1-1-1994

Stem-length requirements for chain folding of periodic polypeptides/

Howard S. Creel
University of Massachusetts Amherst

STEM-LENGTH REQUIREMENTS FOR CHAIN FOLDING OF PERIODIC POLYPEPTIDES

A Dissertation Presented

by

HOWARD S. CREEL

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 1994

Department of Polymer Science and Engineering

STEM-LENGTH REQUIREMENTS FOR CHAIN FOLDING OF PERIODIC POLYPEPTIDES

A Dissertation Presented

by

HOWARD S. CREEL

Approved as to style and content by:

David A. Tirrell, Chair

Maurille J. Fournier, Member

Thomas L. Mason, Member

Richard S. Stein, Member

William J. MacKnight, Department Head
Department of Polymer Science and Engineering

DEDICATION

Since I have already pledged my heart and soul to my wife Susan, it seems a puny thing to offer her a book as a symbol of the love and commitment we share. But her spirit, her laughter and her wisdom were the foundation from which I drew the strength and inspiration to write the words—the entire effort holds no meaning without her.

So I humbly offer the book, firmly believing that while the words are not enduring…

our love is.

ACKNOWLEDGEMENTS

I am deeply indebted to Professor David A. Tirrell for facilitating my growth as a scientist and for making my tenure challenging, exciting and rewarding. By his example, I learned that industry and integrity is as important as intelligence in the conduct of a scientist's career. For that lesson, I am grateful.

I would like to thank my thesis commitee members for their professional guidance. Professor Skip Fournier always lent a ready ear to my concerns and his insight into my research was invaluable (and often uncanny). Professor Thomas Mason had a talent for suggesting the correct experiment to perform; my work in his laboratory was rewarding. Professor Richard Stein was always an island of polymer science while I was awash in molecular biology, and his unique perspective was refreshing. I can only hope that this work reflects their many contributions.

My family has been a constant source of love and support (emotional and financial), even during my nine year so-called "checkered" undergraduate career. My mother, Peggy O'Neill Creel, continues to keep me focused on the important issues in life: science, romance, trees and literature, although her admonishment (that of a librarian) not to clutter my mind with trivial facts (but know where to look them up) may have contributed to my eighteen month cumulative examination odyssey. It is quite likely that she will be the only one outside of my committee that will actually read this book. I will continue to seek the council of my father, Col. Buckner M. Creel III, whose lifetime of professional experience I consider an invaluable asset and advantage in the conduct of my career. His dedication to my education has sustained me for many years. As the only other scientist in the family, my sister Mary Williams was often subjected to long accounts of mundane experiments, a duty she accepted with grace and patience. She continues to be a dedicated confidant and supporter. I also thank her for her graphic design talents, which greatly enhanced the presentation of this work. I love you all.

Few graduate students get an opportunity to participate in a ground-breaking multi-disciplinary research program and work with a talented, motivated group of researchers in a team atmosphere. The first generation Biosynthesis Group members Dr. Mark Krejchi, Dr. Mike Dougherty, Yoshikuni Deguchi, and Eric Cantor were a daily source of insight and motivation: it was a privilege to work with them. In addition, my own work and the Biosynthesis Group as a whole owes a great deal to the efforts of Dr. Kevin McGrath.

I am especially indebted to Professor Mason for allowing a polymer scientist the opportunity to work in his laboratory. The researchers in LGRT 1003 not only offered invaluable advice in the solution of difficult problems in molecular biology, their professional and efficient approach to group research is an ideal to which others should strive. Dr. Karen Sirum-Connally and Dr. Jennifer Pinkham were especially kind and helpful. In addition, my many discussions (not necessary scientific in nature) with Dr. Pinkham were thought-provoking and stimulating.

My collaborations with Professor Ted Atkins, Professor Ronald Beavis, Dr. Brian Chait and Dr. Rong Wong were satisfying and fruitful. The success of this work was assured by their expertise, advice and effort.

The department technicians, Jay Conway, John Domian, and Norm Page were an invaluable resource in the day-to-day operations crucial for a successful research project. Dr. Alan Waddon, Dr. Charles Dickinson, Dr. Janos Borbely, Lou Raboin, Dr. Jane Tirrell, Luisa Cabrera and Cheryl Cote also provided crucial technical support. I am grateful to Dr. Bob Jennings, Dr. Khleantis Koniaris and Ajay Parkhe for their help (mostly involving computers) and friendship.

I am grateful for the department secretaries, Sophie Michelsen, Eileen Besse, Chris Salame and Jan Fournier who solved 100% of the baffling bureaucratic quagmires in which I often found myself. Professor Tirrell's secretaries Sharon Duttlinger and Helen Anglos were always a source of conversation and assistance, and Linda Stryzgowski helped me

cope with the "home stretch" with liberal offerings of goodies and coffee and (more importantly) through friendship and support.

I learned a great deal about organic chemistry from Professor C. Peter Lillya, the details of polymer synthesis from Dr. Brant Kolb and the techniques of molecular biology from Richard Lempicki; all were excellent instructors. The development of my scientific skills was also aided by interactions with a number of outstanding researchers: Dr. Lin-Lin Xhou, Dr. S. Y. Wang, Professor Robert Lenz, Dr. Keith Wilbourne, Dr. Eiji Yashima and especially Dr. Rahim Hani. While none contributed directly to the work found in this thesis, the gifts each gave me, in the form of scientific insight and technique, were invaluable in the effort.

I participated in a number of intramural teams and learned valuable lessons in teamwork and fair play. I am especially grateful to my friends Dr. Jeff Kollodge (who taught me to shoot a basketball with my right hand), Dr. Gregg Bennett (who taught me to turn my wrists over when swinging a bat), and Dr. Nicole Franchina (who supported me in my quest to surround myself with teammates who "relax and dig it"). I also value my friendship and collaboration with Dr. Kevin Schaffer and Dr. Mark Dadmun.

I received valuable insight into perseverence and positive mental attitude from four remarkable people: the only positive aspect of taking nine cumulative examinations was sharing the experience with Dr. Michelle Maden; I count her as a friend because of it. Dr. Cady Coleman breathes life into the tired cliche "reach for the stars but keep your feet on the ground". I'll live my boyhood astronaut dreams through her. I miss Tracie Graff's sage advice and expert hair cuts. And Susan Dawson has taught me countless lessons on coping graciously while managing to smile and maintain that I am the one who helps her!

Finally, this work is a tribute to the late Professor William Bailey, who (by crinkling his tie) nudged a frustrated chemical engineering student towards a career in polymer science. Like him, I have become a scientist who derives great pleasure and satisfaction from creating large molecules. He is missed.

ABSTRACT

STEM-LENGTH REQUIREMENTS FOR CHAIN FOLDING

OF PERIODIC POLYPEPTIDES

FEBRUARY 1994

HOWARD S. CREEL, B.S., UNIVERSITY OF MARYLAND

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor David A. Tirrell

Proteins of uniform sequence and stereochemistry can be produced in bacterial cells for use in fundamental studies of polymer morphology. In earlier work, a copolypeptide with repetitive sequence **1**, designed to adopt a lamellar architecture upon crystallization, was synthesized using genetic engineering techniques. Solid state studies indicated that the material did not adopt the desired architecture, perhaps because the odd number of amino acids in the repeat unit prevented the formation of the appropriate chain trajectory for chain folding, disrupting the normal hydrogen bond pattern associated with ß-sheets (a key component of the desired chain-folded assembly). It was asserted that this energetic liability would be offset by adding hydrogen-bonding pairs to the bulk structure, by insertion of additional alanylglycine dyads in sequence **1**.

$$[(AlaGly)_3ProGluGly]_m \qquad \mathbf{1}$$

To test this premise, solid state analysis of a series of repetitive copolypeptides with increasing numbers of alanylglycine dyads (sequences **2**) was accomplished.

$$[(AlaGly)_nProGluGly]_m \qquad n = 3, 4, 5, 6 \qquad \mathbf{2}$$

In a preliminary study, a copolypeptide with repetitive sequence $[(AlaGly)_4ProGluGly]_{14}$ was synthesized. X-ray diffraction and infrared analysis of this material indicated an increase in solid state order over that of polypeptides with sequence **1**. Mass analysis of this material using matrix-assisted laser-desorption/ionization (MALDI) mass spectrometry prompted the discovery of sequence errors in the DNA code for the protein. This analysis and the analysis of other chain-length variants of sequence **2** (n = 4, m = 10, 12, 16, and 20) highlighted the utility of the MALDI technique for the assessment of protein sequence, molecular weight and purity.

Analysis of proteins with sequence **2** (m = 16, n = 3, 4, 5, 6) using differential scanning calorimetry, infrared spectroscopy, electron microscopy and X-ray diffraction was consistent with an increase in solid-state order with an increase in the number of alanyl-glycine dyads in the repetitive sequence. A model is presented featuring folded chains, which is rationalized on the basis of comparison of the X-ray scattering patterns of the materials with those of poly(L-alanylglycine).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND BACKGROUND

1.1. Introduction

Under normal crystallization conditions, polymer crystal growth is dominated by the tendency of the polymer chains to fold into macroscopic chain-folded lamellae. The chain-folding process is kinetically controlled and as such, it is impossible to predict the architecture ultimately adopted by the folded chains. If such structural control could be gained, the chain-folding process could be used to produce novel nano-scale polymer architectures designed for specific applications.

For most polymers, chain-folding, if it occurs, is the result of relatively weak, non-directional van der Waals interactions. However, control of chain placement is attainable through the use of hydrogen bonds as specific interactions that direct the placement of monomers in relation to each other. In fact, the propensity to chain-fold can be programmed into the chain by judicious selection of monomer sequence. For designed systems, uniformity is the crucial property: the chains must be of precise length, sequence and stereochemistry. More importantly, sequence control must be specific in defining the identity of each monomer of the chain. Currently, no single synthetic scheme affords control over these crucial chain variables.

Such control is achieved routinely in living cells: protein biosynthesis yields complex copolymers chains of specific sequence (defined by the DNA code) and defined stereochemistry (L-amino acids are selected). The cell manufactures these materials for use in a variety of complex biological functions, requiring the amino acid sequence of the protein to direct the organization of the chains in three dimensions. Thus, the controlled use of protein biosynthesis not only gives unprecedented control over chain variables, but

allows the design of amino acid monomer sequences with specific interactions tailored for self-assembly.

The techniques of molecular biology are refined enough to allow the synthesis of practically any sequence of amino acids. However, given the complexity of the interactions of even modest size proteins, do we have enough information to properly design a sequence to adopt the desired architecture? The necessary results come from two sources, structural biology and studies of biomaterials. From structural biology we glimpse the hierarchies of protein structure defined by sequences of amino acids. In fact, the architecture of an ordered domain of a complex enzyme is envisioned as comprised of arrays of defined microstructures—turns, strands and helices—the simple motifs describe the complex assembly. In addition, the analysis of structural proteins can provide design criteria. The structural domains that define the mechanical properties of these biomaterials are generally comprised of repetitive sequences of amino acids: model copolypeptides with short repetitive sequences have been produced that mimic the behavior of the natural materials. Thus, the complex organization of protein chains in space can be reduced to combinations of simple, well-defined structural motifs which can be exploited in the creation of new macromolecular architectures.

These analyses provide the requisite background for the design of new materials that adopt a chain-folded architecture in the solid state. This approach to the synthesis of new materials is fundamentally new: recombinant DNA techniques are traditionally employed for the synthesis of proteins of biological interest. The methods are general and can be used to produce polymers for use in roles requiring stringent sequence and size homogeneity.

The goal of this research is the production, using the techniques of genetic engineering, of materials with a specific chain micro-structure rationally designed to adopt a chain-folded conformation in the solid state. The analysis of these materials should help define the requirements for chain-folding of periodic polypeptides.

1.2. Chain-folding

1.2.1. Description of Polymer Chain-folding

The chain-folded structure of polymer single crystals is well established. Early work by Keller[1], showed that chains in polyethylene single crystals (grown from dilute solution) were oriented normal to the large, flat faces of the lamellar crystals which were only 100-140 Å thick. Keller concluded that the 2000-7000 Å chains could only adopt such a conformation by folding. Since the initial observation, the concept has been continuously reinforced by a wide variety of morphological studies—it has been shown that the chains of numerous polymers chain-fold upon crystallization from solution and the melt (see Wunderlich[2] for a partial list).

Given that from a thermodynamic standpoint the most stable state is that of full extension, why do polymer chains fold? In fact, because polymer chains can access a large number of conformational states and have a sluggish response to change, the observed morphology is almost never a consequence of reaching thermodynamic equilibrium. Rather, kinetic factors direct the chains along the path of the phase change in the most expedient manner. In the formation of crystals from solution, deposition of the chains fully extended along a crystal face is an unlikely event in the time scale of crystallization, despite the fact that this would lead ultimately to the lowest conformational energy state. Local chain-folding occurs more rapidly and thus determines the path of the crystallization, resulting in the formation of a metastable crystalline state.

Crystals obtained from dilute solution generally show the least complicated morphological features; with slow cooling, large crystals with low levels of defects can be grown. The lamellar thickness is roughly an inverse function of supercooling, and is not affected to a large extent by other parameters such as molecular weight and pressure. In addition, the crystals thicken upon subsequent annealing. While there is a wealth of data on the structure of chain-folded polymers, the question of reentry of the chains to form a new

stem remains in dispute. The consensus is a regular reentry model, with the molecule forming a new stem after immediate reentry or after skipping over one or two nearest neighbor sites.[3]

Even with careful control of conditions, the morphology of single crystals can be quite complex (as shown in Figure 1.1) and exhibit disordered packing in the bulk and at the fold surface. There is a composite nature to lamellae: a crystalline core (with some level of defects) and a relatively disordered surface. The level of fold surface disorder increases with crystallization from more concentrated solutions, witnessed by an increase in structural complexity: multilayer crystals, axialites, fibrils, etc.

Chain-folded lamellae are also observed upon crystallization of polymers from the melt. Examination of the structural details is more difficult in these bulk systems, and the connection between chain-folded structures grown from solution and bulk is ambiguous. This discussion is confined to solution properties; the reader is referred to any polymer physics text for a discussion of melt crystallization (for instance, Sperling[4] or Bassett[5]).

1.2.2. Design Criteria for Chain-folded Lamellae

Chain-folded lamellae are unique structures in that the lamellar thickness, a measurable crystalline dimension, corresponds directly to a molecular parameter: the fold length. Thus, control of the fold length translates to control of a macroscopic dimension of the assembly. In order to precisely define the lamellar thickness, chain variables would have to be controlled such that the kinetically favored state is an assembly of chains, each folded periodically at precisely defined locations.

An ideal chain-folded lamellar structure can be envisioned as built up from successive alternation of two basic motifs: a straight stem followed by a reverse turn (Figure 1.2). A chain constructed of periodic repeats of a suitable stem-turn combination could be programmed with the propensity to chain-fold, making the adoption of a regular chain-folded structure of defined thickness the kinetically accessible target. This strategy

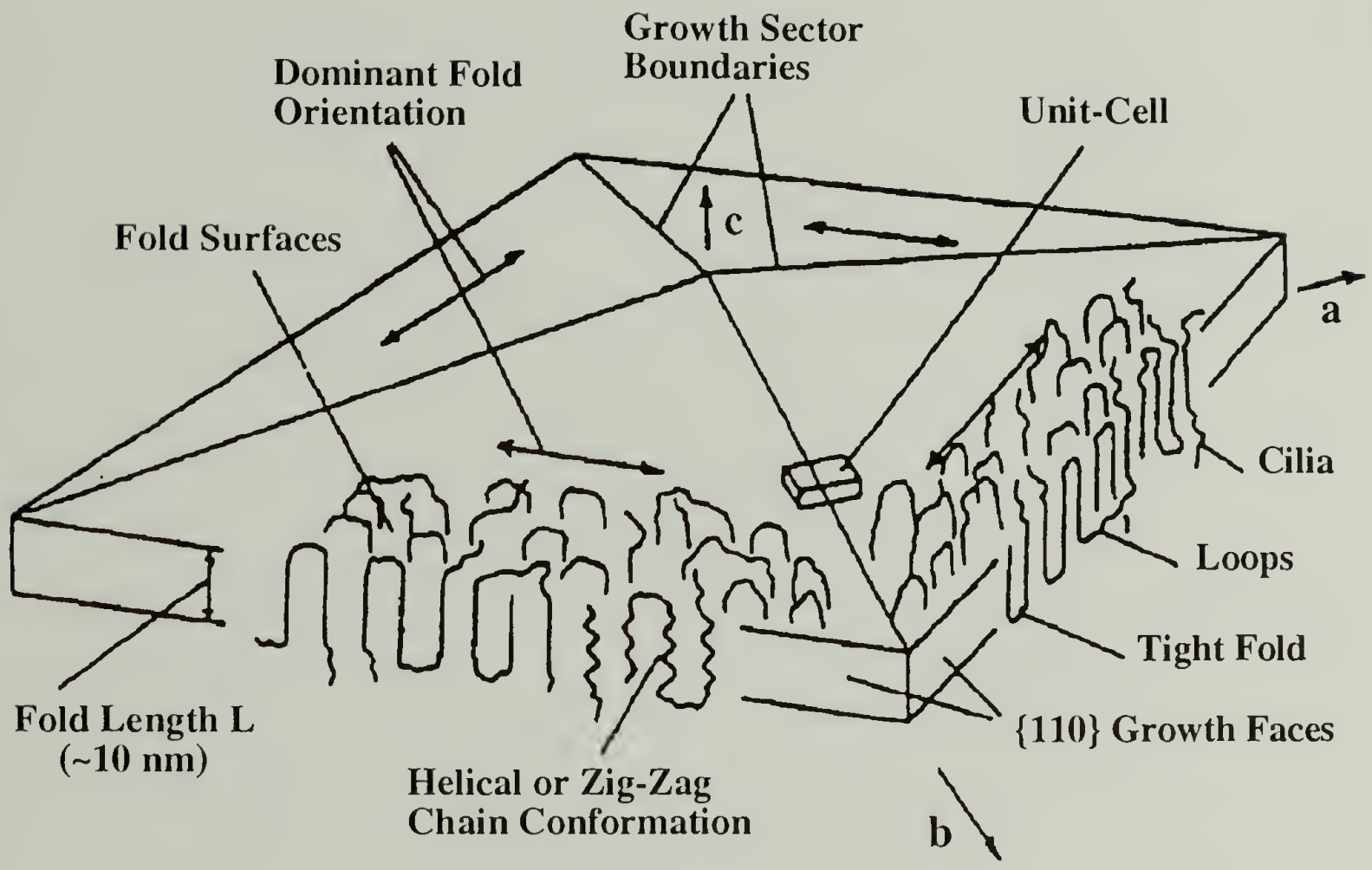


Figure 1.1. Structural details of a polyethylene single crystal. The disorder in the bulk and fold surfaces is emphasized. Reprinted from Lotz[3] with permission of Bernard Lotz.

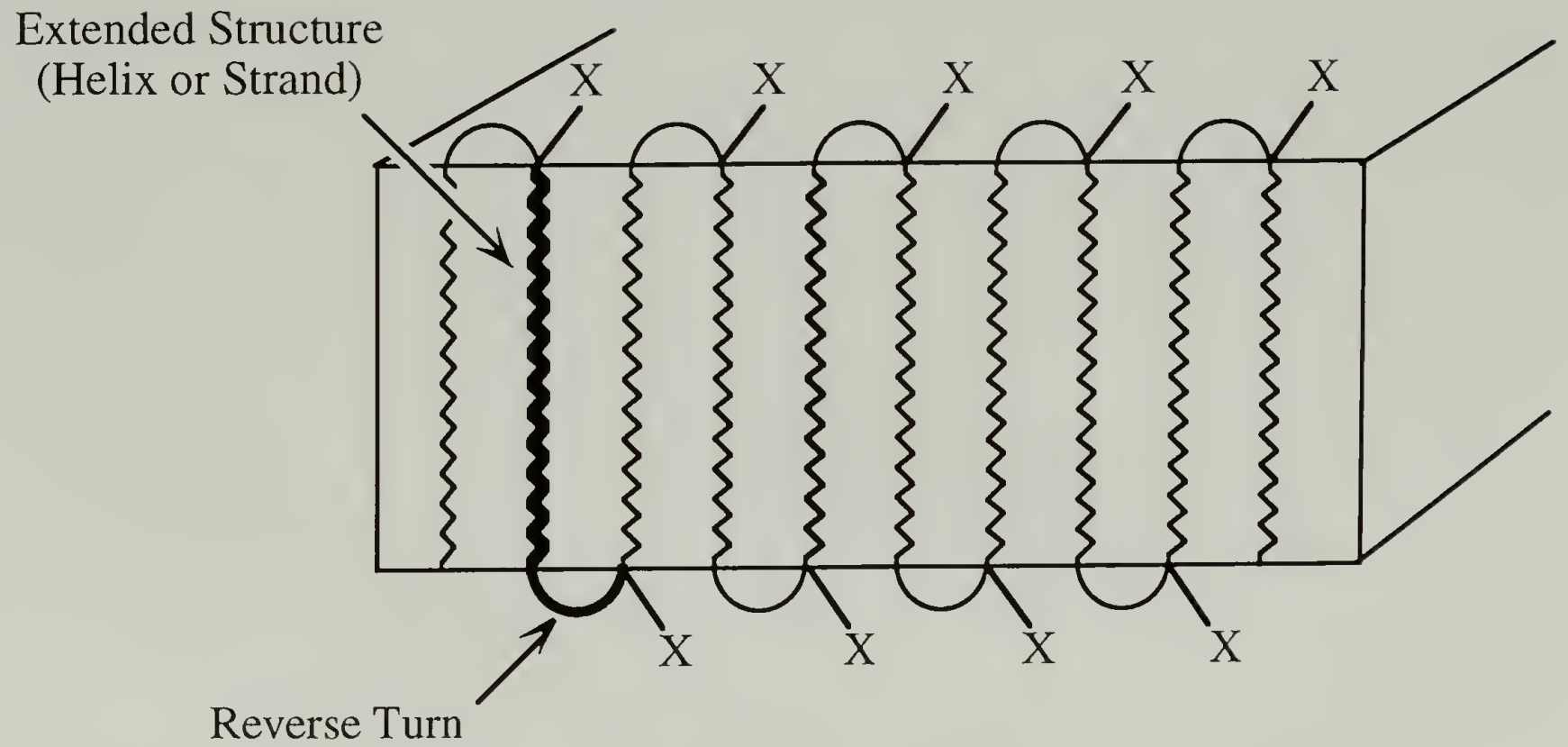


Figure 1.2. Schematic of an idealized lamellar structure. The structure is envisioned as formed *via* regular folding of chains comprised of repeats of a stem-reverse-turn sequence. Adapted from Tirrell and coworkers[6], reprinted with permission of Technomic Publishing Co., Inc.

not only enables the definition of precise crystalline dimensions, but by suitable selection of the monomer composition of the turn component, also allows selective decoration of the fold surface with particular functional groups for subsequent modification of the assembled chains.

1.2.3. Useful Lessons From the Study of Polyamides

Can chain variables be tuned to control lamellar thickness? It is essentially impossible for polyethylene, as the thickness of the metastable lamella is sensitive to crystallization conditions. The methylene repeat unit is small (2.5 Å) with respect to the fold length (~100 Å), and so it is analogous to folding a homogeneous thread: the relatively weak, non-directional dispersion interactions cannot specifically direct the arrangement of the monomers.

However, if as in the polyamide family (nylons), the polymer chain has a regular array of functional groups capable of intrachain hydrogen bonding, and the repeat unit is sufficiently long, chain-fold length comes under a measure of control. It is still a kinetic control, but the most kinetically accessible structures will tend to capitalize on the favorable energetics of the formation of hydrogen bonds. Strategic placement of those bonds promotes the formation of the desired fold length.

The basic structural unit of crystalline nylons is a sheet of extended (zig-zag) chains stitched together by a regular array of linear hydrogen bonds. The chains are generally separated by a distance of 4.8 Å (defined by the geometry of the amide linkage). On a higher order, the sheets pack normal to the hydrogen bonding direction with a separation of 3.7 Å. These features are common to most polyamide single crystals.[7] However, the subtle details of the arrangement of the chains in the sheets and thus the overall sheet packing motif is a direct consequence of the pattern of hydrogen bonding specific to the chemical sequence of the chain.

The nylons can be loosely divided into symmetric and asymmetric structures depending on the relative lengths of the methylene spacers between amide groups in the repeat unit, as illustrated in Figure 1.3. This has important consequences for the placement of the chains as to capitalize on the formation of linear hydrogen bonds. The lamellae of asymmetric nylons 66 and 610 are thought to be comprised of regularly folded chains formed with adjacent reentry. Typically, the chains fold regularly in the methylene spacer between amine groups after every fourth repeat unit.[8] The hydrogen bonds direct the chains in a progressive relative stagger parallel to the chain direction, with the consequence that the sheets align with a stagger to form a structure analogous to a sheared deck of cards. The macroscopic evidence of the formation of this hydrogen bond dependent structure is a measured chain tilt relative to the direction normal to the flat faces of the lamella. These structural details have distinct crystallographic consequences.[9]

The recent study of chain-folded lamellae of nylon 46 by Atkins and coworkers[7] is another important illustration of the hydrogen bonding pattern controlling chain arrangement. It was proposed, based on the observed crystallographic spacings, that the symmetrical arrangement of the short methylene spacers directs chain-folding preferentially at the amide group, with the fold surfaces normal to the chain direction. The amide group eliminated from the hydrogen bonded sheet allows the alignment of the amide groups in the sheet; this alignment is a direct consequence of the symmetrical repeat units (a similar arrangement is impossible for nylon 66). The result is a regular up-down packing of progressive sheets that keeps the chains aligned normal to the flat faces of the lamella (the authors show that a similar sheet arrangement is also seen for specific morphological forms of nylon 6 and nylon 66). The diffraction results seem to support this model, showing a lamellar super-lattice period of 62 Å, in good agreement with the 59 Å value predicted for the tight amide turn model. A comparison of the proposed chain arrangements for nylons 66 and 46 is shown in Figure 1.4.

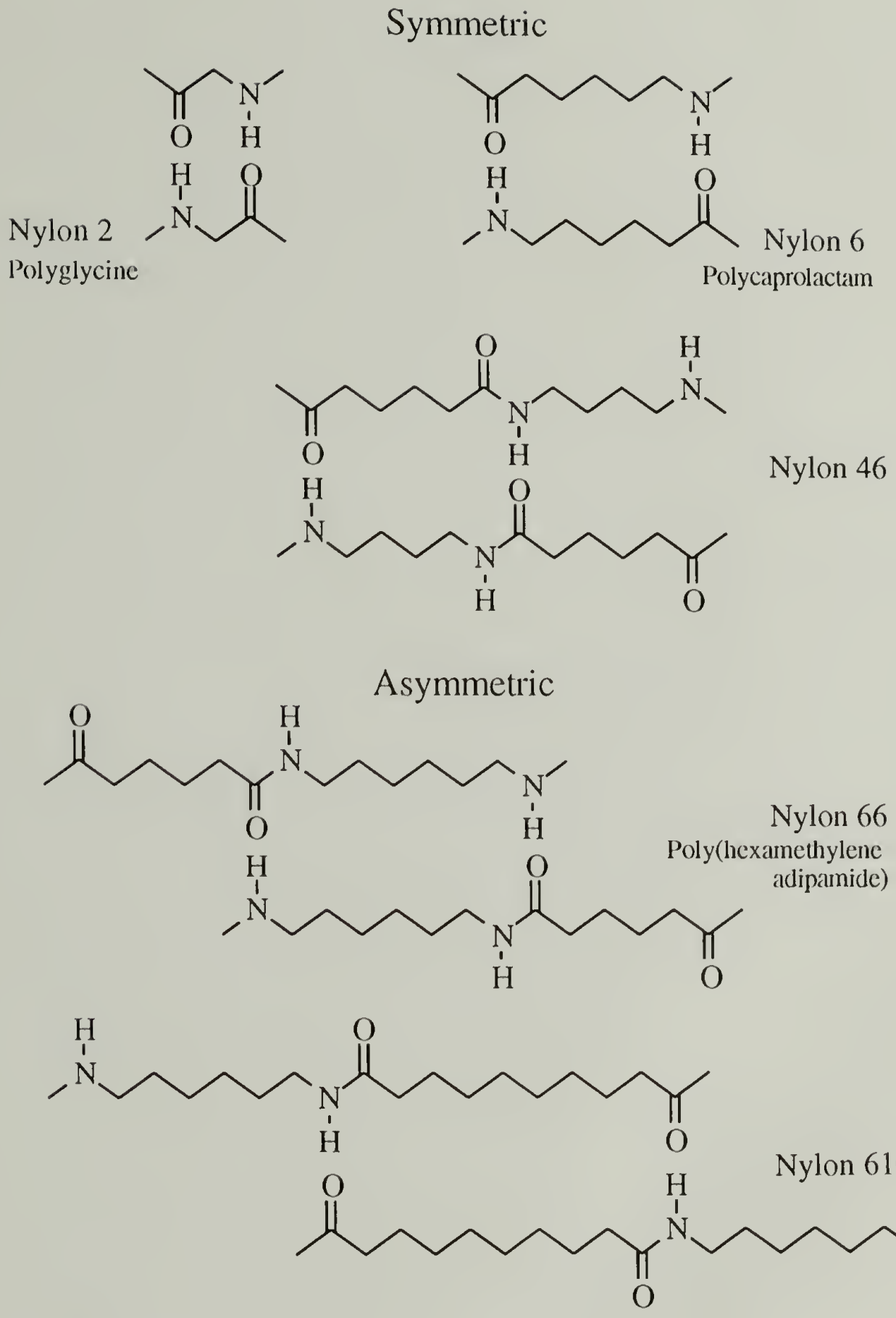


Figure 1.3. Classification scheme for the polyamides. The sequences are termed *symmetric* or *asymmetric*, with equal or alternating numbers of methylene groups between amide groups, respectively. The scheme is presented to emphasize that the arrangement of the amide groups in the chain determines the hydrogen bond pattern of the chain-folded assembly.

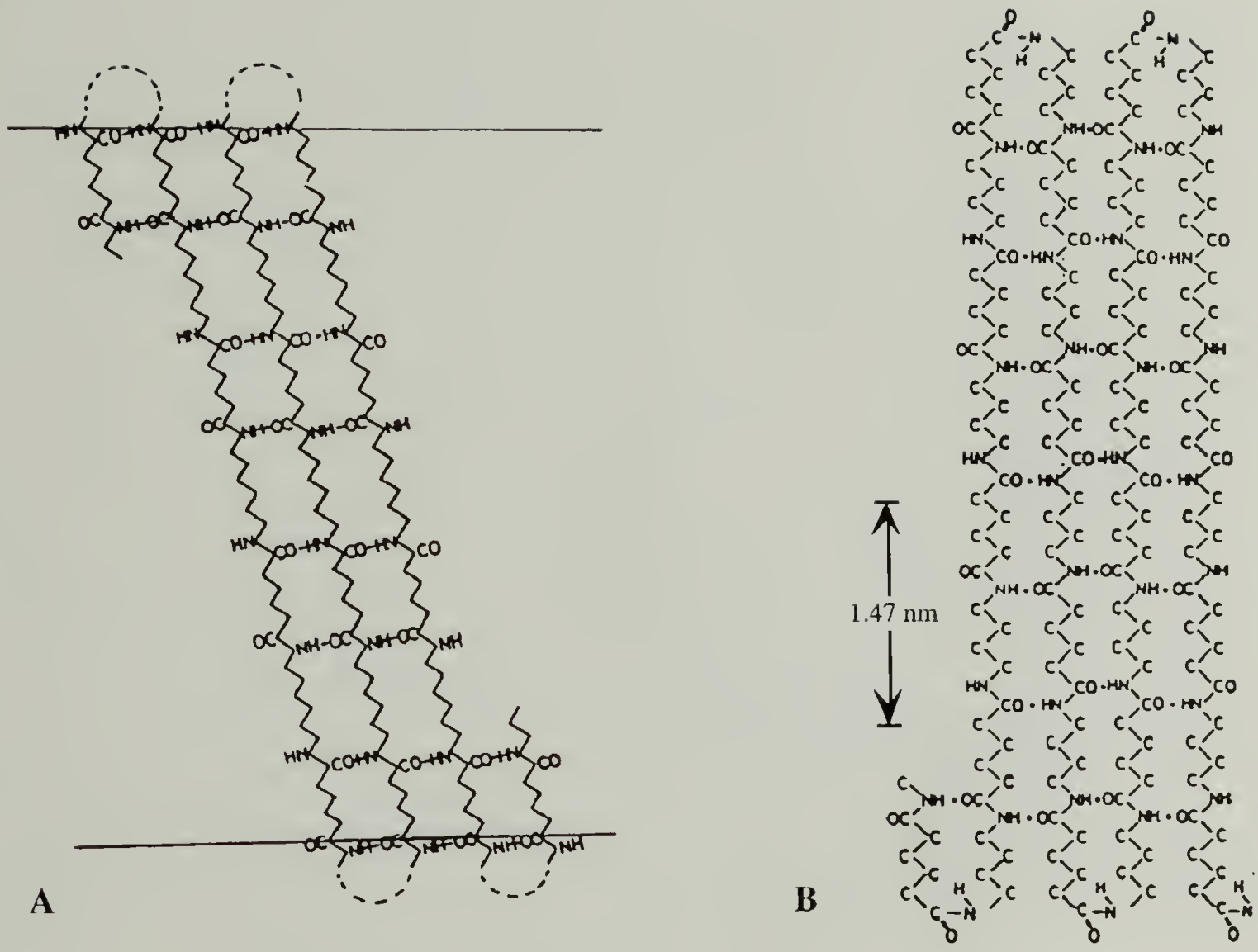


Figure 1.4. Proposed structures of folded polyamide chains. The effect of hydrogen bonding constraints on chain trajectory is emphasized. **A**: Nylon 66 chains folded in the methylene spacer with a chain-tilt relative to the flat faces of the lamella. **B**: Nylon 46 chains folded in the amide group with fold surfaces normal to the chain direction. Adapted from Atkins and coworkers[7]; reprinted with permission from the American Chemical Society.

Despite this well defined structure, nylon 46 can still fold in a methylene spacer, resulting in chain tilt relative to the lamellar normal. However, Atkins and coworkers[7] point out that

> "within the broader family of symmetrical nylons, as the number of CH_2 units between amide groups decreases, a change in the folding mechanism would be expected to occur. Clearly at some stage the alkane segments will be too short for folding anywhere else except in the amide group, which is the case for the proteins (chemically decorated nylon 2). Thus the folding behavior of nylon 46 makes a connection between two separate subject areas."

Therefore, the interest of polymer science in the continuum of materials with regularly spaced amide groups in the chains has crossed over to the structural biology milieu, which seems to validate the premise that the study of protein architecture is useful in the design of novel polymer systems.

