succession of orthogonal bands with respect to the fibril axis. This is generally true for both parallel and antiparallel microfibrils. However, with collagen fibres reconstituted *in vitro* it is possible to find different conformations. In figure 4.17, we present an oblique banding pattern found in reconstituted fibrils of type II collagen from rat chondrosarcoma. The region analysed is outlined in red. It is possible to see that the usual dark and light electron-dense bands corresponding to the gap-overlap regions of the collagen fibrils are diagonal with respect to the fibril axis.

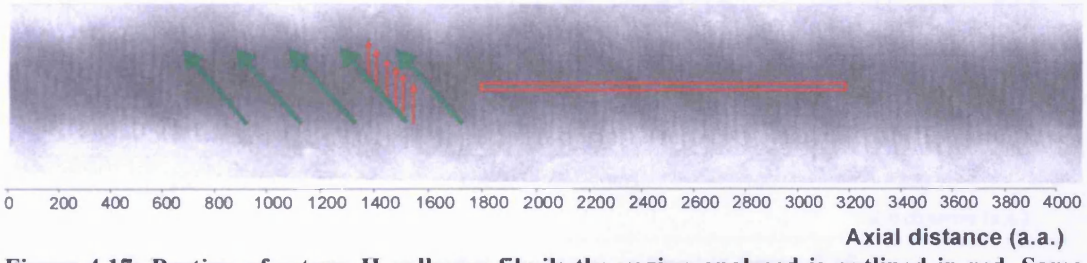


Figure 4.17: Portion of a type II collagen fibril: the region analysed is outlined in red. Some diagonal bands are highlighted by green arrows, and some orthogonal banding is highlighted by vertical red arrows.

It is also possible to see some minor sharper bands (red arrows) interspersed among the major ones (green arrows) that are orthogonal to the fibril axis. This suggests a type of molecular packing where the single microfibrils are made of parallel molecules staggered by 234 a.a. and where the microfibrils join together and are staggered themselves by a different amount.

In figure 4.18, we show the Fourier-filtered density profile of the outlined region in figure 4.17. The noise component of the signal in the Fourier space was filtered out to enhance the periodic structures of the density profile. Due to the low resolution

of the picture, the periodic structures are not well defined. However, for the portion outlined in figure 4.17, the signal is clear enough to allow us to analyse it. To reproduce this particular structure we concentrated at first on the hydrophobic scoring between two parallel microfibrils. In figure 4.19, we show the hydrophobic scoring for two parallel type II modelled collagen microfibrils.

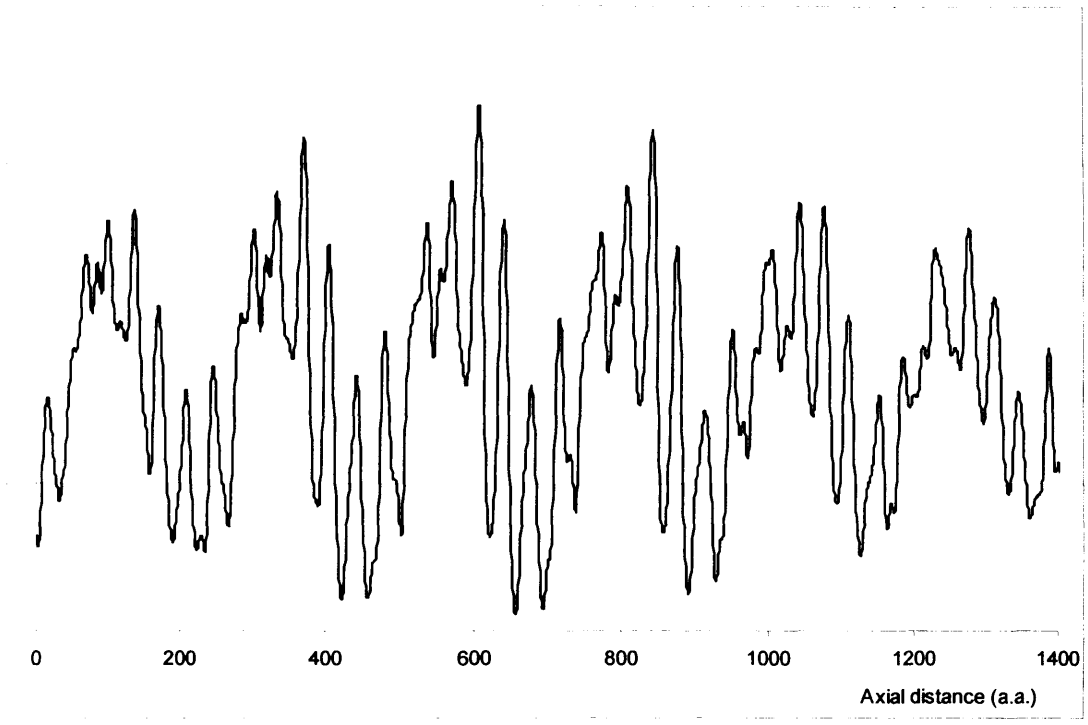


Figure 4.18: Smoothed density profile of the portion outlined in figure 4.17.

The function is 234 a.a. periodic by construction and in addition to the dominant peaks in position 234 a.a. and integer multiples of it, it shows prominent peaks in position 38 and 87 a.a. and also those related to these peaks by mirror-like periodicity relationships. It is thus reasonable to consider these two peaks as possible interaction energy minima that can explain an anomalous staggering between microfibrils.

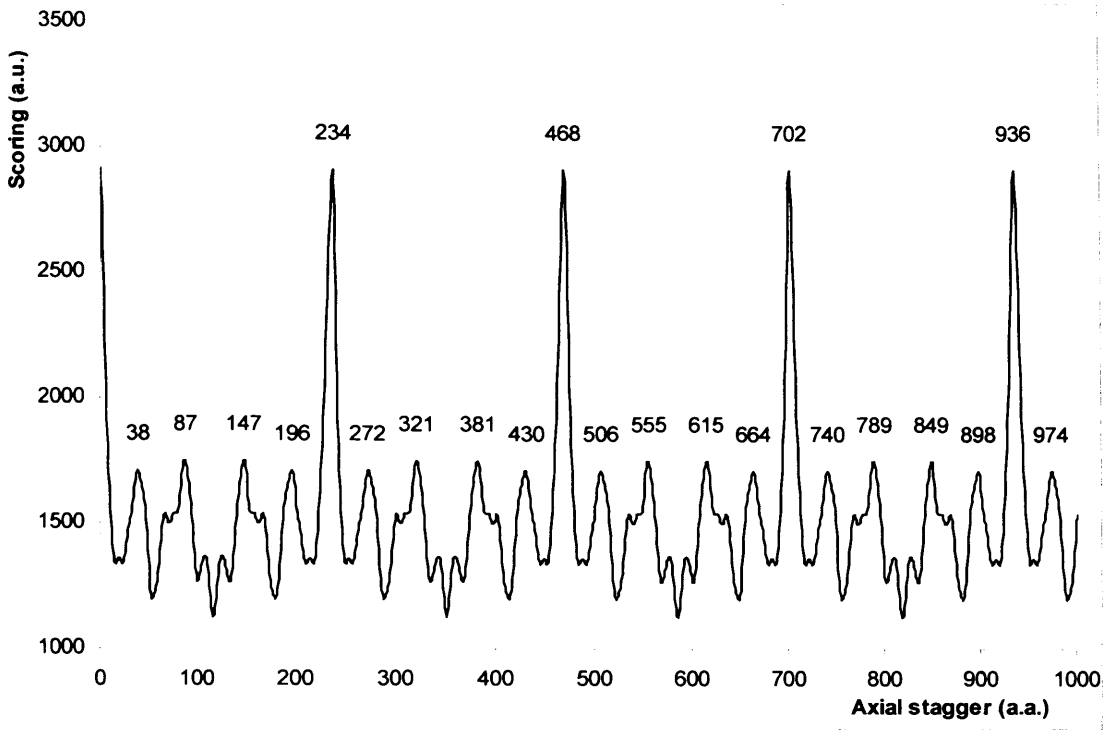


Figure 4.19: Hydrophobic scoring function for two parallel modelled type II collagen microfibrils.

We also considered the effects due to charge. As for type I collagen the attractive electrostatic scoring presents peaks that are less prominent (figure 4.20). It is also possible to see a couple of peaks (31, 46 a.a.) whose average position is at 38 a.a. and that could reinforce the hydrophobic interactions obtained when the microfibrils are displaced by that amount.

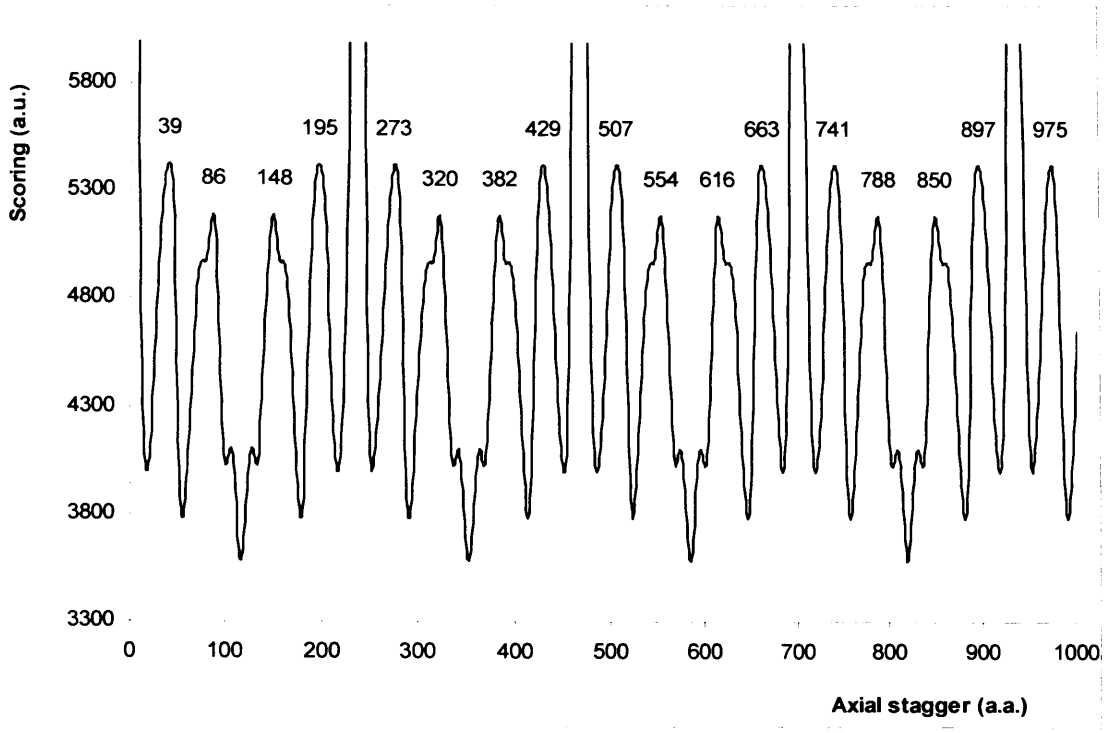


Figure 4.21: Total attractive interaction (scoring function) for two parallel modelled type II collagen microfibrils

Figure 4.22 and 4.23 show portions of modelled microfibrils for type II collagen with a staggering by 39 and 86 a.a. respectively. Both conformations could lead to cross-link formation even though microfibrils with a 86 a.a. stagger have their lysine-hydroxylysines in a more favourable position. We used both staggers to make models of microfibrils to be compared with the experimental density profile shown in figure 4.18. Figures 4.24 and 4.25 show the comparison between the density profiles of the experimental and the simulated density profiles of the two model fibrils with a 39 and a 86 a.a. stagger respectively.

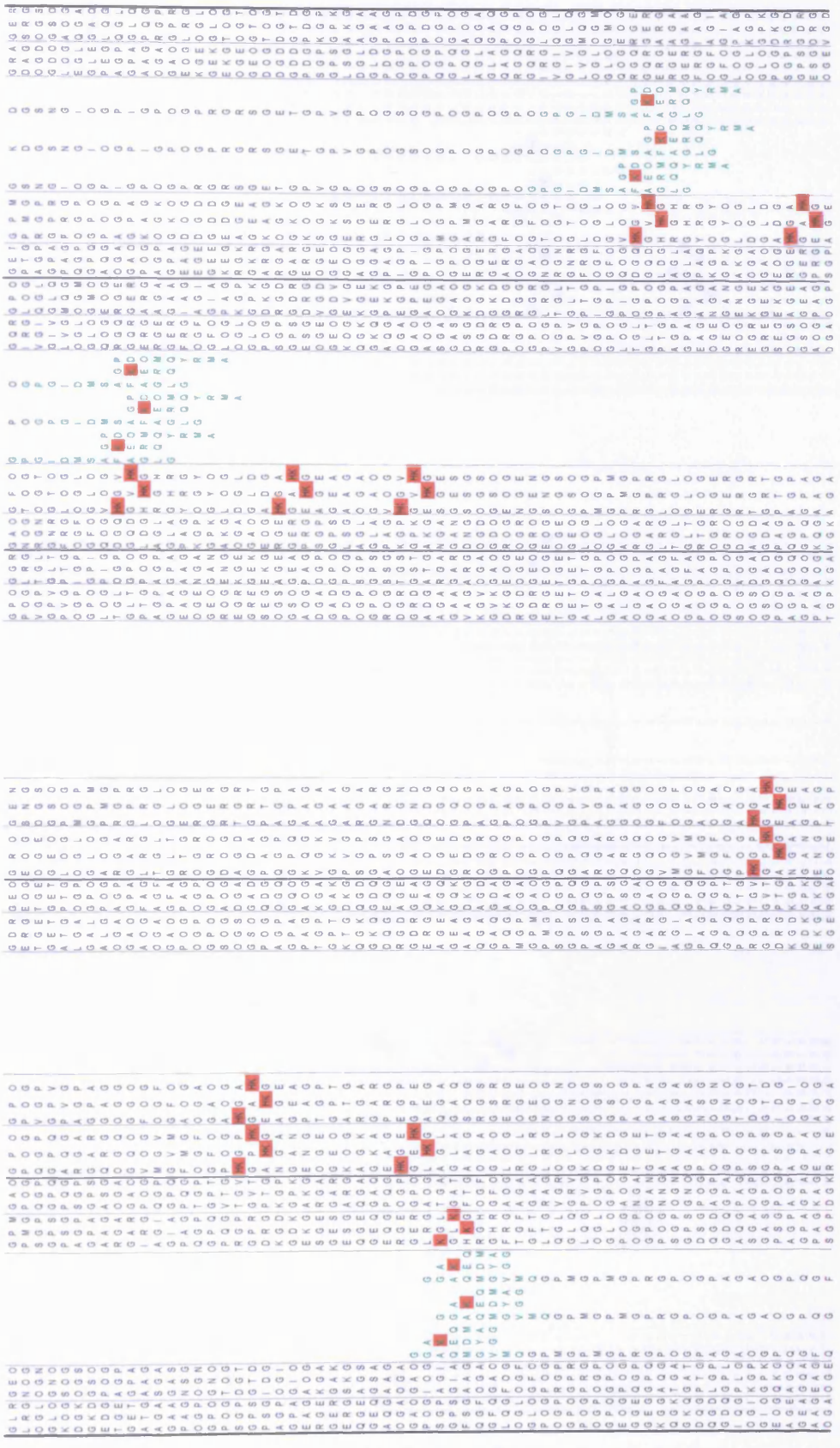


Figure 4.22: Five type II collagen microfibril with a 39 a.a. stagger. Lysines responsible for cross-link formation are in red. Cross-links can arise between lysines belonging to different microfibrils.

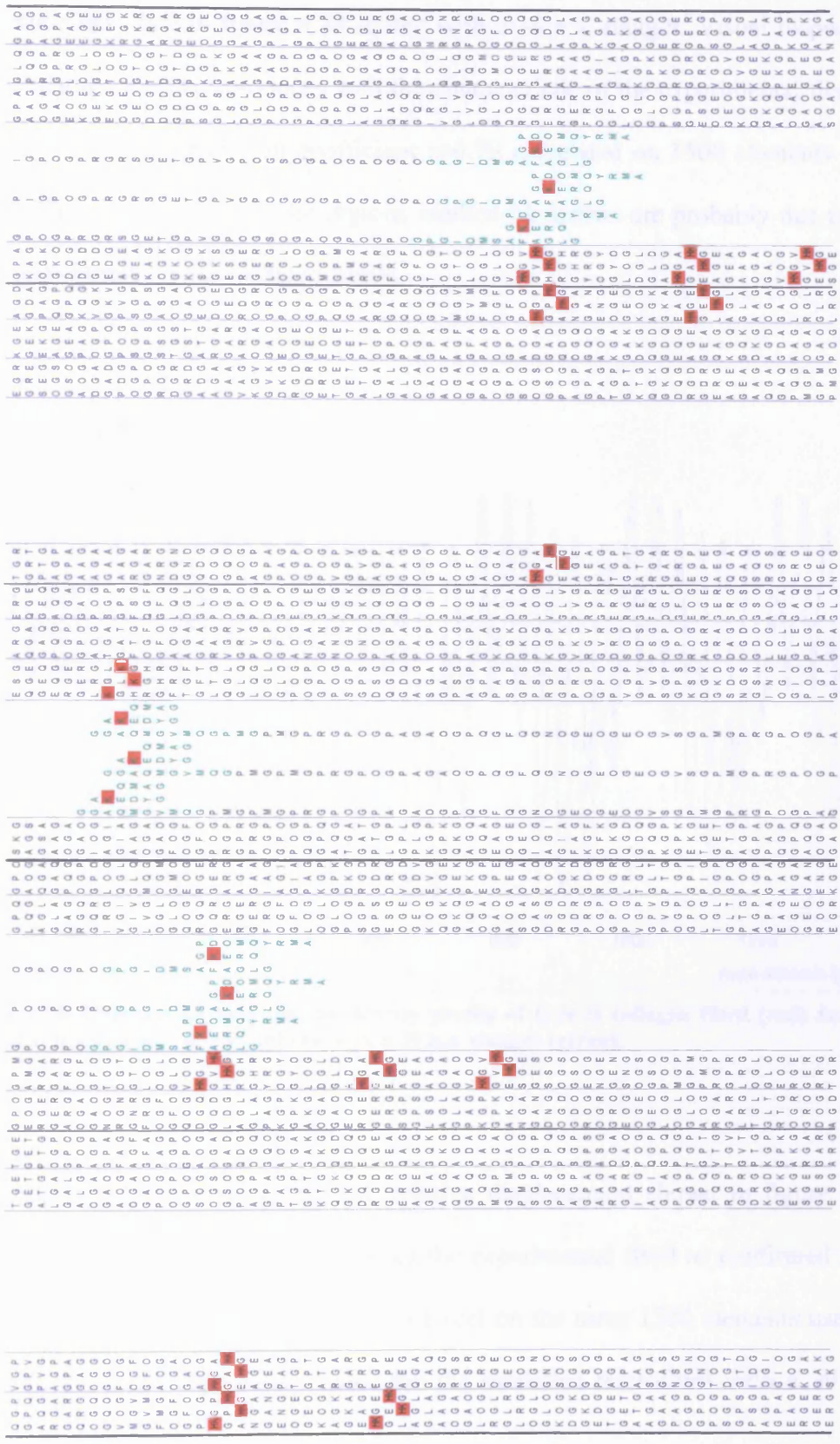


Figure 4.23 Five type II collagen microfibril with a 86 a.a. stagger. Lysines responsible for cross-link formation are in red. Cross-links can arise between lysines belonging to different microfibrils.

The modelled fibril in figure 4.24 is that with a 39 a.a. stagger. Visual inspection shows that the model reproduces well the characteristic of the experimental fibre as confirmed by the correlation coefficient r=0.79 calculated on 1500 elements with Excel. The discrepancies in the regions marked by arrows are probably due to the stain unevenness on the collagen fibrils.