1.2.4. Chain-Folding in Proteins: *Chrysopa flava* Silk

The formation of regularly chain-folded polymers is certainly not a new phenomenon for structural biologists. Chain-folding in natural proteins is illustrated best by a discussion of the "cross-ß" (*vide infra*) orientation seen in a number of insect silk proteins.[10] As in the nylons, the chains align and form sheets with the long axis in the hydrogen bond direction. In addition, the measured dimensions of chain axis relative to the total length of the protein suggests chain-folding. However, the materials are unusual in that shearing the sample preferentially orients the chain axis perpendicular to the direction of shear as opposed to the parallel arrangement of most other chain-folded systems.

The notable example of this structure is egg stalk protein of the green lace-wing fly, *Chrysopa flava*, studied extensively by Geddes and coworkers.[11] The protein, composed primarily of serine (41% by weight), glycine (23.5%) and alanine (20%), is thought to be highly repetitive. The wide angle X-ray pattern of unstretched egg-stalks is consistent with that of a pleated, antiparallel ß-sheet (the details of the structure will be discussed below).

The 9.48 Å interchain (fiber axis direction), and 5.45 Å intersheet (strong equatorial reflection) spacings indicate a preferential alignment of the chains perpendicular to the fiber axis, an orientation supported by infrared dichroism. The fibers were thought to be comprised of flat, ribbon-like filaments with a thickness of 25 Å, a dimension supported by weak equatorial reflections in the X-ray pattern of the native protein and consistent with the proposed chain orientation. The ribbon-like morphology is clearly visible by electron microscopy.[12]

Upon stretching the fiber, the wide angle reflections associated with the cross-ß structure are gradually replaced by the signature of chains aligned parallel with the fiber dimension. In addition, the 25 Å reflections on the equator are replaced by faint meridional reflections corresponding to a periodicity of 55.2 Å. Thus, the major repeat unit of the chains occupies an axial distance of 9.48 Å in the native fiber and 55.2 Å in the stretched material.

Geddes and coworkers[11] presented a model that accounts for the observations in terms of chain-folding, shown in Figure 1.5. It was suggested that the chain folds every eight residues, and the observed change in the small angle reflections arise from the intramolecular rearrangement of a 16 residue stem-turn-stem-turn structural repeat unit, with an axial distance of 9.48 Å in the native protein, which unfolds and chain extends to 55 Å in response to the applied shear. The length of the extended structure corresponds to the normal dimensions of extended ß-sheets (~3.5 Å/residue). This roughly six-fold difference in chain dimension is consistent with the observed macroscopic maximum extension. The length of the proposed eight residue repeat is consistent with the 25 Å low-angle equatorial reflection of the unstretched material. The ribbon-like filaments are comprised of stacked sheets of folded chains, with the long dimension in the hydrogen bond direction and the chain axis perpendicular to the long dimension. The proposed details of the repeat sequence (the authors suggest $[GlyAsp(Ser)_4(Ala)_2]_n$) and turn structure are less compelling, but the evidence for chain-folding is consistent and strong.

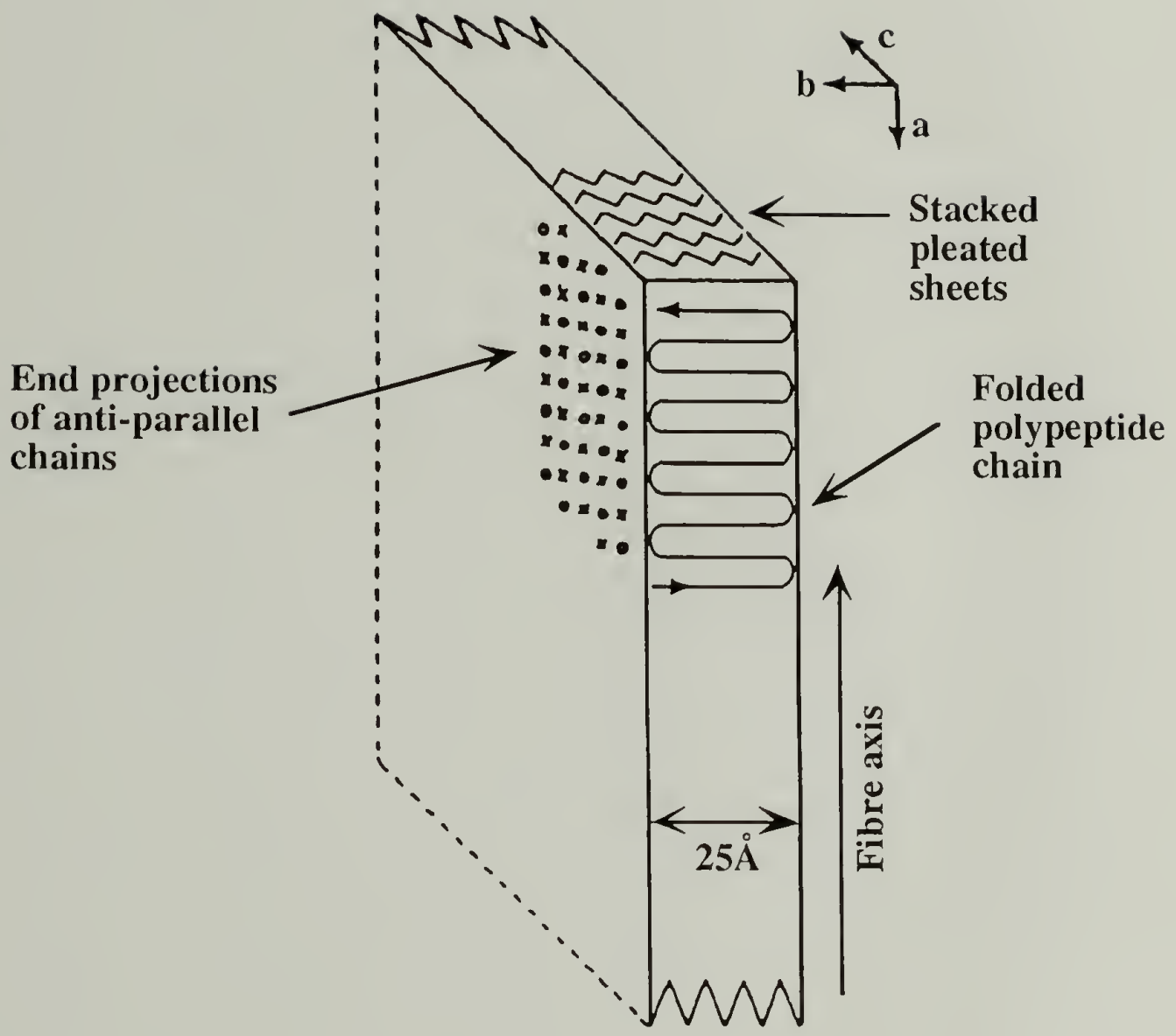


Figure 1.5. Details of the cross-ß structure of *Chrysopa flava* silk. The cross-ß nomenclature refers to the orientation of the chains in the ß-sheet perpendicular to the direction of the fiber axis. The details of the structure are discussed in section 1.2.4. Reprinted from Geddes and coworkers[11] with permission from Academic Press.

In a more recent study, synthetic poly(*O*-benzyl-L-serine) was also shown to adopt the cross-ß pattern.[13] The dominant feature of the X-ray diffraction pattern was an intense 4.7 Å reflection parallel to the orientation direction (attributed to the interchain spacing of a ß-sheet). Once again, this diagnostic reflection allowed the assignment of the chain axis in the cross-ß pattern; the intersheet spacing was measured at 13.8 Å. While there was no indication of small angle reflections associated with stacked lamellae, the wide angle X-ray data and infrared dichroism measurements were consistent with the proposed structure. This study helps to reinforce the diagnostic details of the cross-ß architecture.

These analyses confirm the validity of seeking to explore polymer chain-folding using repetitive polypeptides. However, the structures of the twenty natural amino acids present a seemingly bewildering array of sequence possibilities (Appendix A). The necessary criteria for the rational design of chain-folded architectures can be drawn from two separate disciplines: structural biology, the exploration of the structure and function of proteins of biological interest, and from the study of natural biomaterials, especially the fibrous proteins. Both approaches are similar in that the successful application of criteria gleaned from any structural study requires some understanding of the connection between amino acid sequence and protein chain conformation.

1.3. Basic Protein Chain Variables

The critical protein chain variables are illustrated in Figure 1.6A. The trajectory of the chain through space can be specified by sequentially defining the dihedral angles ϕ, ψ and ω of each residue; for instance, the fully extended all *trans* conformation has $\phi = \psi = \omega = 180°$. The partial double bond character of the peptide bond restricts the dihedral angle to the *trans* conformation ($\omega = 180°$) for most residues in globular proteins (the exception is proline, where the *cis* conformation ($\omega = 0°$) is accessible due to the nature of the pyrrolidine ring).

The restricted amide bond rotation allows the construction of $\phi\psi$ maps (Ramachandran maps)[14] utilizing dihedral angle data obtained from protein crystal structures. For most residues the map resembles that for alanine shown in Figure 1.6B, with the highest density near (ϕ,ψ = -60°, -60°) for residues participating in α-helical structures and a more diffuse maximum near (-90°, 120°) for extended chains and residues in ß-sheets. Glycine shows a highly expanded plot (Figure 1.6C), symmetrical around (ϕ,ψ = 0°, 0°), which reflects increased conformational freedom due to the lack of a side-chain. The structural constraints of proline, on the other hand, restrict the observed conformations (Figure 1.6D) to two narrow regions with $\phi \cong$ -60°.

These analyses underscore the fact that the intra- and inter-chain interactions of even a modest size protein can be quite complex—the free energy expression is also complicated by contributions from hydrogen bonding, electrostatic interactions and the hydrophobic effect. However, because the nature of individual amino acids restricts the allowable chain conformations, the observed protein architecture is generally thought to arise directly from the information encoded in the amino acid sequence. The diversity of global interactions may frustrate the formulation of a precise sequence/structure relationship for globular proteins, but the analysis of sequence dependent behavior of defined micro-domains and folding habits of protein structural components is sufficiently well developed to attempt the design of new chain architectures.

1.4. Sources of Structural Design Criteria

1.4.1. Studies of Globular Proteins

The dependence of protein secondary structure on amino acid sequence is illustrated by the remarkable self-assembly capabilities of many proteins refolded *in vitro*. The actual molecular mechanism of protein folding remains unclear because the vast conformational space accessible to the chains prevents adequate definition of the possible pathways from

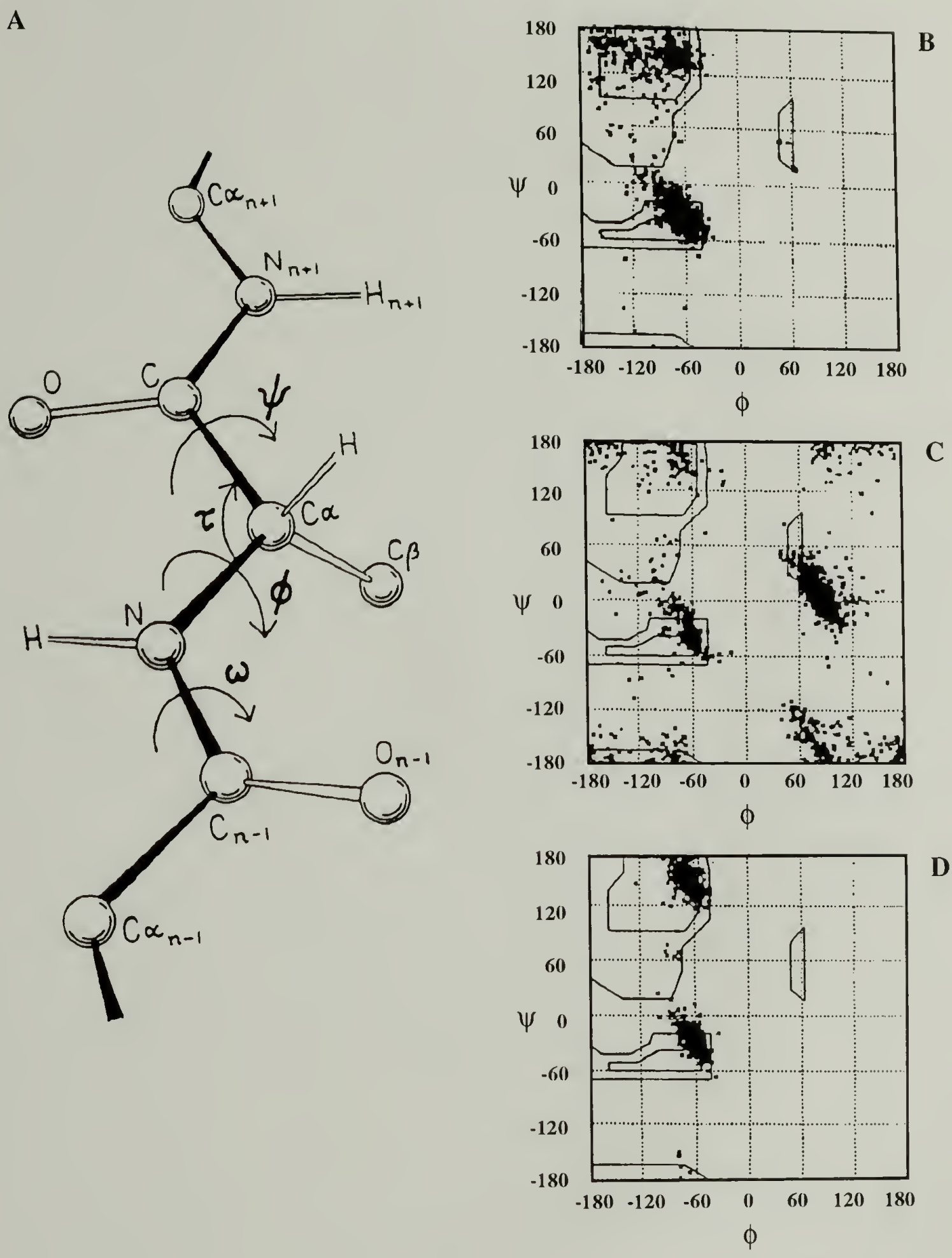


Figure 1.6. Description of polypeptide chain parameters. **A**: Definition of peptide chain dihedral angles ϕ,ψ and ω; from Richardson and Richardson.[15], reprinted with permission from Plenum Press. Ramachadran ($\phi\psi$) plots of **B**: L-alanine, showing the distribution of torsion angles typical of residues with side-chains; **C**: glycine, with more conformational freedom from the lack of a side-chain; and **D**: L-proline, showing the restricted conformational space imposed by the constraints of the pyrrolidine ring. Adapted from Richards[16], reprinted with the permission of W. H. Freeman and Company.

an unfolded to a folded state. The literature concerned with the protein folding problem is vast and well reviewed.[15] However, in his review, Jaenicke builds a strong case for the consensus that the three-dimensional structure adopted by the nascent polypeptide chain is arrived at by the modular assembly of micro-domains.[17] This understanding rests on a large and detailed data base on the structural details of the final, folded state of globular proteins derived from X-ray crystallography and increasingly, nuclear magnetic resonance (NMR) studies.

1.4.2. Polypeptides as Structural Models for Proteins

In the above cited example of the analysis of the cross-ß structure of *Chrysopa flava* silk, consideration of a repetitive polypeptide was used to gain insight into the details of the structure of a complex natural material. In fact, the study of synthetic repetitive polypeptides has enabled the refinement of the structure of many of the fibrous proteins. This is possible because

> "polypeptides have simple, known amino acid sequences, they more often pack in small well-defined unit cells, and their X-ray patterns are unlikely to be complicated by the effects of long-range distortions and extraneous components. Hence, one can calculate interatomic distances and X-ray intensities corresponding to various possible conformations and, by these criteria, establish the structure with far less ambiguity than is possible in direct studies of fibrous proteins. Corresponding features of the X-ray patterns or other physico-chemical observations of the proteins are then interpretable by analogy with the polypeptides.[18]

The process is iterative: the expansion of the structural data base derived from comparison of synthetic polypeptides to their natural counterparts allows the structural details of newly discovered biomaterials to be explored with confidence. Table 1.1 gives a brief overview of some natural structural proteins with analogous synthetic polypeptides. The notable example is the use of poly(L-alanylglycine) and poly(AlaGlyAlaGlySerGly) in the refinement of the structure of the commercially useful *Bombyx mori* silk fibroin proteins (for a recent review, see Kaplan[19]).

The continuing importance of this approach is illustrated by the recent intense interest in the *Nephila clavipes* spider dragline silks which are also thought to have repetitive chemical structures.[20] These materials have comparable extensibility to the silkworm silks, but with higher modulus and strength (a modulus as high as 30 GPa). Analysis of the structure of the proteins has focused on determining repetitive sequences that display similar performance characteristics as the native fibers. A 34 amino acid repeat has been proposed as a consensus sequence, and it seems likely that there are at least two separate protein components[21], but the molecular basis of the excellent material properties of the spider silks remains unclear.[22]

Despite the apparent complexity of the spider silks, the basic premise of structural hierarchy remains valid: repetitive sequences of amino acids define simple structural motifs that in turn direct the often complex behavior of the assembly. We have already defined the

Table 1.1. Structural proteins and model compounds

Material	Sequence	Structural Motif	Model Sequence	Reference
Silk Fibroin *B. mori*	(GAGAG[SG $(AG)_2]_8$GAAGY)	ß-sheet	Poly(AG) Poly(AGAGSG)	23-25
Spider Silk *N. clavipes*	(GNAG) (AAAAAA) (GYGGKG)	ß-sheet and α-helix	(AGRGGNGAG $[A]_6$GGAGNGGY GGLGGNG)	22
Collagen	(PXY)	Triple helix	Poly(PGP)	26
Elastin	(VPGVG) (VPGG) (APGVGV)	ß-spiral	Poly(VPGVG)	27, 28
Mussel adhesive *M. edulis*	(ALPSY-Hyp-Hyp-T-Dopa-L)	Complex composite materials	Poly(ALPSY-Hyp-Hyp-T-Dopa-L)	29, 30

two structural motifs of polymers designed to chain-fold in a precise fashion: the reverse turn and the straight stem. We know the blueprint for the three dimensional structure of protein polymers is encoded in the amino acid sequence. We can now focus on the known behavior of polypeptides for refinement of the design criteria.

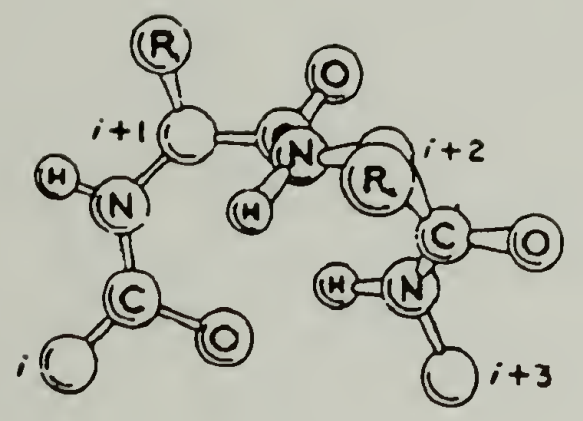
1.5. Turn Structures

Globular proteins, in general, have a compact, roughly spherical shape, with regions of extended structures surrounded by secondary structure often categorized as "random". Examination of such a structure reveals the existence of amino acid sequences that reverse the trajectory of extended structures away from the surface of the protein globule and back into the hydrophobic core. Consideration of the geometrical constraints of chain reversals can help rationalize the identity and geometry of the amino acid constituents of reverse turns in globular proteins.

Venkatachalam[31] first described three types of reverse turns consisting of a four residue corner with a hydrogen bond between the carbonyl of the first residue (the *ith* residue) and the amino proton of the fourth (*i+3*). Chou and Fasman[32] refined the description by defining any four residue region in globular proteins (they looked at the crystal structures of 421 proteins) as a non-helical structure with a distance between the α-carbons of the *i* and *i+3* residues less than 7 Å. In the data base examined, 75% of the structures fell into one of the three categories defined by Venkatachalam. Chou and Fasman determined the positional preference for the placement of amino acids in turn structures and offered a classification scheme designed for use in protein structure prediction. Notable among their observations is the high relative abundance and positional preference of proline in the *i+1* position of the examined turn structures.

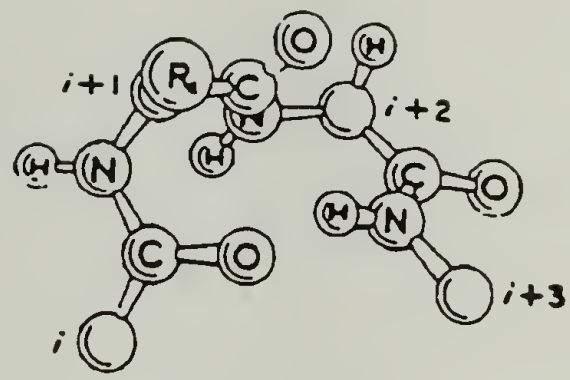
In the search for turn structures for use in a chain-fold repeat, the stringent demands of creating a chain reversal of minimal structure and residue involvement quickly narrow the field to the type I and type II ß-hairpins shown in Figure 1.7A and B. While these

A

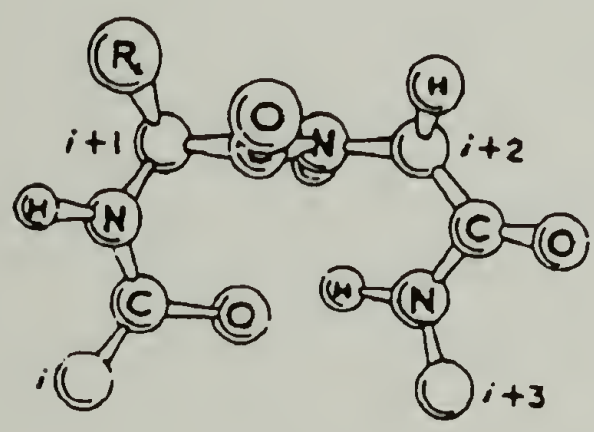


Type I

C

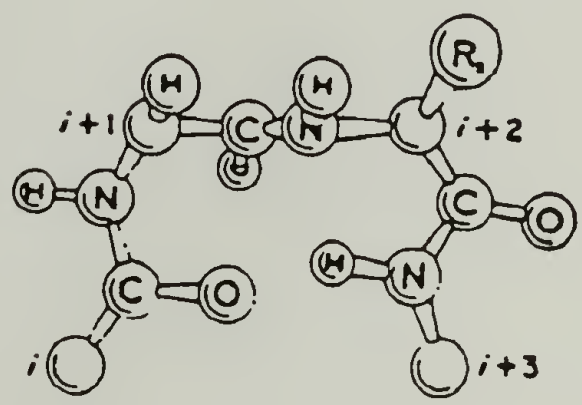


Type I'



B

Type II



Type II'

D

Figure 1.7. Structure of ß-turns.The four types of turns reverse a polypeptide chain trajectory with minimal residue involvement. Types I (**A**) and II (**B**) differ essentially by the direction of the amide bridge between residues *i+1* and *i+2*, with the carbonyl oxygen pointed away from and toward the viewer in Types I and II turns, respectively. The Type I' (**C**) and II' (**D**) turns are the mirror images of the Types I and II turns, respectively. From Wilmot and Thornton[33]; reprinted with permission from Academic Press Ltd.

structures link a variety of extended strands, Sibanda and Thornton pointed out that the tight hairpin connection between antiparallel ß-strands is complicated by the right hand twist of the extended structure (*vide infra*).[34] Considering only those two residue loops connecting two fully hydrogen bonded strands, the majority of structures (27 out of 34) are described by types I' and II" turns (Figure 1.7C and D), which have the correct orientation to match the twist of the ß-strands. These turns represent only 3% and 5%, respectively, of all turns in globular proteins.[32] Although ß-turns exist in a variety of structural environments, the type I' and II' turns are specific (83% and 50%, respectively) to tight ß-hairpins.[34] Sibanda and Thornton also noted the predominance of glycine in these structures, owing to the residue's expanded conformational freedom.

In recent work, Sibanda, Blundell and Thornton have defined a comprehensive classification scheme for the ß-hairpins designed for use in "modeling by homology", where the sequence of a new protein structure is mapped onto the known coordinates of a homologous structure.[35] This data base illustrates the wide variety of turn structures in globular proteins and facilitates the design of new structures for protein engineering. In the present context, the tight hairpins are of the most interest for the control of chain-folding, however the work of Sibanda, Blundell and Thornton is an excellent example of the comprehensive data bases available for use in the design of new architectures.

1.6. Extended Structures

When searching for design criteria of a stem motif for use in chain-folding, the same stringent criteria used in considering turn design must also apply. In the early 1950s, Pauling, Corey and Branson defined two fundamental extended packing arrangements of polypeptide chains comprised of amino acids with planar *trans* amide linkages, which display no close atom contacts and form the maximum number of linear hydrogen bonds.[36] These structures, the α-helix and the ß-sheet, are illustrated in Figure 1.8.

α-helices (Figure 1.8A) have been employed as a stem structure in some well-defined molecular architectures. In fact, chains of poly(L-tyrosine) crystallize as lamellar single crystals comprised of packed arrays of helices, each chain antiparallel to its neighbors and connected by an as yet undefined turn structure.[37] While the study of these structures may be important for defining the constraints of polypeptide chain-folding, in this study, the strands of the ß-sheet were selected for use as an extended stem.

1.6.1. Structural Details of the ß-Sheet

The ß-strand structure, seen in Figure 1.8B, can be considered a helix with two residues per turn and a rise per residue of 3.47 Å. As opposed to the α-helix, the hydrogen bonding is inter-chain with the strands in one of two packing motifs: with the N-termini of adjacent strands aligned (parallel) or opposed (antiparallel).[38] In globular proteins where the sheet domains are of modest size, there is no preference for one orientation (although parallel sheets with less than four strands are rare[39]). However, as will be seen below, the antiparallel chain arrangement predominates in the extended ß-sheets found in fibrous proteins and synthetic polypeptides. Each residue of a ß-strand is rotated 180° with respect to its neighbors, so side-chains alternately extend from opposite faces of the sheet. If possible, the residues display alternating hydrophobic and hydrophilic side-chains, allowing the simultaneous maintenance of polar and non-polar sheet contacts.[38]

The criteria presented above describes an idealized ß-sheet. The hydrogen bonding and steric requirements of side-chain packing induce two consistent deviations from ideality: pleating and twisting. Pleating at C_α carbons occurs with a general contraction of the sheet, and the residue dihedral angles are observed to differ significantly from an all *trans* conformation (ϕ,ψ = -120°, -140°). The sheet is corrugated, with a pronounced displacement of the side-groups normal to the sheet plane. This corrugation is accompanied by a left-handed twist in extended sheets, a consequence of the helical nature of the strands. Models of residue placement in a roughly two-fold helix show that hydrogen bond contacts

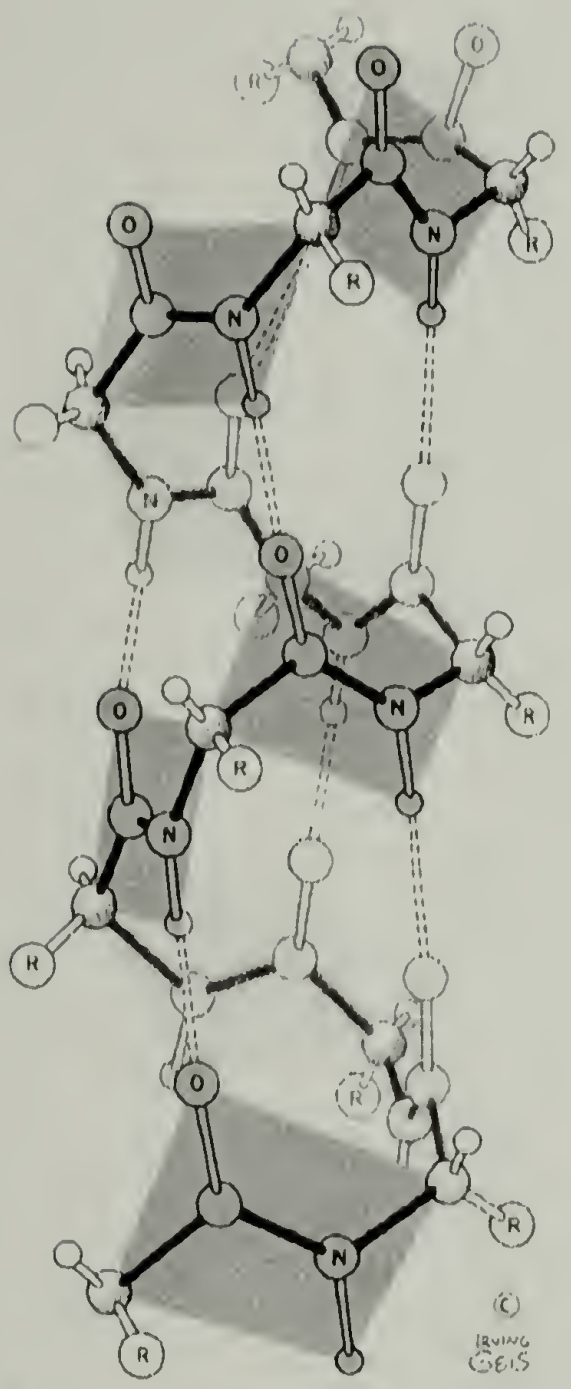


Figure 1.8. Structural details of extended polypeptide structures. **A**: α-helix and **B**: ß-sheet. Details of the ß-sheet structure are discussed in section 1.6.1. From Cantor and Schimmel[38]; reprinted with permission from Irving Geis.

are optimized with a 25° angle between strands, which is translated into the observed twist over the length of the extended sheet.[39] These intricacies of the structural organization of ß-sheets has important consequences in the macroscopic properties of materials that adopt the architecture.

1.6.2. Model Proteins and Polypeptides

Several model materials that adopt ß-sheet structures stand to illustrate the simplicity and subtleties of the ß-sheet architecture. The models discussed are but a subset of a large array of materials that exploit this organizational scheme. Each is presented to introduce the analytical techniques used to probe the solid state structure of these materials and to provide details on the sequence dependent variations in the details of the ß-sheet structure.

1.6.2.1. Polyglycine

The investigation of the structure of form I of polyglycine is an example of the combined use of X-ray diffraction and infrared spectroscopy in the elucidation of the nature of the ß-structure. Because well-oriented samples of polyglycine I were difficult to obtain, the 4.77 Å interchain and 3.44 Å intersheet distances assigned from prominent X-ray reflections were first interpreted in terms of a single chain in the unit cell.[40] However, the predicted parallel chain alignment did not correlate with the weak infrared signal at 1695 cm^{-1} (in addition to the 1630 and 1525 cm^{-1} amide I and amide II vibrations characteristic of the ß-sheet) which is indicative of an antiparallel chain arrangement. This infrared signature arises from frequency splitting of the vibrations of amide groups in antiparallel chains and is easily distinguished from the absorbances associated with an α-helix or random coil. The antiparallel chain arrangement suggested a basic repeat encompassing two chains, and the structure was refined to account for the new data.

It was also determined that the polyglycine sheets were rippled as opposed to pleated.[41] This structure is best described by placing imaginary side groups on the glycine

residues so that alternate strands are comprised of residues of opposite chirality: one chain of L-isomers neighboring a chain of D-isomers. In all residues but glycine, this structure is possible only with an equimolar mixture of D and L chains. Glycine, with no side-chain, apparently prefers this chain packing arrangement.

This study highlights the techniques used to probe the packing of polypeptides in antiparallel ß-sheets. The primary diagnostic is the 1700, 1630 and 1525 cm^{-1} infrared signature, which is sensitive to the structural environment of the residues. Information regarding the packing of the chains however, is obtained from X-ray scattering analysis of oriented samples. The signals arising from the interchain distance of 4.7 Å are prominent and thus the hydrogen bonding direction is generally easy to assign. Analysis of this kind helped define the unique cross-ß structure of the *Chrysopa flava* silks.

As shown above, the interchain and chain axis spacings within ß-sheets are fairly sequence independent. This is a consequence of the fact that the side-chain groups project from the face of the sheet. Therefore the intersheet spacing is critically dependent on and complicated by side-chain architecture and is often the most difficult structural parameter to assign. As Table 1.2 shows, there is a remarkable constancy to the interchain and chain axis parameters, but the dependence of the intersheet spacings on sequence is obvious. Some examples serve to demonstrate the prominent sheet packing motifs.

1.6.2.2. *B. mori* Silk Fibroin and Poly(L-alanylglycine)

The prominent example of a material displaying an antiparallel ß-sheet structure is the silk fiber from the cocoon of the *Bombyx mori* silkworm, examined in detail by Marsh, Corey and Pauling.[42] The structural characteristics of silk fibroin can be approximated by a number of repetitive polypeptides (Table 1.2) with increasingly smaller repeat units. In fact, the ß-structure of silk fibroin was refined upon analysis of poly(L-alanylglycine), and can best be described in terms of the structural characteristics of the synthetic analogue.

Table 1.2. Structural parameters of ß-sheets. Adapted from Fraser and MacRae.[43]

	Structural Parameter (Å)		
	Intersheet distance	Axial rise/residue	Interchain distance
Polyglycine	3.44	~3.48	4.77
B. mori fibroin	3.87, 5.33	3.49	4.70
B. mori fibroin, Cp fraction*	3.87, 5.26	3.44	4.69
Poly(AGAGSG)	3.79, 5.26	3.43	4.70
Poly(L-alanylglycine)	3.79, 5.17	3.47	4.72
Poly(L-alanine)	5.27	3.45	4.73
A. pernyi fibroin	5.30	3.47	4.72

*The precipitate formed after digestion of solubilized silk with chymotrypsin

The structure of poly(L-alanylglycine) was examined rigorously by Fraser and coworkers[25] who described a unit cell structure with $a = 9.44$ Å (hydrogen bond direction), $b = 8.96$ Å (intersheet direction) and $c = 6.94$ Å (chain axis direction). Infrared dichroism supported the proposed antiparallel arrangement and the two-chain unit cell and confirmed X-ray data which indicated a cross-ß chain alignment.

The various sheet packing arrangements of poly(L-alanylglycine) were considered based on the implications of an extended two-fold helix comprised of exactly alternating residues. As described above, the side-chains project on opposite sides of the strands. It was realized early that the sheets were decorated on only one face with alanyl methyls, but two sheet arrangements were considered: a sequential alignment of alanine faces to glycine faces or an alternating structure with alanine-alanine face contacts and glycine-glycine face contacts. Fraser and coworkers modeled the X-ray pattern of the alternating structure which indicated that there should be significant X-ray intensity arising from the (010) and (030) reflections, based on the alternating distance between the two faces. This prediction was confirmed for well-oriented samples. In addition, the modeling confirmed that the (200) and (210) reflections were consistent with an $a/4$ stagger between adjacent sheets.[44] Based

on these data, the model was refined enough to assign the two components of the 9.44 Å inter-sheet distance: b_{Gly} = 3.79 Å and b_{Ala} = 5.17 Å. This chain packing arrangement (shown in Figure 1.9) was termed poly(L-alanylglycine) form I (PLAG I).

The study of this model allowed the refinement of similar structural details of silk in crystalline form II (silk II), rationalized in terms of the X-ray and infrared signals of PLAG I. As opposed to PLAG I, the polypeptide chains are aligned parallel with the fiber axis (direction of orientation). The interchain and chain axis dimensions of silk II and PLAG I were almost identical (for silk II: a = 9.46 Å and c = 6.97 Å; note that Marsh and coworkers assigned the chain axis as the b direction). The X-ray data suggested the same sheet arrangement as shown in Figure 1.9 with an overall spacing of b = 9.20 Å, however the disorder induced by the constraints of packing serine and tyrosine residues between sheets made assignment of the intersheet component distances more difficult. Marsh and coworkers gave values of b_{Gly} = 3.5 Å and b_{Ala} = 5.7 Å. Lotz, in a later treatment suggested the values b_{Gly} = 3.87 Å and b_{Ala} = 5.33 Å, based on consideration of the packing arrangements of polyglycine and poly(L-alanine).[45]

It should be noted that in a recent study based on conformational modeling and new sample preparation techniques, Takahashi and coworkers presented a model for silk II with alanyl methyls decorating both faces of the sheet.[46] While their approach is intriguing, little experimental data was presented to support their model.

Silk and poly(L-alanylglycine) can be isolated in other related structural forms known as silk I and PLAG II, respectively. Poly(L-alanylglycine) in this arrangement (obtained by dialysis of the polypeptide from an aqueous lithium bromide solution) can be indexed on an orthogonal unit cell (a = 4.72 Å, b = 14.44 Å and c = 9.6 Å)[47] with the asymmetric unit a pair of residues with an axial run of 4.8 Å. The chains are hydrogen bonded in the *ac* plane and exhibit a deep pucker by virtue of helix-like dihedral angles for glycine and an almost standard ß-conformation for alanine. The chain trajectory seems to be described best by a crankshaft pattern.

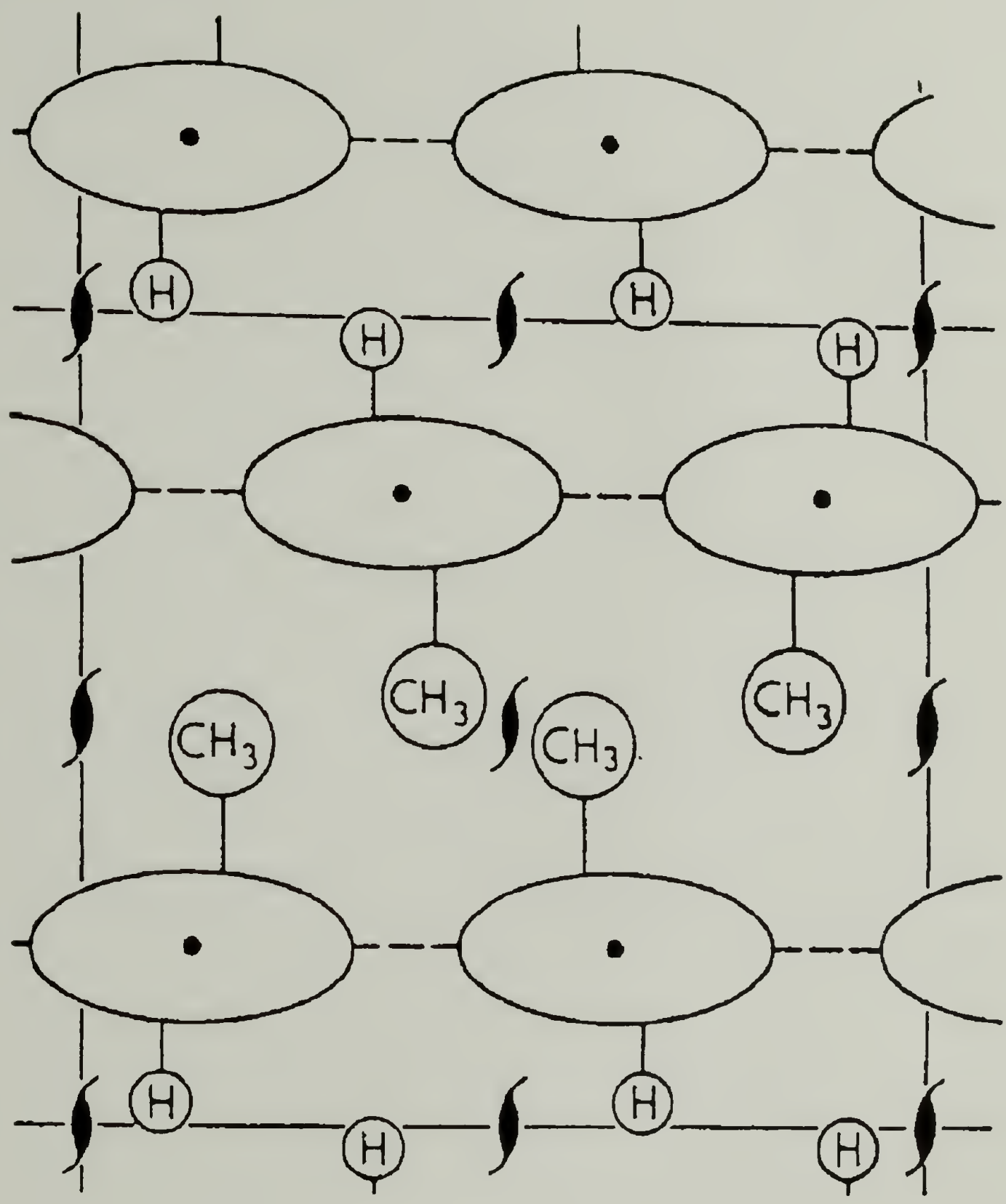


Figure 1.9. Side-chain packing arrangement of poly(*L*-alanylglycine) in form I. The view is down the sheet (along the chain axis) with the packing arrangement of alternating glycine-glycine and alanine-alanine face contacts emphasized. The *a*/4 stagger between sheets is apparent upon examination of the methyl group spacings in adjacent sheets. The details of the structure are discussed in section 1.6.2.2. Adapted from Fraser, MacRae, Stewart and Suzuki[25]; reprinted with permission from Academic Press.

The structure of silk I is less refined than that of the silk II form because chains exhibiting the silk I signature easily deform to a disordered ß-structure under shear, making analysis of oriented patterns difficult. Thus, the assignment of the sheet-packing arrangement is ill-defined. Models by Lotz[45], and more recently by Kaplan and coworkers[19] and Scheraga and coworkers[48], underscore the challenge of refining the details of the structure. For the purposes of identification of chains in this arrangement, a prominent X-ray reflection at 7.21-7.30 Å serves as a useful diagnostic.

1.6.2.3. Tussah Silk Fibroin and Poly(L-alanine)

The structure of poly(L-alanine) is an example of another sheet-packing arrangement. The structural dimensions of 4.73 Å (*a*, interchain), 10.54 Å (*b*, intersheet), 6.89 Å (*c*, chain axis) were assigned on the basis of infrared data showing the presence of a ß-sheet. The overall disorder of the X-ray data did not allow firm assignment of the intersheet structure, but Marsh, Corey and Pauling noticed a similarity in the X-ray signals for poly(L-alanine) with those of the Tussah silk fibroin.[42] This fibrous protein, with a 44% alanine content, has the sheet structure seen in Figure 1.10. The sheets, decorated on both sides with alanyl methyls, are separated by a constant difference of 5.30 Å and have an alternating *a* axis stagger of $+a/4$ and $-a/4$. Arnott and coworkers used these data, especially the close correlation in the 5.30 Å with that seen for poly(L-alanine) to suggest that the inherent disorder in the patterns for the synthetic polypeptide arises from a statistical arrangement of the sheets in a $\pm a/4$ stagger, instead of the exactly alternating arrangement of sheets in the protein.[49] The value of the intersheet distance correlates well with the b_{Ala} distance determined for poly(L-alanylglycine). Once again, the X-ray data were interpreted in terms of specific, sequence dependent packing arrangements of the chains.

The preceding studies demonstrate the key role the analysis of synthetic polypeptides plays in the assignment of structural details of fibrous proteins.

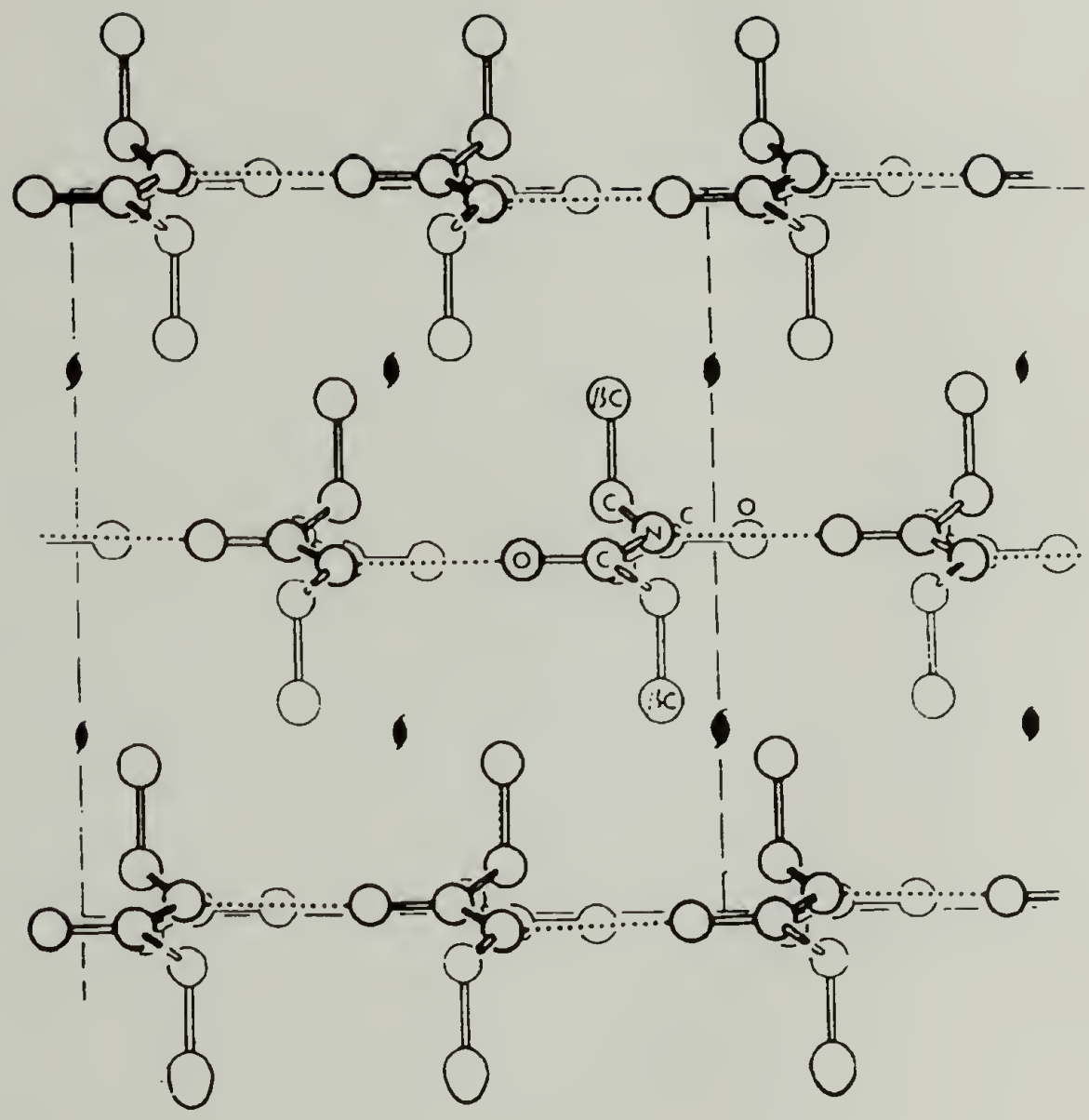


Figure 1.10. Side-chain packing arrangement of *Antheraea pernyi* Tussah silk fibroin. The view is down the sheet (along the chain axis). Both sides of the sheets are decorated with alanyl methyls, with an alternating sheet stagger of +*a*/4 and -*a*/4. The details of the structure are discussed in section 1.6.2.3. From Fraser and MacRae[50]; reprinted with permission from Academic Press.

More importantly, the results serve to supply specific details of polypeptide structure useful in the design of a chain-folded architecture.

1.7. Biosynthesis of Polypeptides

1.7.1. Limitations of Current Synthetic Methods

Supplied with the structural insight necessary for the design of an appropriate polypeptide sequence, the synthesis of the chains with sufficient precision to match the design remains a challenge. There are a variety of synthetic methods currently used for the production of model polypeptides. For instance, homopolymers of amino acids can be synthesized by the ring opening of *N*-carboxyanhydrides (NCAs), but there is no control of molecular weight or distribution.[51] The only feasible route to repetitive copolypeptides is the condensation of oligopeptides of the desired repeat sequence.[52] However, the synthesis of these macromonomer units involves optimizing solvent conditions and side-chain blocking groups and as before, the molecular weight characteristics are dominated by the statistics of the polymerization process. Solid phase techniques aid in the synthesis of monomers of defined length, but yields are limited by the resin functionality and the accumulation of error sequences cap the length of the chains.[53]

On the other hand, the synthesis of amino acid sequences through the use of genetic engineering is a well established technology, until recently exploited mainly for the production of proteins of biological interest. Since there seems to be no theoretical impediment to the production of repetitive protein polymers of high molecular weight[6, 54] the use of this technology represents an important advance in synthetic routes to high polymers. The mechanism of cellular protein biosynthesis enables the precise placement of individual monomers along a chain of defined length, an ability impossible with current chemical synthesis schemes. Thus, successful polymer biosynthesis can yield pure

materials in the classic sense: each chain in the sample described by a single molecular weight, precise sequence and stereochemistry.

1.7.2. General Biosynthetic Strategy

In general, a strategy for the biosynthesis of repetitive polypeptides begins with the design of a short oligonucleotide encoding the target peptide sequence, which is chemically synthesized and enzymatically polymerized to yield a family of DNA multimers, each encoding a discrete number of repeats of the sequence. Multimers of a particular size are isolated and inserted into an expression system consisting of a host cell (generally the bacterium *E. coli*) and the requisite DNA machinery to control the synthesis of the target protein. Once constructed, this system can be maintained as a production facility from which target protein can be produced in the cell, recovered and purified.

The first published report using such a strategy, by Doel and coworkers[55], demonstrated the construction of synthetic polymeric genes encoding repeats of the dipeptide L-aspartyl-L-phenylalanine. While not motivated by the production of polymeric materials, they demonstrated that repetitive DNA sequences of up to 900 base pairs were maintained in *E. coli*. The production of poly(L-aspartyl-L-phenylalanine) was demonstrated by the incorporation of radiolabeled aspartic acid and phenylalanine into the recombinant proteins expressed in *E. coli* and identified by gel electrophoresis.

1.7.3. Biosynthesis of Rationally Designed Architectures

There have been a number of attempts to biosynthetically produce rationally designed materials. Goraj, et al., for instance, expressed in *E. coli* a polypeptide designed to adopt an eight stranded α/ß-barrel motif.[56] The structure of the purified material was probed using circular dichroism spectroscopy and gel electrophoresis and shown to have close to the predicted helical content and to fold in a cooperative, two-state transition. Another structure, Felix (*F*our h*elix*), comprised of a 79 amino acid protein designed to

fold into an up and down four-helix bundle, was expressed in *E. coli* and purified to homogeneity.[57] Preliminary analysis of the protein by circular dichroism and fluorescence spectroscopy indicates that the chains are monomeric and predominately α-helical in solution and the structure is linked by a designed disulfide bond. While not conclusive, these data rule out several alternate structures, and more in-depth analysis may verify this unique architecture.

From the *de novo* design and analysis of new protein architectures comes important new information applicable to studies in polymer science and structural biology. The topic is well reviewed[58], especially DeGrado's "minimalist approach" to protein design[59], where the amino acid sequence of a new protein is determined upon consideration of the sequence dependent structures of small peptide subunits. Whatever the approach, it is increasingly apparent that the ready biosynthesis of defined protein chains in bacteria is the most direct route to precise protein structures.

1.7.4. Biosynthesis of New Materials

A number of important repetitive polypeptides based on the structural proteins have been synthesized *via* biosynthetic routes, including elastin[60] and a structural component of the mussel adhesive protein.[61] Goldberg and coworkers[62] constructed a synthetic 335 base pair sequence encoding repeats of the tripeptide collagen analog GlyProPro. Larger DNA sequences were found to be unstable in *E. coli* and they demonstrated that 80% of the peptide was degraded within 40 minutes of induction.

Despite this and other failures[63, 64], the notable success of Cappello and coworkers in the production of silk-like proteins (SLP) and silk-elastin-like proteins (SELP) demonstrates the power of this technology.[65, 66] The SLP materials, featuring repeats of the hexapeptide GlyAlaGlyAlaGlySer prevalent in silk fibroin, were found to exhibit behavior consistent with the formation of hydrogen-bonded ß-sheets. The SELP materials were designed as modifications of the SLP's with the incorporation of potentially

noncrystallizable sequences GlyAlaAlaGlyTyr, derived from silk fibroin[67] and ValProGlyValGly, derived from mammalian elastin.[68] Variants of these sequences have been produced in multi-gram quantities and form the basis of a commercial product, Pronectin™ F, which incorporates multiple copies of the RGD cell attachment ligand of human fibronectin in structural peptides based on the SLP materials.[69]

In an excellent recent description[69] of the development of Pronectin™ F, Cappello places the present work in a larger context by pointing out that

> "investigators in the field hope that through sequence control, the production and study of protein polymers will ultimately lead to the development of revolutionary new materials that will satisfy some of the complex materials properties needed in such applications as medicine, health care, diagnostics, and sensors."

This unique approach to the synthesis of novel materials will continue to gain importance as it is exploited in polymer science, structural biology and pharmaceutics.

1.8. Background and Context of the Present Work

1.8.1. Initial Effort Towards Controlled Chain-folding

The present study is based on the work of McGrath and coworkers, who demonstrated the production of new materials with the goal of preparing "architecturally well-defined polypeptides of predictable solid-state structures".[70] Using criteria of the type discussed above, a repetitive polypeptide was designed to adopt, upon crystallization, a well-defined lamellar architecture defined by the stem-turn repeat sequence **1.1**.

$$[(AlaGly)_3ProGluGly] \qquad \mathbf{1.1}$$

A straight stem sequence consisting of three alanylglycine dyads was proposed based on structural analysis of extended poly(L-alanylglycine) chains.[25, 71] It was proposed that

each of these nearly fully extended ß-strands would be interrupted by proline, which was shown to break extended secondary structure (helices or sheets) in globular proteins.[72, 73] The proline residue was proposed to initiate the reverse turn; this function supported by the relatively high incidence of proline in turn structures of globular proteins.[32] The glutamic acid residue would also help to force chain reversal with its rather bulky side-chain excluded from the interior of the nascent lamella. Placement of this residue would decorate the fold surface with carboxylic acid functional groups for subsequent modification of the materials. The result was a repetitive polypeptide chain, with repeating structural units designed to extend ~30 Å before reversing direction—the aggregate of chain-folded sheets organized in a well defined lamellar architecture.

McGrath and coworkers[70] demonstrated the synthesis of a polymer with 54 repeats of sequence **1.1** in a controlled fashion in *E. coli* using a synthetic scheme based on that described in section 1.4.4. An oligonucleotide encoding two repeats of the sequence was synthesized and enzymatically ligated in a head-to-tail fashion to yield a statistical multimer population, with each DNA fragment encoding a repetitive polypeptide of discrete size. These DNA multimers were cloned and isolated, and a fragment encoding 54 nonapeptide repeats inserted into pET3-b, an expression vector in which transcription is driven by T7 RNA polymerase.[74] A recombinant plasmid containing the insert in the proper orientation was used to transform *E. coli* strain BL21(DE3) pLysS.[74] Expression of the target sequence was induced by the addition of isopropyl-ß-D-thiogalactopyranoside (IPTG) and monitored by the incorporation of [^{3}H]-glycine into proteins produced after induction. The material was isolated in milligram quantities by a simple purification scheme and the chemical structure confirmed by spectroscopy, amino acid compositional analysis and N-terminal sequencing. Mass spectrometry showed the material to be essentially monodisperse.[75]

Solid state structural analysis of the biosynthetically produced polypeptide revealed no evidence for the proposed lamellar structure. In fact, the wide angle X-ray scattering

pattern consisted of diffuse halos indicating the absence of any solid state order. In addition, the fourier transform infrared (FTIR) spectrum showed only the 1657 cm^{-1} amide I and 1540 cm^{-1} amide II absorbances typical of the peptide chains in a random conformation. These results were obtained for samples prepared as powders and films and seem to indicate that the regular, chain-folded ß-sheet is not a preferred conformation for this material.

Based on recognition of potential design flaws with sequence **1.1** (*vide infra*), a polymer comprised of 36 repeats of sequence **1.2** was produced and the solid state behavior probed by a variety of methods.[76, 77]

$$[(AlaGly)_3GluGly] \qquad \mathbf{1.2}$$

This material was demonstrated to have the X-ray scattering signature indicative of a ß-sheet conformation, which was corroborated using fourier transform infrared (FTIR) and solid state nuclear magnetic resonance (NMR) spectroscopy. More importantly, lower angle X-ray reflections were seen, which were thought to arise from a lamellar superlattice repeat of 38 Å. Upon swelling the materials in glycerol, these signals disappeared, perhaps indicating an increased spacing between lamellae induced by swelling of the interlamellar amorphous regions. These data were taken as a positive indication of the existence of lamellar structures in the sample.

In addition, other sequences with increasing numbers of alanylglycine dyads (n = 4, 5 and 6) have also shown a ß-sheet structure.[78] However, the evidence for the adoption of a lamellar structure in these sequences is not conclusive and a comparison of the small angle reflections between the sequences of increasing stem-length show no consistent pattern of lamellar thicknesses. An oriented sample comprised of similar sequence **1.3** does indicate a meridional reflection consistent with an interlamellar spacing of ~40 Å.[79] In addition, this material can also be isolated in a form with an oriented X-ray

scattering pattern thought to be consistent with a folded structure with chains in a silk I like conformation.

$$[(AlaGly)_3AspGly] \qquad \mathbf{1.3}$$

1.8.2. Apparent Design Flaws in the Proposed Sequence

The analysis of these unique materials was conducted as part of a coordinated program aimed at understanding the constraints of sequence dependent polypeptide architecture. In this context, the present study was prompted by the dramatic difference in solid-state organization of materials with sequence **1.1** and sequence **1.2**, illustrated by comparison of typical X-ray powder patterns in Figure 1.11. As will be seen, this seemingly simple sequence change, the removal of the proline residue from repetitive sequence **1.1**, alters both the hydrogen bond pattern associated with normal antiparallel ß-sheets and the nature of the putative turn structure. Thus, consideration of the pattern of hydrogen bonding associated with the sequences and the role of the proline unit in the turn is crucial to understanding the failure of sequence **1.1** to adopt the desired lamellar structure.

In retrospect, the sequence of **1.1** incorporated a "fatal flaw". The design process resulting in the nonapeptide repeat unit involved the coupling of separate stem and turn sequences which, in turn, were based logically on criteria outlined above. The straight stem comprised of three L-alanylglycine dyads was proposed to bias the structure to the formation of a ß-sheet with a measurable fold length. The proposed use of this six residue stem repeat was well founded on the structural studies of extended poly(L-alanylglycine) chains.

In a separate design process, a turn sequence was proposed to place a proline in the *i+1* position of a four residue corner: the *i*th residue would be the last glycine of the hexapeptide stem. This was a logical choice based on proline's observed role as a "breaker"

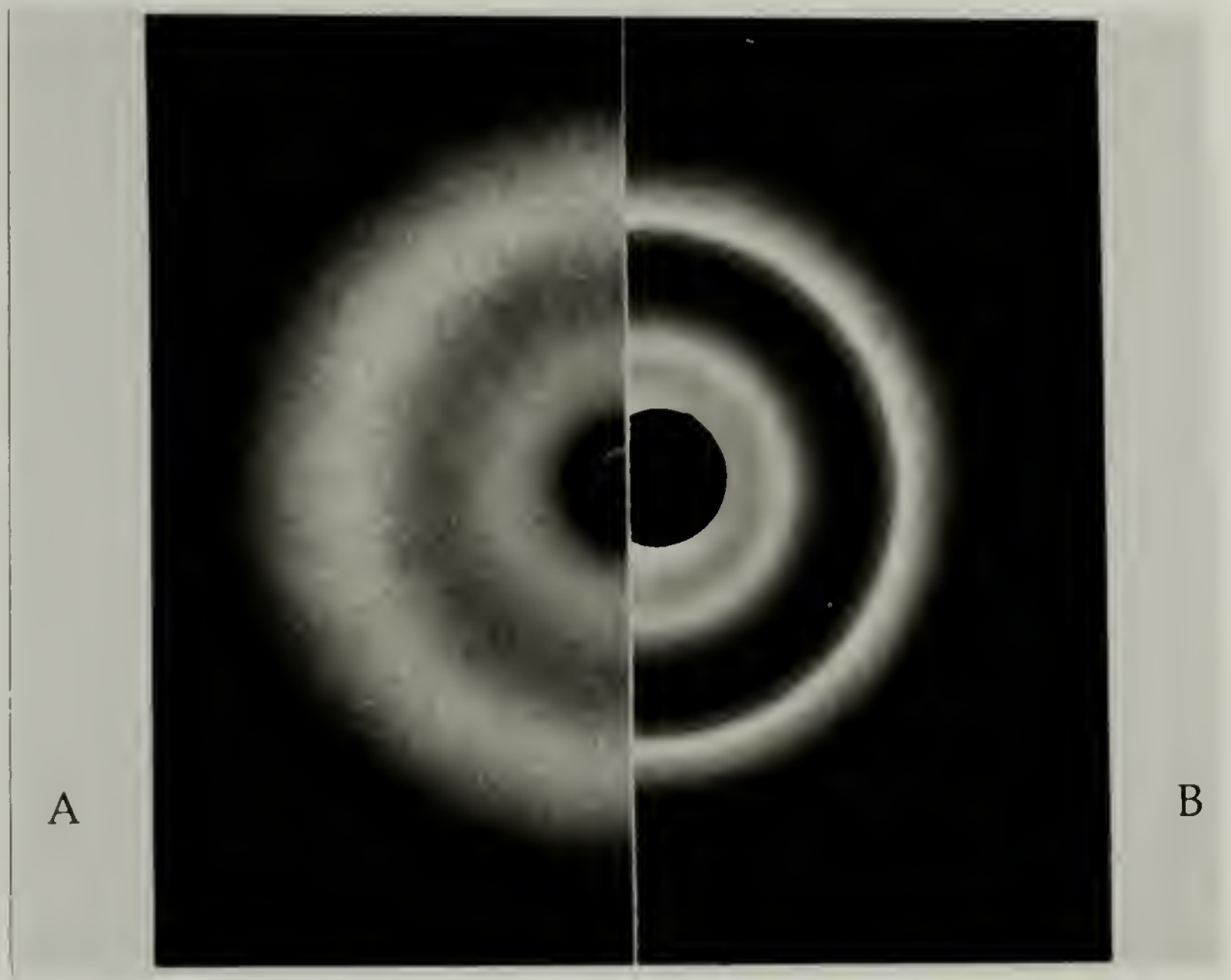


Figure 1.11. Comparison of X-ray powder patterns of repetitive polypeptides. **A**: Repeat sequence $(AlaGly)_3ProGluGly]^{75}$, with amorphous halos indicative of a relatively disordered solid-state structure, and **B**: Repeat sequence $[(AlaGly)_3GluGly]^{77}$, showing reflections associated with a ß-sheet structure. The origins and implications of this apparent increase in solid-state order is discussed in sections 1.8.1 and 1.8.2.

of secondary structure: the residue cannot participate in the regular, hydrogen-bonded structure of ß-sheets or helices (because of its bulk and the absence of the imido hydrogen) and therefore is often found at the end of extended structures.[15] In addition, a survey of turn structures in globular proteins by Chou and Fasman[32] and others[80] showed that proline has the highest positional preference of any residue and often occupies the second position of types I and II ß-turns. The placement of glutamic acid at the *i+2* and glycine at the *i+3* positions of the turns were also consistent with these studies.

The difficulty arose with the coupling of the hexapeptide stem sequence terminated with the first residue of the turn with the tripeptide sequence proposed to complete the the four residue chain reversal. The result of this construction was a repeat sequence comprised of an odd number of amino acids, the effect of which is apparent only upon consideration of molecular models of folded polypeptide chains with odd (sequence **1.1**) and even (sequence **1.2**) numbers of residues, as illustrated in Figure 1.12.

In Figure 1.12A, two repeats of the even number sequence [$(AG)_6EG$] are shown both face and edge on. The extended stem proceeds with a regular hydrogen bonding pattern until reversed by a ß-turn with standard dihedral angles. The next stem proceeds antiparallel with its neighbor, a glycine aligning with an alanine, with no interruption of the hydrogen bonding pattern. In this arrangement, the alanyl methyls of one strand are on the opposite face of the sheet from its neighbors resulting in a decoration of both faces of the extended sheet with methyl groups, reminiscent of the arrangement in poly(L-alanine) and Tussah silk fibroin. At the end of the second strand, another standard ß-turn reverses the chain direction in the correct trajectory, i.e. the same direction as the previous turn, allowing reentry of the next stem.

On the other hand, due to the nature of the hydrogen bonding pattern, a chain comprised of odd numbers of amino acids can accomodate only one normal ß-turn connection between the strands of two stem-turn repeats. As illustrated in Figure 1.12B, equivalent ß-turns on the termini of the strands precludes the alternating turn directions

required for regular chain reentry: equivalent turn structures force the chain to fold back on itself, preventing the formation an extended structure. In addition, the glycine and alanine residues in the sheet align in register, resulting in the placement of all the alanine methyl groups on one side of the sheet as in poly(L-alanylglycine) and *B. mori* silk fibroin. A model constructed with normal ß-turns followed by "forced" reverse turns that allow for regular chain reentry results in a disruption of the normal hydrogen bonding pattern associated with antiparallel ß-sheet, with each stem-turn leaving a hydrogen bond unaccommodated in the sheet.

While the effect of the odd repeat sequence is probably the major impediment to chain-folding of **1.1**, Sibanda and Thornton observed that proline generally does not participate in the requisite type I' and II' tight hairpin turn structures connecting antiparallel ß-strands in globular proteins (see section 1.5).[34] Thus, there is no precedent for the participation of proline in the turn structure critical for repetitive chain-folding of ß-strands.

Apparently then, the energy penalty of the interrupted hydrogen bonding pattern coupled with the role of proline in the proposed turn structure frustrates chain-folding in polymers with sequence **1.1**, favoring the observed conformationally disordered glassy state. The chain-folded state of **1.2**, with a sequence obtained by the removal of a proline from **1.1**, is support for this analysis.

1.8.3. Structural Criteria for Achieving Chain-folding

Given the natural tendency of polymer chains to fold, can the structure of **1.1** be adjusted in any other way (besides the removal of the proline) to bias the chains to adopt the desired folded structure? The energetics of residues in a regular ß-sheet favor its adoption: with sequence **1.1** the turn constraints restrict the necessary conformation. Thus, regular chain-folding can be realized in models of sequence **1.1** by loosening the tight hairpin connection between strands. This "unzipping" of the ß-sheet at the termini can be realized in models by postulating a *five* residue reverse turn that shifts the strands by one

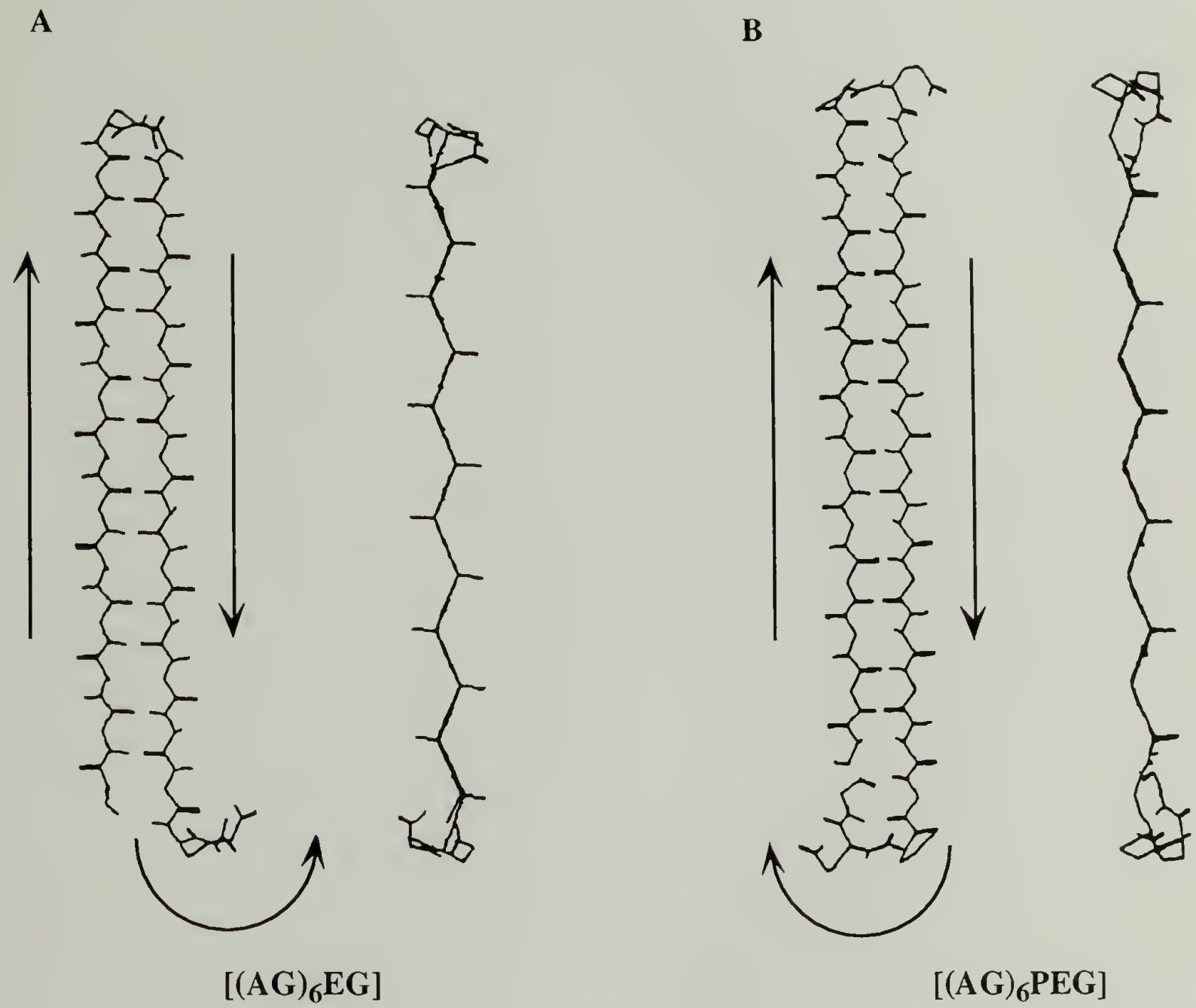


Figure 1.12. Dependence of chain trajectory and side-chain orientation on the amino acid repeat sequence of repetitive copolypeptides. Sheet normal (left) and edge-on (right) views are presented. **A**: Even repeat sequence [$(AlaGly)_6GluGly$]. Termination of the stem-turn-stem structure with an equivalent turn facilitates the adoption of the appropriate chain trajectory for reentry into the nascent lamella. The alanyl residues on adjacent chains alternate, thus decorating both sides of the sheet with alanyl methyls. **B**: Odd repeat sequence [$(AlaGly)_6ProGluGly$]. Termination of the stem-turn-stem structure with an equivalent turn reverses the chain back on itself, in the opposite direction required for reentry. The alanyl residues on adjacent chains align, thus decorating only one side of the sheet with alanyl methyls. The structures were generated on a Silicon Graphics work station using Polygraf software.

amino acid per stem turn and places each unsatisfied hydrogen bond in a bulged turn structure at the fold surface. Such a turn in each repeat allows the appropriate extended chain trajectory and results in the sheet structure seen in Figure 1.11A, with alanyl methyls decorating both faces of the sheet.