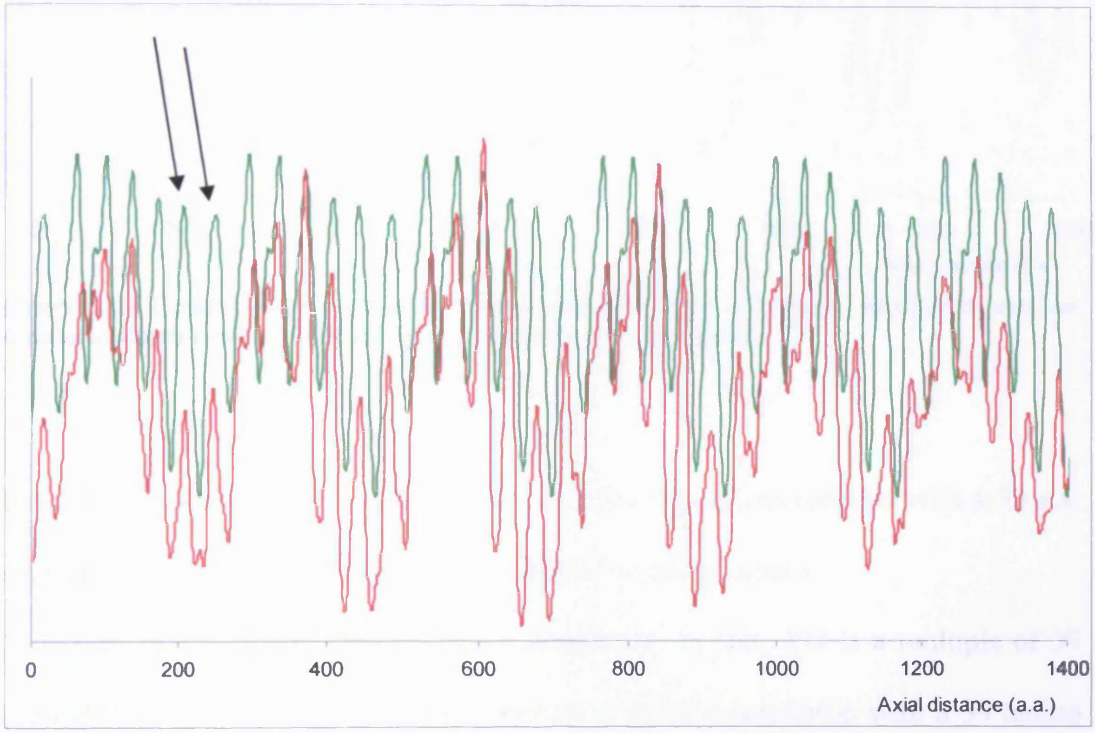


Figure 4.24: Comparison between the density profile of type II collagen fibril (red) and the simulated fibrils made of microfibrils with a 39 a.a. stagger (green).

The modelled fibril with an 86 a.a. relative displacement shown in figure 4.25 agrees poorly with the density profile of the experimental fibril as confirmed by a correlation factor r=0.24 calculated with Excel on the same 1500 elements used to calculate the cross correlation factor for the modelled microfibril with a 39 a.a. stagger.

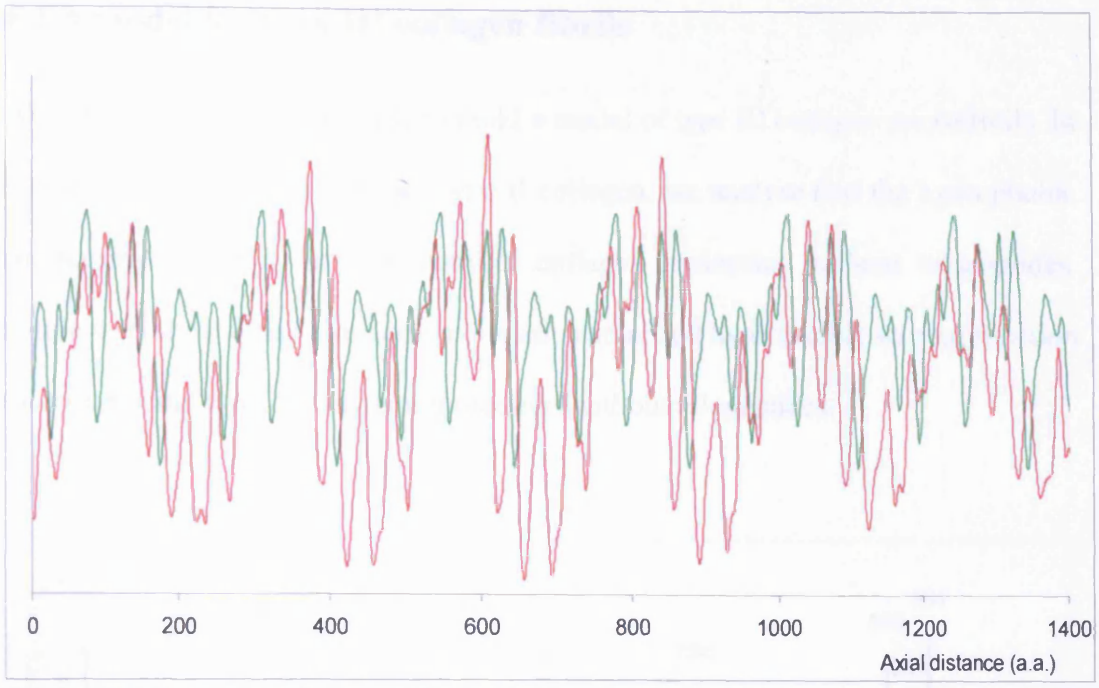


Figure 4.25: Comparison between the density profile of type II collagen fibril (red) and the simulated fibrils made of microfibrils with a 86 a.a. stagger (green).

It seems reasonable to conclude that the model made of microfibrils with a 39 a.a. stagger is the one that can explains the oblique banding pattern.

Note that 39 a.a. for the stagger is not accidental. In fact, 234 is a multiple of 39 (234=39*6). This means that if we consider a set of microfibrils with a 39 amino acids stagger, the seventh microfibril is again in register with the first one.

A possible generalisation of this observation is that in collagen fibrils, alternative staggering conformations with oblique banding would be allowed for microfibrils if they can go back to an in register configuration with a relative small number of microfibrils in the unit cell.

4.4 A model for type III collagen fibrils

We can use our scoring method to build a model of type III collagen microfibrils. In a similar fashion to that used for type II collagen, we analyse first the hydrophobic autoscoring function for two parallel collagen molecules without telopeptides. Figure 4.26 shows the Gaussian-smoothed normalised hydrophobic scoring function for two parallel type III collagen molecules without telopeptides.

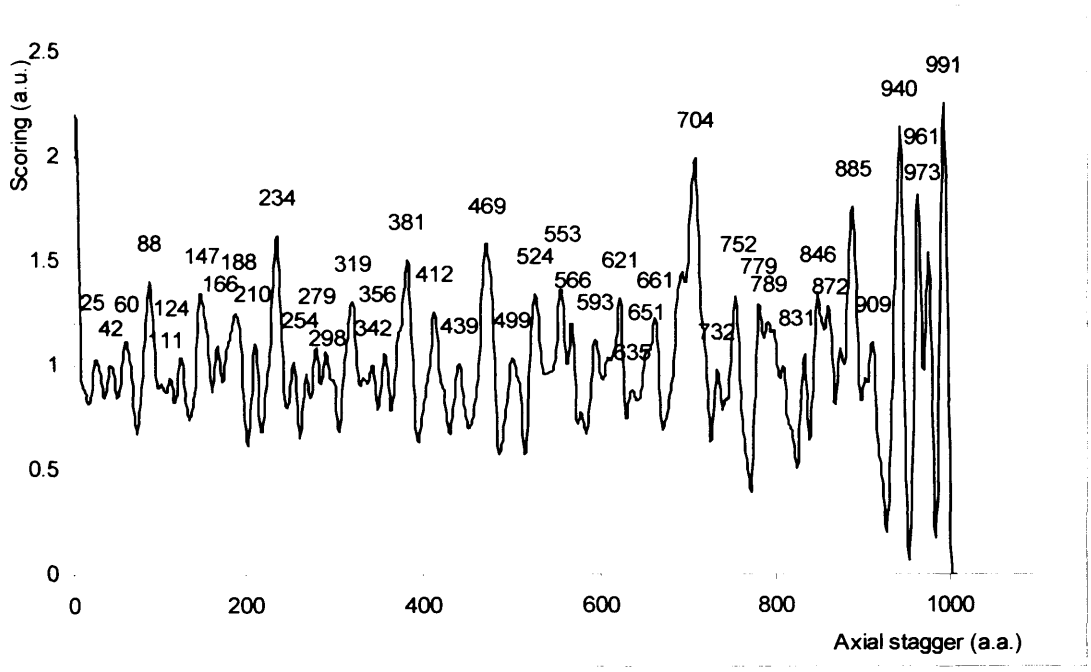


Figure 4.26 Normalised Gaussian-smoothed hydrophobic parallel scoring function for murine type III collagen.

Prominent peaks in positions similar to those found for type I and II collagen are apparent. More specifically, they are those in positions 0, 234, 469, 704, and 940 that are separated by about 234 a.a.. The same holds true for the other prominent peaks related by mirror symmetry. The average distance between corresponding peaks is 233.5 ±2.5 a.a.. The electrostatic attractive scoring for the same pair of molecules produces a graph with prominent peaks in positions 0, 234, 468, 704 and

937 that strengthen the effect of the hydrophobic interactions (figure 4.27).

From these graphs we thus can predict that type III collagen molecules pack with a common stagger of 234 a.a.. For this reason, we build a model of a type III collagen microfibril made of five collagen molecules with a 234 a.a. stagger. A further criterion for the formation of microfibrils is that of a possible formation of cross-links between the lysines of the telopeptides and the corresponding hydroxylysines of the triple helix.

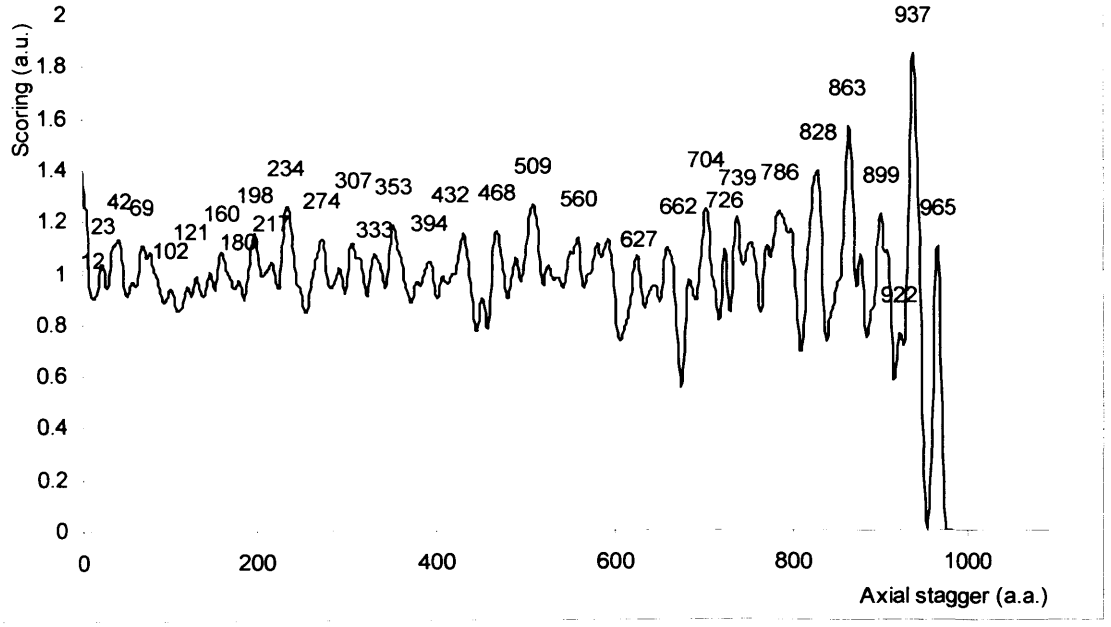


Figure 4.27: Normalised Gaussian-smoothed attractive electrostatic scoring function for two murine type III mouse collagen molecules.

In the type III collagen sequence published by Toman and de Combrugghe (Toman and de Combrugghe, 1994) for murine type III collagen, the N-telopeptides are too short for hydroxylysines to face each other if the telopeptides bend and fold as for type I and type II collagens. Therefore, for the type III collagen N-telopeptides we predict a straight configuration. However, C-telopeptides are long enough to allow

folding and still form hydroxylysine mediated cross-links. We therefore force lysine 16 of the C-terminal telopeptide to face hydroxylysine 93 of the collagen triple helix. This assumption and the 234 a.a. stagger help us to devise a particular 'S' shaped folded conformation for the telopeptides. As already pointed out for the case of type II collagen, glycines and alanines can be used as hinges around which to fold the telopeptides. In figure 4.28, we represent the folded conformation used for the telopeptides to build our modelled microfibril.

The projected length for the molecules with telopeptides folded as described is 1052 a.a. long thus the gap between successive molecules is 118 a.a. long.

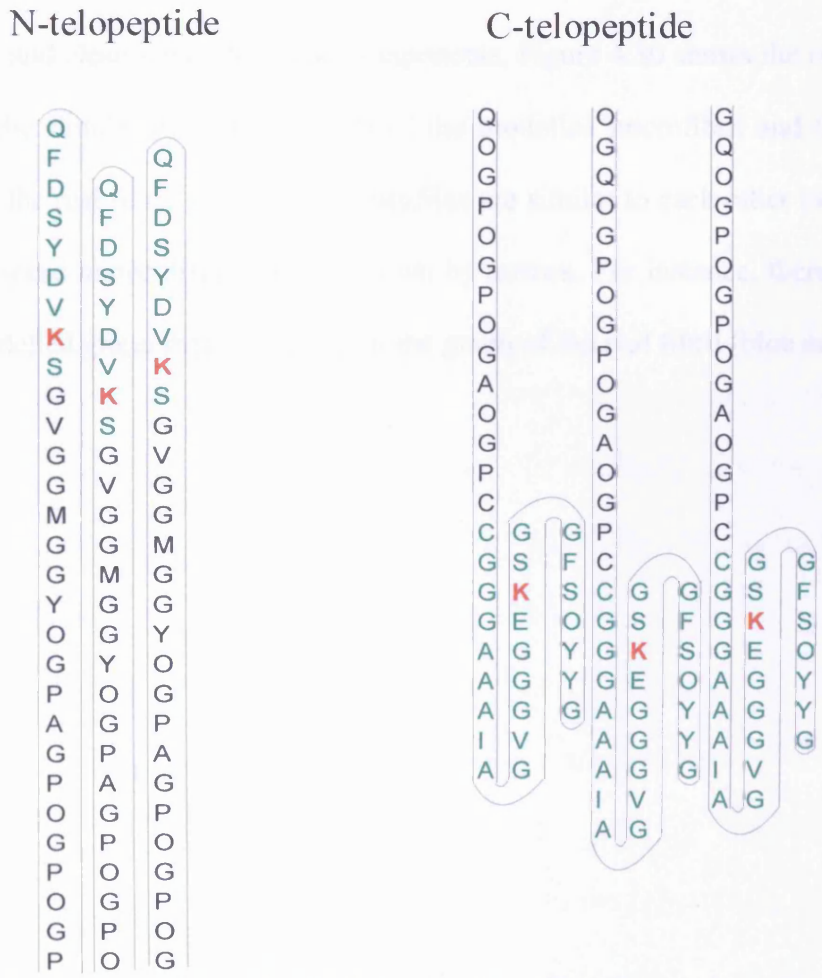


Figure 4.28: Folded conformation used for the type III collagen telopeptides. Residues belonging to the telopeptides are coloured in green, while lysines responsible for cross-link formation are in red. Note that the C-telopeptides fold around an A-G or a G-G pair. G and A are the smallest amino acids so that steric hindrance is reduced to a minimum there.

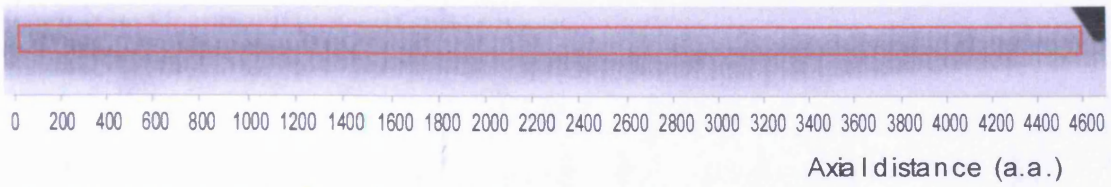


Figure 4.29: Portion of a fibril of type III collagen with the section used for our measurements outlined in red.

We then calculated the simulated density profile for the modelled type III collagen microfibril and compared it with an experimentally obtained fibril of type III collagen. Figure 4.29 show a portion of a reconstituted fibril of type III collagen from rabbit skin where the portion used for the comparison is outlined in red.

We filtered the Fourier transform of the real density profile to enhance the periodic structures and clean away the noise components. Figure 4.30 shows the comparison between the simulated density profile of the modelled microfibril and the density profile of the real fibril. Both density profiles are similar to each other even though there are some minor discrepancies shown by arrows. For instance, there is a peak in the modelled graph that is missing in the graph of the real fibril (blue arrow).

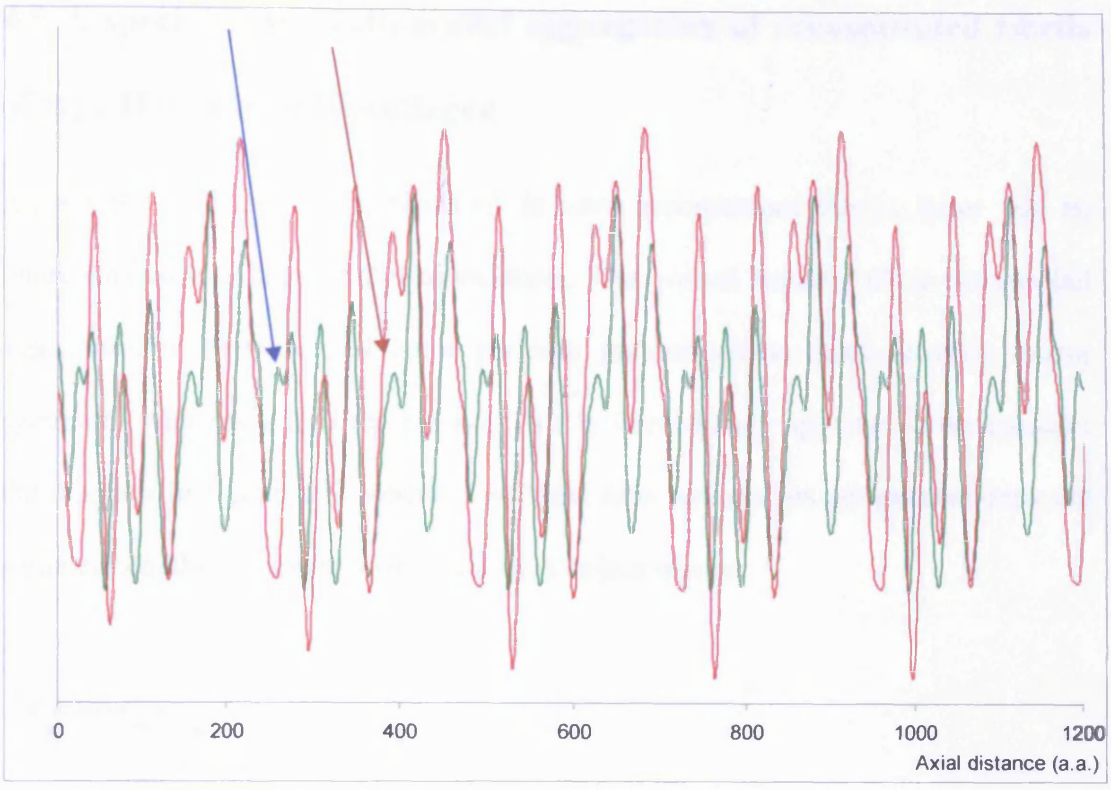


Figure 4.30: Comparison between the simulated density profile of a modelled microfibril of type III collagen (green) and the density profile of the portion of type III collagen fibril outlined in figure 4.26 (red).