Bulged turns are common in globular proteins[35], with torsion values consistent with the constraints of the proline residue. In addition, the bulky pyrrolidine ring can be placed reasonably in such a loosened turn structure. A survey of the turn structures in globular proteins[35] reveal a number of five residue turns, with the structure stabilized by two hydrogen bonds, one between the *i* and *i+5* residues and the other between the *i* and *i+4* residues.

The constraints of a loosened structure, along with the unsatisfied hydrogen bonds in each stem-turn, still represent an energetic liability preventing chain-folding, as witnessed by the disordered solid state structure of **1.1**. Apparently, the chains do not have the freedom to adopt a loose reverse turn and the chain-folded state remains inaccessible. However, the folded state could be favored by building in stabilization energy into the bulk structure in the form of additional hydrogen bonding pairs added to the stem structure. In addition, extending the stems may lend the necessary conformational freedom to match the turn constraints for adjacent reentry.

Based on these criteria, the addition of increasing numbers of alanylglycine dyads in the proposed stem of repeat sequence **1.1** should result in a bias towards a chain-folded architecture by virtue of a progressive dilution of the energetic effects of the inherent hydrogen bond mismatch and the increased likelihood of the adoption of an appropriate turn structure. Thus a series of copolypeptides of general sequence **1.4** would be expected to proceed from the observed glassy state (n = 3) to a chain-folded crystalline state on the basis of the increased stem-length (n = 4, 5, and 6). A comparison of models of the proposed structures is found in Figure 1.13, showing the increased length of the stem

relative to the turns, the loosened turn structures and the decoration of both sides of the proposed sheet with alanyl methyls, a consequence of the equivalent turns in each stem. The turn structure was modeled with torsion angles derived from a five-membered turn formed by residues 35-39 of γ-chymotrypsin (found in the data base of Sibanda, Blundell and Thornton).[35] The effects of the longer stem lengths are evident in Figure 1.13, with the relatively stable ß-sheet beginning to dominate the structure with increasing stem length.

$$[(AlaGly)_nProGluGly]_m \qquad \textbf{1.4}$$

This work describes the synthesis of materials of sequence **1.4** with stem-lengths of n = 3, 4, 5, and 6. It is hoped that the solid state characterization of these materials will help define the determinants of chain-folding in periodic polypeptides.

1.9. Organization of the Thesis

The thesis is organized into five chapters. Chapter 1 is designed to provide the rationale and context for the design, synthesis and analysis of the copolypeptides of sequence **1.4**. Chapter 2 is provided to outline the basic concepts of recombinant DNA techniques. Chapter 3 describes the biosynthesis of the copolypeptides examined in this study; the synthesis of a polymer with fourteen repeats of an undecapeptide repeat (n = 4) is described in some detail. Chapter 4 presents the analysis of this material by matrix-assisted laser desorption/ionization mass spectrometry and demonstrates the power of the technique in the analysis of protein molecular mass, sequence and purity. The preliminary mass spectral analysis prompted an extended study of a series of polymers with sequence **1.4** (n =4 and m = 10, 12, 14, 16, 20), also discussed in Chapter 4. The stem length variants of sequence **1.4**, with n = 3, 4, 5, and 6 and length m = 16 were synthesized and their solid state behavior examined by X-ray diffraction, spectroscopy, microscopy and thermal analysis; the results of this analysis are presented and discussed in Chapter 5.

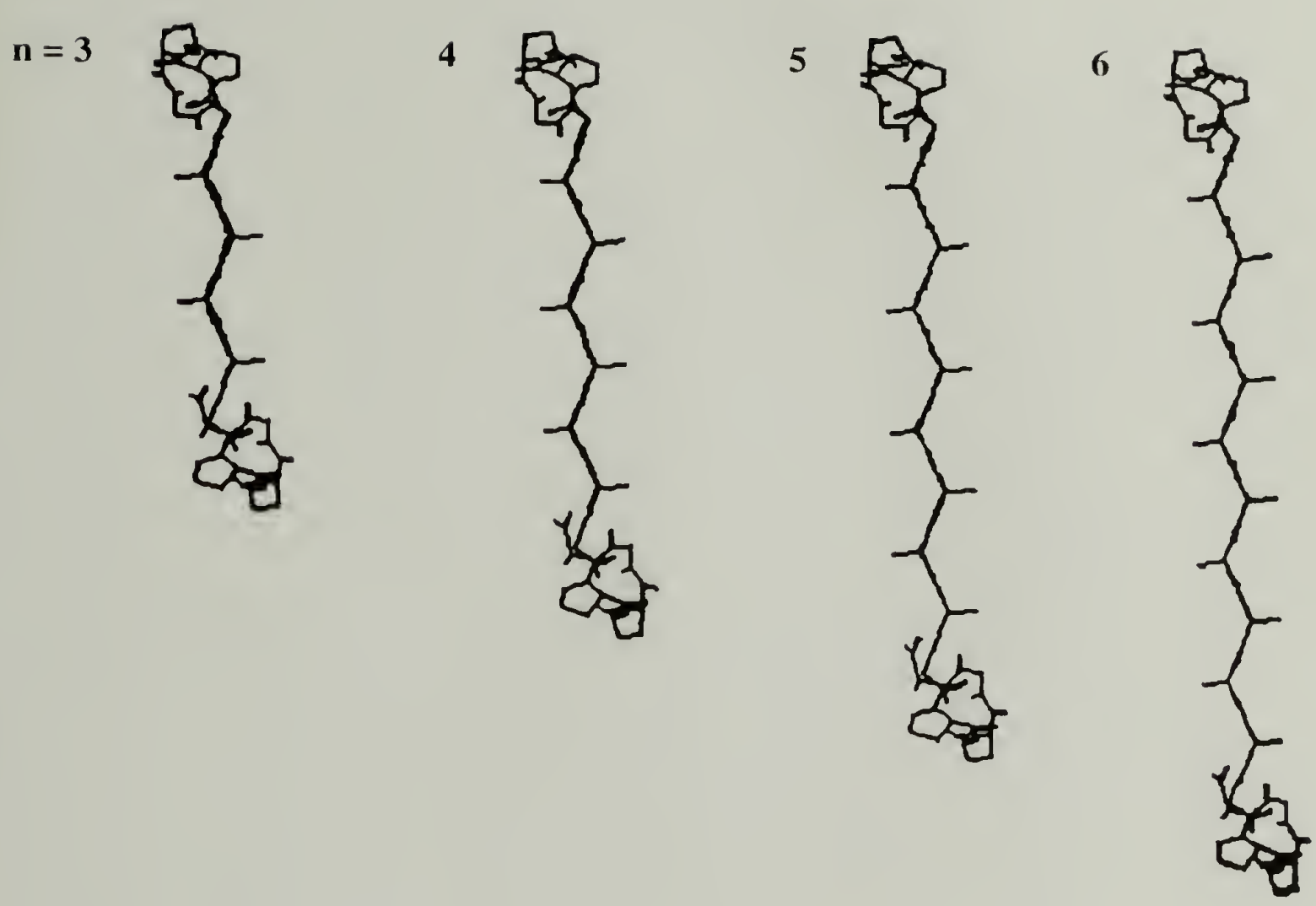


Figure 1.13. Models of polypeptides with sequence [(AlaGly)$_n$ProGluGly]. Left to right, n = 3, 4, 5, and 6. Sheets are viewed edge on. The structures were modeled with "loosened" turn structures using torsion angles derived from a five membered turn formed by residues 35-39 of γ-chymotrypsin (from the data base of Sibanda, Blundell and Thornton[35]). The decoration of both sides of the sheets with alanyl methyls (as in the structure of [(AlaGly)$_6$GluGly] in Figure 1.12A) and the increasing amount of ß-sheet relative to the turn structures with increasing stem-length is emphasized. The structures were generated on a Silicon Graphics work station using Polygraf software.

1.10. References

1. A. Keller, *Philos. Mag.* **2**, 1171 (1957).

2. B. Wunderlich, *Macromolecular Physics* (Academic Press, New York, 1973), vol. 1, Chap. 3.

3. B. Lotz, *Mater. Res. Soc. Symp. Proc.* **255**, 95 (1992).

4. L. H. Sperling, *Introduction to Physical Polymer Science* (Wiley-Interscience, New York, 1986).

5. D. C. Bassett, *Principles of Polymer Morphology* (Cambridge University Press, Cambridge, 1981).

6. M. J. Fournier, et al., *J. Bioact. Compat. Polym.* **6**, 327 (1991).

7. E. D. T. Atkins, M. Hill, S. K. Hong, A. Keller, S. Organ, *Macromolecules* **25**, 917 (1992).

8. P. Dreyfuss and A. Keller, *J. Macromol. Sci., Phys.* **B4**, 811 (1970).

9. E. D. T. Atkins, A. Keller, D. M. Sadler, *J. Polym. Sci., Polym. Phys. Ed.* **10**, 863 (1972).

10. R. D. B. Fraser and T. P. MacRae, *Conformation in Fibrous Proteins and Related Synthetic Polypeptides* (Academic Press, New York, 1973), pp. 326-334.

11. A. J. Geddes, K. D. Parker, E. D. T. Atkins, E. Beighton, *J. Mol. Biol.* **32**, 343 (1968).

12. R. D. B. Fraser and T. P. MacRae, p. 331.

13. Y. Takeda, *Biopolymers* **29**, 1125 (1990).

14. G. N. Ramachandran and V. Sasisekharan, *Adv. Protein Chem.* **23**, 239 (1968).

15. J. S. Richardson and D. C. Richardson, in *Prediction of Protein Structure and the Principles of Protein Conformation* G. D. Fasman, Eds. (Plenum, New York, 1989), p. 48.

16. F. M. Richards, in *Protein Folding* T. E. Creighton, Ed. (W. H. Freeman and Company, New York, 1992), p. 37.

17. R. Jaenicke, *Biochemistry* **30**, 3147 (1991).

18. W. Traub, in *Proceedings 1st Cleveland Symp. Macromol.* A. G. Waltons, Ed., (Elsevier, Cleveland, 1976), p. 23.

19. D. L. Kaplan, et al., *MRS Bull.* **17**, 41 (1992).

20. M. Xu and R. V. Lewis, *Proc. Natl. Acad. Sci. USA* **87**, 7120 (1990).

21. R. V. Lewis and M. B. Hinman, *J. Biol. Chem* **267**, 19320 (1992).

22. R. V. Lewis, *Acc. Chem. Res.* **25**, 392 (1992).

23. R. D. B. Fraser and T. P. MacRae, Chap. 13.

24. R. D. B. Fraser, T. P. MacRae, F. H. C. Stewart, *J. Mol. Biol.* **18**, 580 (1966).

25. R. D. B. Fraser, T. P. MacRae, F. H. C. Stewart, E. Suzuki, *J. Mol. Biol.* **11**, 706 (1965).

26. A. Rich and F. H. C. Crick, *J. Mol. Biol.* **3**, 483 (1961).

27. L. B. Sandberg, et al., *Pathol. Biol.* **33**, 26 (1985).

28. D. W. Urry, B. Haynes, H. Zhang, R. D. Harris, K. U. Prasad, *Proc. Nat. Acad. Sci. USA* **85**, 3407 (1988).

29. J. H. Waite, *J. Biol. Chem.* **258**, 2911 (1983).

30. J. H. Waite, *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **31(1)**, 181 (1990).

31. C. M. Venkatachalam, *Biopolymers* **6**, 1425 (1968).

32. P. Y. Chou and G. D. Fasman, *J. Mol. Biol.* **115**, 135 (1977).

33. C. M. Wilmot and J. M. Thornton, *J. Mol. Biol.* **203**, 221 (1988).

34. B. L. Sibanda and J. M. Thornton, *Nature (London)* **316**, 170 (1985).

35. B. L. Sibanda, T. L. Blundell, J. M. Thornton, *J. Mol. Biol.* **206**, 759 (1989).

36. L. Pauling, R. B. Corey, H. R. Branson, *Proc. Nat. Acad. Sci. USA* **37**, 205 (1951).

37. F. J. Padden and H. D. Keith, *J. Appl. Phys.* **36**, 2987 (1965).

38. C. R. Cantor and P. R. Schimmel, *Biophysical Chemistry Part I: The Conformation of Biological Macromolecules* (W. H. Freeman, New York, 1980).

39. G. E. Schultz and R. H. Schirmer, *Principles of Protein Structure*. C. R. Cantor, Ed., Springer Advanced Texts in Chemistry (Springer Verlag, New York, 1979).

40. B. Lotz, *J. Mol. Biol.* **87**, 169 (1974).

41. F. Colonna-Cesari, S. Premilat, B. Lotz, *J. Mol. Biol.* **87**, 181 (1974).

42. R. E. Marsh, R. B. Corey, L. Pauling, *Biochim. Biophys. Acta* **16**, 1 (1955).

43. R. D. B. Fraser and T. P. MacRae, p. 313.

44. R. D. B. Fraser and E. Suzuki, *Anal. Chem.* **38**, 1770 (1966).

45. B. Lotz and F. Colonna Cesari, *Biochimie* **61**, 205 (1979).

46. Y. Takahashi, M. Gehoh, K. Yuzuriha, *J. Polym. Sci., Part B: Polym. Phys.* **29**, 889 (1991).

47. B. Lotz and H. D. Keith, *J. Mol. Biol.* **61**, 201 (1971).

48. S. A. Fossey, G. Nemethy, K. D. Gibson, H. A. Scheraga, *Biopolymers* **31**, 1529 (1991).

49. S. Arnott, S. D. Dover, A. Elliott, *J. Mol. Biol* **30**, 201 (1966).

50. R. D. B. Fraser and T. P. MacRae, p. 324.

51. E. Katchalski and M. Sela, *Adv. Protein Chem.* **13**, 249 (1958).

52. M. Bodanszky, *Peptide Chemistry* (Springer Verlag, Berlin, 1988).

53. S. B. H. Kent, *Annu. Rev. Biochem.* **57**, 957 (1988).

54. M. J. Fournier, T. L. Mason, D. A. Tirrell, *Mater. Res. Soc. Symp. Proc.* **218**, 45 (1991).

55. M. T. Doel, et al., *Nucleic Acids Res.* **8**, 4575 (1980).

56. K. Goraj, A. Renard, J. A. Martial, *Protein Eng.* **3**, 259 (1990).

57. M. H. Hecht, J. S. Richardson, D. C. Richardson, R. C. Ogden, *Science* **249**, 884 (1990).

58. J. S. Richardson and D. C. Richardson, *Trends Biochem. Sci.* **14**, 304 (1989).

59. W. F. Degrado, Z. R. Wasserman, J. D. Lear, *Science* **243**, 622 (1989).

60. D. W. Urry, et al., *Polym. Prep., Am. Chem. Soc., Div. Polym., Mat. Sci. Engr.* **66**, 399 (1992).

61. A. J. Salerno and I. Goldberg, *Polym. Prep., Am. Chem. Soc., Div. Polym., Mat. Sci. Engr.* **66**, 398 (1992).

62. I. Goldberg, A. J. Salerno, T. Patterson, J. I. Williams, *Gene* **80**, 305 (1989).

63. S. C. Gupta, H. L. Weith, R. L. Somerville, *Bio/Technology*, 602 (1983).

64. J. Biernat and H. Koester, *Protein Eng.* **1**, 353 (1987).

65. J. Cappello, et al., *Biotechnol. Prog.* **6**, 198 (1990).

66. J. Cappello, et al., *Mater. Res. Soc. Symp. Proc.* **174**, 267 (1990).

67. F. Lucas, J. T. B. Shaw, S. G. Smith, *J. Mol. Biol.* **2**, 339 (1960).

68. L. B. Sandberg, N. T. Soskel, J. B. Leslie, *N. Engl. J. Med.* **304**, 566 (1981).

69. J. Cappello, *MRS Bull.* **17**, 48 (November 1992).

70. K. P. McGrath, M. J. Fournier, T. L. Mason, D. A. Tirrell, *J. Am. Chem. Soc.* **114**, 727 (1992).

71. J. M. Anderson, H. H. Chen, W. B. Rippon, A. G. Walton, *J. Mol. Biol.* **67**, 459 (1972).

72. P. Y. Chou and G. D. Fasman, *Biochemistry* **13**, 222 (1974).

73. P. Y. Chou and G. D. Fasman, *Biochemistry* **13**, 211 (1974).

74. F. Studier, A. H. Rosenberg, J. J. Dunn, J. W. Dubendorff, *Methods Enzymol.* **185**, 60 (1990).

75. K. P. McGrath, personal communication, (1991).

76. M. T. Krejchi, M. J. Fournier, T. L. Mason, D. A. Tirrell, *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **32(1)**, 411 (1991).

77. M. T. Krejchi, personal communication, (1992).

78. Y. Deguchi, personal communication, (1992).

79. M. Dougherty, personal communication, (1992).

80. Y. Isogai, G. Nemethy, S. Rackovsky, S. J. Leach, H. A. Scheraga, *Biopolymers* **19**, 1183 (1980).

CHAPTER 2

MATERIALS AND GENERAL METHODS

2.1. Overview

This chapter is designed primarily to introduce the techniques of the gene and protein manipulations used in the synthesis of the copolypeptides described in Chapter 3. No attempt is made to be either comprehensive or complete: the concepts presented are strictly those necessary to aid the presentation of details of the synthetic schemes of Chapter 3. With a few noted exceptions the material is drawn from three sources: *Molecular Cloning: A Laboratory Manual*, 2nd ed. (hereafter referred to as Maniatis)[1], laboratory procedures of the research group of Professor Thomas Mason, LGRT 1003 University of Massachusetts, and manufacturer's suggested protocols.

2.2. Materials

2.2.1. Reagents

The following is a list of reagents (along with the suppliers name and specific grade) used in the synthesis of the various materials described in Chapter 3.

Acetic acid, glacial, ACS grade	Fisher Scientific
Acetone, Spectranalyzed®	Fisher Scientific
Acetonitrile, HPLC grade,	Fisher Scientific
Acrylamide, 97%	Aldrich
Adenosine-5'-triphosphate, disodium salt (ATP)	Sigma
Agarose, electrophoresis grade	Gibco BRL
Ammonium acetate, ACS grade	Fisher Scientific

Ammonium bicarbonate, certified	Fisher Scientific
Ammonium chloride, ACS grade	Fisher Scientific
Ammonium hydroxide, ACS grade	Fisher Scientific
Ammonium persulfate, > 98%	Sigma
Ammonium sulfate, ACS grade	Sigma
Ampicillin, sodium salt	Sigma
Antifoam A concentrate	Sigma
Bacto-agar	Difco
Bacto-tryptone	Difco
Bacto-yeast extract	Difco
Bisacrylamide, 99+%	BRL
Boric acid, ACS grade	Fisher Scientific
5-bromo-4-chloro-3-indolyl-ß-*D*-galactopyranoside (X-gal)	Sigma
Bromophenol Blue, reagent grade	Aldrich
Calcium chloride, anhydrous	Fisher Scientific
Calcium chloride, reagent grade	Fisher Scientific
Casamino acids, technical grade	Difco
Chloramphenicol, 99%	Sigma
Chloroform, Spectranalyzed®	Fisher Scientific
Coomassie Blue G-250, reagent grade	Aldrich
Cyanogen bromide, 97%	Aldrich
Deoxyadenosine-5-[α-^{35}S]-thiotriphosphate, triethylammonium salt ([α-^{35}S]-dATP)	Amersham
Dichlorodimethylsilane, 2% in 1,1-trichloroethane	BDH Chemicals
N,*N*,-Dimethylformamide, OmniSolv®	EM Scientific
Dithiothreitol (DTT), 99%	Sigma

Ethanol, 100% and 95%	Fisher Scientific
Ethylenediaminetetraacetic acid, disodium salt, dihydrate 99+%, Sigma grade	Sigma
Formamide, ACS grade	Fisher Scientific
Formic acid, 88%	Fisher Scientific
Formic acid, 90%, PCS reagent	Baker
Glucose, 99%	Sigma
Glycerol, 99%	Sigma
Glycerol, reagent grade	Fisher Scientific
Glycine, 99%	Sigma
^{3}H-glycine, 1 mCi/mL	Dupont/NEN
Hexaaminecobalt (III) chloride	Alfa Products
Hydrochloric acid, reagent grade	Fisher Scientific
Isopropanol, reagent grade	Fisher Scientific
Isopropyl-ß-D-thiogalactopyranoside (IPTG)	Sigma
L-amino acids, 99%	Sigma
Lithium bromide, purified	Fisher Scientific
Low melting point agarose (LMP agarose)	Gibco BRL
Magnesium acetate, ACS grade	Fisher Scientific
Magnesium chloride, ACS grade	Fisher Scientific
Magnesium sulfate, ACS grade	Fisher Scientific
Magnesium sulfate, heptahydrate, ACS grade	Fisher Scientific
Manganese chloride, tetrahydrate	Sigma
ß-mercaptoethanol, 99%	Fisher Scientific
3-(*N*-morpholino)propanesulfonic acid (MOPS), 99.5%	Sigma
Nonidet P-40	Sigma
NuSieve® GTG LMP agarose	FMC BioProducts

Oyster glycogen	Sigma
Phenol, 99+% redistilled nucleic acid grade	BRL
Poly(ethylene glycol), (PEG) M.W. 3350	Sigma
Poly(ethylene glycol), (PEG) M.W. 8000	Sigma
Polyacrylamide, M.W.>$5x10^6$	BDH Chemicals
Ponceau S concentrate	Sigma
Potassium acetate, ACS grade	Fisher Scientific
Potassium chloride	Fisher Scientific
Potassium hydroxide, ACS grade	Fisher Scientific
Serum albumin, Bovine, Fraction V, 96-99%	Sigma
Sodium acetate, anhydrous, ACS grade	Fisher Scientific
Sodium bicarbonate, anhydrous	Sigma
Sodium carbonate, anhydrous, ACS grade	Fisher Scientific
Sodium chloride, USP grade	Fisher Scientific
Sodium citrate, ACS grade	Fisher Scientific
Sodium dodecylsulfate (SDS), 99%	BDH Biochemicals
Sodium hydroxide, ACS grade	Fisher Scientific
Sodium phosphate, dibasic, anhydrous, ACS grade	Fisher Scientific
Sodium phosphate, dibasic, heptahydrate, ACS grade	Fisher Scientific
Sodium phosphate, monobasic, ACS grade	Fisher Scientific
Sodium sulfate, anhydrous	Fisher Scientific
N,N,N',N'-tetramethylethylenediamine (TEMED), 99%	BRL
Trifluoroacetic acid, 99+%	Aldrich
Tris(hydroxymethyl)aminomethane (Sigma 7-9), 99-99.5%	Sigma
Tris(hydroxymethyl)aminomethane (Tris base), reagent grade	Sigma
Tris(hydroxymethyl)aminomethane hydrochloride (Tris·HCl) 99%+	Sigma

Urea, ultra-pure	ICN Biomedicals
Vitamin B1	Sigma
Xylene cyanol	Sigma
Zinc chloride, reagent grade	Fisher Scientific

2.2.2. Solutions

2.2.2.1. Stock Solutions

Unless otherwise indicated, all solutions were prepared with house distilled water which was passed through a high capacity, mixed-bed ion exchange column (Barnstead D8901) and redistilled. Sterile water was prepared by autoclaving at 15 lb/sq. in and 115°C. Solutions were sterilized by one of two methods: autoclaving [indicated with an asterisk (*)], or filtration through a sterile 0.2 μm filter unit (Nalgene) [indicated with a double asterisk (**)]. No mark indicates that the solution was used without a final sterilization step.

SOLUTION	CONCENTRATION
Adenosine triphosphate	6 mM
Ammonium persulfate	10% (w/v)
Ampicillin, sodium salt	200 mg/mL
Chloramphenicol	25 mg/mL in 95% ethanol
Calcium chloride*	1 M
Dithiothreitol	1 M
EDTA, disodium salt	0.5 M, pH to 8.0 with sodium hydroxide
Ethidium bromide	10 mg/mL
Glucose**	20% (w/v)
Glycerol*	20% (v/v)

IPTG	20 mg/mL in dimethylformamide
Magnesium chloride*	1 M
Magnesium sulfate, heptahydrate*	1 M
Oyster glycogen	2 mg/mL
Poly(ethylene glycol), MW 8000	40% (w/v)
Potassium acetate*	5 M
Potassium hydroxide	1 M
Sodium acetate*	3 M, pH to 4.9 with acetic acid
Sodium chloride*	5 M
Sodium dodecylsulfate (SDS)**	20% (w/v)
Sodium hydroxide	2 M
Tris·HCl pH 8.0*	1 M, pH to 8.0 with hydrochloric acid
Tris·HCl pH 7.6*	2 M, pH to 7.6 with hydrochloric acid
Vitamin B1**	1% (w/v)
Zinc chloride	1 M
X-gal**	40 mg/mL in dimethylformamide

2.2.2.2. Other Solutions

Unless otherwise noted, the following solutions are listed by concentration of the components and were prepared by either dissolution of solid reagents or dilution of stock solutions. Sterilization techniques are indicated where appropriate.

TE:

10 mM Tris·HCl (pH 7.6)

1 mM EDTA (pH 8.0)

TE solutions with different pH values were prepared from the corresponding Tris·HCl solution (filter sterilized).

TFB1:

30 mM potassium acetate

100 mM potassium chloride

10 mM calcium chloride

50 mM manganese chloride

15% (v/v) glycerol

pH to 5.8 with 0.2 M acetic acid (filter sterilized).

TFB2:

10 mM MOPS

75 mM calcium chloride

10 mM potassium chloride

15% (v/v) glycerol

pH to 6.5 with 0.2 M potassium hydroxide (filter sterilized).

GTE:

50 mM glucose

25 mM Tris · HCl (pH 8.0)

10 mM EDTA (pH 8.0)

Sterilized by autoclaving.

KOAc buffer:

The solution was prepared by combining:

60 mL of 5 M potassium acetate

11.5 mL of glacial acetic acid

28.5 mL water

Filter sterilized.

SDS/NaOH solution:

0.2 M sodium hydroxide

1% sodium dodecyl sulfate

This solution was prepared immediately before use.

Phenol solution:

Melted phenol was extracted with equal volumes of 1 M Tris·HCl (pH 7.5) until the pH of the aqueous layer was 7.5. The solution was stabilized with 8-hydroxyquinoline (10 mg per 100 mL of solution) and the TE saturated phenol solution (referred to hereafter as phenol) stored at 4 °C in the dark.

2.2.3. Media

2.2.3.1. M9AA(-G) Minimal Medium

M9AA(-G) medium solution was prepared by adding to 740 mL sterile water:

200 mL 5X M9 salts

20 mL 20% glucose

40 mL 25x AA (-G) solution

5X M9 Salts:

The following salts were dissolved in 1 L water

64 g sodium phosphate, dibasic, heptahydrate

15 g potassium phosphate, monobasic

2.5 g sodium chloride

5.0 g ammonium chloride

Sterilized by autoclaving.

25X AA (-G) solution

50 mg of each amino acid except glycine was dissolved in 85 mL water, the pH was adjusted to 7.0 with 2 M sodium hydroxide, the volume adjusted to 100 mL with water and the solution filter sterilized.

2.2.3.2. Complex Media

The compositions of complex media are shown in Table 2.1. Media solutions were prepared by dissolving the solid material in water and were sterilized by autoclaving.

Table 2.1. Composition of complex media

Medium	Tryptone (g/L)	Yeast extract (g/L)	Sodium chloride (g/L)
LB	10	5	10
1X YT	5	8	5
2X YT	10	16	5
SOB*	20	5	0.5
SOC**	20	5	0.5

*Add 10 mL of 250 mM KCl before and 5 mL of sterile 2 M $MgCl_2$ after autoclaving.
**Add 10 mL of 250 mM KCl before before autoclaving. Add 5 mL of sterile 2 M $MgCl_2$ and 2 mL of sterile 1 M glucose after autoclaving (cool solution before adding glucose).

2.2.4. Strains

2.2.4.1. General Strains

Most routine cloning manipulations were performed using *E. coli* strain HB101 as a bacterial host. HB101 is a highly transformable strain deficient in general recombination activity (rec A^-). DH5α and DH5αF' were used in lieu of HB101 in early work, but it was observed that some DNA sequences stable in HB101 were degraded in DH5 strains. HB101 cells were used exclusively in the synthesis of the stem-length variants subjected to

solid-state analysis. *E. coli* strain JM101 was used for preparation of recombinant single stranded templates for sequencing by transfection with M13 phage. The strain carries the F pillus necessary for transfection. All strains were obtained from frozen stocks maintained in LGRT 1003 University of Massachusetts.

2.2.4.2. Expression Strains

The bacterial host for expression of the copolypeptides was *E. coli* strain BL21(DE3)pLysS.[2] In this strain, the gene encoding T7 polymerase is incorporated into the bacterial chromosome under the control of the *lac*UV5 promoter, inducible by the addition of isopropyl-ß-*D*-thiogalactopyranoside (IPTG). In principle, this system is completely selective because the host cell polymerase and T7 polymerase act at different promoters and there are no other T7 promoters in the cell. Therefore, any DNA introduced into the cell under the control of a T7 promoter can be selectively transcribed by the addition of IPTG: production of T7 polymerase results in the accumulation of RNA message from the target DNA. In fact, T7 transcription is active enough that (after a short lag) most of the cell's resources are quickly given to the production of the target protein.

One potential problem is the toxicity of the recombinant proteins produced in the cell. While transcription of the target sequence is under strict control, production of T7 polymerase is not. Expression of some toxic proteins by low levels of polymerase may inhibit growth of the cell. To prevent this, the BL21(DE3)pLysS cells used in this study contained the plasmid pLysS, which supplies a low level of T7 lysozyme—a natural inhibitor of T7 polymerase—to the cell (the plasmid also confers resistance to chloramphenicol). The presence of pLysS offers an additional advantage in that lysozyme promotes rapid cell lysis once released in the medium. Lysing the cells can be initiated by a freeze/thaw cycle and rapidly completed by the lysozyme supplied by pLysS.

Once produced, foreign proteins may be degraded through the action of proteolytic enzymes in the cell. The BL21 strains have been engineered to reduce some protease

activity—they are *lon* deficient and lack the *omp*T outer membrane protease which can degrade proteins during purification.

The BL21(DE3) strains were obtained from frozen stocks maintained in LGRT 1003 University of Massachusetts.

2.2.5. Plasmids

2.2.5.1. Cloning Plasmids

The cloning vector pUC18 (Figure 2.1A) was used for manipulation of DNA monomers and double stranded DNA sequencing. pUC18 is a small, high copy number vector which confers ampicillin resistance to the host bacterial strain.[3] The polylinker region contains multiple restriction sites—the *Bam*H I and *Eco*R I sites were used for cloning of monomer sequences. The plasmid was obtained from the LGRT 1003 University of Massachusetts DNA stocks.

p937.51 is a small, very high copy number vector used as a "transfer vector" for isolation of DNA multimers for insertion into the expression plasmid (Figure 2.1B).[4] The vector confers resistance to chloramphenicol to the host bacterial strain. The important feature of the vector is the unique *Ban* I site flanked by codons for methionine and *Bam*H I sites. This construction allows the cloning and characterization of *Ban* I multimers which can be isolated by *Bam*H I digestion and inserted into pET-3b (*vide infra*). The flanking methionines allow the cleavage of the N- and C-terminal fusion proteins by treatment with cyanogen bromide. The plasmid was obtained as a gift from Joe Cappello at Protein Polymer Technologies, San Diego, CA.

2.2.5.2. Expression Plasmid

Proteins were produced in pET-3b[2] (Figure 2.2), a 4637 base pair expression vector in which transcription is driven by T7 RNA polymerase *via* the gene 10 promoter.

The pET (**p**lasmid for **E**xpression by **T**7 polymerase) plasmids are high copy number plasmids that confer resistance to ampicillin to the host cell. Transcription is directed counterclockwise and the initiation elements are upstream of a *Bam*H I cloning site. Downstream from this site is a Tϕ transcription termination sequence. Thus, sequences cloned into the *Bam*H I site of pET3-b come under control of the gene 10 promoter and are terminated at a specific site downstream. The sequences flanking the cloning site are such that 11 and 18 amino acids are added on the N- and C-termini, respectively. The plasmid was obtained from the LGRT 1003 University of Massachusetts DNA stocks.

2.2.6. Enzymes

2.2.6.1. Restriction Enzymes

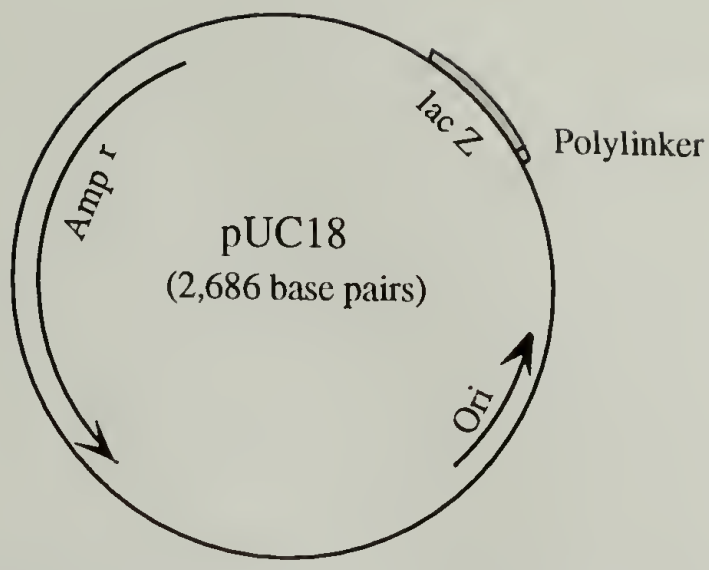
Enzyme	Sequence	Concentration	Source
Apa I	5'-GGGCC↓C-3'	4 Units*/µL	New England Biolabs
Ava I	5'-C↓PyCGPuG-3'	10 Units/µL	New England Biolabs
*Bam*H I	5'-G↓GATCC-3'	20 Units/µL	New England Biolabs
Ban I	5'-G↓GPyPuCC-3'	20 Units/µL	New England Biolabs
*Eco*R I	5'-G↓AATTC-3'	20 Units/µL	New England Biolabs

*Values of "Units" are defined by and specific to the manufacturer.