This is in the proximity of the N-telopeptide. Possibly, it indicates that the conformation for the N-telopeptide we used could be improved, or maybe that there are genetic differences between mouse, whose type III collagen sequence was used to create the model, and rabbit, whose type III collagen was imaged in the microscope. The other major discrepancy (red arrow) is the low intensity of a peak of the modelled graph. This peak corresponds to the central plane of the gap region of the microfibril and could be once again due to the genetic difference between mouse and rabbit. Visual inspection of the graph shows an acceptable correspondence between model and experimental fibril as confirmed by the correlation coefficient r=0.69 obtained applying equation 2.3 as given by the Excel cross correlation formula: a significant value.

4.5 A special case: antiparallel aggregation of reconstituted fibrils of type II and type III collagen

Among the different micrographs of *in vitro* reconstituted fibrils, there was an interesting case of antiparallel association. The typical banding of an antiparallel association is characterised by a periodic pattern whose bands possess mirror symmetry with respect to the period. This is immediately apparent if we consider the diagram in figure 4.31 where a periodic structure and its antiparallel copy are summed together to create a structure with mirror planes.

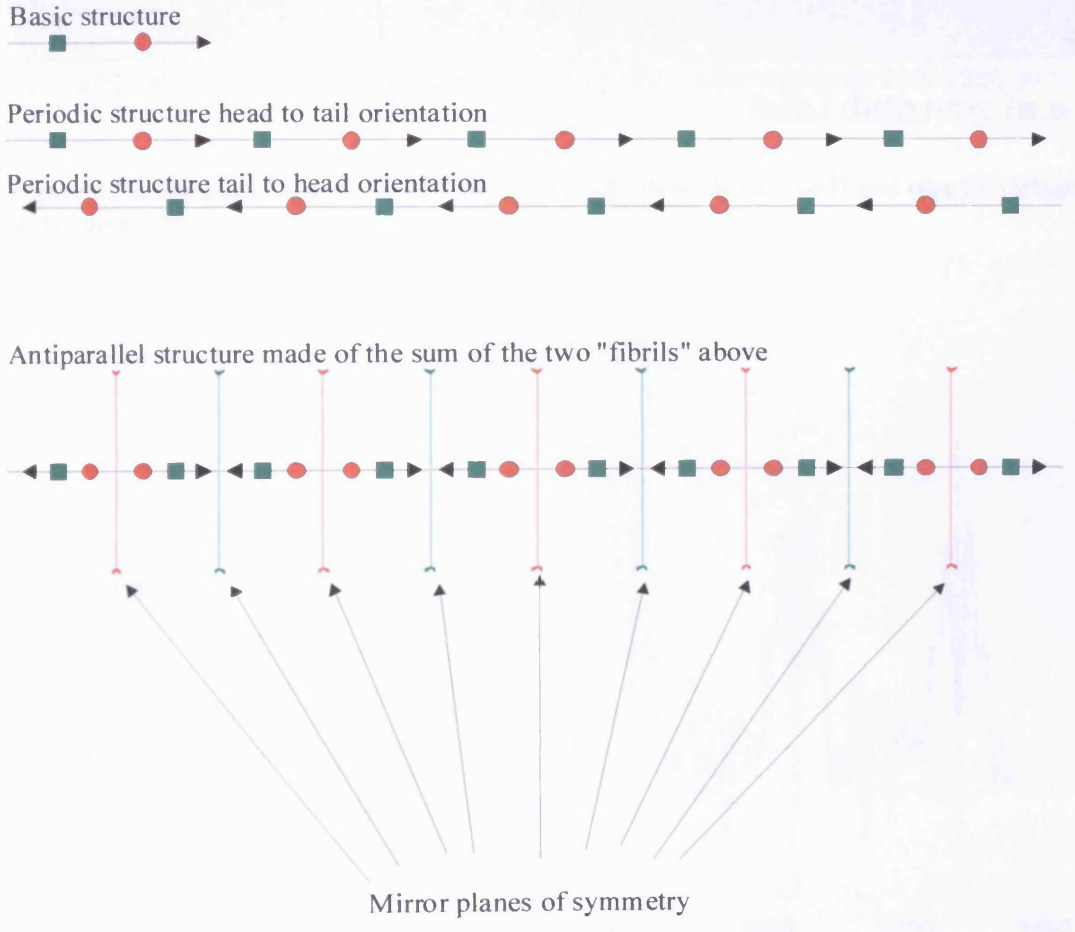


Figure 4.31 Diagrammatic representation of a periodic structure associated in an antiparallel manner

Our specimen is special because it is made of antiparallel fibrils obtained from a mixture of type II and type III collagens in 1:1 proportions in stock. Figure 4.32 shows a portion of a fibril made of collagen microfibrils aggregated in an antiparallel manner. The portion outlined is the one actually analysed.

Figure 4.33 is the Fourier-filtered density profile of the portion outlined in figure 4.32. The repetition of a mirror symmetrical pattern about the red planes drawn in figure 4.33 is apparent.

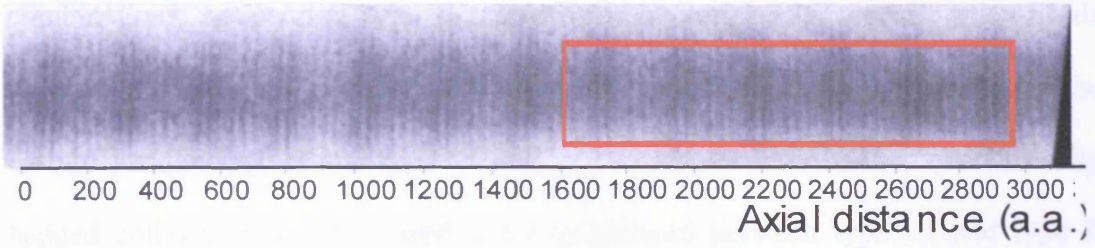


Figure 4.32: Portion of an antiparallel fibril made of a mixture of type II and type III collagen in 1:1 proportion in stock.

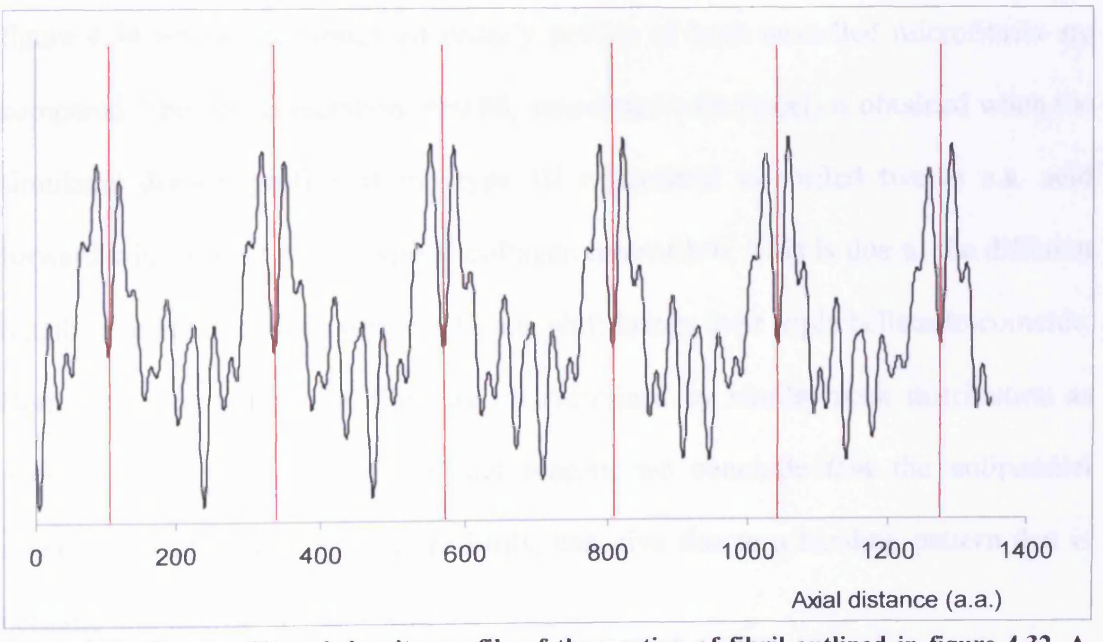


Figure 4.33: Fourier filtered density profile of the portion of fibril outlined in figure 4.32. A family of mirror-planes is drawn in red

For our model representation, we assumed that the antiparallel conformation clearly visible in figure 4.32 and 4.33 is the result of the repetition of a basic sub-unit made of one microfibril of type II collagen and of one microfibril of type III collagen oriented in an antiparallel fashion. We used a heterotypic combination of the two collagens because the fibril was obtained from a mixture of collagens. In addition, graph 4.34 is not made of regions exactly symmetrical allowing us to conceive a model made of mixed antiparallel microfibrils. Of all the alternative models, this is the simplest and the one that gives the best results. In fact the other possible combinations: antiparallel type II microfibrils or antiparallel type III microfibrils gave results that were inferior. We used the microfibrils modelled above for the case of parallel type II and type III collagen fibres, to describe this symmetrically banded collagen fibril. We used a 1:1 proportion between type III and type II modelled microfibrils to respect the proportion used to build the real fibril *in vitro*. Even though type II and type III collagen microfibrils are not identical, the sequences of their charged amino acids closely resemble each other as shown in figure 4.34 where the simulated density profile of both modelled microfibrils are compared. The best correlation ($r=0.88$, calculated with Excel) is obtained when the simulated density profile of the type III microfibril is shifted twelve a.a. acid forward with respect to the type II collagen microfibril. This is due to the different length of the telopeptides where a 12 a.a. shift brings their triple helices to coincide. Both simulated density profiles are characterised by similar peak distribution as well as similar intensities. For this reason, we conclude that the antiparallel arrangement of two similar microfibrils, can give rise to a banding pattern that is periodic and mirror-symmetrical.

We could therefore apply the scoring method to probe antiparallel aggregations of type III and type II collagen microfibrils. Figure 4.35 and 4.36 present the hydrophobic and the attractive electrostatic scoring functions for this conformation. Graph 4.35 shows how the peaks in position 115 or 198 correspond to energy wells for antiparallel aggregation. They are also reinforced by peaks in positions 116 and 201 from the electrostatic attractive scoring (figure 4.36). If we combine the effect of both attractive interactions multiplying together these graphs point-by-point we obtain the graph in figure 4.37 that is characterised by prominent peaks at positions 115 and 200.

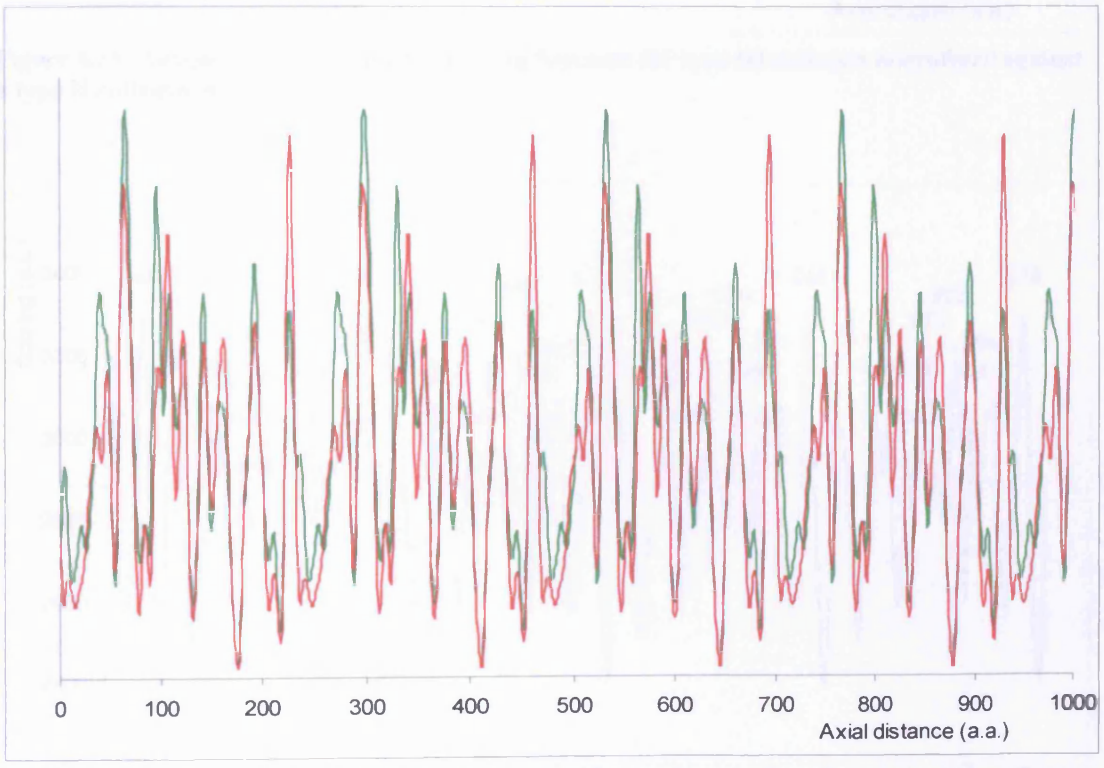


Figure 4.34: Comparison between the simulated density profile of a type III collagen microfibril (red) and a collagen type II microfibril (green). Best agreement is reached when type II microfibril density profile is shifted twelve a.a. forward with respect to type III microfibril.

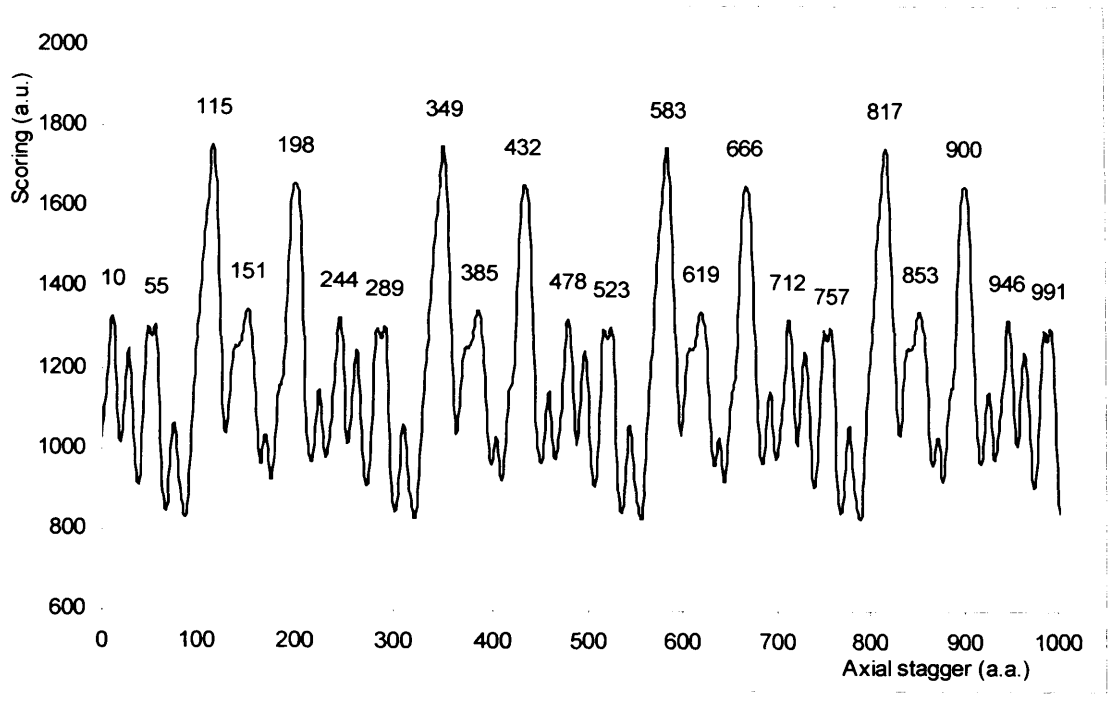


Figure 4.35: Antiparallel hydrophobic scoring function for type III collagen microfibril against a type II collagen microfibril.

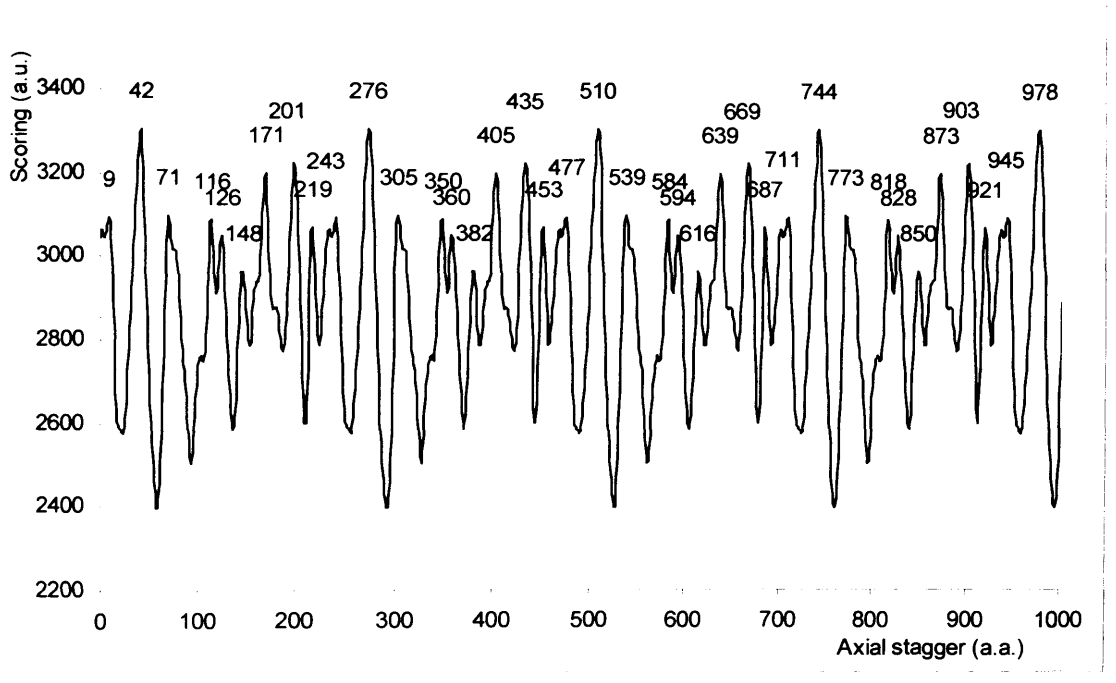


Figure 4.36: Attractive electrostatic scoring function for a type III collagen microfibril against a type II collagen microfibril.

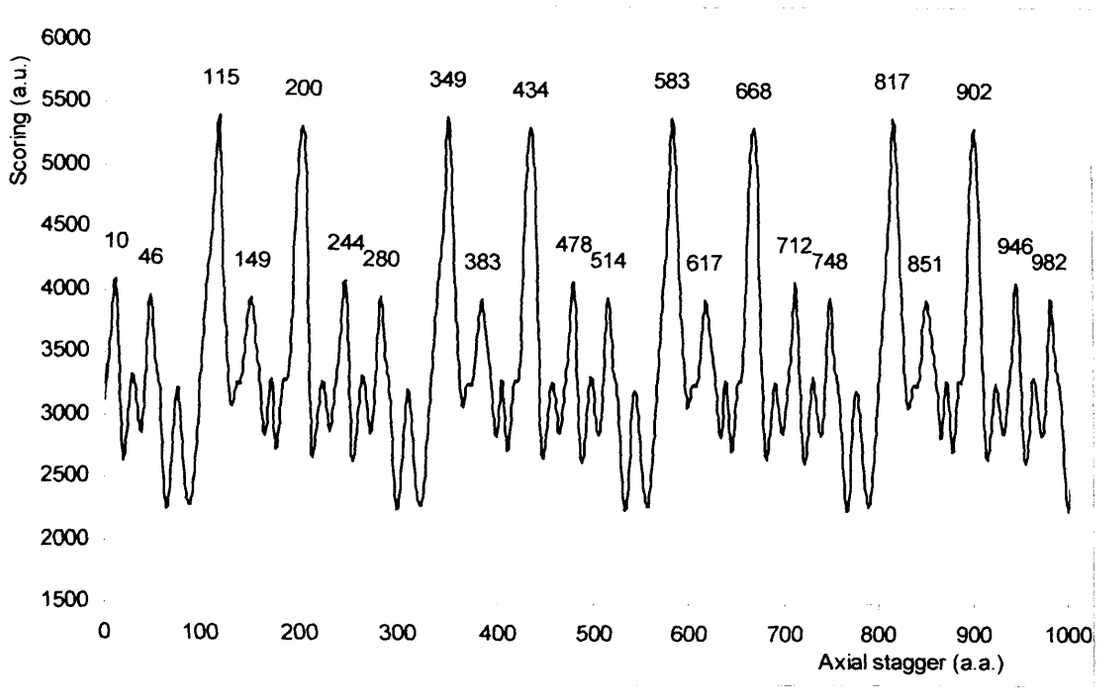


Figure 4.37: Total attractive interaction (scoring function) for type II and type III collagen microfibril in an antiparallel orientation.