2.2.6.2. Other Enzymes

Enzyme	Concentration	Source
T4 polynucleotide kinase	10 Units/µL	New England Biolabs
T4 DNA ligase	1 Unit/µL	Gibco BRL
Alkaline phosphatase	1 Unit/µL	Boehringer Mannheim
Ribonuclease A		Sigma

A



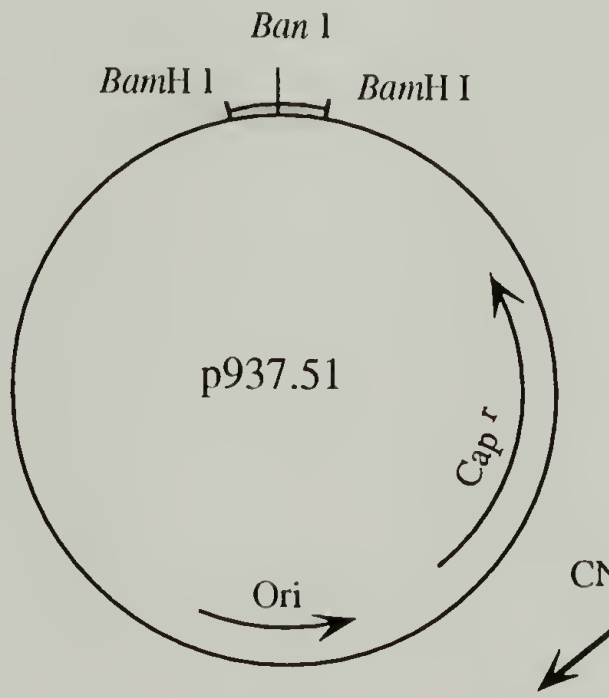
400 450

CCAAGCTTGCATGCCTGCAGGTCGACTCTAGAGGATCCCCGGGTACCGAGCTCGAATTCG

Hind III *Bam*H I *Eco*R I

pUC18 Polylinker Sequence

B



CNBr cleavage sites

Asp Pro Met Phe Lys Tyr Ser Arg Asp Pro Met Gly Ala Arg Met

G GAT CCT ATG TTT AAA TAT TCT CGC GAT CCG ATG GGT GCC CGC ATG...

*Bam*H I *Ban* I

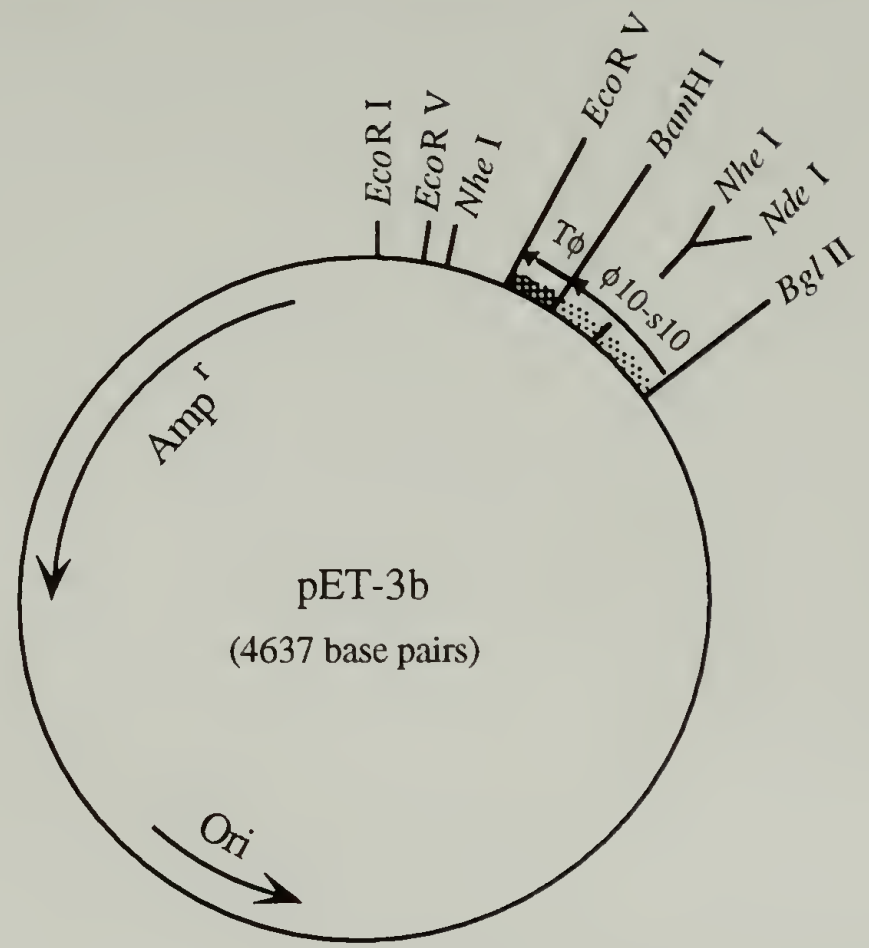
His Ile Arg Pro Gly Arg Tyr Gln Leu Asp

...CAC ATC CGC CCG GGT CGA TAT CAG CTG GAT CC

*Bam*H I

p937.51 Linker Sequence

Figure 2.1. Schematic of cloning vectors pUC18 and p937.51. **A**: The details of pUC18[3]; the sequence of the polylinker region containing the *Bam*H I and *Eco*R I sites used for cloning the DNA monomer is shown. **B:** The details of p937.51[4]; the sequence of the linker, featuring the *Ban* I site for multimer insertion and *Bam*H I sites used for subsequent insertion of the multimer in pET3-b, is shown. The methionines flanking the repetitive sequence (sites for cleavage of the N- and C-terminal fusion sequences by treatment with cyanogen bromide) are shown. Details of the plasmids are discussed in section 2.2.5.1.



(75 bp)

Start Ala Ser Met Thr Gly Gly Gln Gln Met Gly Arg Asp Pro

AGATCT•••TTAAGAAGGAGATATACATATG GCT AGC ATG ACT GGT GGA CAG CAA ATG GGT CGG GAT CCG...

Bgl II — S.D. — *Nde* I — *Nhe* I — **φ10 – s10 Initiation Region** — *Bam* HI

Ala Ala Gln Lys Ala Arg Lys Glu Ala Glu Leu Ala Ala Ala Thr Ala Glu Gln Stop

...GCT GCT AAC AAA GCC CGA AAG GAA GCT GAG TTG GCT GCT GCC ACC GCT GAG CAA TAA•••GATATC

Tφ Termination Region — (76 bp) — *Eco* RV

Figure 2.2. Schematic of expression vector pET3-b.[2] The gene 10 promoter and Tφ termination regions flanking the *Bam*H I multimer cloning site (between the indicated *Bgl* II and *Eco*R V sites) is shown. Details of the plasmids are discussed in section 2.2.5.2.

2.3. Bacterial Culture Methods

2.3.1. General Considerations

The primary concern in manipulating bacterial strains was sterility. All procedures were carried out with safeguards against contamination. Vessels or plates containing bacteria were only exposed in the vicinity of the flame of a bunsen burner (the area surrounding the flame is a sterile micro-environment created by convection currents). The flame also served to sterilize equipment (spreaders, loops etc.). The instrument was either heated directly in the flame or dipped in ethanol, and the ethanol ignited and allowed to burn away. More extensive discussion of these techniques can be found in basic cloning manuals. All procedures described in Chapters 2 and 3 were conducted under the appropriate sterile conditions.

2.3.2. Culturing Bacteria

Bacteria were manipulated both on agar media plates and in liquid media. Plates were prepared by adding 15 g of agar to 1 L of medium before autoclaving. The hot mixture was cooled to ~50-60°C, antibiotics added if necessary and the molten mixture (25-50 mL) poured into a petri dish and allowed to solidify. Excess moisture was removed by incubation at 40°C for an hour and the plates stored inverted at 10°C. To plate cells, an aliquot (no more than 300 µL) of liquid culture was added to the surface of the plate, and the liquid spread evenly with a glass spreader. After the liquid was entirely absorbed, the plates were incubated inverted at the proper temperature.

Bacteria were also cultured in liquid media. For large volume manipulations, the medium was sterilized in the vessel; for smaller volumes, sterile medium was transferred aseptically to a sterile vessel. In a typical procedure, appropriate antibiotics were added as a solution to the medium and cells introduced *via* sterile wooden stick, loop or liquid culture.

Vessels (tubes and flasks) were fitted with loose-fitting caps to allow aeration of the medium and incubated with rotary agitation.

2.3.3. Measuring Bacterial Cell Density

Cell growth was measured routinely using a Klett-Summerson Photoelectric Colorimeter. The measured value will be referred to as optical density or OD and reported as Klett units (KU), which are arbitrary units specific to the colorimeter. No attempt was made to rigorously calibrate the instrument to specific bacterial strains, but in general, mid-log growth was roughly 60-80 KU and the saturation point was reached at 300-400 KU. In the early to mid-log cell growth phase, the Klett unit was considered roughly equivalent to ten times the absorption values measured at a wavelength of 600 nm on a Hitachi U2000 Spectrophotometer (values so recorded will be noted as OD_{600}).

2.3.4. Preparation of Competent Cells

Bacterial cells can be rendered permeable to uptake of plasmid DNA by treatment with divalent cations followed by a brief heat treatment. All such competent cells [*E. coli* strains DH5αF', HB101 or BL21(DE3)pLysS] were either obtained from frozen laboratory stocks or prepared in the following manner:

A 1 mL aliquot of an overnight saturated culture of the appropriate bacteria was used to inoculate 50 mL of LB media and the culture incubated at 37°C to an OD of ~55-65 KU. The cells were incubated on ice for 5 min and 40 mL of the culture transferred to a centrifuge tube. The cells were pelleted by centrifugation (3000 rpm, 4°C, 5 min), gently resuspended (no vortexing) in 16 mL TFB1 and incubated on ice for 5 min. The cells were again pelleted by centrifugation as above, gently resuspended in 1.6 mL TFB2 and incubated on ice until use (at least 5 min). Cells so prepared were used for transformations within 24 hr or frozen at -70°C for later use (with some loss of transformation efficiency).

2.3.5. Transformations of Cells with Plasmid DNA

The efficiency of bacterial transformations with plasmid DNA was measured by the the number of bacteria colonies observed per μg of DNA used in the transformation. Typical efficiencies of cells prepared by the following procedure were in the range of 10^6-10^7 colonies/μg.

Plasmid DNA (5-20 ng in 5 μL solution) was mixed with an ice-cold solution of competent cells (100-200 μL) in a Falcon 2059 polypropylene tube and incubated on ice for 30 min. The tubes were then incubated at 42°C for 90 sec, placed on ice for 2 min, 900 μL of SOC medium was added and the mixture incubated at 37°C for 1 hr with shaking. A portion of the solution (up to a maximum of 300 μL) was plated on agar medium plates containing the appropriate antibiotic. If the entire volume of cells were used, they were either plated on multiple plates (recommended for plasmids conferring resistance to ampicillin) or the cells were collected by a brief, low speed centrifugation, resuspended in 200 μL of SOC and plated. The plates were incubated inverted at 37°C. Typical incubation times were 12-30 hrs, depending on the strain.

2.4. General DNA Methods

2.4.1. Gel Electrophoresis

2.4.1.1. Description

DNAs were analyzed and purified by electrophoresis on both agarose and polyacrylamide gels. The appropriate technique was selected depending on the length of the DNA fragments to be analyzed: above 700-800 bp, 0.8 to 1.5% agarose gels were used. Smaller oligonucleotides (20-700 bp) were analyzed on 2% agarose gels, polyacrylamide (8-15%) or 4% NuSieve® GTG LMP agarose. The selection of medium was made based on knowledge of the resolving range (obtained from Maniatis or manufacturer's data).

Agarose solutions in TAE buffer containing ethidium bromide (5 µL of a 10 mg/mL solution per 100 mL agarose solution) were heated to boiling and the gels poured as horizontal slabs and allowed to solidify. Wells were formed by insertion of a comb of appropriate size. The gels were submerged in TAE buffer in an appropriate apparatus and DNA samples in gel loading buffer were loaded in the wells. The DNAs were separated by electrophoresis at 50-150 V constant voltage. Migration was monitored by noting the position of the bromophenol blue and xylene cyanol dye bands. After electrophoresis, the ethidium bromide stained DNA bands were visualized by irradiation with UV light.

Polyacrylamide gels were prepared between glass plates by polymerizing an acrylamide solution (8-15%) in TBE buffer using ammonium persulfate (700 µL of 10% APS per 100 mL acrylamide solution) and TEMED (35 µL per 100 mL acrylamide solution) as initiators. Wells were formed by insertion of a comb of the appropriate size. After gelation, the comb was removed, the glass plates affixed to a vertical electrophoresis apparatus and the exposed ends of the gel submerged in TBE buffer contained in reservoirs on the top and bottom of the apparatus. DNA samples in loading buffer were loaded into the wells and electrophoresed at 100-150 V constant voltage. After electrophoresis, the gel was removed from the glass plates and the DNA stained by incubation in 200 mL of 0.5 µg/mL ethidium bromide solution. After destaining in water, the DNA bands were visualized under UV light.

2.4.1.2. Solutions for Gel Electrophoresis of DNA

30% Acrylamide

The solution was prepared by dissolving in 100 mL water:

29 g acrylamide

1 g *N*,*N*'-methylenebisacrylamide

Dissolution of the components was aided by gentle heating (40°C).

TBE buffer

45 mM Sigma 7-9

45 mM boric acid

1 mM EDTA

TAE buffer

40 mM Sigma 7-9

40 mM glacial acetic acid

1 mM EDTA

DNA gel loading buffer

0.25% (w/v) bromophenol blue

0.25% (w/v) xylene cyanol FF

30% (v/v) glycerol

2.4.2. Isolation of Plasmid DNA from Liquid Culture

2.4.2.1. Plasmid Mini-preparation

All plasmids were isolated from liquid bacterial culture on a small scale (1.5-5 mL saturated cell culture) using a modified version of the alkaline lysis method of Birnboim and Doly.[5] Saturated liquid cell culture (1.5 mL) was placed in an Eppendorf tube and the cells pelleted by centrifugation in a microcentrifuge for 2 min. The supernatant was removed and the cells were resuspended in 100 µL of GTE. The cells were lysed by the addition of 200 µL of freshly prepared SDS/NaOH solution. After incubation on ice for several minutes, 150 µL of KOAc buffer was added to precipitate chromosomal DNA. After incubation at room temperature for 10 min, the contaminants were pelleted by

centrifugation in a microcentrifuge for 10 min and the supernatant (typically 450 µL) removed to a new tube.

RNA was digested by adding 2 µL of a 10 mg/mL DNase free ribonuclease A (RNase) solution and incubation at 37°C for 30 min. Contaminants were extracted from the solution with an equal volume of a 1:1 phenol/chloroform solution. After vigorous agitation, the layers were separated by a brief centrifugation and the top, aqueous layer removed to a new tube. This process was repeated using 100% chloroform. An equal volume (typically 400 µL) of 100% ethanol was added and the solution incubated at -20°C for 30 min to precipitate the plasmid DNA.

The DNA was pelleted by centrifugation in a microcentrifuge for 30 min at 4°C, the supernatant removed and the pellet washed with cold 70% ethanol. The solvent traces were collected by centrifugation and removed by aspiration with a drawn-out Pasteur pipette. The pellet was dried *in vacuo* and resuspended in TE or water (generally 20-50 µL).

The purity of the DNA obtained using this procedure was sufficient for most manipulations (restriction digests, ligations etc). If higher purity was needed, the DNA was reprecipitated using poly(ethylene glycol). The dry pellet was resuspended in 35 µL water, 10 µL of a solution of 25% PEG containing 2.5 M sodium chloride added and the mixture incubated on ice for 30 minutes. The solution was centrifuged and all traces of the supernatant removed by aspiration with a drawn-out pasteur pipette. The pellet was dried *in vacuo* and resuspended in 20-50 µL water or TE for further analysis or use.

2.4.2.2. Large Scale Plasmid Isolation

Plasmids were isolated from large liquid culture volumes using either a modification of the plasmid miniprep procedure described above or using Qiagen column procedures. In a typical procedure, 1 L of saturated culture in 2X YT medium was centrifuged in four 250 mL centrifuge bottles (2000x g, 4°C, 20 min), the supernatant removed and the cells resuspended in a total volume of 40 mL GTE. The cells were lysed by the addition of

80 mL, freshly made SDS/NaOH solution, incubated on ice for 10 minutes and the chromosomal DNA precipitated by the addition of 60 mL KOAc buffer. After a brief incubation on ice, the chromosomal DNA was pelleted by centrifugation (27,300x g, 4°C, 30 min) and the supernatant carefully removed to a new bottle by pipette (generally it was necessary to repeat this step to ensure removal of all chromosomal DNA). Plasmid DNA and RNA contaminants were precipitated by addition of 300 mL 100% isopropanol, pelleted by centrifugation (27,300x g, 4°C, 30 min), washed with 2 mL 70% ethanol, resuspended in minimal (generally ~4 mL) TE solution and transferred to a 40 mL centrifuge tube. RNAase A (20 µL of a 10 mg/mL solution) was added and the reaction incubated at 37°C for 45 minutes. The solution was extracted with 4 mL 1:1 phenol/chloroform mixture and 4 mL chloroform. The plasmid DNA (in a typical volume of 3 mL) was precipitated at -10°C after addition of 300 µL 3 M sodium acetate (pH 4.8) and 3 mL 100% isopropanol. The DNA was isolated by centrifugation at 13,500x g (-20°C, 30 min), and the pellet was washed with 200 µL 70% ethanol, dried in vacuo and resuspended in 1 mL water. A solution (300 µL) containing 25% poly(ethylene glycol) and 2.5 M sodium chloride was added and the mixture incubated on ice for 30 min. The solution was centrifuged as before and the supernatant removed by aspiration with a drawn-out pasteur pipette. The pellet was dried *in vacuo* and resuspended in 400 µL water.

Better yields of higher purity DNA were obtained using Qiagen high surface area anion exchange columns. In a typical procedure (the manufacturer's protocols were followed) cells were collected, resuspended and treated with RNase, and lysed using the solutions provided by the manufacturer. After removal of the chromosomal DNA, the supernatant was loaded on a prewashed column. Contaminants were removed from the column by allowing a wash solution to move through the column by gravity flow. The pure plasmid DNA was eluted from the column, precipitated by the addition of isopropanol, washed in 70% ethanol and dried.

2.4.2.3. Quantitation of DNA

The amount of DNA in most preparations was estimated by separating the DNA by agarose gel electrophoresis in the presence of ethidium bromide and comparing the intensity of the visible DNA bands (upon UV irradiation) with the intensity of a band from a known amount of a molecular weight standard. Equal masses of linear DNA give rise to equal intensities, so a direct comparison of masses can be made knowing the number of base pairs of each fragment. Intensities were compared by eye so the technique gave only a crude estimate of the yield.

Quantities of DNA were also estimated from UV absorption measurements of DNA solutions in water at 260 and 280 nm. An OD_{260} of 1 corresponds to a concentration of 50 μg/mL for double stranded DNA, 40 μg/mL for single stranded DNA and RNA and ~20 μg/mL for single stranded oligonucleotides. The purity of the preparation can be estimated by the ratio of readings at 260 and 280 nm: an OD_{260}/OD_{280} ratio of 1.8 to 2.0 indicates a low level of protein contamination.

2.4.3. DNA Sequencing

The sequences of DNAs were determined using the dideoxy method of Sanger.[6] In all cases, commercially available kits (either ^{T7}Sequencing™ from Pharmacia LKB or Sequenase™ from United States Biochemical) were used and manufacturer's protocols followed exactly.

The method involves the *in vitro* synthesis of a DNA strand from a single stranded template by a DNA polymerase. Synthesis is initiated at a double stranded priming site created by annealing a short oligonucleotide complementary to a sequence upstream of the region of interest. Chains are extended by incorporation of 2'-deoxynucleoside-5'-triphosphates (dNTPs). The chains are terminated after the polymerase incorporates a 2',3'-dideoxynucleoside-5'-triphosphate (ddNTP), which lacks the 3' hydroxyl necessary of further reaction. With chain extension in the presence of dNTPs and one of the four

ddNTPs in the proper ratio, a fraction of the total population of extended chains will be terminated at all of the sites where the ddNTP can be incorporated.

The complete sequence is obtained by four separate reactions, each incorporating one of the four ddNTPs. The chains are tagged by incorporation of a radioactively labeled dNTP (usually [α-^{35}S]-dATP), fractionated by gel electrophoresis and analyzed by autoradiography. The four reaction mixtures are analyzed in adjacent lanes so that the sequence can be read directly by interpreting the sequential banding pattern on the autoradiograph.

One potential difficulty in the analysis of sequencing reaction products is the presence of band compressions during electrophoresis. These often arise in the analysis of G and C rich DNA sequences when guanine residues form secondary structures that are not fully denatured during electrophoresis and thus cause anomalous migration of DNAs of similar length. This difficulty was minimized by replacement of the dGTPs with analogs containing altered bases such as 7-deaza-dGTP or 7-deaza-2'-deoxyinosine-5'-triphosphate (7-deaza-dITP). The use of each analog has a drawback: 7-deaza-dGTP does not resolve all compressions and 7-deaza-dITP produces a high background. The comparison of DNAs incorporating both analogs allowed the successful determination of most sequences.

2.4.4. Enzymatic Manipulation of DNA

Four types of enzymes were employed for manipulation of DNA: restriction endonucleases, T4 DNA ligase, calf intestine alkaline phosphatase and T4 polynucleotide kinase. The functions of these enzymes are illustrated in Figure 2.3 and discussed below.

2.4.4.1. Cutting and Joining DNA

The enzymes used for cutting and joining DNA fragments are restriction endonucleases and ligases, respectively. Most type II restriction endonucleases (the type used in this study) recognize and cleave DNA at a site within a specific sequence (usually a

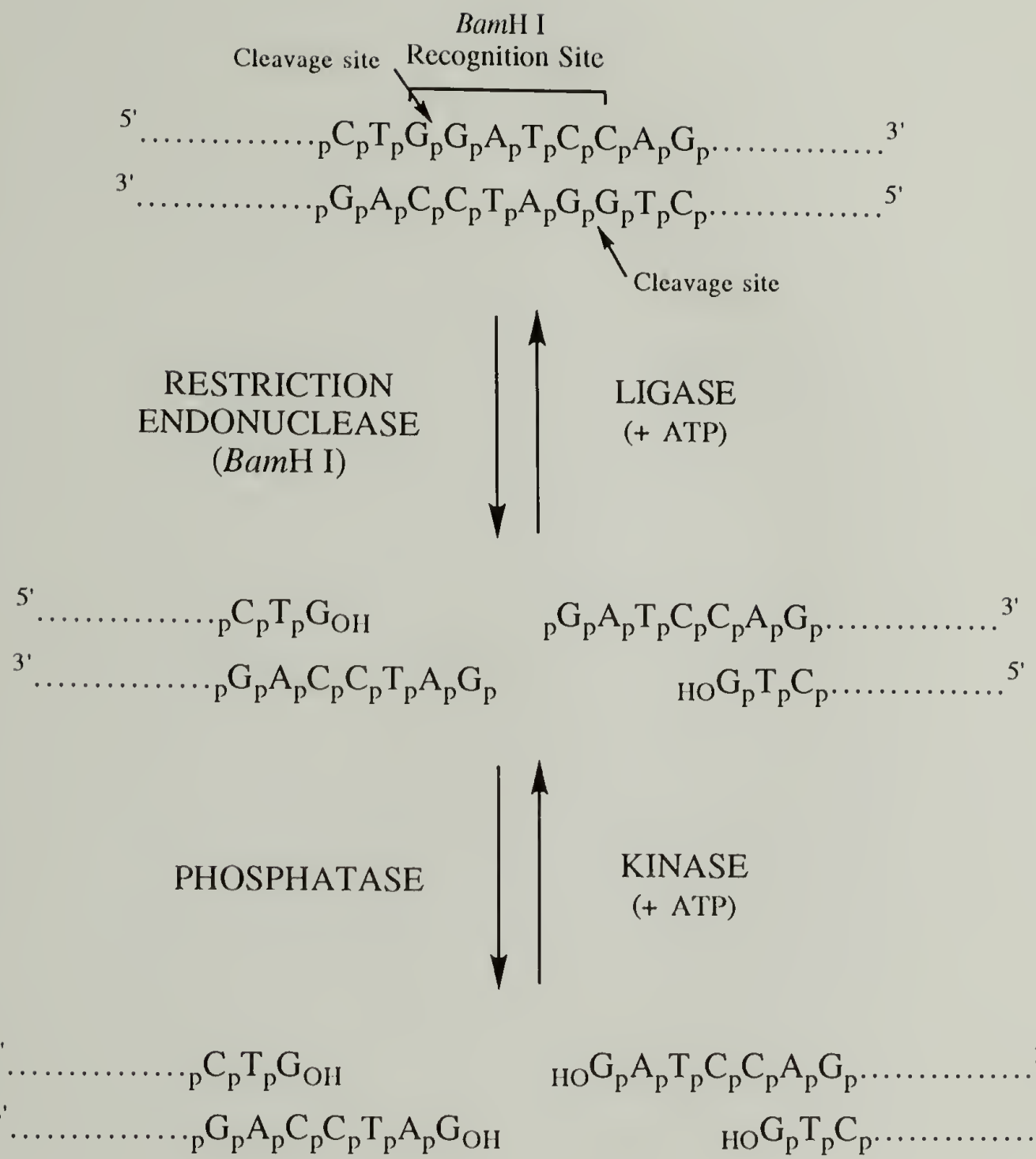


Figure 2.3. The function of and relationship between DNA modification enzymes. DNAs are cut at specific sites with *restriction endonucleases* and ligated (after annealing of termini) with *ligases*. *Phosphatases* are used to remove the phosphate from the 5'-terminus of linear DNA; conversely, *kinases* catalyze the transfer of a phosphate from ATP to the 5'-terminus. All of the indicated enzymes were used in the generation of the expression vectors encoding the target polypeptides.

tetra-, penta- or hexanucleotide). Staggered cleavage (e.g. *Bam*H I : 5'-G↓GATCC-3') of the double strands yields "sticky ends", most with protruding 5' and recessed 3' single strands. Even cutting leaves "blunt ends". Cleavage of DNA with most of the endonucleases used in this study yields staggered cuts with four base pair overhangs; the 3' ends bear a hydroxyl group and the 5', a phosphate. Reactions were performed in a buffer (supplied by the manufacturer) at an ionic strength compatible with the particular enzyme. Microgram quantities of DNA were cleaved in 20-100 μL buffer volumes with 20-50 U of enzyme. Typical incubation times at 37°C were 2-12 hours.

Restriction enzymes can be used to isolate specific DNA fragments with ends capable of annealing to strands from disparate sources—a critical step in the assembly of DNAs for specific applications. Once isolated, the strands can be joined in a fashion dictated by the nucleotide sequence recovered upon cleavage with the restriction enzyme. Upon annealing of the strands, the ends are allowed to associate by hydrogen bonding and a covalent phosphodiester bond is formed between adjacent 3'-hydroxyl and 5'-phosphate termini in a reaction catalyzed by T4 DNA ligase. The reaction requires magnesium ions and ATP as cofactors. The critical variable is the concentration of compatible DNA termini. Polymerization of complementary oligomers is favored by high concentration of ends. Insertion of fragments into linearized vectors requires a large excess of insert ends to suppress recircularization of the plasmid. The reaction is enhanced by the presence of condensing agents such as poly(ethylene glycol). Typical insertion reactions were performed in 10 μL of reaction buffer containing a magnesium salt and ATP. The total amount of DNA was generally 10 ng, with a ten-fold molar excess of insert over plasmid. Reactions were performed at 16°C for 4-20 hours.

2.4.4.2. Modification of DNA Termini

T4 polynucleotide kinase catalyzes the transfer of a γ-phosphate from ATP to the 5'-terminus of DNA. In the forward reaction, the phosphate is transferred directly to the

dephosphorylated 5' terminus of a DNA strands such as those of oligonucleotides recovered after solid phase synthesis. This reaction can also be used to label DNA with a radioactive phosphate group through the use of [γ-^{32}P]ATP. [Note that there is also an exchange reaction where, in the presence of [γ-^{32}P]ATP and excess ADP, the enzyme first transfers the terminal 5' phosphate of phosphorylated DNA to ADP, then rephosphorylates the ends with the radioactive label]. In general, picomole quantities of double or single-stranded DNA with 5' hydroxyls were phosphorylated by a T4 polynucleotide kinase catalyzed reaction with a slight excess of labeled or unlabeled ATP in an appropriate buffer. The reaction mixtures were incubated at 37°C for 30 min, the reaction terminated by addition of EDTA and phenol/chloroform extraction and the DNA purified for futher use.

The complementary DNA terminus modification reaction, dephosphorylation, was used primarily to suppress self-ligation of vectors. Calf intestine alkaline phosphatase (CIP) catalyzes the removal of 5' phosphates from DNA. Treatment of linearized vectors with a phosphatase places hydroxyls on all termini. Thus, since ligase requires the presence of a 5' phosphate to form a covalent bond, the vector termini cannot be self-ligated, favoring insertion over recircularization. The phosphate groups on the insert ensure covalent connection to two strands of the vector. The nicked vectors can be used to transform competent bacteria, and the missing covalent bonds are formed *in vivo*. The enzyme can dephosphorylate picomoles of 5'-termini with an incubation of 30 min at 37°C in the presence of 1 mM zinc. The enzyme is removed by extensive phenol/chloroform extraction and the DNA recovered from the solution.

2.4.5. Isolation and Purification of DNA Fragments

The purification of DNA fragments begins with the separation of the fragments from unwanted DNAs by gel electrophoresis followed by excision of a gel band containing the DNA of interest. In the initial stages of this study, large fragments (e.g. linearized vectors) were purified from agarose gels by electroelution and smaller fragments were

isolated from polyacrylamide gels using the "crush and soak" method. Both methods suffered from low yields and operational difficulties and were replaced with methods involving low melting point agarose. All the methods are discussed below.

Electroelution involves elution of DNA from gels by application of an electric field. In general, DNA containing agarose gel slices were loaded an IBI Model UEA Electroeluter, submerged in TBE and the DNA eluted from the gel into a 7.5 M ammonium acetate buffer by application of a 100 V constant voltage for 30 min. The ~500 µL of salt buffer containing the DNA was removed by syringe and the DNA precipitated by the addition of 100 µg oyster glycogen and 500 µL 100% isopropanol at -10°C. The DNA was recovered by centrifugation in a microcentrifuge, washed with 70% ethanol, dried *in vacuo* and resuspended in TE. In the crush and soak method, polyacrylamide gel slices were chopped or crushed and the fragments soaked in gel elution buffer at 37°C overnight. The mixture was loaded on a 1 mL syringe barrel plugged with siliconized glass wool, centrifuged and the supernatant recovered in an Eppendorf tube. The gel fragments were washed with gel elution buffer, and the wash collected by centrifugation. The DNA was recovered from solution as described above.

In the latter stages of the synthesis, DNAs were almost exclusively purified using low melting point (LMP) agarose. This is a high purity agarose that has a defined melting point at ~65°C. Two types of agarose were employed, depending on size of the fragment to be separated: for larger fragments (>200 bp), standard LMP agarose was used with slabs prepared at 1-1.8%. Smaller fragments (<150 bp) were separated using on 4% NuSieve® GTG LMP agarose (FMC BioProducts) gels. In both cases, the gel slab was formed as described in section 2.4.1.1 and the DNA of interest separated and visualized. The gel slice containing the desired DNA was excised, placed in a tube and melted by incubation at 65°C. The volume of the liquid agarose solution was estimated, an equal volume of phenol added and the mixture vortexed for 3 min. The samplewas centrifuged at room temperature for 10 min and the top aqueous layer removed *via* pipette, avoiding the layer of precipitated

gel material. The supernatant was then extracted with a 1:1 phenol/chloroform mixture, and again with chloroform. The DNA was precipitated from the solution by addition of 3 M sodium acetate (pH 4.9) and 100% ethanol and incubation at -10°C overnight, recovered by centrifugation at 4°C for 30 min, washed with 70% ethanol and dried *in vacuo*.

Gel elution buffer:

0.5 M ammonium acetate

0.1% (w/v) SDS

10 mM magnesium acetate

1 mM EDTA (pH 8.0)

2.5. SDS-PAGE Analysis of Proteins

2.5.1. Description

Protein samples were routinely analyzed using the sodium dodecylsulfate (SDS) discontinuous polyacrylamide system introduced by Ornstein[7] and Davis[8] and modified by Laemmli.[9] A schematic of the system is shown in Figure 2.4.

The upper (stacking) and lower (resolving) gels are cross-linked polyacrylamide, the stacking gel containing Tris·HCl (pH 6.8) and the resolving gel containing Tris·HCl (pH 8.8). The buffers in both the upper and lower reservoirs contain Tris·HCl (pH 8.3); all components contain 0.1% SDS. Upon application of an electric field across the gel, the samples enter the stacking gel with chloride ions forming the leading edge of the mobile front, and glycine molecules trailing. The proteins are maintained and swept along by the steep voltage gradient between the moving boundaries until being deposited upon the surface of the resolving gel. The higher pH of the resolving gel favors the ionization of glycine and the glycine molecules migrate ahead of the protein samples, allowing the separation of the proteins by sieving in the gel under the application of constant voltage.

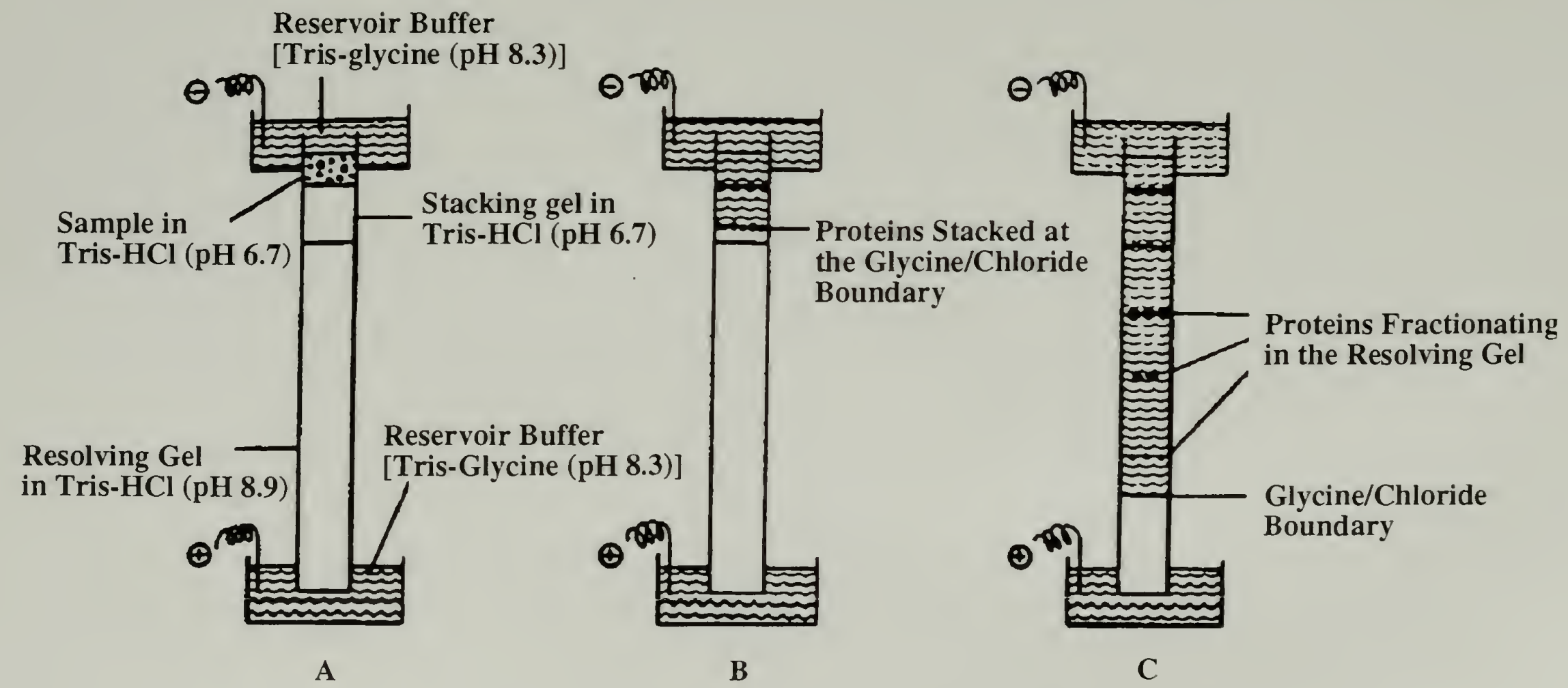


Figure 2.4. Overview of the experimental set-up for discontinuous SDS-PAGE. The system was introduced by Ornstein[7] and Davis[8] and modified by Laemmli.[9] **A**: The protein samples are placed in a stacking gel containing Tris·HCl (pH 6.7). The buffers in both the upper and lower reservoirs contain Tris·HCl (pH 8.3) and all the components contain 0.1% SDS. **B**: Upon application of an electric field across the gel, the samples migrate through the stacking gel in a compact band between the leading chloride ion front and the trailing glycine molecules. **C**: The proteins are deposited upon the surface of the resolving gel containing Tris·HCl (pH 8.9). The higher pH of the resolving gel favors the ionization of glycine, causing the ionized glycine molecules to migrate ahead of the protein sample; the SDS-protein complexes are separated by the sieving action of the gel. Reprinted from Hames[10], with permission from the Oxford University Press.

The samples were prepared in the presence of a reducing agent and SDS and were heated before loading, to ensure that the chains were denatured upon entering the gel. During electrophoresis, the proteins bind SDS and form negatively charged protein-SDS complexes. Because, in general, the proteins bind a number of SDS molecules proportional to their molecular weight (and not sequence) the complexes migrate through the gel and are separated on the basis of the molecular weight of the protein.

After electrophoresis, the gel slab is removed from the apparatus and the proteins visualized by staining with Coomassie Brilliant Blue G-250; 0.1% in a water, methanol, acetic acid solution (5:5:2 by volume). This treatment also serves to fix the proteins to the gel. Excess stain is removed by soaking in a 30% methanol/10% acetic acid solution.

2.5.2. Solutions and Buffers

2X Sample buffer

A stock solution was prepared by combining:

10 mL 0.5 Tris·HCl pH 6.8

20 mL 10% SDS

1 mL 0.2 M EDTA

10 mL glycerol

0.5 mL 1% bromophenol blue

7.5 mL water

ß-mercaptoethanol (20 µL) was added to 980 µL of the stock solution just prior to sample preparation.

1% Polyacrylamide

Polyacrylamide (1 g) was added slowly (over several hours) to 100 mL water with stirring. Complete dissolution took 6-10 hours.

SDS-PAGE Solution F: Running buffer

The solution was prepared by combining:

230.4 g glycine

80 mL 0.2 M EDTA

32 mL 25% SDS

The mixture was diluted to 7 liters with water, the pH adjusted to 8.6 with Sigma 7-9 and volume then adjusted to 8 liters with water.

SDS-PAGE Solution A: Polyacrylamide solution

The solution was prepared by dissolving in 100 mL water:

30 g acrylamide

0.8 g N, N' methylenebisacrylamide

The solution was filtered and stored at 4°C in the dark.

SDS-PAGE Solution B: Running gel buffer.

The solution was prepared by combining:

90.75 g Sigma 7-9

20 mL 0.2 M EDTA solution

8 mL 25% SDS solution

The solution was adjusted to pH 8.8 with HCl and 500 mL with water.

SDS-PAGE Solution C: Stacking gel buffer

The solution was prepared by combining:

6 g Sigma 7-9

4.0 mL 0.2 M EDTA

1.6 mL of 25% SDS

The solution was adjusted to pH 6.8 with HCl and 100 mL with water.

2.6. Instrumentation

A variety of centrifuges were used in the synthesis of the materials. Most routine high volume centrifugations were performed using a Sorvall RC2-B; a variety of microcentrifuges were also employed. In Chapter 3, specific centrifugation conditions were reported where appropriate, with no reference to a particular instrument.

Nuclear magnetic resonance spectra were recorded on either a Varian XL-200 or XL-300 spectrometer. Routine infrared spectra were recorded on an IBM 38 spectrometer. Thermal analyses were performed using a Perkin Elmer DSC-4 differential scanning calorimeter and Perkin-Elmer TGA-7 thermogravimetric analyzer. A number of other instruments were used for specific applications and described (where appropriate) in Chapters 4 and 5.

2.7. References

1. J. Sambrook, E. F. Fritsch, T. Maniatis, *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Press, Cold Spring Harbor, 1989).

2. F. Studier, A. H. Rosenberg, J. J. Dunn, J. W. Dubendorff, *Methods Enzymol.* **185**, 60 (1990).

3. C. Yanisch-Perron, J. Vieira, J. Messing, *Gene* **33**, 103 (1985).

4. F. A. Ferrari, C. Richardson, J. Chambers, S. C. Causey, T. J. Pollack, U.S. Patent Appl. 927,258 (1985).

5. H. C. Birnboim and J. Doly, *Nucleic Acids Res.* **7**, 1513 (1979).

6. F. Sanger, S. Niklen, A. R. Coulson, *Proc. Natl. Acad. Sci. USA* **74**, 5463 (1977).

7. L. Ornstein, *Ann. N. Y. Acad. Sci.* **121**, 321 (1964).

8. B. J. Davis, *Ann. N. Y. Acad. Sci.* **121**, 404 (1964).

9. U. K. Laemmli, *Nature (London)* **227**, 680 (1970).

10. B. D. Hames, in *Gel Electrophoresis of Proteins: A Practical Approach* B. D. Hames, D. Rickwood, Eds. (IRL Press Limited, London, 1981) p. 9.

CHAPTER 3

SYNTHESIS OF COPOLYPEPTIDES

3.1. Introduction and Overview

Three distinct groups of materials were synthesized for use in this study. The initial synthesis of HC4-14 was accomplished to test the viability of the hypothesis regarding stem-length control of chain folding. This synthesis is presented in some detail in section 3.2. Anomalies in the mass analysis of HC4-14 (Chapter 4) prompted the synthesis of other chain-length variants of sequence **3.1** with n = 4, materials designated HC4-m (*vide infra*). An improved scheme was employed in the latter synthesis; the improvements are outlined in section 3.3.

$$[(AlaGly)_nProGluGly]_m \qquad \mathbf{3.1}$$

Using methods refined in these preliminary syntheses, stem-length variants of **3.1** with a specific chain-length (m = 16 and n = 3, 4, 5, 6; designated HCn-16) were synthesized. The objective of this synthesis was the controlled production of materials with specified chain lengths; procedural modifications to accomplish this are outlined in section 3.4.

3.1.1. Gene Design Criteria

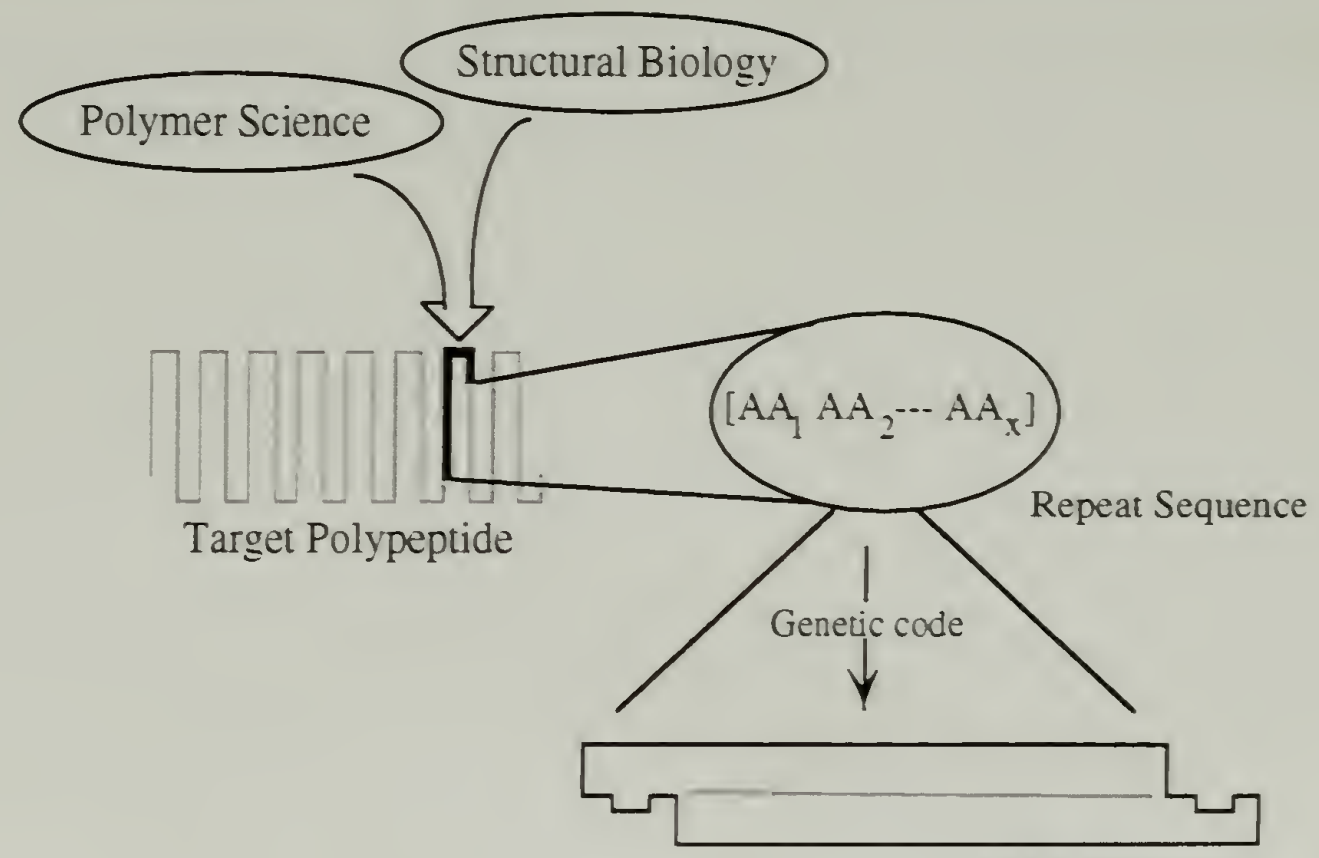
The synthetic strategy begins with the selection of the target polypeptide repeat structure [$(AlaGly)_4$ProGluGly] based on the criteria discussed in Chapter 1. Two repeats of the undecapeptide sequence were encoded in a short DNA fragment, as shown in Figure 3.1A. This 82 base pair oligonucleotide (Figure 3.1B) was designed to take advantage of well established techniques for the cutting and joining of DNA. Terminal overhangs compatible with ends generated by *Eco*R I (5' terminus, AATT) and *Bam*H I (3'

terminus, GTAC) digestion were incorporated to allow the facile cloning of the 82 base pair sequence into the cloning vector pUC18 for subsequent purification and amplification. An in-frame stop codon (TAA) was included that serves to inactivate ß-galactosidase activity when cloned into the polylinker region of pUC18 plasmids. The disruption of ß-galactosidase synthesis both by the presence of the stop codon and the shift in the reading frame of the gene by the insertion of the synthetic sequence insures blue/white selection of colonies containing recombinant plasmids.[1]

Perhaps the most crucial engineering feature of the sequence is the incorporation of two *Ban* I restriction sites near each end of the oligonucleotide. These sites (GGTGCC), which code for a GlyAla dyad in the peptide, allow the coding sequence to be cut out of the recombinant pUC plasmid after amplification to yield the 66 base pair monomer. Digestion at the nonpalidromic *Ban* I restriction site yields termini (5'-GTGC and 3'-CACG) that are not self complementary (non-palindromic), allowing only "head to tail" ligation upon multimerization and thus the exclusion of inverted sequences that disrupt the coding of the polypeptide. In addition, a unique *Apa* I site was included in the sequence which serves, upon digestion, as an additional verification of the presence of insert in recombinant plasmids.

Codon selection is another important parameter in the design of synthetic genes constructed for use in *E. coli*. Choice of codons reflected optimal codon usage in *E. coli*[2] constrained by the need to engineer *Ban* I sites in precise locations in the sequence. To that end, well represented codons GGT and GGC were used extensively to code for the glycine residues, and the rarely seen GGG sequence was only included once. Similarly, well represented codons GCT and GCG were the primary coding sequences for alanine, but the need to engineer two *Ban* I sites forced the inclusion of GCC, the least observed alanine codon. With the selection of CCG for proline and GAA for glutamic acid, codon usage for all residues was thus adjusted to aid in minimizing difficulties in translation.

A



B

*Eco*R I
5'–(G)AATTCG
GC

Stop	Gly	Ala	Gly	Ala	Gly	Ala	Gly	Ala	Gly	Pro	Glu
TAA	GGT	GCC	GGC	GCT	GGT	GCT	GGG	GCC	GGT	CCG	GAA
ATT	CCT	CGG	CCG	CGA	CCA	CGA	CCC	CGG	CCA	GGC	CTT

Ban I

Gly	Ala	Gly	Ala	Gly	Ala	Gly	Ala	Gly	Pro	Glu	Gly	Ala	
GGT	GCA	GGC	GCT	GGC	GCG	GGC	GCG	GGC	CCG	GAA	GGT	GCC	G–3'
CCA	CGT	CCG	CGA	CCG	CGC	CCG	CGC	CCG	GGC	CTT	CCA	CGG	CCTAG(G)

Apa I — *Ban* I — *Bam*H I

Figure 3.1. Sequence design criteria for the DNA monomer used in the synthesis of repetitive polypeptides. **A**: Schematic of the design process. A repetitive amino acid sequence is designed using specific structural criteria derived from biology and polymer science. A DNA sequence encoding the target protein repeat is then determined using the genetic code; this sequence is referred to as the DNA monomer. **B**: Sequence of the 82 base pair oligonucleotide duplex containing the DNA monomer encoding two repeats of [(AlaGly)4ProGluGly]. The terminal *Eco*R I and *Bam*H I restriction sites used for cloning the monomer in pUC18 and the *Ban* I sites used for multimerization and insertion in p937.51 are shown. The details of the sequence design are discussed in section 3.1.1.

3.1.2. Brief Description of the Synthetic Scheme

With the sequence design established, the synthesis of the protein expression system proceeded as described in section 1.7.2 and illustrated in Figure 3.2. To summarize the strategy briefly, two oligonucleotides were first synthesized using an automated, solid phase strategy. The complementary strands were annealed to form the 82 base pair DNA duplex shown in Figure 3.1A, which was cloned into the *Eco*R I and *Bam*H I restriction sites in the polylinker region of pUC18. After transformation of *E. coli* strain DH5α with the ligation mix, a recombinant vector was identified and the sequence of the insert verified. The plasmid was then digested with *Ban* I and the 66 base pair *Ban* I DNA monomer recovered and purified. This *Ban* I monomer, with non-palindromic sticky ends, was enzymatically polymerized to yield a family of DNA multimers coding for polypeptide sequences of various lengths. The multimer population was cloned into the single *Ban* I site of p937.51, the ligation mix transformed in *E. coli* strain HB101 and the transformants screened for recombinant plasmids by colony hybridization. DNA multimers of various lengths were recovered by *Bam*H I digestion, and a fragment containing seven repeats of the DNA monomer selected and inserted into the pET3-b expression vector. A recombinant plasmid with the insert in the correct orientation was used to transform the *E. coli* strain BL21(DE3)pLysS and the expression of the protein demonstrated by *in vivo* labeling of proteins produced after induction with IPTG. The protein was produced on a 10 L batch scale and purified for further analysis.

The basic strategy outlined above was employed in the synthesis of the other materials described in this chapter. With the synthesis of the longer stem-length variants, there were some improvements in the technical details of the strategy which are highlighted where appropriate.

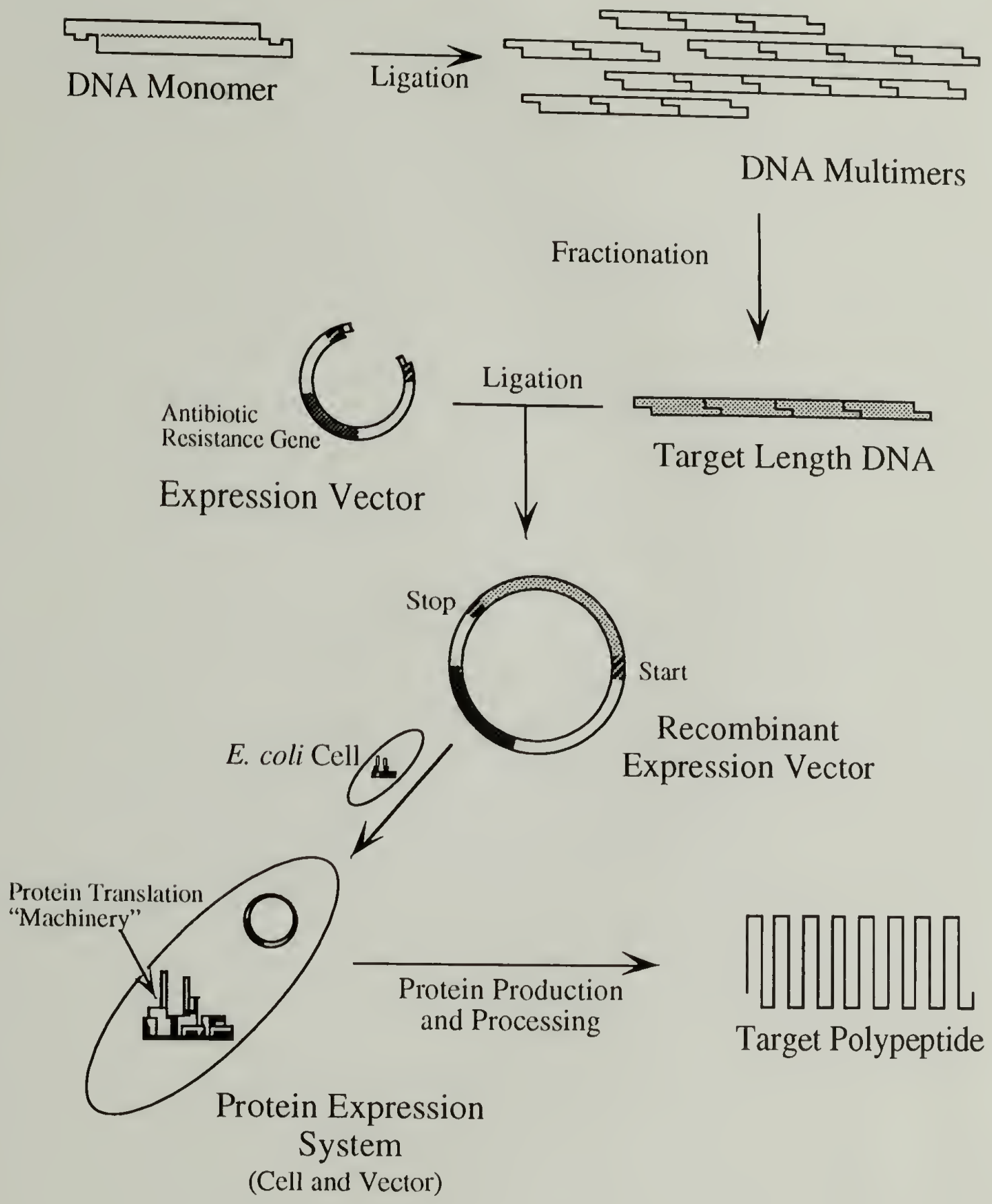


Figure 3.2. Overview of the biosynthesis of repetitive copolypeptides. The purified DNA monomer is enzymatically polymerized to yield a family of DNA multimers each encoding different lengths of the target polypeptide. A multimer of the desired length is selected and cloned into an expression vector. In the present study, the multimer fractionation and expression vector construction were facilitated by an intermediate cloning step using the "transfer vector" p937.51. The recombinant expression vector, containing the DNA encoding the target polypeptide and appropriate transcription sequence elements, is used to transform an *E. coli* strain containing the requisite protein translation machinery. This vector/strain construction becomes an expression system that is used for production of the polypeptide. The details of the synthesis scheme are discussed in section 3.1.2.

3.1.3. Nomenclature

A variety of DNAs and proteins are discussed in the following chapters. The nomenclature employed is unique to this study and is based on the target sequence **3.1**. All recombinant vectors retain the designation of the parent plasmid with specific designations appended using suffices defined in Table 3.1.

Table 3.1. Summary of plasmid DNA and protein nomenclature

Monomer Cloning Vector	pUCn-x
Multimer Transfer Vector	p937.51n-m
Expression Vector	pETn-m
Fusion protein	HCn-m
Cleaved protein	HCn-m*c*

x = Number of $[(AlaGly)_nProGluGly]$ repeats coded in oligonucleotide
n = Number of (AlaGly) dyads in repeat sequence of polypeptide
m = Number of repeat sequences in polypeptide
The suffix *c* designates a polypeptide with fusion ends removed

3.2. Synthesis of HC4-14

3.2.1. Synthesis of pUC4-2

3.2.1.1. Synthesis of Oligonucleotides

Two 82 base pair oligonucleotides, coding strand **3.2** and complement **3.3** were synthesized on a Biosearch Model 8700 DNA synthesizer using CED-phosphoramidite chemistry.[3] Syntheses were carried out on a 0.25 μM scale using glass support columns functionalized with the 3' terminal base.

After synthesis, the column material was flushed into a microfuge tube with 1 mL of ammonium hydroxide and the crude oligonucleotides cleaved from the support by incubation at 65°C for 19 hours. After removal of the glass support by centrifugation for 2

min, the solution was dried *in vacuo*, the pellet suspended in 500 μL water and insoluble material removed by centrifugation for 5 min. UV absorption measurements were taken on diluted samples at 260 nm, where 1 OD_{260} = 33 μg oligonucleotide/mL solution.[4] The results of the syntheses are summarized in Table 3.2. Yields were based on a theoretical yield for an 82 bp oligonucleotide of 6.64 mg and the average stepwise coupling efficiency (S) was calculated using the relation S = exp[lnY/(n-1)] where Y is the overall yield and n is the number of coupling steps.

5' - AATTCG TAA GGT GCC GGC GCT GGT GCT **3.2**
GGG GCC GGT CCG GAAGGT GCA GGC GCT
GGC GCG GGC GCG GGC CCG GAA GGT GCC G - 3'

5' - GC ATT CCA CGG CCG CGA CCA CGA CCC **3.3**
CGG CCA GGC CTT CCA CGT CCG CGA CCG
CGC CCG CGC CCG GGC CTT CCA CGG CCTAG - 3'

Table 3.2. Summary of oligonucleotide synthesis results

	Oligo **3.2**	Oligo **3.3**
Yield	1.68mg	1.49mg
%Yield	25.3%	22.4%
Coupling efficiency	98.9%	98.7%

Crude oligomers were purified using denaturing polyacrylamide gel electrophoresis. A 20 μL portion of each oligonucleotide solution (OD_{260} = 2.10) was mixed with 20 μL of formamide, heated to 55°C for 5 min and loaded into each of four wells of a 10% polyacrylamide gel containing 50% urea (8.3 M). A 50:50 mixture of formamide and aqueous solution containing 0.05% xylene cyanol and 0.05% bromophenol

blue was used as a marker. The DNAs were separated by electrophoresis at 500 V constant voltage for 4 hours, after which the product bands were visualized with ultraviolet shadowing and excised from the gel. The gel slices were chopped finely and the DNA eluted overnight at 37°C in 400 µL of gel elution buffer. The gel residue was pelleted by centrifugation for 15 min, the supernatant was removed and the pellet washed with 125 µL elution buffer. The combined supernatants (~500 µL) were treated with oyster glycogen to 200 µg/mL and sodium chloride to 0.25 M, 1 mL of 100% ethanol was added and the mixture incubated overnight at -10°C to precipitate the oligonucleotides. The precipitated oligomers were pelleted by centrifugation for 30 min at -20°C, the pellets washed twice with 200 µL cold 70% ethanol and dried *in vacuo*. The dry pellet was resuspended in 22 µL water, centrifuged for 5 min to remove insoluble material and the supernatant (~20 µL) recovered. The recovery of the purified oligonucleotides was calculated based on optical absorbance measurements made at 260 nm and summarized in Table 3.3.

Table 3.3. Summary of oligonucleotide purification results

	Oligo **3.2**	Oligo **3.3**
OD_{260} units loaded	8.40	7.40
OD_{260} units recovered	0.40	0.20
Percent purified yield	4.8%	3.8%
Amount recovered	10.0 µg	7.0 µg

3.2.1.2. Phosphorylation and Annealing of Purified Oligonucleotides

A 3 µg (124 pmol) portion of each purified oligonucleotide was suspended in 20 µL of a solution containing 80 mM Tris·HCl (pH 7.5), 20 mM dithiothreitol, 10 mM magnesium chloride and 1 mM ATP. T4 polynucleotide kinase (20 U) was added and the solutions incubated at 37°C for 45 minutes. Each reaction mixture was heated to 95°C for 2 minutes, the solutions were combined and sodium chloride added to 100 mM. The

mixture was placed in an 80°C water bath and the oligonucleotides allowed to anneal as the solution slowly cooled to room temperature over three hours. The solution was brought to a volume of 100 µL with water and extracted once with 100 µL of 1:1 phenol/chloroform and once with 100 µL of chloroform. A 10 µL portion of 3 M sodium acetate (pH 4.8), 3 µL of 10 mg/mL oyster glycogen and 300 µL 95% ethanol were added and the annealed duplex precipitated at -10°C overnight. The precipitated duplex was recovered by centrifugation at -20°C for 30 minutes, and the pellet washed twice with 200 µL cold 70% ethanol, dried *in vacuo* and suspended in 50 µL water. A portion of this solution was run on a 2% agarose gel and quantified by comparison to λ/*Hin*d III digestion fragments (New England Biolabs) as a DNA molecular weight standard. A total of 10 µg of the duplex was recovered, which represents a 71% purified yield.

3.2.1.3. Preparation of pUC18 and Insertion of 82-mer Duplex

A 3.1 µg portion of pUC18 (Bethesda Research Laboratories) was digested overnight at 37°C with restriction enzymes *Eco*R I (20 U) and *Bam*H I (50 U) in 50 µL of a solution containing 100 mM sodium chloride, 10 mM Tris·HCl (pH 7.5) and 10 mM magnesium chloride. Analysis of the digest on a 2% agarose gel using λ/*Hin*d III DNA digestion fragments as a marker allowed quantitation of the pUC18 digest solution at a concentration of 50 ng/µL. A 20 ng (2.3 pmol) portion of this digest was incubated at 70°C for thirty minutes to inactivate enzymes and mixed with the 82 bp duplex (200 ng, 7.5 pmol) in 50 µL of a solution containing 80 mM Tris·HCl (pH 7.5), 1mM ATP, 20 mM dithiothreitol, and 10 mM magnesium chloride. After incubation in icewater for 10 minutes, T4 DNA ligase (400 U) was added and the reaction incubated overnight at 4°C.

3.2.1.4. Transformation of *E. coli* Strain DH5α

Competent *E. coli* strain DH5α cells were obtained from Kevin McGrath; (their preparation is reported elsewhere[5]). A 10 µL portion of the pUC18/82 bp duplex ligation

mix was diluted to 50 µL with water, chilled on ice, mixed gently with 400 µL ice-cold competent cell solution and incubated on ice for 4.5 hours. A 20 ng portion of uncut pUC18 was used as a transformation control. The cells were incubated in a 43°C water bath for 1 minute and placed on ice for 15 minutes. YT medium (600 µL) was added and the cells incubated at 37°C for 45 minutes to express antibiotic resistance. The cells were plated on a YT agar plate containing 200 µg/mL ampicillin, 20 µg/mL X-gal and 20 mg/mL IPTG and incubated at 37°C overnight. Insertion of the 82 bp duplex in the polylinker of pUC18 disrupts normal α-complementation and thus colonies containing recombinant plasmids were identifed by their white color against a background of blue colonies containing plasmids with no insert. Of a total of 338 colonies, 106 were white: a blue/white colony ratio of 31%. The transformation control indicated an approximate transformation efficiency of $3x10^6$ transformed cells/µg pUC18 DNA.

3.2.1.5. Isolation of Recombinant Plasmids

Cells from ten white colonies on the original transformation plate were streaked on a YT agar plate containing 200 µg/mL ampicillin, 20 µg/mL X-gal and 20 mg/mL IPTG. Only 4 of these restreaked colonies were found to be white. Cells from the four colonies were used to inoculate 5 mL YT medium containing 200 µg ampicillin in a 10 mL culture tube and allowed to grow to saturation at 37°C overnight. Plasmids were isolated from these colonies using the alkaline lysis plasmid miniprep procedure described in section 2.4.2.1. One of two 5 µL portions of the resuspended plasmid solution (of a total of 20 µL) was incubated at 37°C overnight with restriction enzymes *Ban* I (25 U) and *Apa* I (20 U), respectively, in a 50 µL solution containing 10 mM Tris·HCl (pH 7.5), 10 mM magnesium chloride and 50 mM sodium chloride. Analysis of the *Apa* I digestion products on a 1% agarose gel indicated the expected linearized vector bands (insertion of the 82 bp duplex creates a single *Apa* I site in the recombinant vector). The *Ban* I digestion products were separated on a 10% polyacrylamide gel, and the expected banding pattern observed

for the 200 and 110 bp vector and the 66 bp *Ban* I monomer fragments. The additional 1274 and 1097 bp vector bands from *Ban* I digestion were observed on a 1% agarose gel.

3.2.1.6. Sequencing of Recombinant Plasmids

The nucleotide sequence of the recombinant pUC18 plasmids containing the 82 base pair inserts were determined using the Sanger dideoxy sequencing method described in section 2.4.3. The sequencing was accomplished using the ^{T7}Sequencing™ Kit supplied by Pharmacia LKB Biotechnology AB.

To denature the double stranded recombinant plasmids, 2 µL of 2 M sodium hydroxide solution was added to 2 µg of plasmid in 8 µL of water. The solution was incubated at room temperature for 10 minutes and the denatured template precipitated by the addition of 3 µL of 3 M sodium acetate pH 4.5, 7 µL of water and 60 µL of 100% ethanol. After incubation at -20°C for 20 minutes the precipitated denatured template was collected by centifugation at 4°C for 30 min. The supernatant was discarded and the pellet washed with 200 µL of cold 70% ethanol, dried *in vacuo* and redissolved in 10 µL of water.

The manufacturer's protocol for the sequencing reactions was followed explicitly; in addition to the nucleotide mixes, enzyme and solutions included in the ^{T7}Sequencing™ Kit and the double stranded templates, the following components were used: -40 M13 sequencing primer (5'-GTAAACGACGGCCAGT-3') supplied by New England Biolabs (20:1 primer/template molar ratio), 7-deaza dGTP and 7-deaza dITP nucleotide analogs supplied in the Deaza ^{T7}Sequencing™ Kit (Pharmacia), and [α-^{35}S]-dATP 10 mCi/mL, supplied by Amersham.

Three sets of four sequencing reactions were performed for each template; one each utilizing normal nucleotides (dGTP) and the two nucleotide analogs 7-deaza dGTP and 7-deaza dITP. The reaction products were analyzed on a 0.2 mm thick, 6% polyacrylamide gel containing 50% urea which was poured between two glass sequencing gel plates fitted for use in a BRL Model S2 vertical gel apparatus. A 3 µL portion of each of the

sequencing reactions was heated to 90°C for two minutes then loaded onto the gel. The samples were separated by electrophoresis at 58 W constant power for 1.5 hours, and the gel was removed onto a sheet of Whatman #1 filter paper, and dried *in vacuo* at 85°C. The dried gel was used to expose an autoradiograph for 2 days.

The sequence was determined for the four templates by comparison of all reactions using the three nucleotides. As described above, band compressions were evident in the sequence analysis. Although the use of 7-deaza dGTP did not resolve all band compressions and there were troublesome background bands when 7-deaza dITP was employed, the use of a combination of both analogs enabled an unambiguous confirmation of the sequence of two of the four templates. One of these plasmids, designated pUC4-2, was selected for further analysis.

3.2.1.7. Large-scale Preparation of pUC4-2

pUC4-2 plasmid was isolated from the selected strain using the large-scale alkaline lysis plasmid preparation procedure described in section 2.4.2.2. After isolation, the pellet was dried *in vacuo* and resuspended in 400 µL of water. Two 1 µL portions of the solution were digested with *Apa* I and *Ban* I and the digestion products analyzed as described in section 3.2.1.5. Analysis of the banding pattern was as expected and no RNA or chromosomal DNA was evident. The concentration of the DNA solution was estimated to be 4 µg/µL by comparison with molecular weight standards. This estimation was consistent with the value of 3.5 µg/µL determined for diluted samples by UV absorption measurements at 260 nm. The OD_{260}/OD_{280} value of 1.7 indicated a low level of protein contamination. Subsequent to the discovery of sequence errors in the recombinant pET plasmid (described in Chapter 4), the sequence of the insert in this preparation of pUC4-2 was verified using the techniques described in the preceding section.