However, both these positions present problems when used to explain antiparallel packing. A shift of 115 a.a. between antiparallel microfibrils produces a molecular packing where the gap and overlap regions are adjacent. In this way, there is no overlapping between molecules of the different microfibrils so that long antiparallel fibrils can not be formed. Moreover, the comparison between experimental and modelled staining is poor as visible in figure 4.38 and confirmed by a correlation factor r=0.4 calculated on 1800 elements.

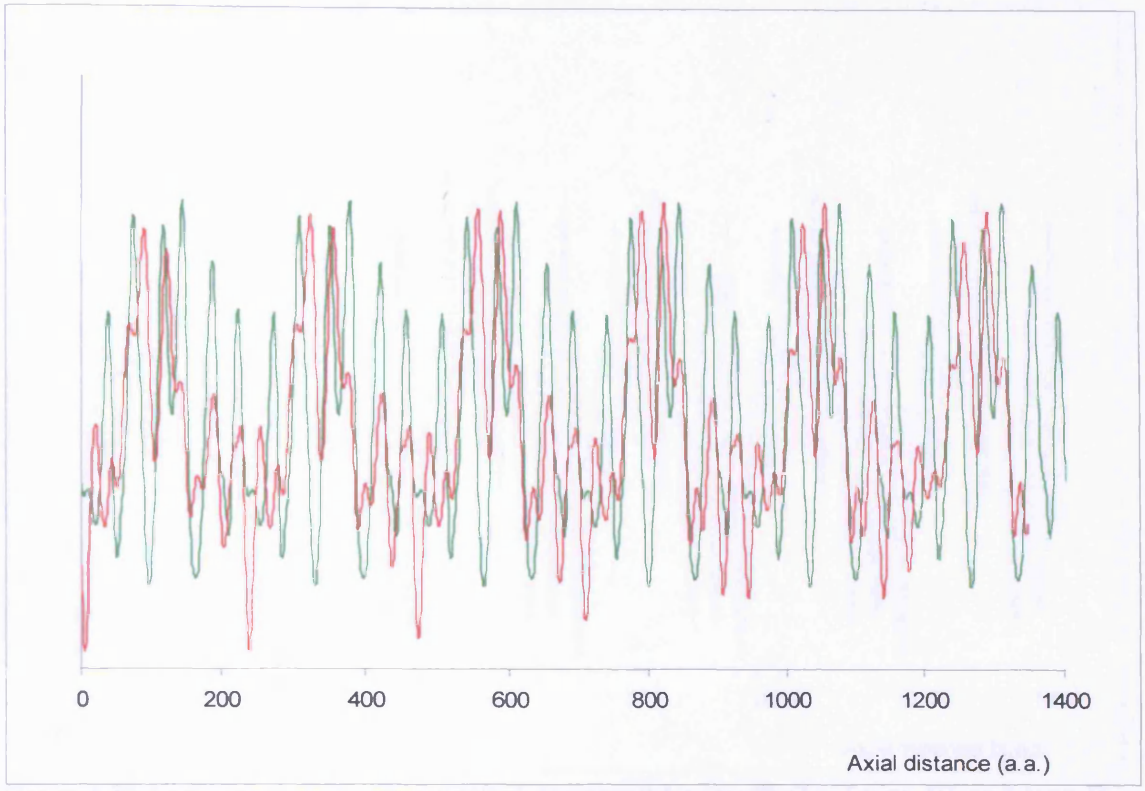


Figure 4.38: Comparison between modelled antiparallel microfibrils of type III and type II collagen in 1:1 proportion with a 115 a.a. stagger (green) and the real microfibril (red)

A shift by 200 a.a. would bring the C-telopeptide of type III collagen to face the C-telopeptide of type II collagen possibly allowing the formation of hydroxylysine mediated cross-links. However when the density profile of this modelled microfibril is compared to the experimental fibril the result is also poor as shown in figure 4.39 where the correlation coefficient r=0.56 calculated on 1800 elements.

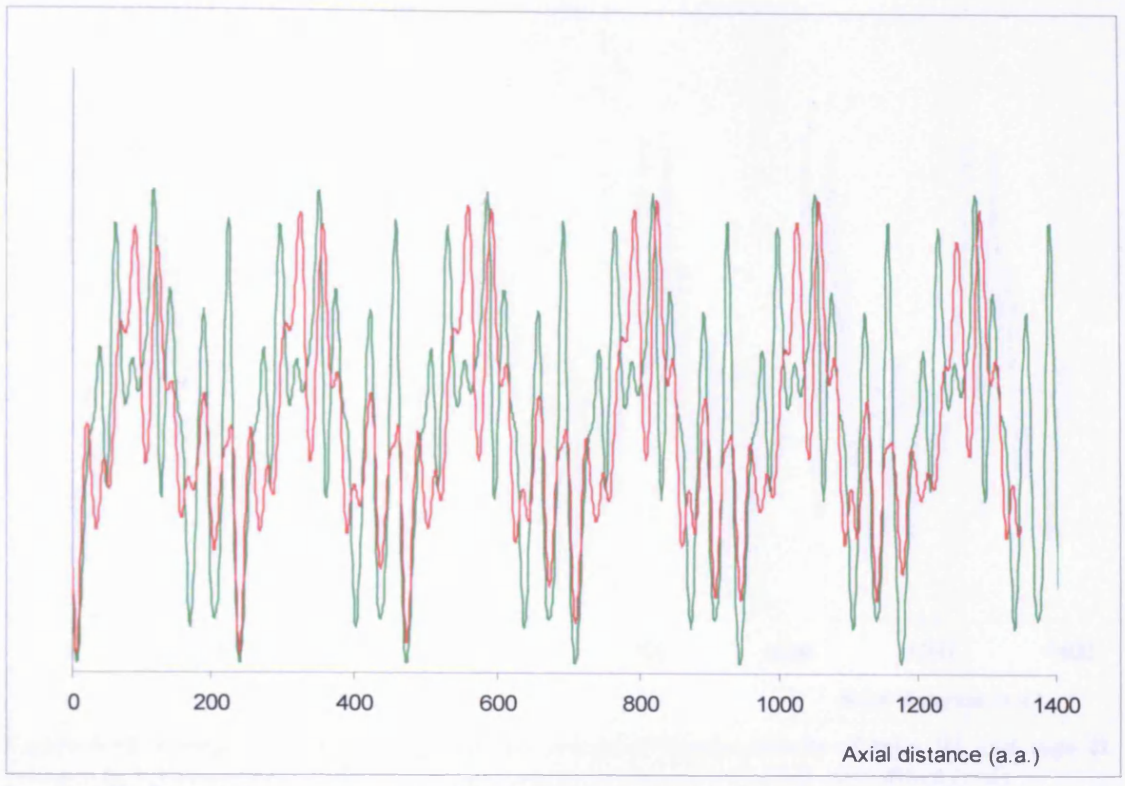


Figure 4.39 Comparison between modelled antiparallel microfibrils of type III and type II collagen in 1:1 proportion with a 200 a.a. stagger (green) and the real microfibril (red)

If instead of using the 200 a.a. displacement to build the antiparallel fibril we used a 178 a.a. displacement we obtain the graph in figure 4.40 where the agreement between simulated and real staining is apparent as confirmed by the cross correlation factor r=0.9 calculated on 1800 elements. Since graph 4.41 does not present any peak at position 178 a.a., it thus seems that the hydrophobic and electrostatic interactions are not sufficient in this case to explain this particular aggregation. A possible explanation could be that the collagen mixture contains proteoglycans or other molecules that act as mediators on the surface of the molecules forcing them to combine in a way different from that predicted by the scoring method alone.

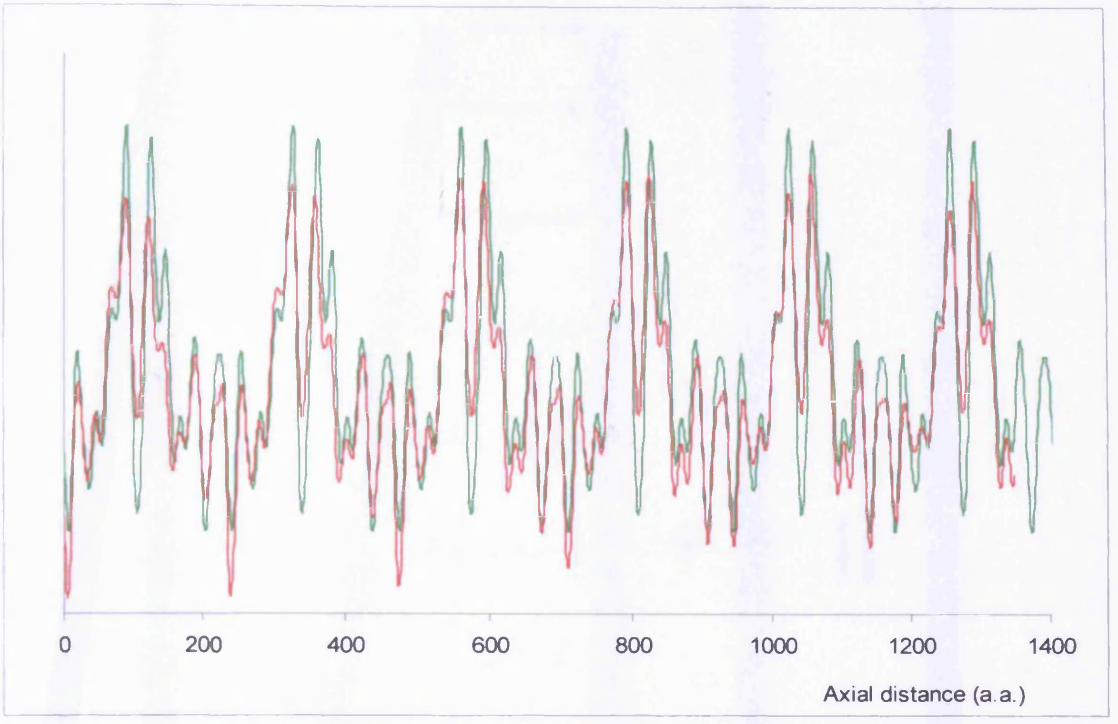


Figure 4.40: Comparison between modelled antiparallel microfibrils of type III and type II collagen in 1:1 proportion with a 178 a.a. stagger (green) and the real microfibril (red)

Another possible explanation for the *failure* of our method could be that the collagen microfibrils may not assemble with a simple lateral association, but they could in fact supercoil around each other. In this case, the fibrils would be made of microfibrils that form a superhelical structure. Such structure can be pictured in two dimensions as microfibrils that are at an angle with respect to each other as shown in figure 4.41 where a type II collagen microfibril is drawn at an angle (the coiling angle) with respect to the type III microfibril that is set horizontally with respect to the page.

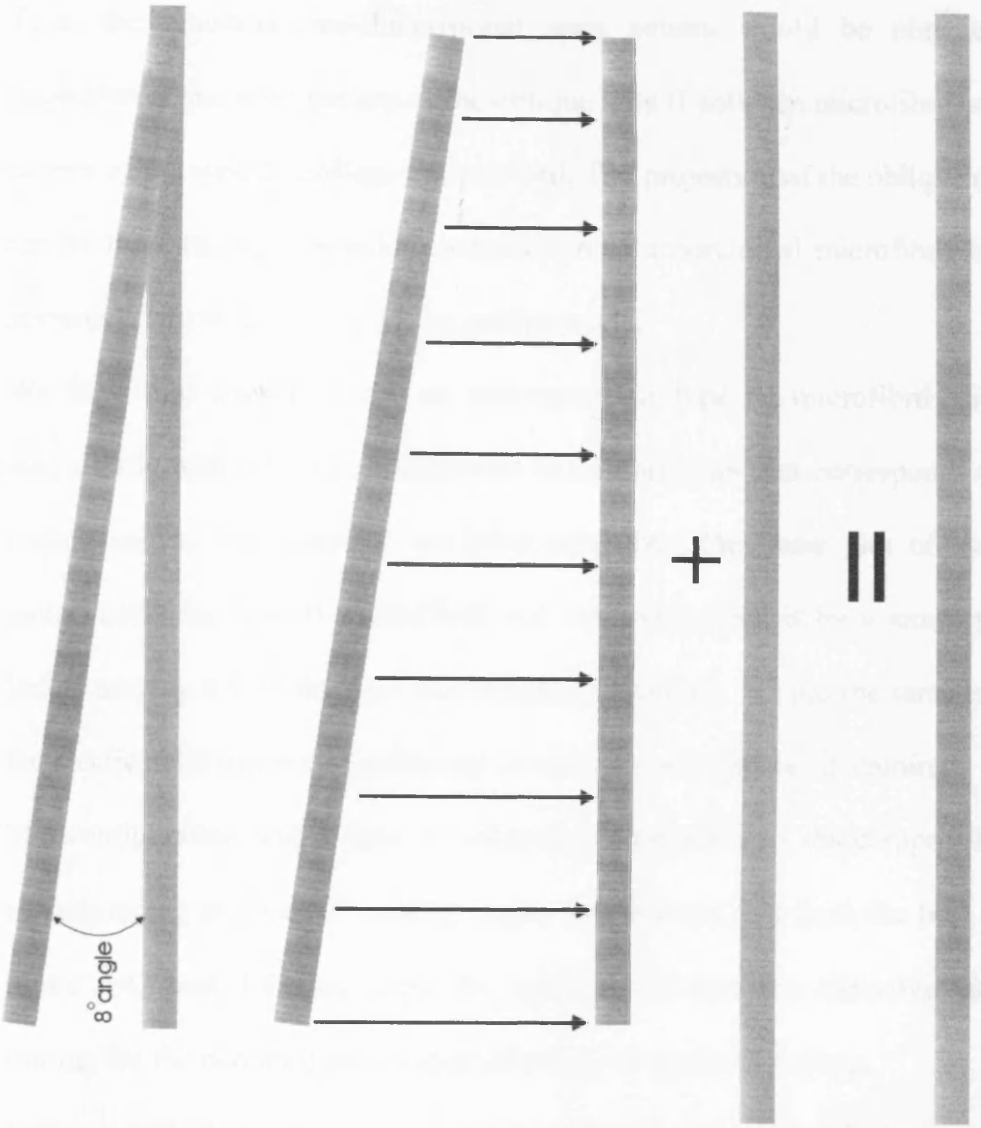


Figure 4.41: Representation of the supercoiling between type III and type II antiparallel collagen fibrils. Type II collagen winds around type III collagen with an 8° angle. The density profile of the experimentally obtained antiparallel fibril is given by the density profile of the type III collagen microfibril and that of the axially projected type II collagen microfibril combined together.

Thus, the resulting one-dimensional stain pattern would be obtained as the projection of the stain pattern of the oblique type II collagen microfibril on the stain pattern of the type III collagen microfibril. The projection of the oblique microfibril can be thought of as an axial compression of a horizontal microfibril by a factor corresponding to the cosine of the coiling angle.

We have thus tried to score an uncompressed type III microfibril with type II microfibrils with different coefficients of compression that corresponds to defined coiling angles. For example, we have considered the linear plot of hydrophobic amino acids for type II microfibril and we compressed it by a small percentage before scoring it with the type III collagen microfibril. We did the same to calculate the electrostatic interaction plots and to represent the simulated staining.

The configuration with a type II collagen microfibril with 1% compression factor, corresponding to about 8° coiling angle, was the one that gave the best results. In figure 4.42 and 4.43 we show the hydrophobic and the attractive electrostatic scoring for the two antiparallel microfibrils with about 8° coiling.

In figure 4.42, the peaks in position 120 and 205 are to be the ideal energy wells for antiparallel aggregation. The attractive electrostatic scoring (figure 4.45) has peaks in the proximity of those just mentioned, in position 130 and 206.

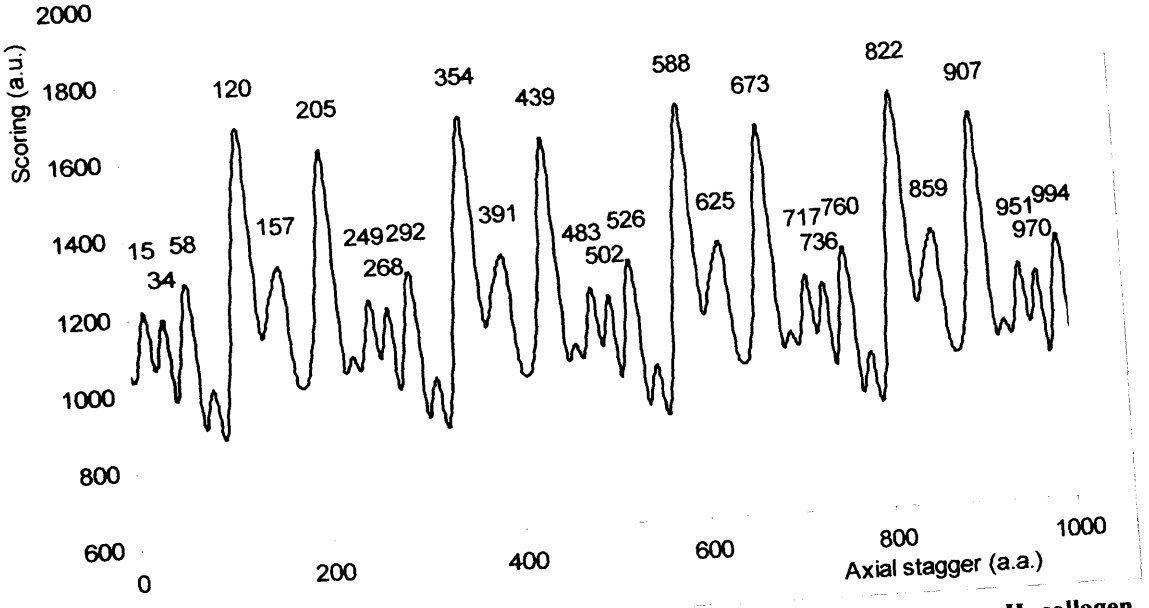


Figure 4.42: Antiparallel hydrophobic scoring function for type III and type II collagen microfibril with an 8° coiling angle.

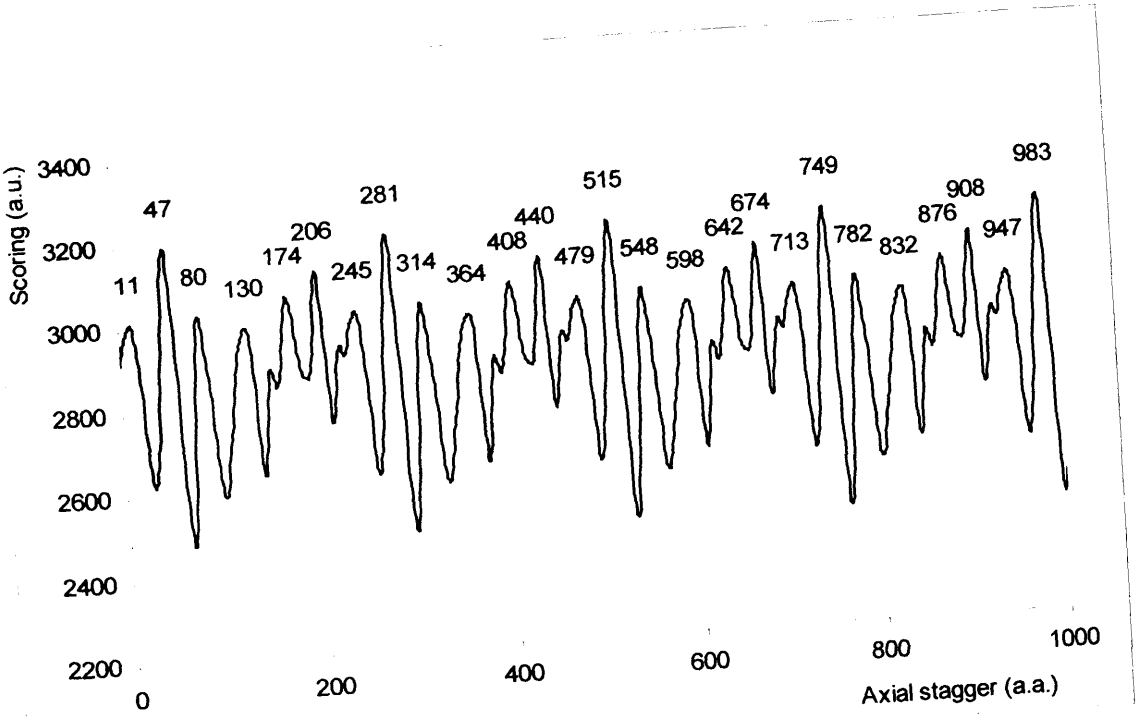


Figure 4.43: Antiparallel attractive electrostatic scoring for type III and type II collagen microfibril with an 8° coiling angle.

If, as done above, we consider the point by point multiplication of both graphs 4.42 and 4.43 we obtain the graph 4.44 that represents the total attractive interaction when both interactions act together. It has prominent peaks in positions 120 and 205 a.a. as the graph for the hydrophobic interaction. We choose a 120 a.a. shift for the antiparallel supercoiled aggregation of type II and type III collagen antiparallel fibrils.

In figure 4.45, we show the comparison between the simulated staining for the antiparallel supercoiled fibril obtained and the staining of the experimental fibril shown in figure 4.33.

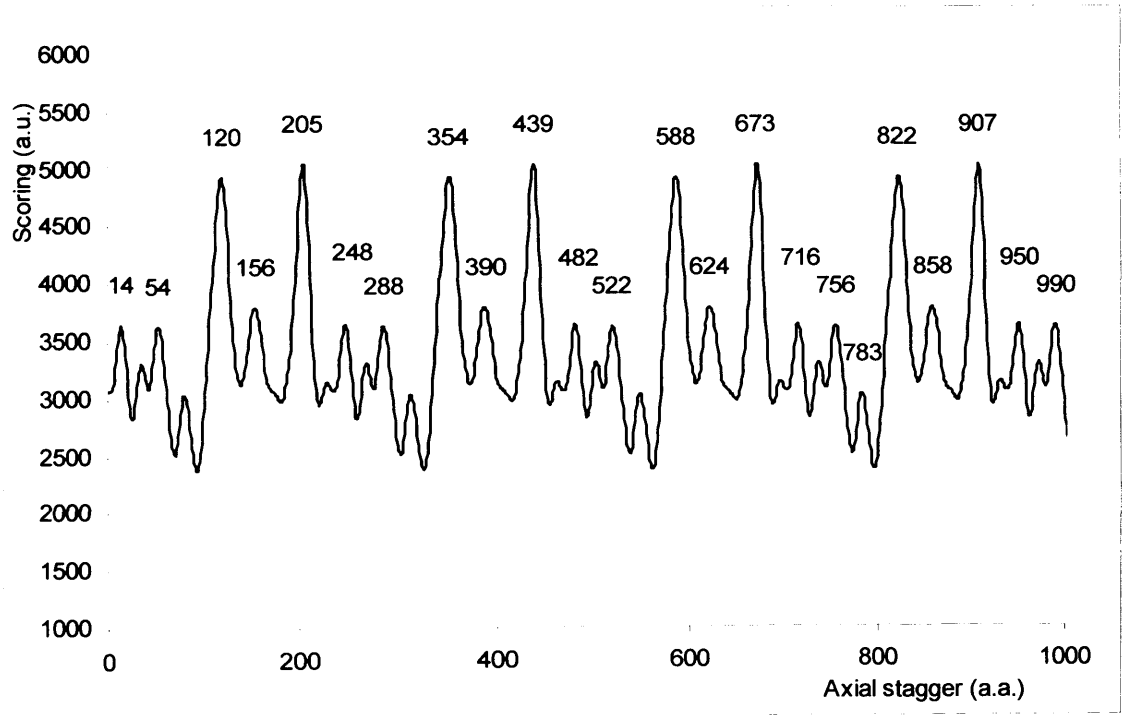


Figure 4.44: total attractive interaction for an antiparallel microfibril with type two collagen wound around type III collagen with an 8° coiling angle.

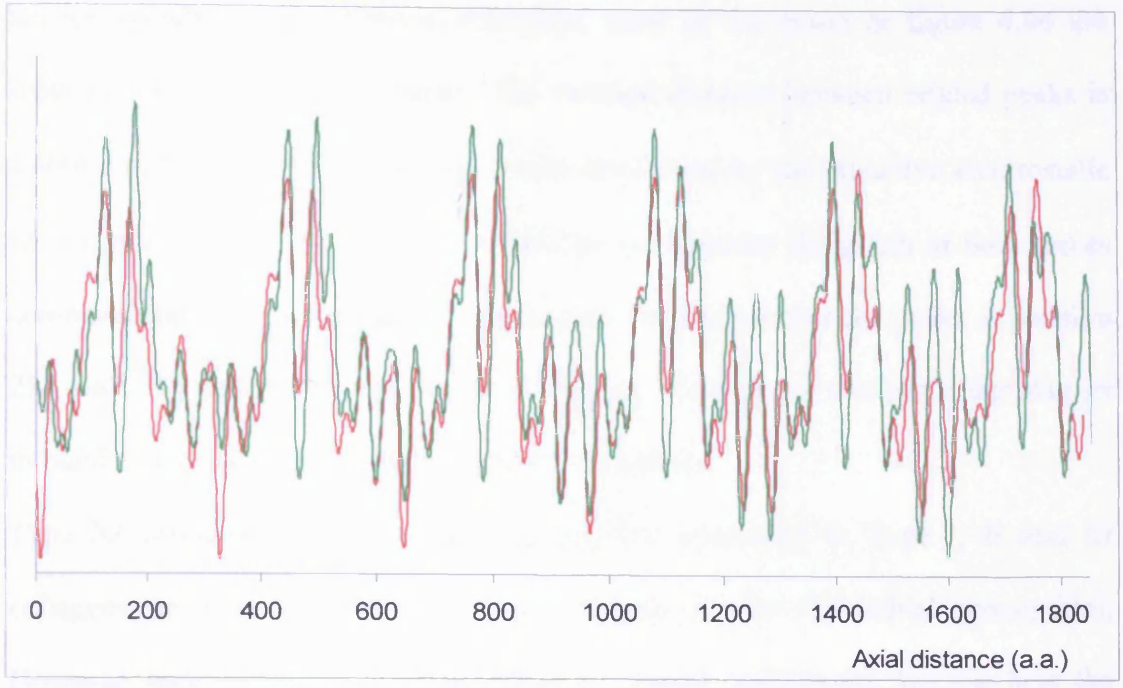


Figure 4.45: Comparison between simulated staining for an antiparallel supercoiled fibril with an 8° angle (green) density profile and the density profile of a real antiparallel fibril (red)

The accordance between simulated and real density profile is highly significant (r=0.83 calculated on 1800 elements) although slightly lower than the one obtained previously when the two antiparallel microfibril are uncoiled and separated by 178 a.a..

4.6 Comparison between model and experimental fibrils of type XI collagen

We studied the hydrophobic scoring between two parallel type XI collagen molecules to identify their stagger during fibril formation. In figure 4.46, we present the hydrophobic interaction of two parallel type XI collagen triple helices. The prominent peaks common to all the other fibrillar collagens are well visible also in this case. More specifically, they are at positions 0, 235, 467, 703 and 935. In a

fashion common to all fibrillar collagens, most of the peaks in figure 4.46 are separated by the 234 a.a. distance. The average distance between related peaks is 234±4 a.a. This general behaviour is also reinforced by the attractive electrostatic scoring shown in figure 4.47. If in addition we consider the graph of both forces combined together represented in figure 4.48, we can see that the peaks at position 234, 467, 703 and 926 are the most prominent. We thus use this particular stagger to build a microfibril made of five parallel molecules.

Type XI collagen is an anomalous collagen if compared to Type I, II and III collagens because *in vivo* it retains a portion of the N-terminal propeptides. However, their extent and conformation is not well understood. We use here the PDB sequence were the N-telopeptides and the C-telopeptides are 17 and 21 residues long for both α_1(XI) and α_2(XI) chains while they are 19 and 27 a.a. long for the α_3(XI) chain. To maximise the stability of the microfibril we fold the telopeptides in such a way that their lysines, when present, face the corresponding hydroxylysines of the triple helix. We choose the particular folding configuration where Lys 5N of the first chain faces Lys 930 of the first chain (figure 4.49) belonging to a molecule with a 936 a.a. stagger. In this case, there are no lysines in the remaining N-telopeptides, for this reason, we folded them in such a way that they occupy the smallest volume. C-telopeptides folding was chosen so that Lys 17C of the first α-chain faces Lys 87 of the first α-chain of another molecule. Similarly, we made Lys 11C in the second and third α-chain to face Lys 84 of the second and third α-chain respectively of another molecule in the same microfibril (figure 4.49). In figure 4.50 we show a portion of the microfibril built in this manner. Since the projected axial length is 1038 a.a. the gap used to build the microfibril is 132 a.a. long. We calculated the density profile of this model.

In figure 4.51 we show a portion of a type XI collagen microfibril reconstituted in vitro, the part used for comparison is outlined in red. The density profile was filtered in Fourier space to reduce its resolution and enhance the periodic structures. In figure 4.52, we reproduce the density profile obtained in this way where it is possible to notice some periodic structures.

In figure 4.53, we present the comparison between the simulated density profile of the 234 a.a. stagger model and the density profile of the experimentally obtained fibril. The correspondence between most features of both profiles is apparent and confirmed by the correlation coefficient r=0.8 calculated on 1170 elements. It is nonetheless clear that some peaks that are well defined in the simulated density profile (shown by blue arrows) are only moderately visible in the density profile of the real fibril. However, this can be ascribed to the difficulty in defining a proper length and conformation for the N-terminal propeptides and to the resolution of the micrograph.

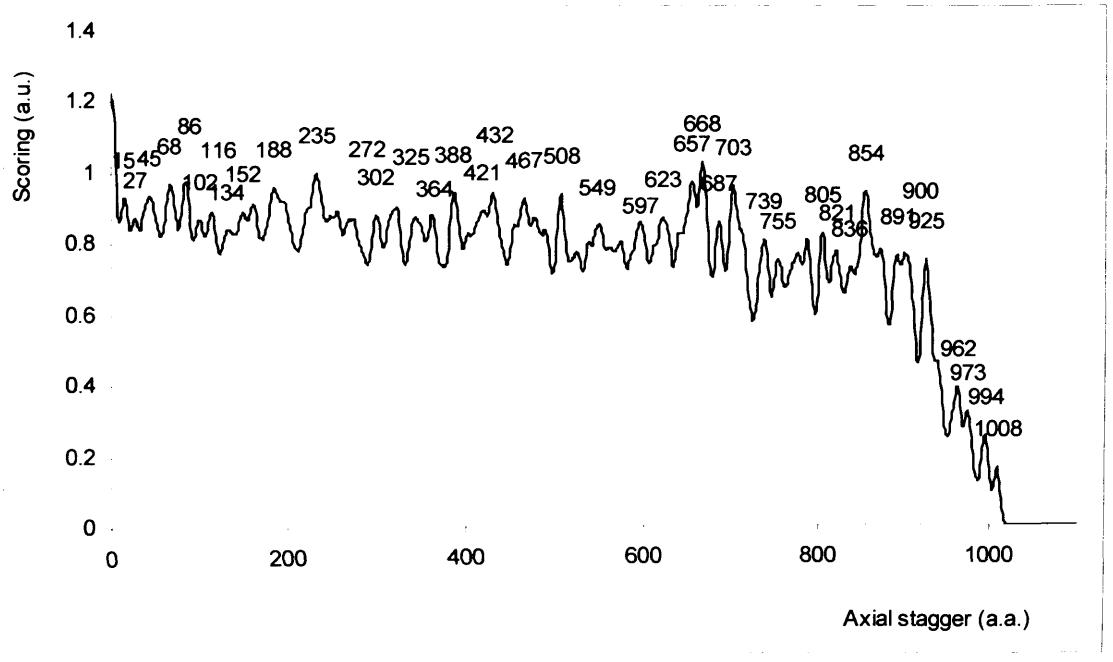


Figure 4.46: Hydrophobic scoring function for two parallel type XI collagen molecules without telopeptides.

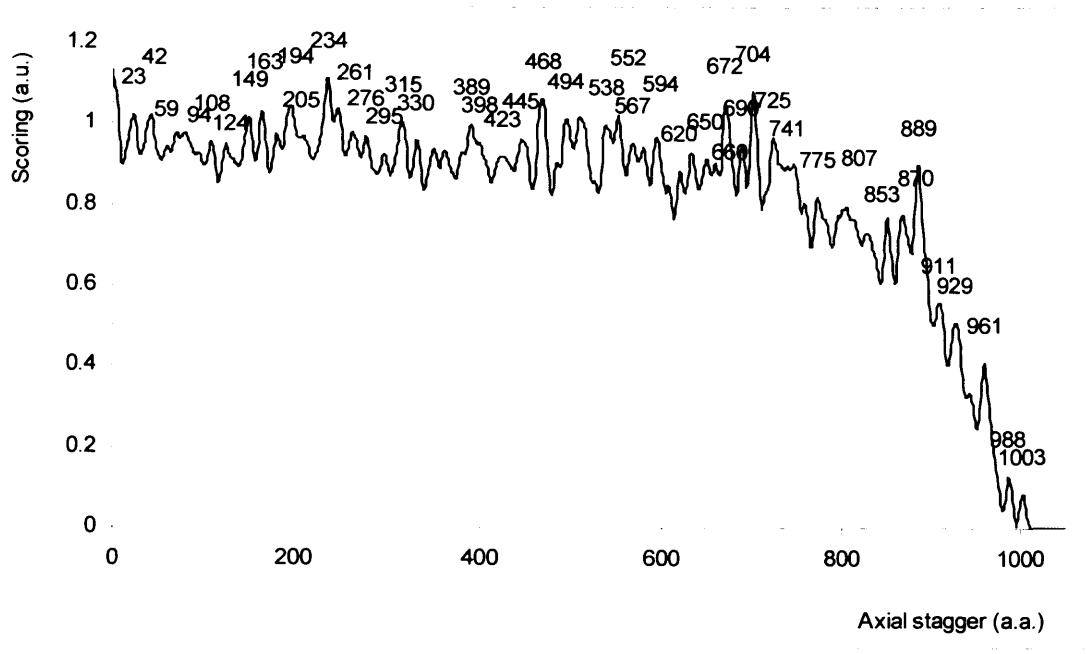


Figure 4.47: Attractive electrostatic scoring function for two type XI parallel collagen molecules without telopeptides.

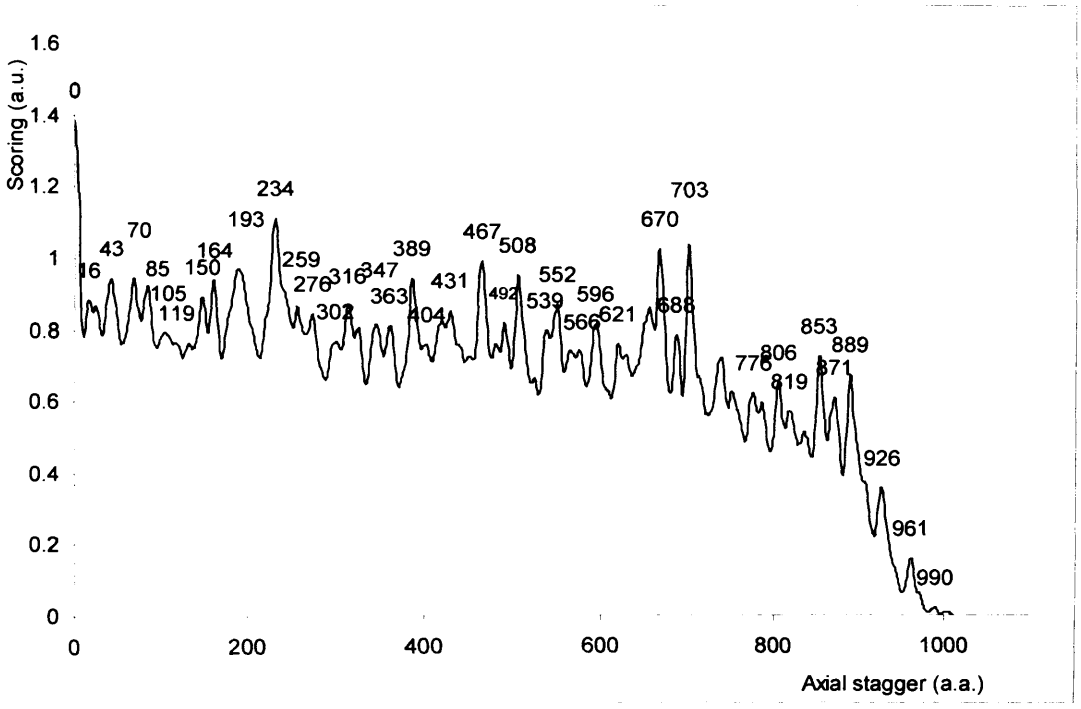


Figure 4.48 Combined attractive interaction (scoring function) for two type XI parallel collagen molecules without telopeptides.

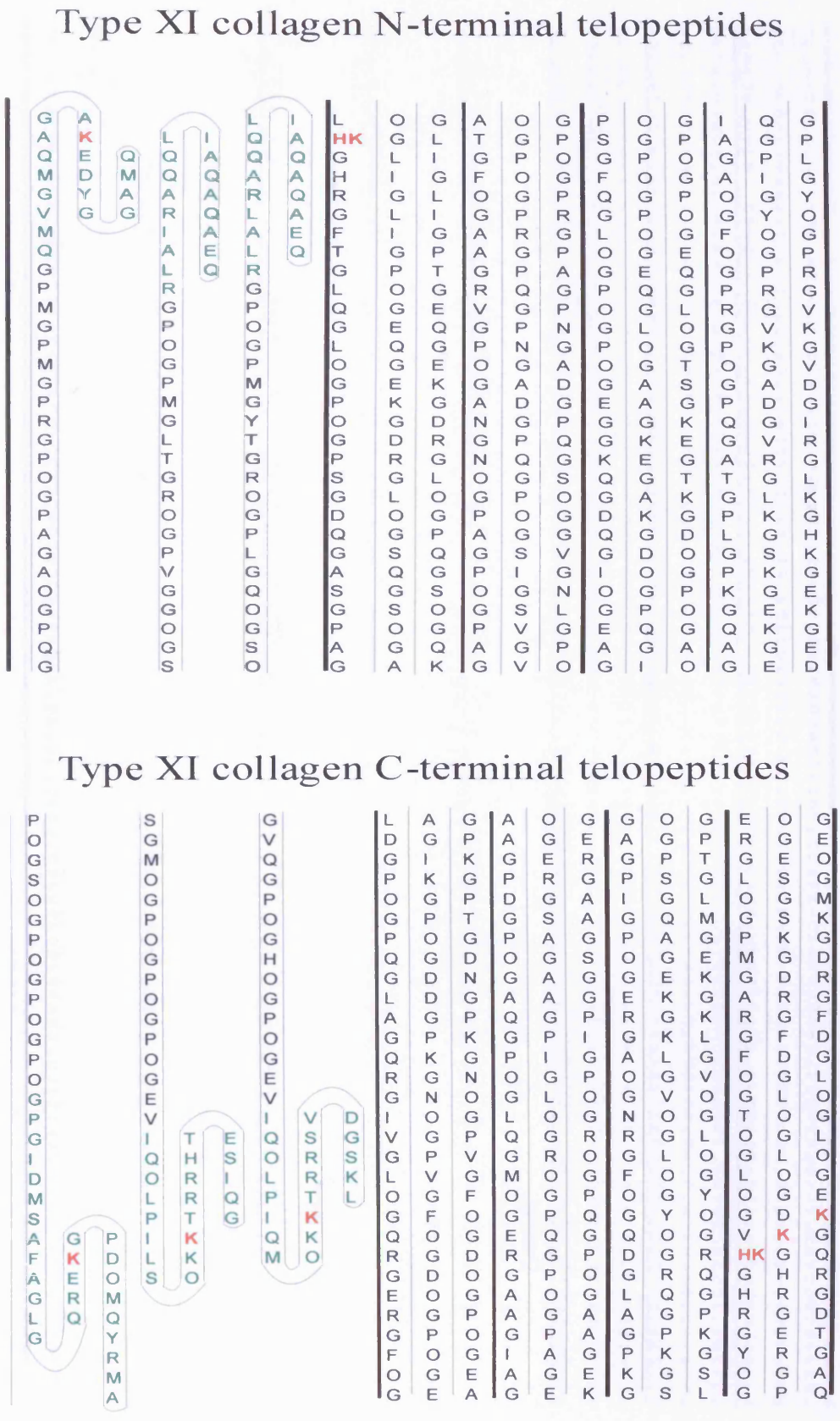


Figure 4.49: Folding configuration used for murine type XI collagen telopeptides. Lysines responsible for cross-link formations are coloured in red. Residues belonging to the telopeptides are in green.