3.2.2. Synthesis of p937.514-14

3.2.2.1. Preparation of 66 Base Pair *Ban* I DNA Monomer

A 100 µg portion of pUC4-2 was suspended in 300 µL of a solution of 10 mM Tris·HCl (pH 7.5) and 10 mM magnesium chloride containing 1000 U of *Ban* I (BRL) and incubated at 37°C overnight. Additional enzyme (100 U) was added and incubated for two hours at 37°C. A 30 µL portion of the solution was loaded on each of 8 lanes of a 20 cm long 2% agarose gel in TBE and the DNA separated at 180 V for 2 hours. The ethidium bromide stained DNA bands were visualized under UV light and the appropriate band containing the 66 base pair *Ban* I digestion fragment excised from the gel with a razor blade. The DNA was eluted from the gel by electroelution on an IBI Model UEA electroelution apparatus at 90 V for 30 minutes into a 7.5 M ammonium acetate buffer as described in section 2.3.5. The 66 bp monomer was precipitated from the ~500 µL solution by the addition of 100 µg of a 10 mg/mL oyster glycogen solution and 300 µL 100% isopropanol and incubation at -10°C overnight. The DNA was pelleted by centrifugation at 4°C for 30 min, the supernatant was discarded and the pellet washed with 70% ethanol and dried *in vacuo*. The dry pellet was resuspended in 10 µL of water. Analysis of this solution on a 10% polyacrylamide gel showed a single band migrating at the correct size relative to the pBR322/*Msp* I digestion fragments used as molecular size markers. The concentration of the DNA solution was estimated to be 130 ng/µL: a total yield of 1.3 µg.

3.2.2.2. Preparation of p937.51

p937.51 (100 µg) was digested with 200 U of *Ban* I (New England Biolabs) in 100 µL of 10 mM Tris·HCl (pH 7.5), 10 mM magnesium chloride at 37°C overnight. The digestion products separated by electrophoresis on a 1% agarose gel at 100 V for 1.5 hours, the plasmid containing gel bands excised and the DNA eluted into 3 M sodium acetate on an IBI Model UEA electroelution apparatus by electrophoresis at 100 V for 15

min. After recovery of ~300 µL of buffer, 600 µL cold 100% ethanol was added and the DNA precipitated at -20°C overnight. After centrifugation at 4°C for 20 min, the pelleted DNA was washed with cold 70% ethanol, and dried *in vacuo*.

The 5' ends of the linearized vector were dephosphorylated to suppress self-ligation in subsequent cloning steps. The purified vector pellet was resuspended in 50 µL of a solution containing 0.5 M Tris·HCl (pH 8.0), 1 mM zinc chloride, 1 mM magnesium chloride and 10 mM spermidine. Calf intestine alkaline phosphatase (BRL, 1 U) was added and the reaction incubated at 37°C for 30 min. A second portion of enzyme was added and the incubation continued for an additional 30 min. The reaction was terminated by the addition of 40 µL of water, 10 µL of 10X STE solution (1 mM sodium chloride, 100 mM Tris·HCl (pH 8.0), and 10 mM EDTA) and 5 µL of 10% SDS. The solution was extracted twice with 100 µL 1:1 phenol/chloroform, twice with 100 µL chloroform and the DNA precipitated by the addition of 200 µL ethanol and incubation at -20°C overnight. The DNA was pelleted by centrifugation at 4°C for 30 min, the pellet washed with 70% ethanol and dried *in vacuo*. The dry DNA pellet was dissolved in 20 µL water and placed on a Millipore 0.025 µm filter floated on 40 mL of water in a petri dish and then dialyzed for 4 hours. A portion of the purified, dephosphorylated vector was analyzed on a 1% agarose gel. Visualization of the ethidium bromide stained DNA indicated a single band migrating at the expected molecular weight. The concentration of the DNA solution was estimated to be 1.5 µg/µL by comparison to known amounts of DNA standards. The total yield of purified vector was 30 µg (30% overall yield).

3.2.2.3. Multimerization and Insertion in p937.51

A solution containing 1 µg of the 66 bp DNA monomer was adjusted to 50 mM Tris·HCl (pH 8.0), 20 mM dithiothreitol, 10 mM magnesium chloride and 1 mM ATP in a total volume of 19 µL and incubated at 0°C for ten minutes. T4 DNA ligase (400 U) was added and the reaction incubated at 14°C for 10 hours. The digestion products in a 15 µL

portion of the ligation solution were separated by electrophoresis on a 1.5% agarose gel at 100 V for 1.5 hours. Analysis of the ethidium bromide stained DNA under UV light (Figure 3.3) showed that ligation of the monomer yielded a multimer population with fragments from dimers to large DNA fragments of ~8000 base pairs.

A portion of the solution containing the multimers (estimated to contain a total of 50 ng of DNA), was mixed with 25 ng of *Ban* I digested, dephosphorylated p937.51 in 19 uL of 50 mM Tris·HCl (pH 8.0), 20 mM dithiothreitol, 10 mM magnesium chloride and 1 mM ATP and the solution incubated on ice for 10°C. T4 DNA ligase (400 U) was added and the reaction incubated at 14°C overnight. Two control reactions were performed: *Ban* I digested p937.51 with no insert was used as a ligation control and *Ban* I digested, dephosphorylated p937.51 with no insert was used to verify the suppression of self-ligation of the plasmid.

3.2.2.4. Transformation of DH5αF'

The solution containing the ligation products was diluted to 100 µL with water, and 10 µL of this solution, as well as the control ligation mixtures and 20 ng of pUC18 (as a transformation control) was mixed with 100 µL of an ice-cold solution of competent cells of *E. coli* strain DH5αF' obtained from frozen stocks maintained in LGRT 903 University of Massachusetts and the mixture incubated on ice for two hours. The cells were incubated at 42°C for 90 sec, 300 µL YT medium was added and the mixtures were incubated at 37°C for 30 min to express antibiotic resistance. A 300 µL portion of this solution was plated onto YT agar plates containing 25 µg/mL chloramphenicol and incubated at 37°C. After 14 hours, 260 colonies were observed growing on the sample plate containing recombinant plasmids. The transformation efficiency was approximately $7x10^5$ transformed cells/µg DNA. The religated p937.51 with no insert was transformed with an efficiency of $3x10^4$ cells/µg DNA. No colonies were observed with cells transformed with religated *Ban* I cut, dephosphorylated p937.51, indicating the success of the phosphatase treatment.

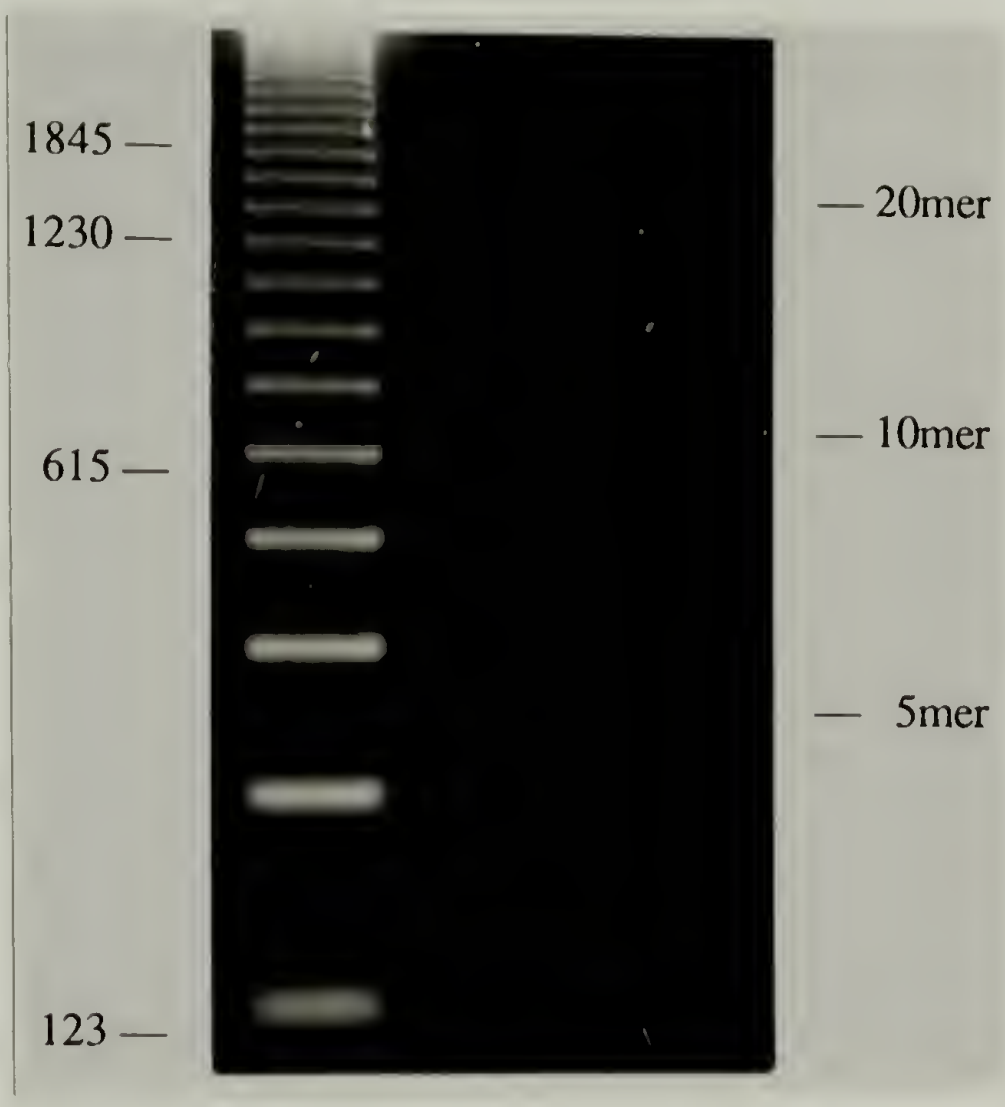


Figure 3.3. Agarose gel electrophoresis analysis of DNA multimers. Left lane: DNA standards, each differing by 123 base pairs. Right Lane: Separated DNA multimers, with fragments from dimers to dodecamers resolved on the gel. Signals from ethidium bromide stained DNA fragments as large as ~8000 base pairs were evident by direct UV irradiation.

3.2.2.5. Analysis of Recombinant Plasmids

Transformants were screened for the presence of recombinant plasmids by colony hybridization using as a probe the 82 base pair oligonucleotide (sequence **3.2**) labelled at the 5' terminus with [γ-^{32}P] ATP.

The oligonucleotide (60 pmol) was mixed in 20 µL of 80 mM Tris·HCl (pH 8.0), 20 mM dithiothreitol and 10 mM magnesium chloride containing [γ-^{32}P] ATP (ICN, 170 µCi). T4 polynucleotide kinase (50 U) was added and the reaction incubated at 37°C for 30 min. The reaction solution was adjusted to 50% formamide, loaded onto a 14 cm 10% polyacrylamide gel containing 50% urea, the DNA separated by electrophoresis at 20 mA constant current for 2 hours and eluted using the crush and soak method described in section 2.3.5. The recovered dry pellet was resuspended in 22 µL of water, centrifuged briefly to remove insoluble material and the supernatant (~20 µL) recovered. Scintillation analysis of a portion of the solution indicated that 3 µCi (706,000 CPM/µL) of radioactive ATP was incorporated.

Selected transformants were plated in a grid (50 transformants/plate) on YT agar plates containing 25 mg/L chloramphenicol and incubated at 37°C until the colonies had grown to approximately 1 mm in diameter. The colonies were transferred to Whatman #3 filter paper, and fixed by sequentially soaking the filters in petri dishes containing 10 mL of the following solutions for the indicated times: 0.5 M sodium hydroxide for 7 min, twice in 0.5 M Tris·HCl (pH 7.4) for 5 min and 0.5 M Tris·HCl (pH 7.4)/1.5 M sodium chloride for 5 min. The filter paper was air-dried and soaked in a petri dish containing 10 mL of a solution of 0.5 M sodium chloride, 6 mM EDTA, 75 mM dibasic sodium phosphate, and 25 mM monobasic sodium phosphate. A 0.5 µL portion of the solution containing the labeled probe (~3x10^6 CPM) was added and the filters incubated at 65°C for 8 hours. The filters were washed for 15 min at 65°C in petri dishes containing the following solutions: twice in 10 mL of 5X SSC (SSC is 750 mM sodium chloride and 75 mM sodium citrate)

containing 0.2% SDS, twice in 2X SSC, and twice in 1X SSC. The filters were air dried and analyzed by autoradiography.

Of the 100 colonies analyzed, at least 50 yielded spots on the autoradiograph with intensity sufficiently above the background to warrant analysis. The background intensity was established using p937.51 without insert as a negative control which yielded a faint spot on the autoradiograph. Based on the results of the colony hybridization procedure, twenty of the colonies were selected for further analysis.

3.2.2.6. Isolation of Multimers from Recombinant p937.51 Plasmids

The recombinant plasmids from the selected colonies were isolated using the plasmid miniprep solution described in section 2.4.2.1 and the recovered DNA resuspended in 20 µL of water. A 10 µL portion of each of the isolated plasmids was digested with 25 U of *Bam*H I in 50 µL of a solution containing 10 mM Tris·HCl (pH 7.5), 10 mM magnesium chloride and 50 mM sodium chloride at 37°C overnight. The digestion products were separated by electrophoresis on a 1.5% agarose gel, stained with ethidium bromide and visualized under UV light.

Of the twenty plasmids analyzed, six contained inserts of appreciable size. Of the remaining samples, nine contained monomer or dimers and five showed multiple bands. The six plasmids containing relatively large inserts were isolated from 50 mL of saturated culture in 2X YT medium using the large scale plasmid preparation described in section 2.4.2.2. Approximately 2 µg of the recovered plasmids were digested with 200 U of *Bam*H I in 50 µL of the NEBuffer supplied with the enzyme. Analysis of the digestion products on a 1.5% agarose gel showed single bands. The molecular size of the inserts was estimated by comparison to λ/*Hin*d III and pBR322/*Msp* I digestion fragments used as molecular weight markers. The results are summarized in Table 3.4.

The DNA multimer fragments were recovered from the gel by electroelution and precipitated as described in section 2.3.5. The pellets were resuspended in 20 µL of water

and the solutions were dialyzed against water on a Millipore 0.025 μm filter. Analysis on a 1.8% agarose gel showed the DNA multimers as single bands (see Figure 3.4). The DNA solution concentrations were estimated to be ~40 ng/μL by comparison to size markers.

Table 3.4. Summary of isolated DNA multimers.

Sample number	Length (base pairs)*	Number of DNA monomers	Number of undecapeptide repeats
10, 14	339	4	8
11	405	5	10
19	471	6	12
16	537	7	14
17	735	10	20

*Calculated value consistent with observed length

3.2.3. Construction of Expression System

3.2.3.1. Insertion of the Multimer in pET3-b

Purified, *Bam*H I digested, dephosphorylated pET3-b was obtained from Kevin McGrath[6] and dialyzed against water on a Millipore 0.025 μm filter. A 25 ng portion of the vector was mixed with 25 ng of each insert (~9:1 insert/vector molar ratio) in T4 DNA ligase buffer (supplied by New England Biolabs) and 1 mM ATP. T4 DNA ligase (400 U) was added and the fragments ligated at 16°C for ten hours. A portion of the vector was ligated without insert as a phosphorylation control. After incubation, the ligation mixtures were diluted to 100 μL with water and 10 μL of the solution used to transform competent *E. coli* strain HB101 cells (obtained from frozen stocks maintained in LGRT 903 University of Massachusetts) as described in section 3.2.1.4. The cells were plated on YT agar plates containing 200 μg/mL ampicillin and incubated overnight at 37°C. The results of the transformation are summarized in Table 3.5.

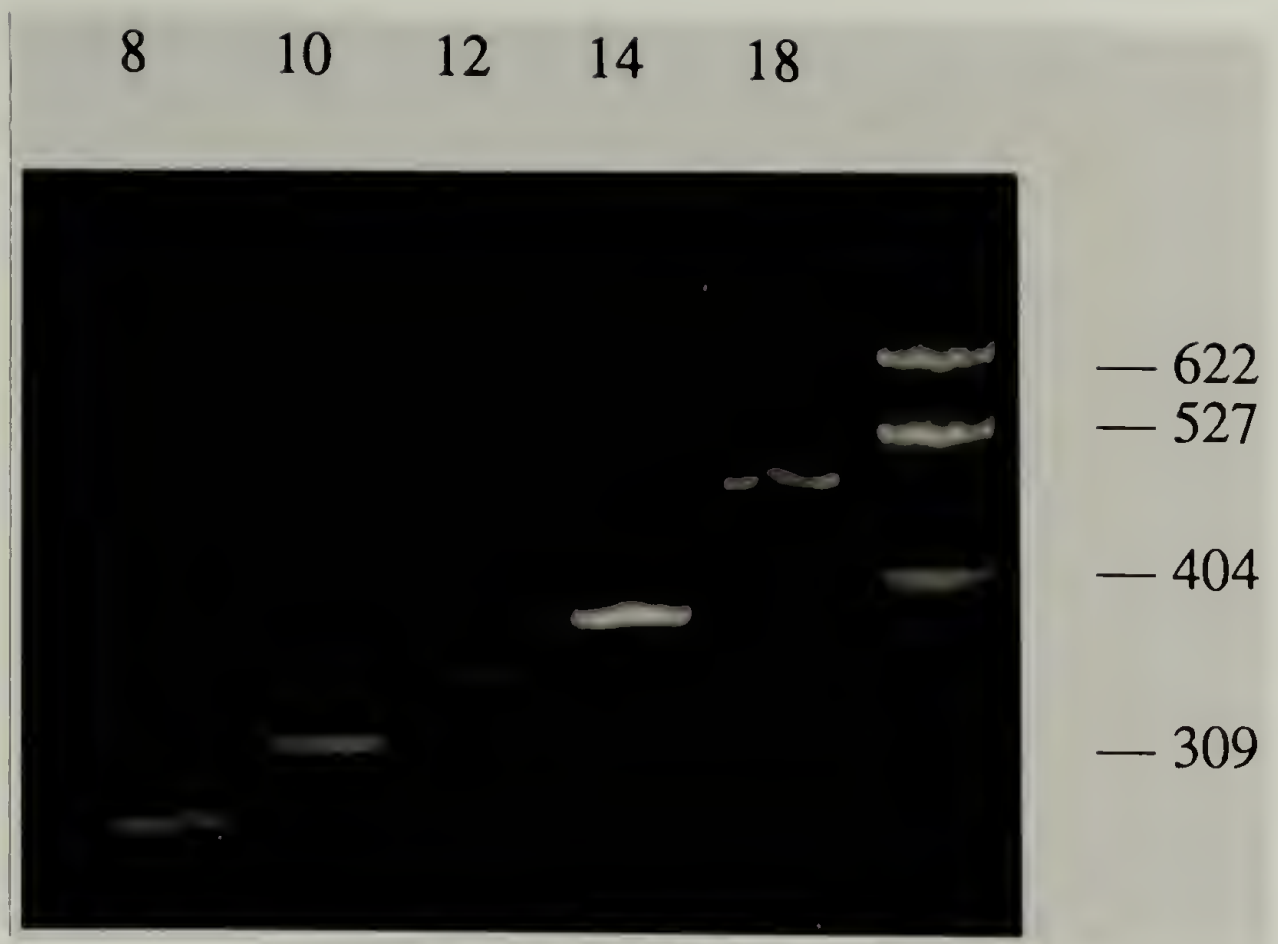


Figure 3.4. Agarose gel electrophoresis analysis of DNA multimers. DNA fragments were isolated by *Bam*H I digestion of recombinant p937.51 plasmids containing *Ban* I multimer inserts. Lanes progress from left to right showing *Bam*H I multimers encoding the indicated number (8, 10, 12, 14, and 18) of undecapeptide repeats. The size of the inserts was estimated by comparison to λ/*Hin*d III digestion fragments used as molecular weight markers.

Selected transformants (Table 3.4) were used to inoculate 5 mL of YT medium and the cells grown to saturation at 37°C. Plasmid DNAs were isolated from the transformed cells as described above and digested with *Ava* I (N.E. Biolabs, 20U) in 20 µL of a 10 mM Tris·HCl pH 7.5 solution with 10 mM magnesium chloride at 37°C overnight. Digestion of recombinant plasmids containing inserts in the proper orientation yields a vector fragment of 3458 bp and an insert fragment with size 1175+X bp, where X is the number of base pairs in the insert. For instance, *Ava* I digestion of the recombinant plasmid containing the insert with 7 repeats of the monomer yields fragments of 3458 and 1712 bp. Analysis of the digestion products on a 1.5% agarose gel showed a number of recombinant plasmids with the expected pattern; the results are summarized in Table 3.5. A plasmid, containing an insert with 7 repeats of the monomer, (encoding 14 repeats of the undecapeptide sequence) was selected for further analysis. The plasmid was designated pET4-14.

Table 3.5. Summary of DNA multimer inserts in pET3-b

Sample Number	Number of DNA monomers	Number of transformants	Colonies examined	Colonies w/ insert	Inserts w/ correct orientation
10	4	1	1	0	—
11	5	110	5	3	1
19	6	3	3	3	2
16	7	13	9	6	2
17	10	1	1	0	—

3.2.3.2. Transformation of *E. coli* Strain BL21(DE3)pLysS

Frozen competent *E. coli* strain BL21(DE3)pLysS cells were obtained from Kevin McGrath[6], thawed, and placed on ice. A 20 ng portion of pET4-14 in 10 µL water was mixed with 100 µL of ice-cold competent cell solution and incubated on ice for 1.5 hours. The cells were heated to 42°C for 90 sec, 300 µL YT medium was added and the mixture incubated at 37°C for 30 min. A 200 µL portion of the cell mixture was plated onto YT agar

plates containing 200 µg/mL ampicillin and 25 µg/mL chloramphenicol and incubated at 37°C overnight. Approximately 100 colonies were obtained from the transformation. Plasmids were isolated from ten colonies using the procedure outlined in section 2.4.2.1 and digested with *Bam*H I. Analysis of the digestion products showed the 537 bp insert and the 631 bp *Bam*H I fragment from pLysS. There was no evidence of DNA instability upon inspection of the gel.

3.2.4. Expression of HC4-14

3.2.4.1. *In vivo* Labeling of Recombinant Proteins

Production of the synthetic polypeptide (designated HC4-14) was monitored by the incorporation of ^{3}H-glycine into proteins after induction of protein synthesis. BL21(DE3)pLysS cells transformed with the pET4-14, pET3-b and pET3-27 (obtained from Kevin McGrath) were used to inoculate 5 mL of M9AA(-G) medium (containing all amino acids except glycine) with 200 µg/mL ampicillin and 25 µg/mL chloramphenicol and incubated overnight at 37°C. Cultures of BL21(DE3)pLysS/pET3b were used as a negative control and BL21(DE3)pLysS/pET3-27 cells (see McGrath and coworkers[7]) were used as a positive control for the expression system. [Note that the designation pET3-27 employed by McGrath corresponds to pET3-54 in the scheme used here]. Approximately 300 µL of the saturated culture was used to inoculate 10 mL of M9AA(-G) medium with a starting optical density of 10 KU (measured using a Klett-Summerson Colorimeter). The cultures were incubated at 37°C with vigorous aeration, cell growth was monitored (Figure 3.5) and 75 µCi of ^{3}H-glycine added when the cultures reached an optical density of 60 KU. At 80 KU, isopropyl-ß-D-thiogalactopyranoside (IPTG) was added to a final concentration of 0.4 mM to induce protein production. Portions of the cultures representing approximately equal numbers of cells (as calculated from the observed optical density) were removed at

0, 15, 30, 45 and 60 min after induction. The cells were recovered by centrifugation at room temperature for 2 min, the supernatant discarded and the cells washed in 1 mL YT medium. After a second centrifugation, the pelleted cells resuspended in 100 µL of TE, lysed by the addition of 100 µL 2X sample buffer and prepared for SDS-PAGE analysis.

3.2.4.2. SDS-PAGE Analysis of Proteins

The lysates were analyzed on a discontinuous SDS 10% polyacrylamide gel using the methods described in section 2.4. After heating to 90°C for two minutes, 10 µL portions of cell lysate samples were loaded onto the gel and the proteins were separated by electrophoresis at 20 mA constant current for 9 hours. The gel was stained with Coomassie Brilliant Blue G-250 solution to visualize separated proteins and destained in 30% methanol containing 10% acetic acid. The gel was soaked in Enlightning™ autoradiography enhancer (duPont NEN) for 3 hours, transferred to Whatman #3MM blotting paper, dried *in vacuo* at 60°C for 5 hours, and analyzed by fluorography.

The results of the analysis are shown in Figure 3.6. Some *E. coli* protein bands are visible in all lanes. The recombinant protein (HC4-14) appears as a prominent new band as early as 30 min after induction and is seen to achieve and maintain a constant intensity over the course of the analysis. The banding patterns for the positive and negative controls are as expected. This analysis was used as evidence of the high-level expression of the target protein.

3 2.5. Large-scale Production of HC4-14

A 5 mL saturated culture of BL21(DE3)pLysS/pET4-14 was used to inoculate 250 mL of fresh medium containing 25 µg/mL chloramphenicol and 200 µg/mL ampicillin and the culture was incubated at 37°C for 6 hours. This culture, in turn, was used to inoculate 10 L of YT medium with antibiotics in a New Brunswick Scientific Inc.

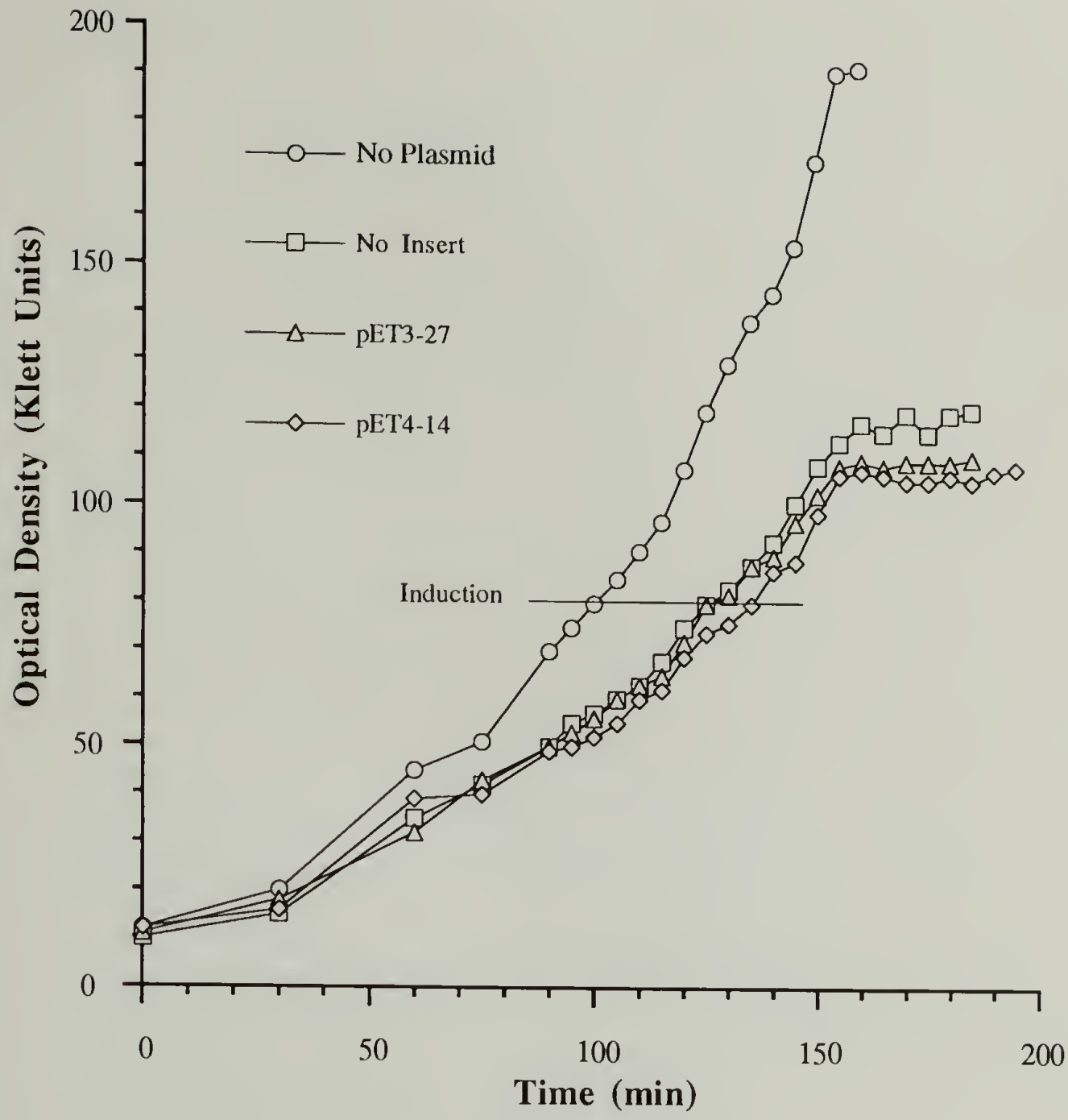


Figure 3.5. Cell growth plot of HC4-14 in BL21(DE3)pLysS *E. coli*. Cells with no pET plasmid (no induction control), pET3-b (no insert control), pET3-27 (positive control) and pET4-14 were grown in M9AA(-G) medium containing antibiotics and the growth monitored using a Klett-Summerson Colorimeter. Cells containing pET plasmids are seen to grow at a slower rate than the no pET plasmid control. Protein production was induced by addition of IPTG at a cell density of 80 KU. Cell growth is retarded as early as 25 min after induction in the strains with the pET plasmids relative to the no pET plasmid control. Expression of HC4-14 is discussed in section 3.2.4.

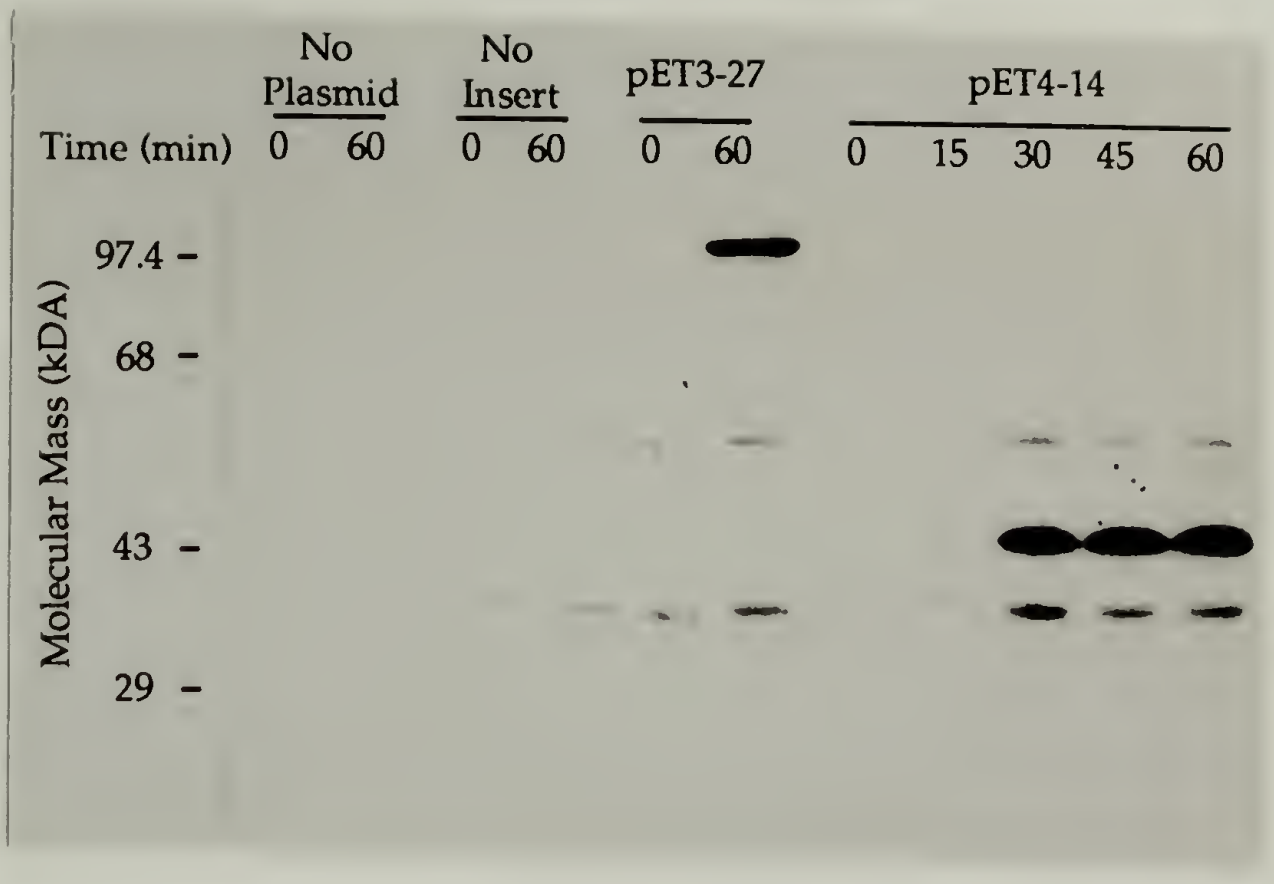


Figure 3.6. SDS-PAGE Analysis of ^{3}H labeled proteins expressed in BL21(DE3)pLysS *E. coli*. Labeled proteins produced at induction with IPTG (t = 0) and sixty minutes post-induction (t = 60) and obtained from approximately equal numbers of cells were separated on a 10% SDS-polyacrylamide gel at 20 mA constant current and stained with Coomassie Brilliant Blue G-250 to visualize the indicated molecular mass markers. The gel was soaked in Enlightning™ autoradiography enhancer (duPont NEN) for 3 hours, transferred to Whatman #3MM blotting paper, dried *in vacuo* at 60°C for 5 hours, and analyzed by fluorography. Lanes 1-4 of the autoradiograph are negative controls: BL21(DE3)pLysS with no pET plasmid (No Plasmid control) and containing pET3-b (No Insert control). Only faint background bands are visible in these lanes. A prominent band appears at t = 60 (Lane 6) for positive control BL21(DE3)pLysS/pET3-27 (see McGrath and coworkers[7]). Production of labeled HC4-14 is seen as early as 15 minutes after induction (Lane 8); a prominent new band of significant intensity appears at ~43,000 Da 30, 45, and 60 minutes post-induction. While the intensity of the bands on the autoradiograph appears constant in these lanes, there is an increase in the intensity of the Coomassie Blue stained protein bands (not shown), indicating accumulation of the protein with time. The apparent molecular mass of the protein (~43,000 Da) differs significantly from the calculated value of 17,207 Da. This is also true for the KM3-27 positive control (observed ~97,400, calcd 42,123 Da). The significance of these observations is discussed in section 4.1. Expression of HC4-14 is discussed in section 3.2.4.

Microferm Fermenter to an $OD_{600} = 0.10$ (measured at 600 nm). The temperature was equilibrated to 37°C and 1 mL of Antifoam A concentrate added to suppress foaming. The cells were grown to $OD_{600} = 1.0$ with vigorous aeration (15 psi house compressed air at a rate of 9,000 cc/min with 400 rpm agitation). Protein synthesis was induced by the addition of IPTG to 0.4 mM (0.95 g in 10 mL of dimethylformamide). After 2.5 hours, the cells were harvested in four 1 L bottles by centrifugation at 4000 rpm (Sorvall RC-3 centrifuge, HG-4L rotor) for 20 min at 4°C. The cells were resuspended in 100 mL water placed in a 250 mL centrifuge bottle and frozen at -70°C. After thawing, the lysate was homogenized by sonication with a Branson Sonifier at 40 units of power (arbitrary units). The lysate was centrifuged at 16,300x g at 4°C for 30 min to remove insoluble cell debris. Cycles of sonication and centrifugation were repeated several times to obtain a clarified cell extract.