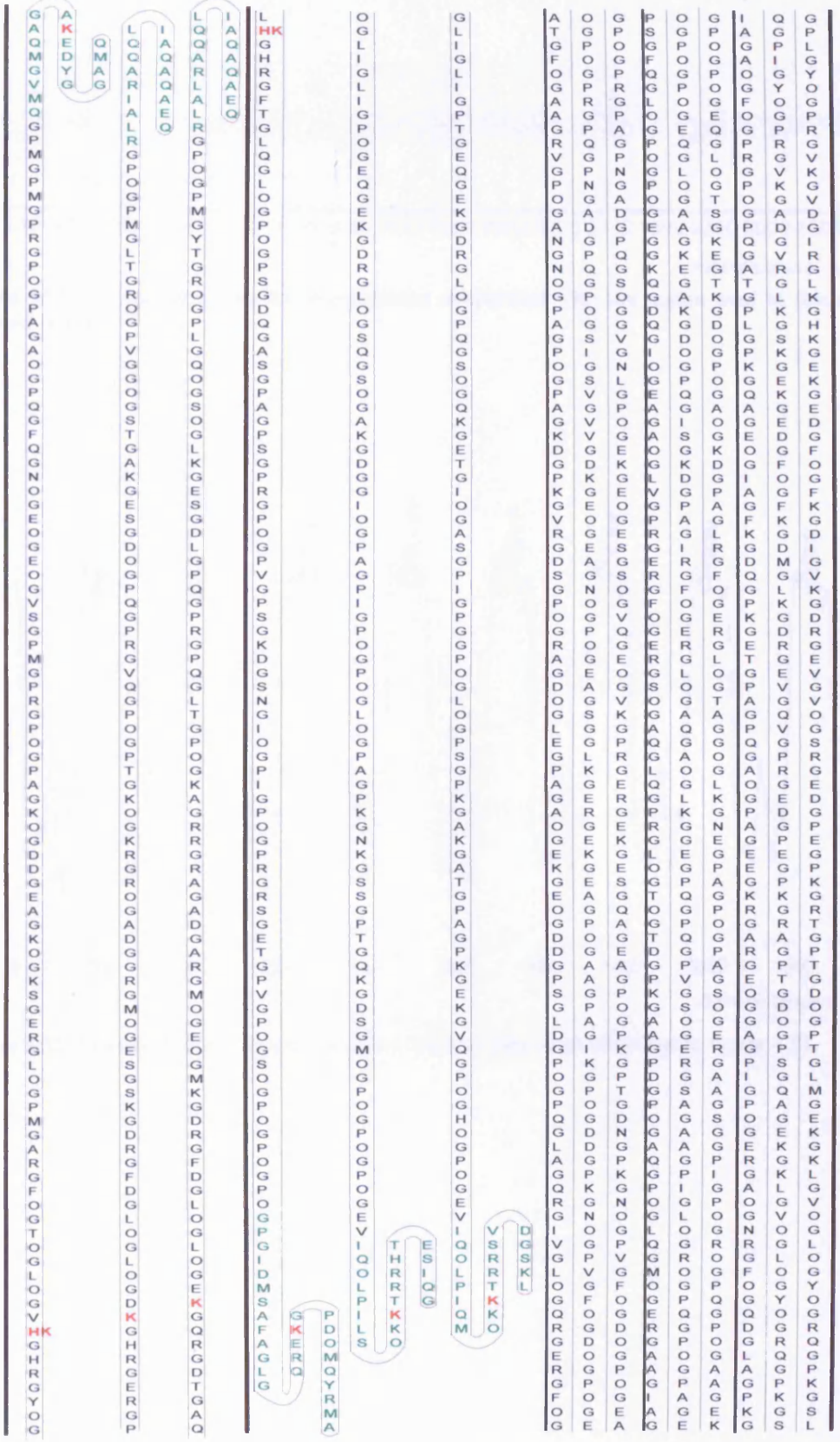


Figure 4.50: Portion of the modelled murine type XI collagen microfibril containing telopeptides. Residues belonging to the telopeptides are in green, lysines responsible for cross-links formation in red.

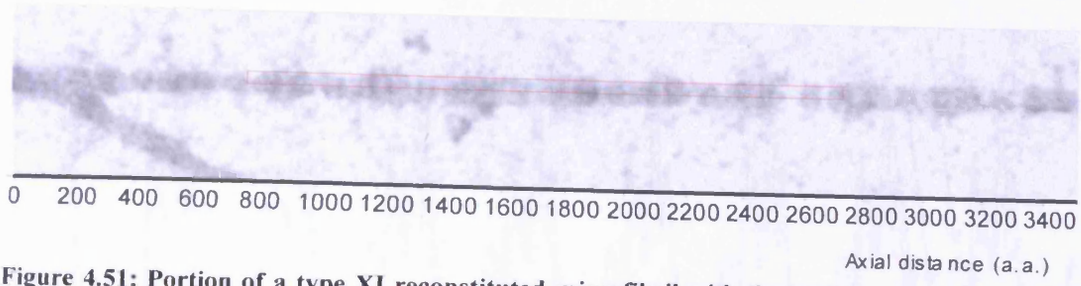


Figure 4.51: Portion of a type XI reconstituted microfibril with the region used in this study outlined in red.

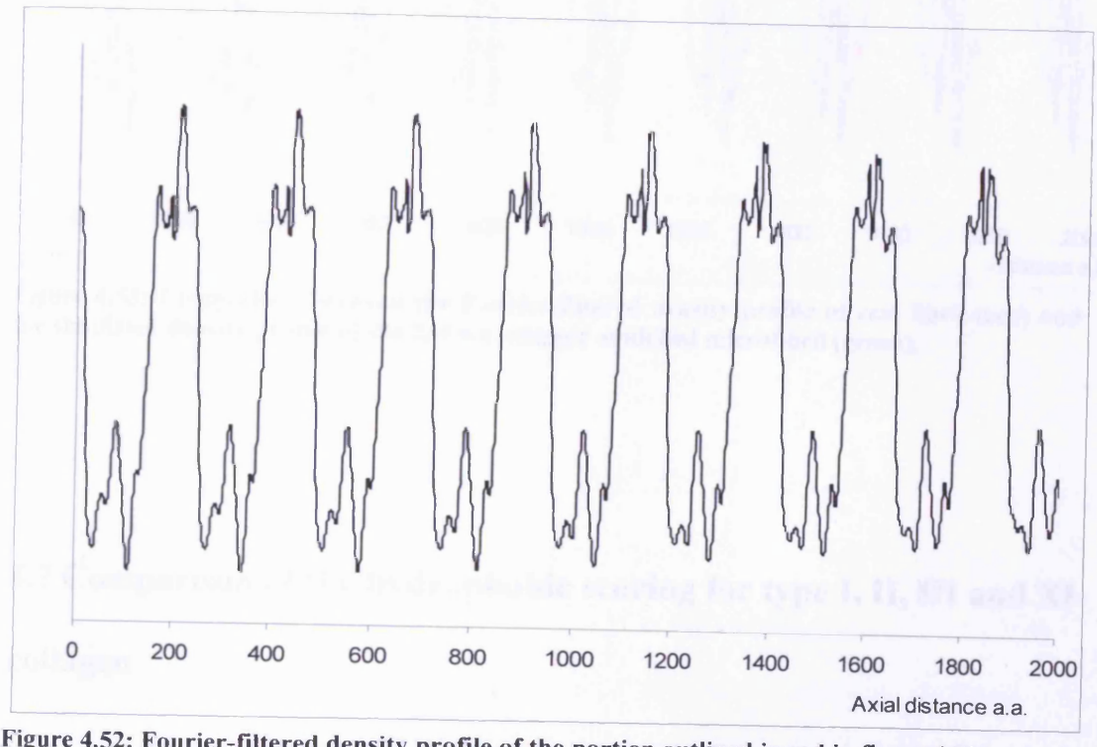


Figure 4.52: Fourier-filtered density profile of the portion outlined in red in figure 4.51.

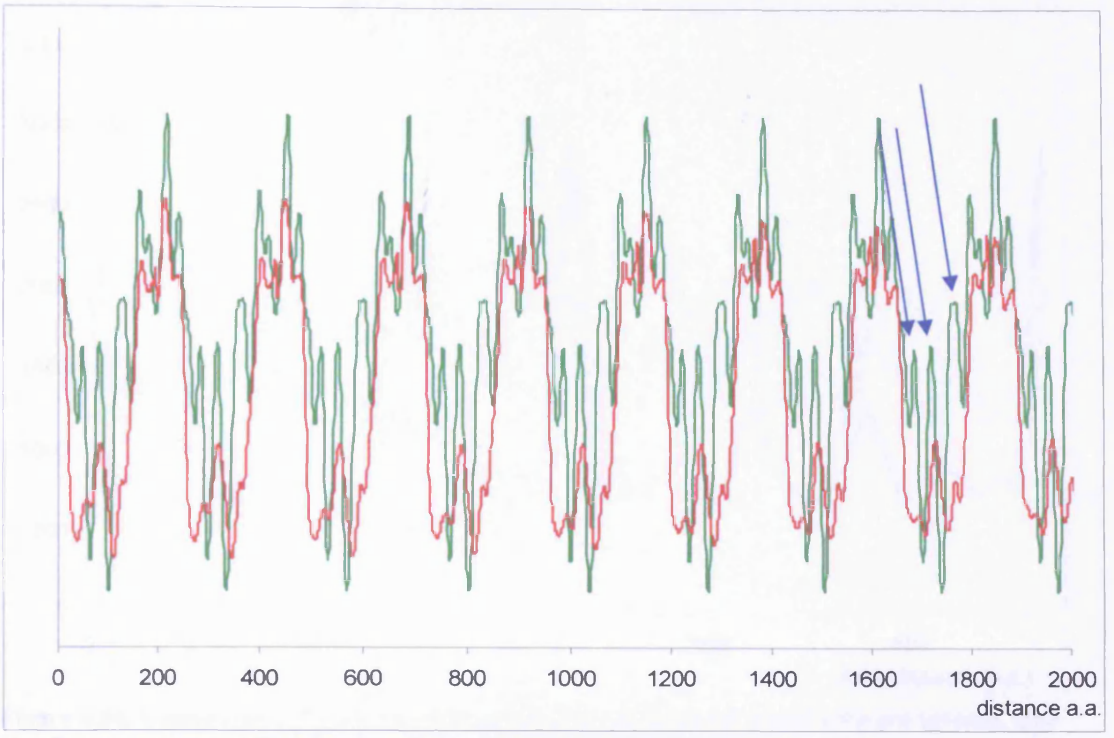


Figure 4.53: Comparison between the Fourier-filtered density profile of real fibril (red) and the simulated density profile of the 234 a.a. stagger modelled microfibril (green).

4.7 Comparison of the hydrophobic scoring for type I, II, III and XI collagen

In figure 4.54, we present the hydrophobic scoring function for type I (green), type II (red), type III (black) and type XI (blue) collagen. In this figure there is no shift along the Y-axis for the different scorings, thus the curves can be associated to the *hydrophobic energies* felt by the different collagen types during fibril formation. The scoring is 234 a.a. periodic by construction, while the general shape of the function is similar for all collagen types

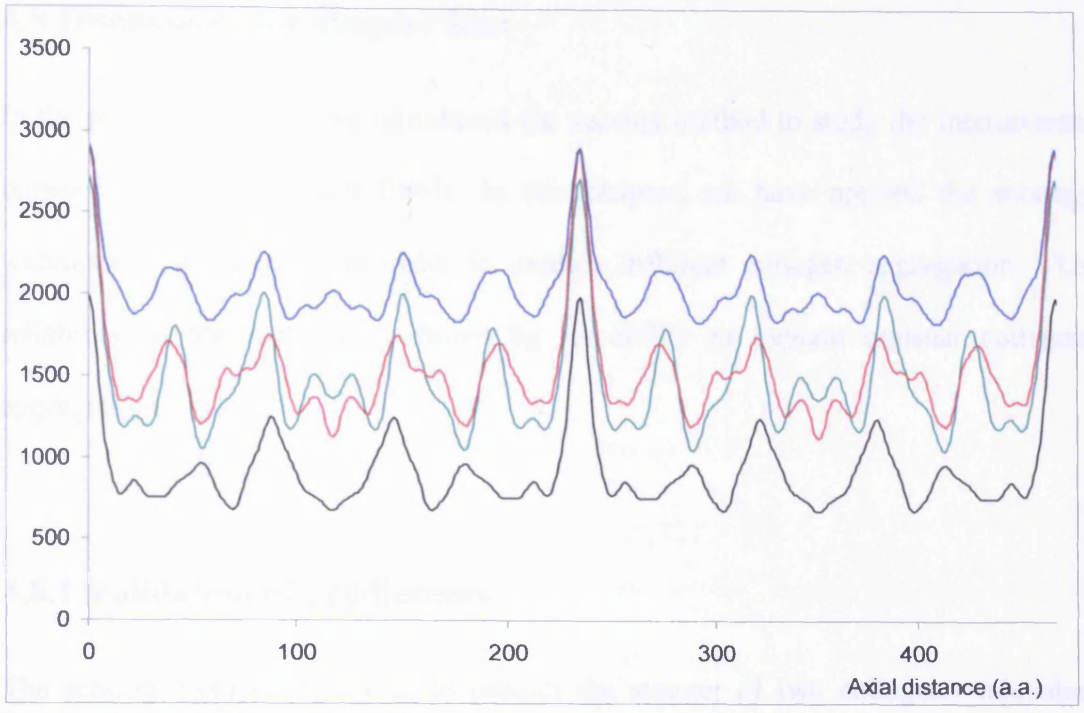


Figure 4.54: Comparison of the hydrophobic scoring function for type I collagen (green), type II collagen (red) type III collagen (black) and type XI collagen (blue).

A feature that is immediately apparent is how type XI collagen is the one with the highest scoring profile. This is somewhat unexpected because collagen I is the one expected to have the highest scoring values because of its role in the formation of fibrils (tendons, ligaments, bones) that have to endure highly stressful events such as compression, stretching and torsion. Type XI collagen is instead associated with cartilage that works as shock absorber (compression). It thus seems that the load of work that type XI collagen has to withstand is less demanding. However, type XI collagen is thought to work as a scaffold around which type II collagen aggregates and build fibres (Gelse et al., 2003). In this way, it could constitute a central rigid core that gives support to type II collagen hence the high values for the scoring function. Type I and II collagen have scoring functions that are comparable, while type III collagen, mainly found in blood vessels and skin were a certain elasticity is needed, is the one with the lowest hydrophobic scoring function.

4.8 Discussions for chapter four

In the previous chapter, we introduced the scoring method to study the interactions between adjacent collagen fibrils. In this chapter, we have applied the scoring technique systematically in order to explain different collagen aggregation. The reliability of the method is shown by its ability to explain unusual collagen aggregations

4.8.1 Validation of predictions

The scoring method allows us to predict the stagger of two collagen molecules when they pack in both parallel and antiparallel fashions. However, *in vivo* we find mainly fibrils. Thus, we used the scoring method in two ways: first, we used it to study the interactions among molecules that drive the formation of microfibrils and second, we applied the same methods to microfibrils to study how they pack together. The validation of the method was obtained by comparing the predicted fibrils with real fibrillar aggregations produced *in vitro*. This is achieved with the simulated staining method described above.

At present, it is a widely accepted idea that the fibrillar collagens pack with about 234 a.a. stagger in a parallel way (Hulmes et al., 1973; Meek et al., 1979; Parry, 1987; Suzuki et al., 1999; Wess, 2005) but no explanation, to our knowledge was put forward for antiparallel aggregations. The explanation of this is one of the major outcomes of this chapter. Another very interesting result is the explanation of an unusual oblique banding that is sometimes found in fibrils reconstituted *in vitro*. However, the more obvious result relative to 234 a.a. periodic parallel packing is

still important because it described parallel aggregation for type II, III and XI collagens that were not previously explored.

4.8.2 Antiparallel packing

To verify our type I antiparallel model, we compared it with micrographs of type I collagen antiparallel fibrils published by Bruns (1976). We built type I collagen microfibrils according to the results from the scoring method and we scored such fibrils in an antiparallel way. The hydrophobic scoring is characterised by well defined prominent peaks (figure 4.3) while the electrostatic attractive scoring is characterised by less prominent peaks (figure 4.4). However, the electrostatic attractive scoring modulates the hydrophobic scoring and enhances its peaks (figure 4.7). We chose to represent the combined interactions with a multiplication of the scoring function rather than a sum because of two reasons: 1), the scoring method tells us where the interaction is more intense but can not quantify it exactly. Thus, it does make no sense to sum two scoring functions that are not comparable. 2) We were interested to know those scoring positions where the two forces act together and this is better represented with a multiplication. Figure 4.10 confirms the validity of our model since model and real microfibril correspond to each other so well to have a cross-correlation r=0.86. In figure 4.55, we represent a scheme of the antiparallel collagen packing for type I collagen fibrils. In it, collagen molecules are yellow while the gap regions are black. The relative shift between microfibrils is that predicted by our analysis. We used the same approach to explain antiparallel aggregation in a heterotypic fibril made of type II and type III collagen in a 1:1 proportion.

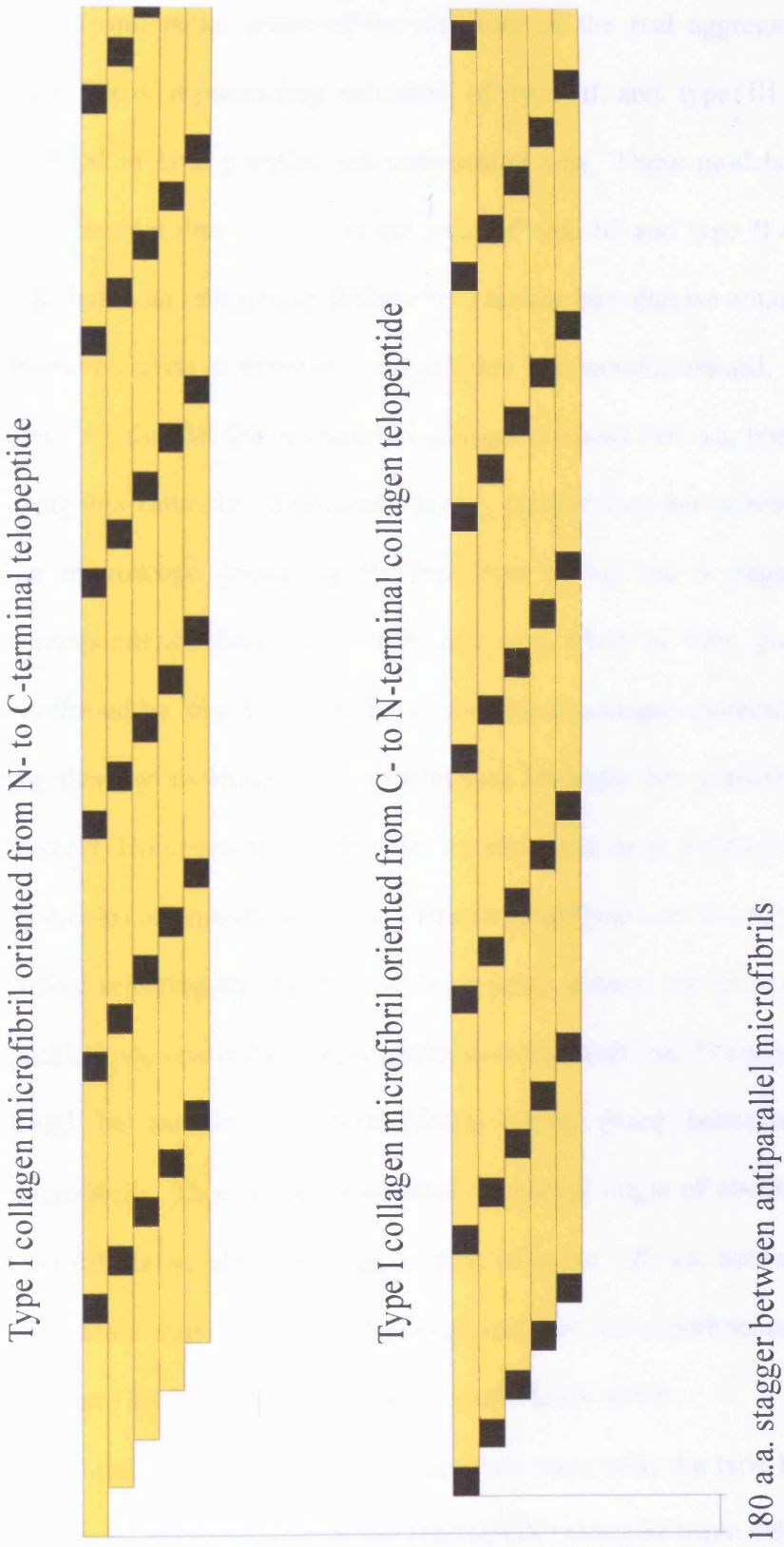


Figure 4.55: Diagram for type I collagen antiparallel aggregation.

To try and make sense of the structure of the real aggregates we built several microfibrils representing mixtures of type II and type III collagen molecules oriented in both parallel and antiparallel way. These models did not explain the experimental data, except in the case of type III and type II collagen microfibrils oriented in an antiparallel fashion that are the best that we obtained.

However, even in this case, things were not straightforward. The scoring function gives a value for the antiparallel stagger of about 200 a.a. but if we build a model using this value, its calculated staining pattern does not correspond to what seen in the microscope (figure 4.39). However if we use a stagger of 178 a.a., the correspondence between model and real fibril is very good (figure 4.40) as confirmed by its r=0.9. This could mean that collagen molecules are forced to pack together by molecules (proteoglycans) that are not considered by our scoring method. However, the microfibrils were made from purified preparations, and no molecules or proteins should be present apart from type II and type III collagen.

Before rejecting the validity of the scoring method for its inability to explain this packing, we saw if there were alternative explanations. We postulated that the offset could be ascribed to supercoiling taking place between the two different microfibrils. Thus if we hypothesise a coiling angle of about 8° between the two microfibrils we obtain a stagger value of about 120 a.a. between microfibrils. This produces a staining pattern that explains well the experimental data. As confirmed by figure 4.45 and by a correlation coefficient r=0.83.

In addition, it is also worth noticing that unusually, the type III collagen molecule has twenty imperfections of the kind (GGY) along its triple helix, most of which are concentrated in the centre of the molecule. This feature could influence the length, possibly through kinks, of the type III collagen molecule. This would alter the

results from our scoring method because of an altered packing structure. However, these are all hypotheses that should be tested by other methods.

4.8.3 Diagonal banding

During *in vitro* fibril formation, collagen aggregates can create fibrils with bands that are oblique with respect to the fibril axis. These bands have a D periodicity and are interspersed with minor bands orthogonal to the axis, this is shown in figure 4.17 for type II collagen. This suggests that the fibril is likely to be made by the usual elemental structures (the microfibrils or groups of microfibrils) that are displaced one with respect to the other by a distance less than D. The combined hydrophobic and electrostatic interactions can easily explain this unusual state of aggregation. The graph in figure 4.21 shows the combined interactions as a function of the distance. We have the usual prominent peak at 234 a.a. and a peak at 39 a.a.. The 234 a.a. stagger is the one that explains the dark bands that are D periodic.

If we pack the microfibrils, which are built with an intrinsic 234 a.a. periodicity, with a 39 a.a. stagger, the 234 a.a. periodic dark bands acquire an oblique appearance. This is clearly seen in figure 4.24 where the model and the real fibril have cross correlation coefficient r=0.79. In addition, this stagger allows cross-link formation (figure 4.22). Moreover, a 39 a.a. stagger is such that molecules belonging to the first and the seventh microfibrils with such a stagger are again in register. This allows the usual cross-links that we find among the 234 a.a. staggered parallel molecules. It seems thus natural to postulate that oblique banding can be formed only between microfibrils whose stagger is a submultiple of 234 a.a..

Figure 4.56: Comparison between a real collagen fibril with an oblique banding and the modelled fibril made of parallel microfibril with a 39 a.a. stagger.

Figure 5.56 shows an idealised representation of type II collagen oblique banding. In it, every molecule is yellow and the gap region is black. Five molecules join to form microfibrils that have 39 a.a. stagger with respect to each other. It is possible to notice the formation of diagonal bands with 234 a.a. stagger by construction. The seventh microfibril is again in register with the first one reinforcing the interaction among microfibrils. A portion of the real fibril with oblique banding is also shown for comparison. The bands of the real and the modelled fibrils do not have the same inclination only because they are not to scale.

4.8.4 Parallel fibril aggregation

We applied the scoring method combined with the simulated staining technique, to explain parallel fibril formation of type II, III and XI collagen. The main characterising feature is that they have hydrophobic scoring graph (figure 4.11, 4.26 and 4.46) where the dominant peaks are allocated at about 0 a.a. plus integer multiple of about 234 a.a. In addition, it is possible to notice a mirror-like distribution of the peaks within sub-regions that are 234 a.a. long. Another striking feature is that all the main peaks have a corresponding peak 234 a.a. away in an adjacent region. This characterises the typical 234 a.a. periodicity of the fibrillar collagens. Thus, we built fibrils that explain the experimental data. In particular, the correlation factor for the type II collagen fibril and its model is r=0.82, while the correlation factor for type XI collagen and its model is r=0.8. It is however necessary to point out that some features of the calculated staining pattern, for type XI collagen, are different from those found in the real fibril. This could be due to

the poor resolution of the image we analysed or to the difficulty in defining the correct folding for the N-terminal propeptide.

The comparison between the model and the real microfibril for type III collagen has a correlation r=0.69 (figure 4.30). This is still acceptable but it is the lowest obtained in our modelled fibrils. This could be due to the contribution of the telopeptides. In fact, the main incongruence between model and real microfibril, shown by arrows in figure 4.30, corresponds to the positions of the telopeptides. However at this point the model can not be improved for a lack of amino acids sequence data.

4.9 Concluding remarks for chapter 4

Chapter 4 validated the scoring method by comparing collagen fibrils models built from scoring functions predictions and real collagen fibrils by means of the simulated staining technique. These two methods combined together are a powerful tool to explain collagen packing but they can be extended to explain the packing of any fibrillar protein. The main results of this part were:

1. Collagen molecules aggregate with a 234 a.a. stagger to form parallel fibrils. This forces the telopeptides to have a limited set of configurations.
2. The scoring method can predict antiparallel packing stagger for homotypic collagen fibrils.
3. The stagger for heterotypic antiparallel fibrils (type III/type II collagen) can be explained if we postulate a 8° supercoiling angle between type III and type II collagen fibrils.
4. The scoring method can explain diagonal banding aggregations during parallel fibril formation.
5. The scoring method combined with the simulated staining describes the packing of microfibrils.

CHAPTER 5

CONCLUSIONS

Final remarks

In this thesis, we presented a study on fibrillar collagens packing. Collagen is the most abundant protein in vertebrates constituting more than 25% of the total protein mass (Knupp and Squire, 2003), and fibrillar collagens represent 90% of all collagens (Gelse et al., 2003). They are found in bone, dermis, tendon (type I collagen), cartilage, intervertebral disc (type II and type XI collagen) skin, blood vessel (type III collagen), lung, cornea and bone (type V collagen). Fibrillar collagens have a further property that is essential for the work presented here: from a structural point of view, they are relatively simple. By this, we mean that they can be represented as linear structures and this is fundamental for the analytical methods we have presented here. The scoring method, in principle, can be applied not only to fibrillar collagens but also to all proteins that can be represented as linear structures. In fact, the hydrophobic and electrostatic interactions arising between two such proteins can be represented by means of a correlation. This simple representation is effective because it allows a concise description of the major forces that contribute to the formation of fibrils as a function of the mutual displacement. The results are very interesting and could be used to improve our understanding of the role of hydrophobic and electrostatic forces during protein folding in general.

We used murine type I collagen as model protein to test the method. One important result is that the role of telopeptides resulted to be fundamental. In fact, once they were taken into account they proved to be determinant in explaining the D-stagger between collagen molecules. The hydrophobic residues in the telopeptides reinforce the regularity with which hydrophobic clusters are distributed along the collagen

molecule. This distribution is D-periodic and each period is made of five clusters that are equally spaced with a 42 a.a. interval between them followed by a gap about 66 a.a. long. This distribution is the same whether it is read right to left or left to right. The immediate consequence is that one can postulate an antiparallel aggregation in collagen fibrils. This can be confirmed with the scoring method that is able to predict the correct stagger for antiparallel aggregations. In addition, the hydrophobic cluster sequence can be used to identify those amino acids that contribute the most to collagen packing and, in the case of type I collagen, they are methionine, tyrosine and phenylalanine.

We wish to point out here that once a linear sequence of amino acids representing a protein is created, the method can be applied systematically giving results that can be interpreted immediately. The steps that need to be implemented to carry out an analysis are the following:

1. Represent the protein as a linear sequence of values associated to hydrophobic or electrostatic properties of the amino acids
2. Small non-linear domains, like the telopeptides in the case of collagen, can be projected onto the linear sequence to make the results more realistic
3. Calculate the correlation function of the sequences representing the proteins
4. Smooth and normalise the cross-correlation function to enhance the most prominent peaks
5. Periodic patterns within the cross-correlation function tell us about possible periodicities in the fibrillar aggregation
6. Find the periodic distribution of scoring function. These are responsible for fibrillar aggregation

7. Use of the periodic patterns to build models of microfibrils that represent the basic structure of the fibril

When we applied this algorithm to type I collagen antiparallel molecules, a 234 a.a. periodic pattern was found. In addition, the most prominent peak was at 87 a.a.. This means that antiparallel collagen molecules aggregate with a 234 a.a. stagger and with an initial offset of 87 a.a. between molecules (or 181 a.a. between collagen microfibrils). To validate the predictions we used a simulated staining technique (Meek et al., 1979; Ortolani et al., 2000) to represent models of collagen microfibrils arranged according to the predicted stagger. The comparison ($r \approx 0.86$) between the modelled antiparallel microfibrils and the real antiparallel type I collagen microfibril confirmed the validity of the method.

The scoring algorithm was applied also to heterotypic (type II/type III) collagen. In this case, the results were not straightforward. Two interpretation could be put forward: 1) there is a third set of molecules (e.g. proteoglycans) that contribute to collagen packing but that are not taken into account in the scoring method; 2) there is supercoiling between antiparallel type II and type III collagen microfibrils at an about 8° angle. The fact that we can suggest this second hypothesis means that the scoring method can give insights not only on linear collagen aggregations, but also on its three-dimensional packing; a remarkable result from a method based only on one-dimensional linear sequences.

The fitness of the algorithm was proved once more by explaining an unusual collagen aggregation. The case of type II collagen diagonal banding was explained very well by our method that predicts an alternative 39 a.a. stagger between parallel collagen microfibrils. In addition, it suggests that collagen can pack using stagger

values that are submultiples of 234 a.a. because molecules can go back in register reinforcing the attractive interactions. Finally, we applied the algorithm to a more "classical" parallel fibril packing such as that in type II, type III and type XI collagens.

Our approach was restricted here only to the study of collagen fibrils. However, it allowed us to put forward models and to validate them and to describe, explain and predict different collagen packing schemes. We think that the method can be extended to any protein that can be represented linearly. This is another step toward the explanation of protein packing and of protein folding.

Further steps

Since the process we presented here is simple, in the near future it would be interesting to write software application that receives as an input some sequences of amino acids of linear proteins and gives as an output the ideal packing configuration for them.

The application of this software to heterotypic collagen fibrils in parallel and antiparallel configurations could help us to improve the method and to extract more information on the three-dimensional structure of fibril packing.

Fibrillar collagens type V, XXIV and XXVII were not considered in this thesis for lack of experimental data. It would be therefore interesting to extend our method to these collagens to see how well it can describe their association properties.

APPENDIX

Appendix

In this appendix, we present two programs that were written for this thesis: cross_correlation.cpp and interpolation.cpp. In addition, a program written by Dr. Knupp (Fast_Fourier_Transform.cpp) is presented because it was used to calculate the Fourier Transform of the scoring functions analysed in this work. The classes fft.h and arctan.h are necessary for its operation.

The operation of the cross_correlation.cpp program is simple. It is necessary to give as an input the elements of the sequences in txt format (a column of elements of the type 0, 1, 0, 0, 1,..., for example). The program works sliding a sequence (shifting_sequence.txt) along a stationary sequence (stationary_sequence.txt) and calculates the sum of the values obtained multiplying two facing elements. The output (output.txt) is the cross correlation as a function of the shift between the two sequences.

The interpolation.cpp program is more precisely a *compression program*. It permits to express a sequence of elements (sequence.txt) previously written as a function of integer numbers, whose difference between two consecutive numbers is 1, such as 0, 1, 2, 3, ..., as a function of numbers whose difference between two consecutive numbers is less than 1 (compressed_abscissae.txt). The result is a sequence (compressed_sequence.txt) as a function of the compressed abscissae.

The Fast_Fourier_Transform.cpp program is an interactive program that asks the user the number of points to calculate the Fourier Transform on. It then asks an input sequence and gives as an output the power of the Fourier Transform of the sequence and its real and imaginary components.

Cross_correlation.cpp

```
/* scoring algorithm calculates the cross-correlation algorithm of a shifting sequence c[N]
with a stationary sequence d[N] and gives the output y[N] as a function of the shift j
between the two sequences.*/

#include <iostream>
#include <fstream>
using namespace std;

const int N =1014;  //length of the sequence analysed

int main(){

int j;
int i;
double  c[N]={0};   // in1    mobile sequence shifting top
double  d[N]={0};   // in2    stationary sequence

double  y[N]={0};  // output


ifstream in1("shifting_sequence.txt"); //c[]
ifstream in2("stationary_sequence.txt"); //d[]

        ofstream out("output.txt");


  if(!in1){cout << "Cannot open file:in1\n"; return 1;}
  if(!in2){cout << "Cannot open file:in2\n"; return 1;}


for(i=0;i<N;i++){ in1  >> c[i];} in1.close();

for(i=0;i<N;i++){in2  >> d[i];} in2.close();


////////////////////////////////

for(j=0;j<N;j++){

        for(i=0;i<N;i++){

                        if ((i-j) >= 0 )
                                y[j]=y[j] + c[i]*d[i-j];

                }
        }
```

```
for(j=0;j<N;j++)
        out<<y[j]<<endl;
out.close();


 return 0;}
```

Interpolation.cpp

/*This algorithm takes a sequence as a function a[N] of integer numbers (0, 1, 2, 3, ...) and expresses it as a function y[N] of compressed abscissae (0, 0.8, 1.6, 2.4, ...) c[N]. The sequence of compressed abscissae must be given by the program's user */

```
#include <iostream>
#include <fstream>
#include <iomanip>
using namespace std;

const int N =1000;  Number of the elements of the sequence

int main(){


int i,j;

double a[N]={0};   // in1
double c[N]={0};   // in3
double y[N]={0};


   ifstream in1("sequence.txt");  // sequence to be compressed
   ifstream in3("compressed_abscissae.txt");  // compressed abscissae to be used as
                                              // new abscissae for the compressed sequence
   ofstream out("compressed_sequence.txt");   // compressed sequence.

  if(!in1){cout << "Cannot open file:in1\n"; return 1;}

  if(!in3){cout << "Cannot open file:in3\n"; return 1;}

  for(i=0;i<N;i++){in1  >> a[i];} in1.close();
  for(i=0;i<N;i++){in3  >> c[i];} in3.close();

////////////////////////////////
  for(i=1;i<N;i++){
          y[0]=a[0];

          y[i]=a[(int)c[i]]+(a[(int)c[i]+1]-a[(int)c[i]])*(c[i]-(int)c[i]);}


for(j=0;j<N;j++)
        out<<setw(10)<<c[j]<<'\t'<<y[j]<<endl;
out.close();

  return 0;}
```

Fast_Fourier_Transform.cpp

```
#include<iostream.h>
#include<fstream.h>
#include<math.h>
#include <string.h>
#include"arctan.h"
#include"fft.h"

char filetarg[80];


int main(int argc, char* argv[]){

        int m=13;
        cout << "Please enter m. 2^m points will be calculated in the transform: ";
        cin >> m;

        double const pi = 3.14159265359;

        const int N= int(pow(2,m));
        double *A , *B;


        A=new double [N];
        B=new double [N];




        int i=0, counter=0;

        for(i=0;i<N;i++){
                A[i]=0;
                B[i]=0;

        }


        if(argc != 2){
                cout << "Please type filename: ";
                cin >> filetarg;
        }else{
                strcpy(filetarg, argv[1]);
        }

        ifstream datain;
        datain.open(filetarg);
```

```
        strcat(filetarg,"_FFT.txt");
        ofstream dataout_plot;

        dataout_plot.open(filetarg);


        while(!datain.eof() && counter < N){
                datain >> A[counter];
    // datain >> B[counter];
                B[counter]=0;
                ++counter;
        } counter--;
        datain.close();

        FFT(1,m,A,B);

        for(i=0; i<N; i++){

                //dataout_plot << A[i] * A[i] + B[i] * B[i]<< '\t' << arctan(A[i],B[i]) <<
endl;
    dataout_plot << A[i] * A[i] + B[i] * B[i]<< '\t' << A[i] << '\t' << B[i] << endl;

        }



//      FFT(-1,m,A,B);

//      for(i=0; i<N; i++){

//              dataout_plot << A[i] << '\t' << B[i] << endl;
//      }




        dataout_plot.close();

        delete [] A;
        delete [] B;



        return 0;
}
```

Fft.h

```
#ifndef FFT_H
#define FFT_H
/*
  This computes an in-place complex-to-complex FFT
  x and y are the real and imaginary arrays of 2^m points.
  dir =  1 gives forward transform
  dir = -1 gives reverse transform
*/
short FFT(short int dir,long m,double *x,double *y)
{
  long n,i,i1,j,k,i2,l,l1,l2;
  double c1,c2,tx,ty,t1,t2,u1,u2,z;

  /* Calculate the number of points */
  n = 1;
  for (i=0;i<m;i++)
    n *= 2;

  /* Do the bit reversal */
  i2 = n >> 1;
  j = 0;
  for (i=0;i<n-1;i++) {
    if (i < j) {
      tx = x[i];
      ty = y[i];
      x[i] = x[j];
      y[i] = y[j];
      x[j] = tx;
      y[j] = ty;
    }
    k = i2;
    while (k <= j) {
      j -= k;
      k >>= 1;
    }
    j += k;
  }

  /* Compute the FFT */
  c1 = -1.0;
  c2 = 0.0;
  l2 = 1;
  for (l=0;l<m;l++) {
    l1 = l2;
    l2 <<= 1;
    u1 = 1.0;
    u2 = 0.0;
    for (j=0;j<l1;j++) {
```

```
      for (i=j;i<n;i+=l2) {
        i1 = i + l1;
        t1 = u1 * x[i1] - u2 * y[i1];
        t2 = u1 * y[i1] + u2 * x[i1];
        x[i1] = x[i] - t1;
        y[i1] = y[i] - t2;
        x[i] += t1;
        y[i] += t2;
      }
      z =  u1 * c1 - u2 * c2;
      u2 = u1 * c2 + u2 * c1;
      u1 = z;
    }
    c2 = sqrt((1.0 - c1) / 2.0);
    if (dir == 1)
      c2 = -c2;
    c1 = sqrt((1.0 + c1) / 2.0);
  }

  /* Scaling for forward transform */
  if (dir == 1) {
    for (i=0;i<n;i++) {
      x[i] /= n;
      y[i] /= n;
    }
  }

  return(0);
}

#endif
```

Arctan.h

```
#include<cmath>


double arctan(double cos, double sin); //returns arctan(sin / cos)- careful cos comes before
sin in the function call
int sign(double x); // returns sign of x

double arctan(double C, double S)
{

        double Tn, Ra;



                if (S == 0 && C > 0)  Tn = 0;
                if (S == 0 && C < 0)  Tn = 3.14159265359;
                if (S < 0 && C == 0)  Tn = 3 * 3.14159265359 / 4;
                if (S > 0 && C == 0)  Tn = 3.14159265359 / 2;
                if (S == 0 && C == 0)  Tn = -99999;
                        else{

                                Ra = (S / C);


                                if (sign(S) > 0 && sign(C) > 0) Tn = atan(Ra);
                                if (sign(S) > 0 && sign(C) < 0) Tn = atan(Ra) +
3.14159265359;
                                if (sign(S) < 0 && sign(C) < 0) Tn = atan(Ra) +
3.14159265359;
                                if (sign(S) < 0 && sign(C) > 0) Tn = atan(Ra) + 2 *
3.14159265359;
                        }

        return  180 / 3.14159265359 * Tn;
}


int sign (double x)
{
        if (x>=0)
                return 1;
                else return -1;
}
```

References

Articles and books

Adzhubei A.A. and Sternberg M.J.E. Left-handed Polyproline II Helices Commonly Occur in Globular Proteins. J. Mol. Biol. (1993), 229: 472-493.

Bailey A.J. Molecular mechanisms of ageing in connective tissues. Mech. Age. Develop. (2001) 122: 735-755.

Banyard J., Bao L., and Zetters B.R. Type XXIII Collagen a New Transmembrane Collagen Identified in Metastatic Tumor Cells. J Biol. Chem. (2003), Vol. 278, No. 23: 20989-20994.

Beck K. and Brodsky B. Supercoiled Protein Motifs: The Collagen Triple-Helix and the α-helical Coiled Coil. J. Struct. Biol. (1998) 122: 17-29.

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

Bender E., Silver F.H., Hayashi K. and Trelstad R.L. Type I Collagen Segment Long Spacing Banding Patterns. Evidence that the α2 chain is in the reference position. J. Biol. Chem. (1982) Vol. 257, No. 16: 9653-9657.

Birk D.E. Type V collagen: heterotypic type I/V collagen interactions in the regulation of fibril assembly. Micron (2001) 32: 223-237.

Bozzola JJ and Russell LD. Electron Microscopy. Principles and Techniques for

Biologists. (1992) Jones and Bartlett Publishers. Boston

Bracewell RN. The Fourier Transform and its Applications 3rd ed. (2000) McGraw-Hill International Editions

Branden C. and Tooze J. Introduction to Protein Structure 2nd ed (1995) Garland Publishing Inc. New York.

Brodsky B. and Ramshaw J.A.M. The collagen Triple-Helix Structure. Matrix Biol. (1997), Vol. 15: 545-554.

Bruns R.R. Supramolecular Structure of Polymorphic Collagen Fibrils. J. Cell Biol. (1976) Vol. 69: 521-538.

Cantor C.R. and Schimmel P.R. Biophysical Chemistry Part I: The conformation of biological macromolecules, (1980) W. H. Freeman and Company.

Chan V.C., Ramshaw J.A.M., Kirpatrick A., Beck K. and Brodsky B. Positional Preferences of Ionizable Residues in Gly-X-Y Triplets of the Collagen Triple-helix. J Biol. Chem. (1997) Vol. 272, No. 50: 31441-31446.

Chapman J.A., Holmes D.F., Meek K.M. and Rattew C.J. Electron-Optical Studies of collagen fibril assembly. in: Structural Aspects of Recognition and Assembly in Biological Macromolecules. (1981) M. Balaban, J.L. Sussman W. Traub and A. Yonath. Balaban ISS, Rehovot and Philadelphia pp. 387-401.

Chen M., Keene D.R., Costa F. K., Tahk S.H. and Woodley D.T. The Carboxyl Terminus of Type VII Collagen Mediates Antiparallel Dimer Formation and Constitutes a New Antigenic Epitope for Epidermolysis Bullosa Acquisita Autoantibodies. J Biol. Chem. (2001) Vol. 276, No. 24: 21649-21655.

Creighton T.E. Proteins. Structure and Molecular Properties 2nd ed. (1997). W. H. Freeman and Company.

Dalgleish R. The human type I collagen mutation database. Nucleic Acids Res. (1997) Vol. 25, No. 1:181-187.

Darnell J., Lodish H., Baltimore D. Molecular Cell Biology 2nd ed. (1990) Scientific American Books.

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

Eyre D.R. Collagen of articular cartilage. Arthtitis Res. (2002) 4: 30-35.

Eyre D.R. Collagens and Cartilage Matrix Homeostasis. Clin. Orthop. Relat. Res. (2004) No. 427s: s118-s122

Eyre D.R., Weis M.A. and Wu J. Articular Cartilage Collagen: an Irreplaceable Framework? Eur. Cell. Mater. (2006) Vol. 12: 57-63.

Gelse K., Poschl E., Aigner T. Collagens-structure, function, and biosynthesis. Adv. Drug Deliver. Rev. (2003) 55: 1531-1546.

Guex, N. and Peitsch, M.C. Swiss-Model and the Swiss-PdbViewer: An environment for comparative protein modelling. Electrophoresis (1997) 18: 2714-2723.

Gustavsson K.M. The Architecture and Formation of Collagen in: Chemistry and Reactivity of Collagen. Academic Press (1955) New York.

Hayat MA. Principles and techniques of Electron Microscopy. Biological

Applications 3rd ed. (1989) The Macmillan Press LTD.

Hashimo T., Wakabayashi T., Watanabe A., Kowa H., Hosoda R., Nakamura A., Kanazawa I., Arai T., Takio K., Mann D.M.A. and Iwatsubo T. CLAC: a novel Alzheimer amyloid plaque component derived from a transmembrane precursor, CLAC-P/collagen type XXV. EMBO J. (2002) Vol. 21 No. 7: 1524-1534.

Henkel W. Cross-link analysis of the C-telopeptide domain from type III collagen. Biochem. J. (1996) 318: 497-503.

Hessa T., Kim H., Bihlmaier K., Lundin C., Boekel J., Andersson H., Nilsson I, Hite S.H. and Heijne G. Recognition of transmembrane helices by the endoplasmic reticulum translocum. Supplementary data. Nature (2005) 433: 377-381.

Hulmes D.J.S., Miller A., Parry D.A.D., Piez K.A. and Woodhead-Galloway J. Analysis of the Primary Structure of Collagen for the Origins of Molecular Packing. J. Mol. Biol. (1973) 79: 137-148.

Jones E.Y. and Miller A. Analysis of Structural Design Features in Collagen. J. Mol. Biol. (1991) 218: 209-219.

Kadler K.E., Holmes D. F. Trotter J.A., Chapman J.A. Collagen fibril formation. Biochem. J. (1996) 316: 1-11.

Knupp C., Chong N.H.V., Munro P.M.G., Luthert P.J. and Squire J.M. (2002(a)) Analysis of the collagen VI assemblies associated with Sorby's Fundus Distrophy. J. Struct. Biol. (2002) 137: 31-40.

Knupp C., Amin S. Z., Munro P.M.G., Luthert P.J. and Squire J.M. (2002(b)) Collagen VI assemblies in Age related Macular Degeneration. J. Struct. Biol.

(2002)139, 181-189.

Knupp C. and Squire J.M. Molecular Packing in Network-Forming Collagens. The scientific World JOURNAL (2003) 3: 558-577.

Knupp C. Pinali C. Munro P.M. Gruber H.E. Sherratt M.J. Baldock C. and Squire J.M. Structural correlation between collagen VI microfibrils and collagen VI banded aggregates. J. Struct. Biol. (2006) Vol. 154, Issue 3: 312-326.

Kramer R.Z., Bella J., Mayville P., Brodsky B. and Berman H.M. Sequence dependent conformational variations of collagen triple-helical structure. Nat. Struct. Biol. (1999) Vol. 6, No. 5: 454-457.

Kramer R.Z., Venugopal M.G., Bella J., Mayville P. Brodsky B. and Berman H.M. Staggered Molecular Packing in Crystals of a Collagen-like Peptide with a single Charged Pair. J. Mol Biol. (2000) 301: 1191-1205.

Kramer R.Z., Bella J., Brodsky B. and Berman H.M. The Crystal and Molecular Structure of a Collagen-like Peptide with a Biologically Relevant Sequence. J. Mol. Biol. (2001) 311: 131-147.

Kvansakul M., Bogin O., Hohenester E., Yayon A. Crystal structure of the collagen α1 (VIII) NC1 trimer. Matrix Biol. (2003) 22: 145-152.

Kyte J. and Doolittle R.F. A Simple Method for Displaying the Hydropathic Character of a Protein. J. Mol. Bio. (1982) 157: 105-132.

Latvanlehto A., Snellman A., Tu H. and Pihlajaniemi T Type XIII and Some Other Transmembrane Collagens Contain Two Separate Coiled-coil Motifs, Which May Function as Independent oligomerisation Domains. J. Biol. Chem.(2003) Vol. 278,

No. 39: 37590-37599.

Levy Y. and Onuchic J.N. Water Mediation in Protein Folding and Molecular Recognition. Annu. Rev. Biophys. Biomol. Struct. (2006) 35: 389-415.

Linsenmayer, T. F. Collagens. in: Cell Biology of Extracellular Matrix. 2nd ed. (1991) Hay, E.D., Ed. Plenum Press, New York. pp 6-44.

Mallinger R., Kulnig W. and Böck P. Symmetrically Banded Collagen Fibrils: Observations on a New Cross Striation Pattern in Vivo. Anat. Record. (1992) 232: 45-51.

Malone J.P. and Veis A. Type I collagen N-telopeptides adopt an ordered structure when docked to their helix receptor during fibrillogenesis. Proteins (2004) 54: 206-215.

Meek K.M., Chapman J.A. and Hardcastle R.A. The Staining Pattern of Collagen Fibrils. Improved Correlation with Sequence Data. J. Biol. Chem. (1979) Vol. 254 No. 21: 10710-10714.

Metsäranta M., Toman D., de Combrugghe B. and Vuorio E. Mouse type II collagen gene. Complete nucleotide sequence, exon structure, and alternative splicing. J. Biol. Chem. (1991) Vol. 266, No.25: 16862-16869.

Moore J.M. Physical Chemistry 5th ed. (1972). Longman Group Limited.

Nakamura H., Takeshi S. and Akiyoshi W. A theoretical study of the dielectric constant of protein. Protein Eng. (1988) Vol. 2, No. 3: 177-183.

Nykvist P., Tu H., Ivaska J, Kapyla J, Pihlajaniemi T and Heino J. Distinct

Recognition of Collagen Subtypes by $\alpha_1\beta_1$ and $\alpha_2\beta_1$ Integrins. $\alpha_1\beta_1$ Mediates cell adhesion to type XIII collagen. J. Biol. Chem. (2000) Vol. 275, No. 11: 8255-8261.

Oh S.P., Warman M.L., Seldin M.F., Cheng S., Knoll J.H.M., Timmons S. and Olsen B.R. Cloning of c-DNA and Genomic DNA Encoding Human Type XVIII Collagen and localization of the α1(XVIII) Collagen to Mouse Chromosome 10 and Human Chromosome 21. Genomics (1994) 19: 494-499.

Okuyama k., Okuyama K.,Arnott s. Takayanagi M. and Kakudo M. Crystal and Molecular Structure of a Collagen-like Polypeptide $(Pro\text{-}Pro\text{-}Gly)_{10}$. J. Mol. Biol. (1981) 152: 427-433.

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

Ortolani F. Giordano M. and Marchini M. A model for type II collagen Fibrils: Distinctive D-band Pattern in Native and Reconstituted Fibrils Compared with Sequence Data for Helix and Telopeptide Domains. Biopolymers (2000) Vol. 54: 448-463.

Otter A., Scott P.G. and Kotovych G. Type-I collagen α-1 chain C-telopeptide-Solution structure determined by 600 MHz proton NMR spectroscopy and implications for its role in collagen fibrillogenesis. Biochemistry (1988) 27: 3560-3567.

Parry D.A.D. The molecular and fibrillar structure of collagen and its relationship to the mechanical properties of connective tissue. Biophys. Chem. (1988) 29: 195-209.

Petterson E.F., Goddard T.D., Huang C.C., Couch G.S., Greenblatt D.M., Meng E.C. Ferrin T.E. (2004) UCSF Chimera—A Visualization System for Exploratory Research and Analysis. J. Comput. Chem. (2004) 25: 1605-1612.

Phillips C.L., Pfeiffer B.J., Luger A.L. and Franklin C.L. Novel collagen glomerulopathy in a homotrimeric type I collagen mouse (oim). Kidney Int. (2002) Vol 62: 383-391.

Poole C.A., Ayad S. and Gilbert R. Chondrons from articular cartilage. V. Immunohistochemical evaluation of type VI collagen organisation in isolated chondrons by light, confocal and electron microscopy. J. Cell Sci. (1992) 103: 1101-1110.

Prockop, D.J. and Kivirikko, K.I. Collagens: Molecular Biology, Diseases, and potentials for therapy. Annu. Rev. Biochem. (1995) 64: 403-34.

Ramachandran G.N. Structure of Collagen at the Molecular Level in: Treatise on Collagen Vol. 1 Chemistry of Collagen. (1967) Academc Press, London and New York.

Ramachandran G.N. Bansal M. and Bhatnagar R.S. A hypothesis on the role of hydroxyproline in stabilizing collagen structure. Biochim. Biophys. Acta (1973) 322: 166-171.

Ramachandran G.N. Stereochemistry of collagen Int. J. Pept. Prot. Res. (1988) 31: 1-16.

Ramshaw J.A.M., Shah N.K. and Brodsky B. Gly-X-Y Tripeptide Frequencies in Collagen: A Context for Host-Guest Triple-Helical Peptides. J. Struct. Biol. (1998)

122: 86-91.

Rasband, W.S. ImageJ, U. S. National Institutes of Health, Bethesda, Maryland, USA, http://rsb.info.nih.gov/ij/, 1997-2007

Reale E., Groos S., Luciano L., Eckardt C., Eckardt U. In the mammalian eye type VI collagen tetramers form three morphologically different aggregates. Matrix Biol. (2001) 20: 37-51.

Ricard-Blum S. and Ruggiero F. The collagen superfamily: from the extracellular matrix to the cell membrane. Pathologie Biologie (2005) 53: 430-442.

Rich A. and Crick F.H.C. The Molecular Structure of Collagen. J. Mol. Biol. (1961) 3: 483-506.

Sato K., Yomogide K., Wada T., Yorihuzi T., Nishimune Y., Hosokawa N. and Nagata K. Type XXVI Collagen, a New Member of the Collagen Family, Specifically Expressed in the Testis and Ovary. J. Biol. Chem. (2002) Vol. 277, No. 40: 37678-37684.

Salem G. and Traub W. Conformational implications of amino acids sequence regularities in collagen. FEBS lett. (1974) Vol. 51, No.1: 245-249.

Schacke H., Schumann H., Hammami-Hauasli N., Raghunath M. and Bruckner-Tuderman Two Forms of Collagen XVII in Keratocytes. A Full Length Transmenbrane Protein and a Soluble Ectodomain. J. Biol. Chem. (1998) Vol. 273, No. 40: 25937-25943.

Shah N.K., Ramshaw J.A.M., Kirkpatrick A., Shah C. and Brodsky B. A Host—Guest Set of Triple-Helical Peptides: Stability of Gly-X-Y Triplets Containing

Common Nonpolar Residues. Biochemistry (1996), 35: 10262-10268.

Sharp K.A. and Honig B. Electrostatic Interactions in Macromolecules: Theory and Applications. Annu. Rev. Biophys. Biophys. Chem. (1990), 19: 301-332.

Shuttleworth C.A. Molecules in Focus Type VIII collagen Int. J. Biochem. Cell. Biol. (1997) 29: 1145-1148.

Stapley B.J. and Creamer T.P. A survey of left-handed polyproline II helices. Protein Sci. (1999) 8: 587-595.

Toman P.D. and de Combrugghe B. The mouse type-III procollagen-encoding gene: genomic cloning and complete DNA sequence. Gene (1994) 147: 161-168.

Tomono Y., Naito I., Ando K., Yonezawa T., Sado Y., Hirakawa S., Arata J., Okigaki T. Ninomiya Y. Epitope-defined Monoclonal Antibodies against Multiplexins Collagens Demonstrate that type XV and XVIII Collagens are Expressed In Specialized Basement Membranes. Cell Struct.funct. (2002) 27: 9-20.

Traub W. and Fietzek P.P. Contribution of the α-2 chain to the molecular stability of collagen. FEBS lett. (1976) Vol. 69 No. 2: 245-249.

Traub W. Molecular assembly in collagen. FEBS lett. (1978) Vol. 92 No. 1: 114-120.

Van der Rest M.and Garrone R. Collagen family of proteins. FASEB J. (1991) 5: 2814-2823.

Van der Rest M. and Mayne R Type IX collagen Proteoglycan from Cartilage is Covalently Cross-linked to Type II collagen. J. Biol. Chem. (1988) Vol. 263, No. 4:

1615-1618.

Veit G., Kobbe B., Keene D.R., Paulsson M., Koch M. and Wagener R. Collagen XXVIII, a Novel von Willebrand Factor A Domain-containing Protein with Many Imperfections In the Collagenous Domain. J. Biol. Chem. (2006) Vol. 281, No. 6: 3493-3504.

Von der Mark H., Aumailley M., Wick G., Fleischmayer R. and Timpl R. Immunochemestry, genuine size and tissue localisation of collagen VI. Eur. J. Biochem. (1984) 142: 243-502.

Wess T.J. Collagen Fibril Form and Function. Adv. Protein. Chem. (2005) Vol. 70: 341-378.

Wess T.J. and Orgel J.P. Changes in collagen structure: drying, dehydrothermal treatment and relation to long term deterioration. Termochim. Acta (2000) 365: 119-128.

Williams B.R., Gelman R.A., Poppke D.C. and Piez K. Collagen Fibril Formation. Optimal in vitro conditions and preliminary kinetic results. J. Biol. Chem. (1978) Vol. 253, No. 18: 6578-6585.

Wimley W.C. and White S.H. Experimentally determined hydrophobicity scale for proteins ay membrane interface. Nature Struct. Biol. (1996) 3: 842-848.

Wu J. and Eyre D.R. Structural Analysis of Cross-linking Domains in Cartilage Type XI collagen. J Biol. Chem. (1995) Vol. 270, No. 32: 18865-18870.

Young R.D., Lawrence P.A., Duance V.C. Aigner T. and Monaghan P. Immunolocalization of Collagen Types II and III in Single Fibrils of Human

Articular Cartilage. J Histochem. Cytochem. (2000) Vol. 48(3): 423-432.

Yurchenco P.D. and Schnitty J.C. Molecular architecture of basement membranes. FASEB J. (1990) 4: 1577-1590.

Zhang Y. and Cremer P.S. Interactions between macromolecules and ions: the Hofmeister series. Curr. Opin. Chem. Biol. (2006) 10: 658-663.

Websites

Computer Graphics Laboratory, University of California, San Francisco: http://www.cgl.ucsf.edu/chimera/

GlaxoSmithKline R&D and the Swiss Institute of Bioinformatics: http://us.expasy.org/spdbv

Tokyo University of Agriculture and technology:

http://www.tuat.ac.jp/~x-ray/Research/R_2-models.html

UniProt: http://www.ebi.uniprot.org/index.shtml

University of Iowa: www.uiowa.edu/~cemrf/methodology/tem/index.htm