3.2.6. Purification of HC4-14

After the initial clarification, the cell extract was acidified to pH 6 by the addition of glacial acetic acid, then centrifuged at 16,300x g at 4°C for 30 min to remove precipitated contaminants. The supernatant was removed to a new bottle, acidified to pH 5 with glacial acetic acid and precipitated contaminants removed by centrifugation 16,300x g at 4°C for 30 min. This treatment was repeated with adjustment of the pH to 4.5 and then to pH 4.0. The presence of the target polypeptide and the absence of contaminating proteins in the pH 4.0 supernatant was verified by SDS 10% polyacrylamide gel electrophoresis. The pH 4 supernatant (~100 mL) was then adjusted to 40% ethanol by the addition of 67 mL of 100% ethanol and incubated for 6 hours at 10°C to precipitate more contaminants. The suspension was centrifuged at 16,300 x g at 4°C for 30 min, the supernatant removed to a 1 L centrifuge bottle, adjusted to 80% ethanol by the addition of 334 mL of 100% ethanol and the recombinant protein allowed to precipitate at 10°C overnight. The precipitated material was recovered by centrifugation at 16,300x g at 4°C for 30 min, washed with several portions of 80% ethanol and dried *in vacuo*. The dried pellet was dissolved in

20 mL of 0.1 M ammonium bicarbonate, the solution was placed in a 10-12,000 molecular weight cut off dialysis bag and dialyzed against water for two days. The solution was frozen at -78°C and lyophilized under high vacuum. The recovered material was dried over P_2O_5 at 80°C to constant weight. Typical yields of the purified material were 200-400 mg.

3.2.7. Chemical Analysis of HC4-14

Nuclear magnetic resonance spectra were taken of samples prepared in deuterium oxide (Aldrich, 99.996 atom% D) at a concentration of 5-15 mg/mL. The spectra were referenced to the appropriate water peak or to sodium 3-(trimethylsilyl)-1-propanesulfonate (DSS, Aldrich). The spectrum is shown in Figure 3.7 and the results tabulated in Table 3.6. There is no indication of the presence of significant amounts of organic contaminants. All visible peaks can be rationalized with the expected resonances of the residues in the sequence. There was significant overlap of the proline and glutamic acid beta and gamma protons, preventing accurate assignment. The overall spectral resolution is poor, but the majority of the signals can be assigned to the protein and not contaminants.

Samples for infrared analysis were prepared as pressed KBr pellets and the absorbance measured under a nitrogen purge at 4 cm^{-1} resolution with 300 scans. The full

Table 3.6. ^1H NMR assignments for HC4-14.

	Chemical Shift (ppm)			
Residue	α H	ßH	γH	δH
Gly	4.22			
Ala	4.75*	1.49		
Pro	4.39*	2.28, 2.02**	2.03, 2.03**	3.20, 3.65
Glu	4.51*	2.09, 1.97**	2.31, 2.28**	
Unassigned	0.65, 1.00, 2.91, 3.33, 4.90, 5.14			

*Doublet peaks unresolved on spectrum.
**Peaks not resolved on spectrum and no direct assignment possible. Values are typical as reported by Wüthrich[8] and consistent with the observed signals.

spectrum is seen in Figure 3.8 to contain all the prominent resonances associated with the peptide bonds in polypeptides.[9] The absorptions at 1703 and 1632 cm^{-1} appear as shoulders on the main 1657 cm^{-1} peak. The significance of these peaks will be discussed in more detail in Chapter 5. A summary of the peak assignments is found in Table 3.7.

Table 3.7. Assignment of prominent infrared absorbances of HC4-14.

	Absorbance (cm^{-1})	
	HC4-14	*B. mori* Silk*
Amide A	3290(s)	3292(s)
Amide B	3080(w)	3084(s)
Amide I	1703(w), 1653(s), 1632(m)	1701(m), 1657(m), 1632(s)
Amide II	1529(s)	1525(s)

Intensities: Strong (s), medium (m) and weak (w).
*From Fraser and coworkers[9]

Amino acid compositional analysis of HC4-14 was performed by Protein Polymer Technologies in San Diego, California. The sample was hydrolyzed in two separate batches: analysis of a concentrated sample (~30 μg) to resolve the minor amino acid constituents, and a more dilute sample (~5 μg) for accurate determination of the composition of the four primary residues. The results are tabulated in Table 3.8. The concentrated form shows the amino acids in the fusion ends and indicates their presence in the expected amount. The presence of valine is unexpected. In the initial analysis, the detector sensitivity was adjusted such that the values obtained for the prominent residues (A, G, E, P) were not obtained accurately. However, analysis of a dilute sample with the detector in the proper calibration range indicates the correct composition for all residues except for glutamic acid/glutamine (glutamine is hydrolyzed to glutamic acid so the two cannot be analyzed separately).

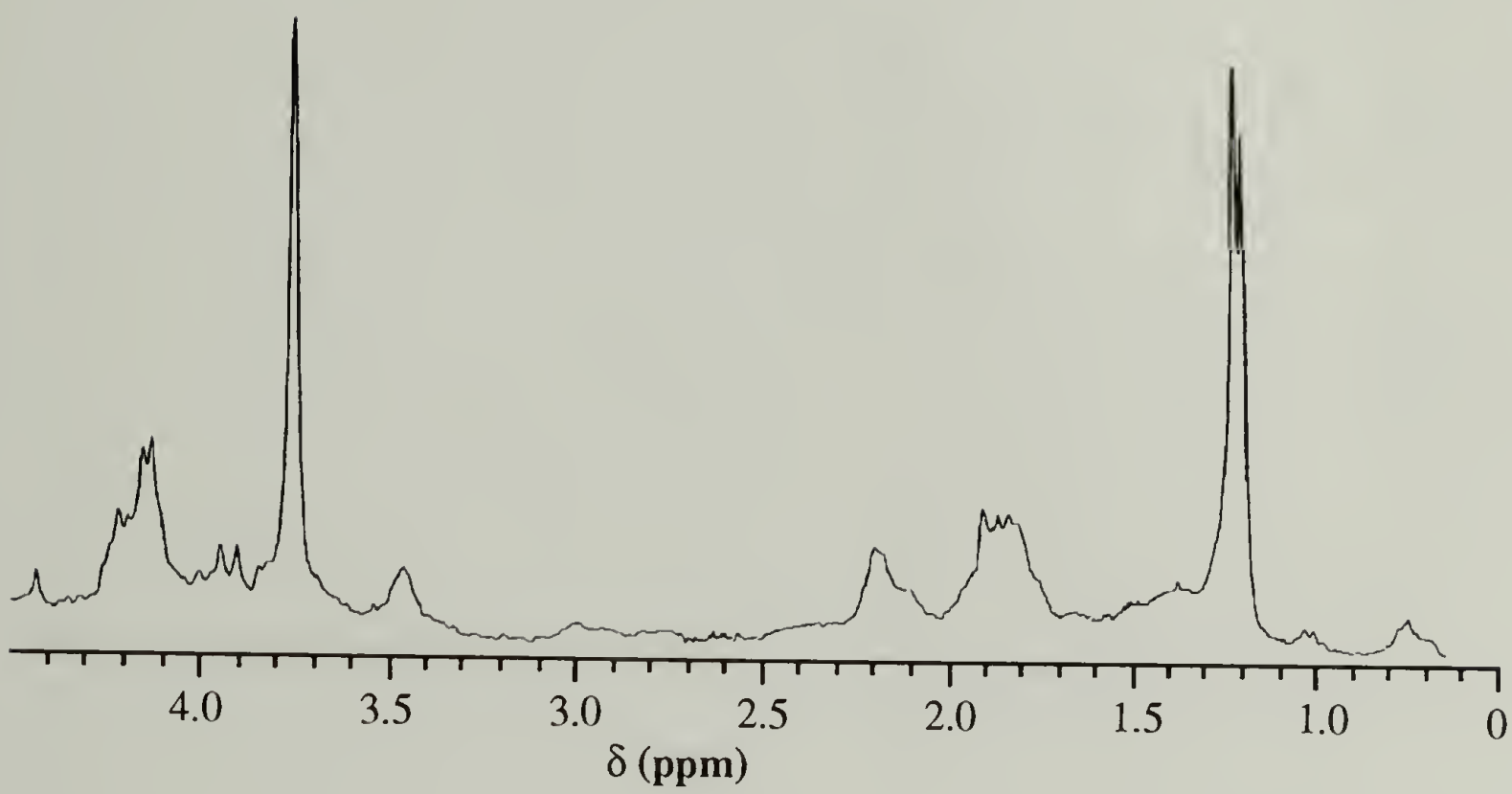


Figure 3.7. ^{1}H Nuclear magnetic resonance spectrum of HC4-14. The sample was prepared in deuterium oxide (Aldrich, 99.996 atom% D) at a concentration of 15 mg/mL.The peaks were referenced to sodium 3-(trimethylsilyl)-1-propanesulfonate (DSS, Aldrich) at $\delta = 0$. No signals for the amide protons were visible at higher ppm. Peak assignments are listed in Table 3.6.

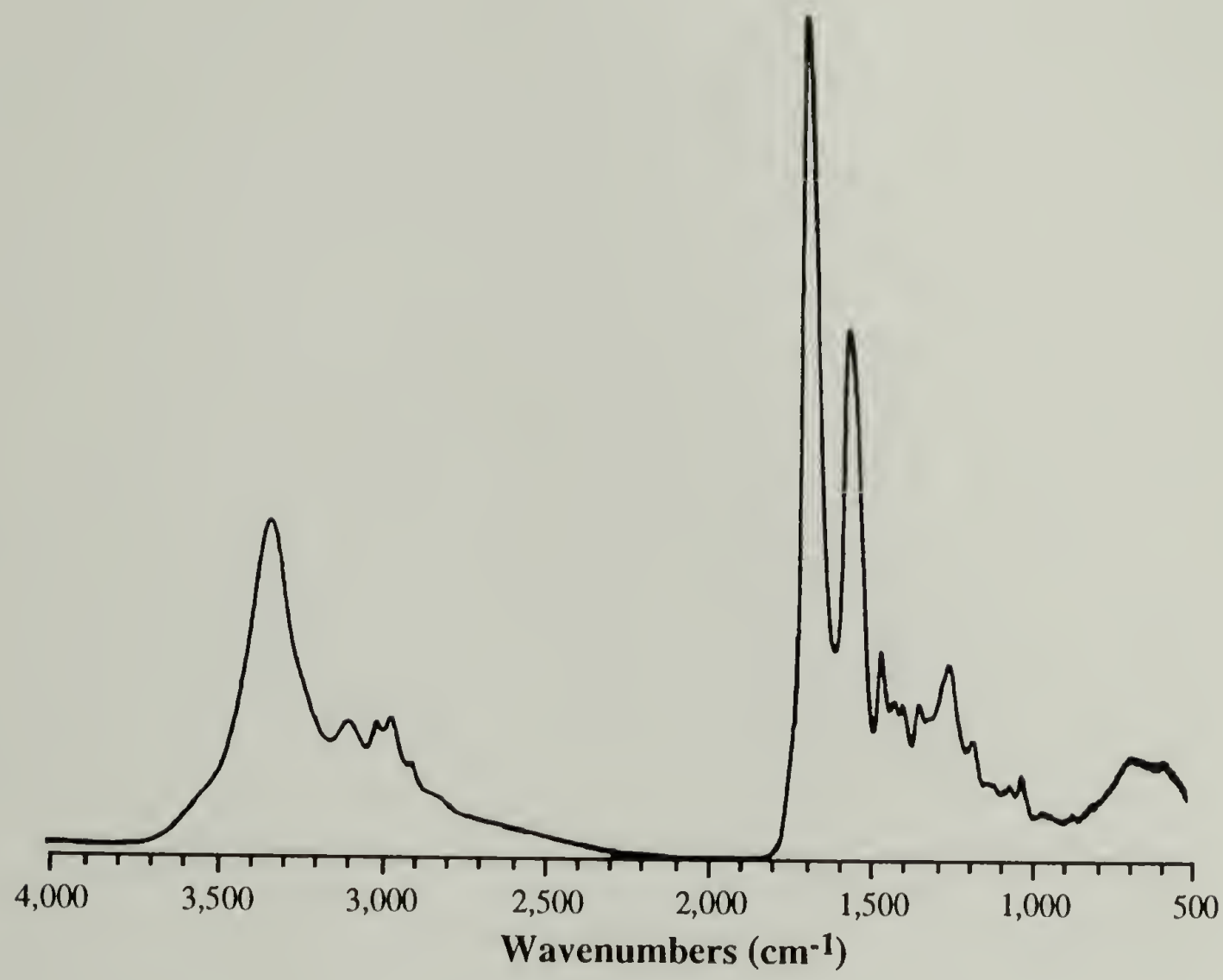


Figure 3.8. Fourier transform infrared spectrum of HC4-14. The samples were prepared as pressed KBr pellets. Assignment of prominent peaks are listed in Table 3.7.

Table 3.8. Amino acid composition analysis of HC4-14.

Residue	Calcd mol%	Conc* mol %	Dilute** mol %
Asx	1.90	1.59	
Glx	10.00	10.14	8.49
Ser	0.95	1.61	
Gly	35.71	40.71	35.99
His	0.48	0.14	
Arg	2.86	2.21	
Thr	0.95	0.91	
Ala	31.43	25.90	30.55
Pro	8.57	9.53	8.91
Val	0	0.86	
Met	2.86	3.04	
Ile	0.48	0.13	
Leu	0.95	0.41	
Phe	0.48	0.74	
Lys	1.43	1.15	

*Concentrated sample (~30 μg) to resolve the minor amino acid constituents.
**Dilute sample (~5 μg) for accurate analysis of the four primary residues.

A sample (4.107 mg) of the polypeptide was heated under a nitrogen purge at 20°C/min in a Perkin-Elmer TGA-7 thermogravimetric analyzer and the mass loss recorded from 30 to 600°C (Figure 3.9). With an increase in temperature, a 10% loss corresponding to the release of water is seen, followed by the onset of degradation at ~220°C. A steady loss is recorded to a constant value of 30% residue at 600°C.

Combustion analysis of HC4-14 was performed by the University of Massachusetts Microanalysis Laboratory. The results are seen in Table 3.9. The values can be rationalized by the presence of 10% water, consistent with the early mass loss seen in the thermogravimetric analysis. The measured ash residue is below the limits of detection, indicating a low level of inorganic contaminants.

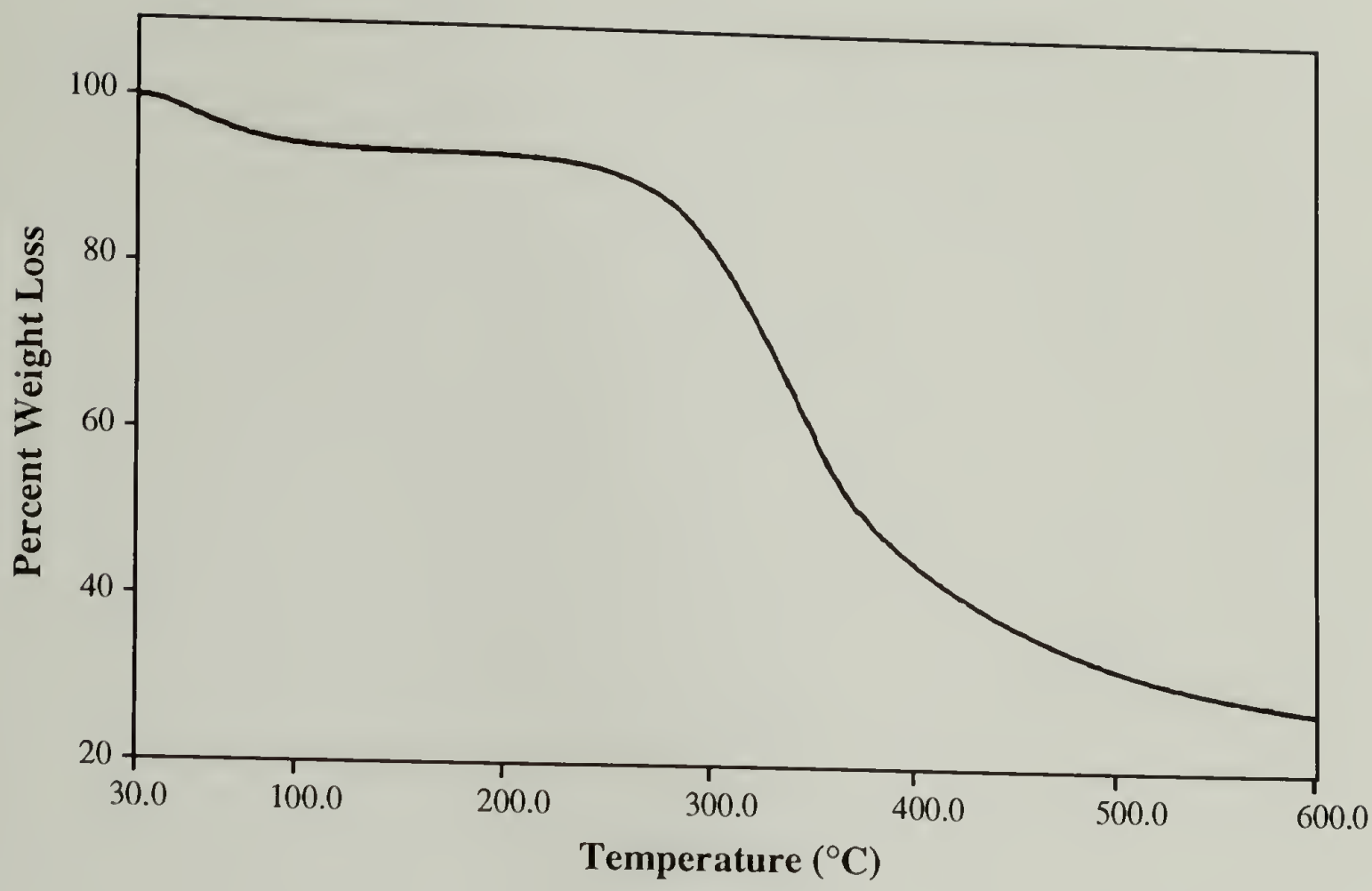


Figure 3.9. Thermogravimetric analysis of HC4-14. A 10 % weight loss is seen from 90-110°C, attributed to water loss. The onset of degradation is seen at 225°C, continuing to constant residue of 30% at 600°C.

Table 3.9. Results of elemental analysis of HC4-14.

Analysis	Calcd* (mol%)	Exptl (mol%)
Carbon	42.90	43.80
Hydrogen	7.16	6.44
Nitrogen	16.71	16.72
Ash	—	<0.1%

*Calculated on the basis of the presence of 10% (w/w) water.

The chemical analysis of HC4-14 can be summarized as follows:

•The NMR and FTIR spectra were consistent with a polypeptide chain containing predominately glycine, alanine, proline and glutamic acid. The NMR spectrum revealed no significant organic contaminants.

•The results of the amino acid composition analysis were consistent with the expected values. The analysis of a concentrated sample confirmed the composition of the low fraction residues of the fusion ends; the presence of valine was unanticipated. The values obtained for a dilute sample were consistent with the calculated composition of the prominent residues.

•The results of mass analysis using matrix-assisted laser-desorption/ionization mass spectrometry, presented in detail in Chapter 4, revealed the presence of two errors in the sequence expected for HC4-14. The protein sample was also found to contain low-mass protein fragments derived from the target molecule. These fragments could be removed by dialysis, yielding a monodisperse sample with the molecular weight expected for the altered sequence.

•Thermogravimetric and elemental analysis indicated the presence of 10% water by weight and the absence of detectable inorganic contaminants.

3.2.8. Cleavage of Fusion Ends

Cyanogen bromide cleavage of HC4-14 was accomplished using the method of Smith.[10] A 100 mg sample of HC4-14 was dissolved in 70% formic acid and crystals of cyanogen bromide (~100 mg) were added. The reaction mixture was incubated at room temperature for 4 hours and the solvent removed *in vacuo*. The residue was dissolved in

distilled water, dried and redissolved in distilled water. The solution was dialyzed in an 8000 molecular weight cut-off dialysis membrane against distilled water overnight. The cleaved protein (53 mg, 80% yield) was recovered by lyophilization and dried *in vacuo*.

Analysis by matrix-assisted laser desorption/ionization mass spectrometry was consistent with the chemical structure expected for HC4-14c with a mass calculated assuming the presence of a homoserine lactone C-terminus, the usual result of cyanogen bromide cleavage.[10] The spectrum is shown in Appendix B. The experimentally determined mass (11,589±2 Da) was slightly higher than the calculated value 11,583 Da and the mass analysis revealed a small amount of both low and high molecular mass fragments, implying that the cyanogen bromide treatment was not optimized to completely cleave the material while suppressing random cleavage and degradative processes. As this sample was used only in crude, preliminary structural analysis, additional chemical analysis was not performed.

3.3. Synthesis of HC4-m (m = 10, 12, 16, 20)

3.3.1. Preparation of 66-mer *Ban* I Monomer

A 5 ng portion of pUC4-2 was used to transform competent HB101 cells prepared as described in sections 2.3.4 and 2.3.5. The presence of the plasmid was verified by agarose gel analysis of *Bam*H I/*Eco*R I and *Ban* I fragments obtained through digestion of plasmids isolated from selected transformants. One transformant was used to inoculate 5 mL of LB medium containing 200 µg/mL ampicillin. The cell culture was grown to saturation at 37°C and then used to prepare a 500 mL saturated cell culture. Plasmid was isolated from this culture using a Qiagen-tip 500 column (Qiagen Inc.). The manufacturer's protocol was followed exactly (see section 2.4.2.2. for a description). After isolation, the concentration of the resuspended plasmid solution was determined to be ~4 µg/µL by agarose gel analysis (400 µg total yield). A 200 µg portion of the plasmid was digested

with *Ban* I, the digestion products were separated by electrophoresis on a 10% polyacrylamide gel and the 66 base pair DNA monomer isolated as described in section 3.2.2.1. After purification, the concentration of the 10 µL DNA solution was estimated (by agarose gel electrophoresis) to be 0.7 µg/µL .

3.3.2. Preparation of DNA Multimers

The DNA monomer (1.4 µg) was ligated as described in section 3.2.2.3 and the entire ligation mixture separated by electrophoresis on a 1.5% low melting point agarose gel. The ethidium bromide stained multimers were visualized under UV light and a gel slice containing the fraction of DNA multimers with between 5 and 30 monomer repeats excised from the gel. The gel slice was placed in a microfuge tube and melted by incubation at 65°C. The volume of the liquid agarose solution was estimated (~400 µL), an equal volume of phenol added and the mixture was mixed by vortex for 3 min. The sample was centrifuged at room temperature for 10 min and the top aqueous layer removed *via* pipette, avoiding the layer of precipitated gel material. The supernatant (~350 µL) was then extracted with a 1:1 phenol/chloroform mixture and again with chloroform. The DNA was precipitated from the solution by addition of a 35 µL of 3M sodium acetate (pH 4.9) and 700 µL of 100% ethanol and incubation at -10°C overnight. The DNA was recovered by centrifugation at 4°C for 30 min, washed with 70% ethanol and dried *in vacuo*. The recovered DNA was resuspended in 10 µL water and used without further analysis.

3.3.3. Preparation of *Ban* I Cut, Dephosphorylated p937.51

A 1 µg portion of p937.51 was mixed with 40 U of *Ban* I in NEBuffer #4 supplied by New England Biolabs and incubated at 37 °C overnight. The digestion products were separated by electrophoresis on a 1% low melting point agarose gel and the ethidium bromide stained DNA bands visualized with UV light. A gel slice containing the linearized plasmid was excised from the gel and the DNA isolated as described in the previous

section. The purified vector was treated with calf intestine alkaline phosphatase as described in section 3.2.2.2., isolated and dissolved in 70 µL TE at a concentration of 10 ng/µL.

3.3.4. Isolation of Multimers

The objective of the experiments outlined in this section was the isolation of a multimer comprised of 7 repeats of the DNA monomer (heptamer). This process took the form of four analyses of multimers cloned into p937.51 using a progression of strategies, with each succeeding strategy based on results obtained from the previous analysis. The analyses are outlined below and a summary of the results is found in Table 3.10.

In Analysis 1, a 4 µL portion of the isolated multimer fraction was mixed with 5 ng of *Ban* I cut, dephosphorylated p937.51 (section 3.2.2.2) in 10 µL of New England Biolabs ligation buffer, T4 DNA ligase (400 U) was added and the mixture was incubated at 16°C overnight. Half of the ligation mix was used to transform competent *E. coli* strain HB101. Twenty transformants (of 43) were selected and used to inoculate 5 mL saturated cultures. The plasmids were isolated, digested with *Bam*H I and the fragments analyzed on a 1% agarose gel. Only two of the selected transformants contained inserts, and the target multimer was not recovered.

In Analysis 2, the remaining ligation mix from Analysis 1 was used to transform competent HB101 cells and 20 of the 31 colonies obtained were selected for analysis as described above. Only four transformants contained plasmids, none with the target multimer.

In Analysis 3, a 4 µL portion of the fractionated multimer ligation solution used in Analysis 1 was ligated into freshly prepared *Ban* I digested, dephosphorylated p937.51 (section 3.2.2.2) and half of the ligation mixture (10 µL) was used to transform competent HB101 cells. A total of 61 colonies were isolated and forty were selected for analysis. Gel analysis of the *Bam*H I digestion fragments showed that 26 of 30 transformants contained

plasmids with inserts and that most of the fragments were monomers. The largest insert recovered was comprised of 6 repeats of the DNA monomer. The DNA fragments were observed to be discrete bands with no indications of instability. However, it is interesting to note that the majority of the recovered multimers were of a size smaller than that of the DNA fragments fractionated from the ligation mixture. No explanation is offered for this observation.

In Analysis 4, a 1 µg portion of the DNA monomer isolated as described in section 3.3.1 was used to synthesize a new DNA multimer population. These multimers were cloned into p937.51 (prepared in Analysis 4) without fractionation and the ligation mixture used to transform HB101 cells. Plasmids from thirty transformants (of 53 colonies) were isolated and the *Bam*H I digestion products analyzed by agarose gel electrophoresis. Once again, the majority of the inserts were hexamers and smaller (mostly dimers and trimers). Two inserts of larger size (an octamer and a decamer) were recovered.

The results of these multimer analyses are summarized in Table 3.10. The search for the multimer with seven monomer repeats was terminated after the completion of the fourth analysis. It is possible that there is some unknown impediment to the isolation of the DNA heptamer, perhaps some biological constraint that prevents the growth of cells transformed with recombinant plasmids containing the heptamer or a recombination process that reduces the DNA to smaller fragment sizes. Based on the overall results of the analyses, multimers comprised of 5, 6, 8, and 10 monomer repeats (transformants 85, 84, 103, and 91 respectively) were selected for further analysis

3.3.5. Construction of Expression System

The selected multimers were excised from the 1% low melting point agarose gels used in the analyses and recovered as described in section 3.3.2.1. The multimers (~1 µg) were isolated and resuspended at a concentration of 0.1 µg/µL in TE. A 10 ng portion of each multimer was mixed with 5 ng of *Bam*H I digested, dephosphorylated pET3-b,

Table 3.10. Summary of isolated DNA multimers encoding HC4-m proteins.

Analysis	Number of colonies	Colonies analyzed	Plasmids w/inserts	Summary of recovered inserts (# of plasmids/insert size)
1	43	20	2	1/4, 1/2
2	31	20	4	1/2, 2/3, 2/6
3	61	40	40	12/1, 6/2, 10/3, 2/4, 4/5, 6/6,
4	53	30	26	9/1, 5/2, 5/3, 3/4, 1/5, 1/6, 1/8 1/10

prepared as described in section 3.3.3 in 10 µL New England Biolabs ligase buffer containing 1 mM ATP, T4 DNA ligase (400 U) was added and the mixtures incubated at 16°C for four hours.

A 5 µL portion of the ligation mixture was used to transform competent HB101 cells, which were plated on LB agar plates containing 200 µg/mL ampicillin and incubated at 37°C overnight. Plasmids from several transformants were isolated, digested with *Bam*H I to verify the presence of an insert and digested with *Ava* I to screen for inserts in the proper orientation as described in 3.2.3.1.

Purified plasmids containing the target DNA multimers in the correct orientation (designated pET4-10, 12, 16, and 20) were used to transform competent BL21(DE3)pLysS cells, which were plated on LB agar plates containing 200 µg/mL ampicillin and 17 µg/mL chloramphenicol and incubated at 37°C for 8 hours. All pET plasmids were successfully transformed, and the presence of the plasmids verified for selected colonies by analysis of *Bam*H I digestion products on a 1% agarose gel (see Figure 3.10). The selected strains were stored for use in protein expression experiments.

3.3.6. Expression and SDS-PAGE Analysis of HC4-m Polypeptides

It was observed in repeated experiments that the recombinant proteins of general sequence **3.1** were easily visualized after SDS-PAGE analysis by staining with Coomassie

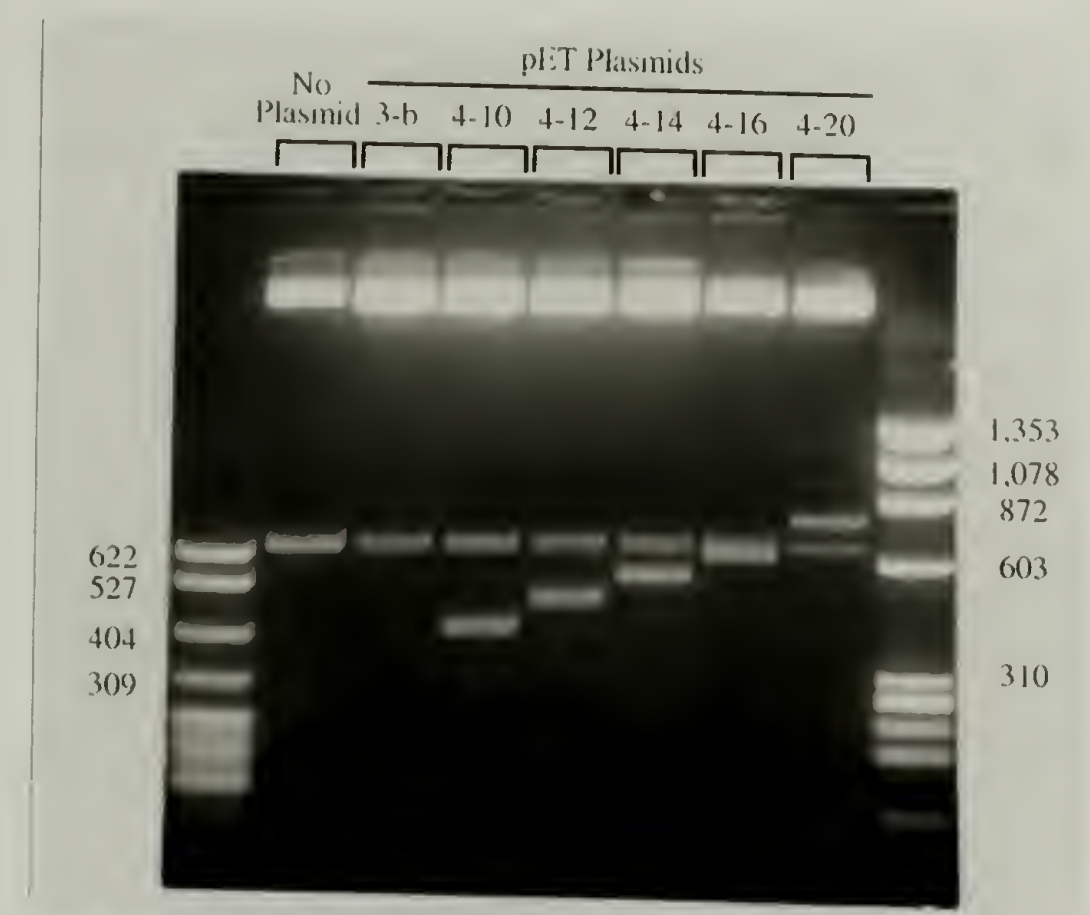


Figure 3.10. Analysis of pET4-m plasmids isolated from BL21(DE3)pLysS *E. coli.* Samples were analyzed on a 1% agarose gel and separated DNA visualized with ethidium bromide. Lane 2 (BL21(DE3)pLysS: no pET plasmid control) shows the 621 bp *Bam*H I fragment from pLysS as well as the 4244 bp vector band. In lane 3 (BL21(DE3)pLysS/pET3-b: no insert control) the expected *Bam*H I cut pET3-b vector band (4637 bp) is unresolved from the pLysS linear vector band. Lanes 4-8 show bands corresponding to multimer inserts from pET4-m (m = 10, 12, 14, 16 and 20) at 405, 471, 537, 603, and 735 base pairs, respectively. The *Bam*H I fragment from pLysS (631 bp) is seen in all lanes. All insert sizes were referenced to DNA standards in lanes 1 and 9. Construction of the pET plasmids is discussed in section 3.3.5.

Blue, obviating the use of *in vivo* labeling experiments of the kind outlined in section 3.2.4.1. In turn, the ability to readily stain the proteins enabled the use of rich media which simplified protein expression experiments. The general protocol is outlined below.

To verify expression of the target proteins in BL21(DE3)pLysS, a 500 µL portion of each saturated cell culture was used to inoculate 50 mL of LB medium containing 200 µg/mL ampicillin and 17 µg/mL chloramphenicol in 250 mL Klett flasks. Cells containing no pET plasmid and pET3-b with no insert were used as controls. Normal starting optical densities were 8-15 Klett units (KU) as measured on a Klett-Summerson Photoelectric Colorimeter (section 2.3.3). The flasks were incubated at 37°C with shaking and the cell density monitored periodically.

Protein production was induced at 50-200 KU by the addition of 24 µL of a 20 mg/mL IPTG solution. At this time (t = 0 min), a 1 mL portion of culture was removed to a microfuge tube, the cells pelleted by centrifugation at room temperature for 2 min and the supernatant discarded. The cells were resuspended in 100 µL TE and lysed by the addition of 100 µL of 2X sample buffer. The optical density was measured at 60, 90, 120, and 180 min after induction and cells from volume fractions of cell culture containing approximately equal number of cells were isolated, resuspended and lysed. For instance, if the $OD_{t=0}$ = 200 KU and the $OD_{t=60}$ = 250 KU, then the appropriate cell culture volume was [(200/250)x(1 mL)] = 0.8 mL. The growth curves for the cell cultures during expression are shown in Figure 3.11.

A 10% discontinuous SDS polyacrylamide gel was prepared as described in section 2.4 and the proteins in 20 µL portions of the lysed cell samples taken at t = 0 and t = 180 were separated by electrophoresis at 12 mA constant current for fourteen hours. The gel was soaked in Coomassie Blue stain solution for 1 hour and then destained for several hours. As seen in Figure 3.12 the appearance of a prominent new band is seen at 180 min for each of the samples against a background of normal *E. coli* proteins. The observed masses are approximately 2.5 times the expected values (discussed in section 4.1).

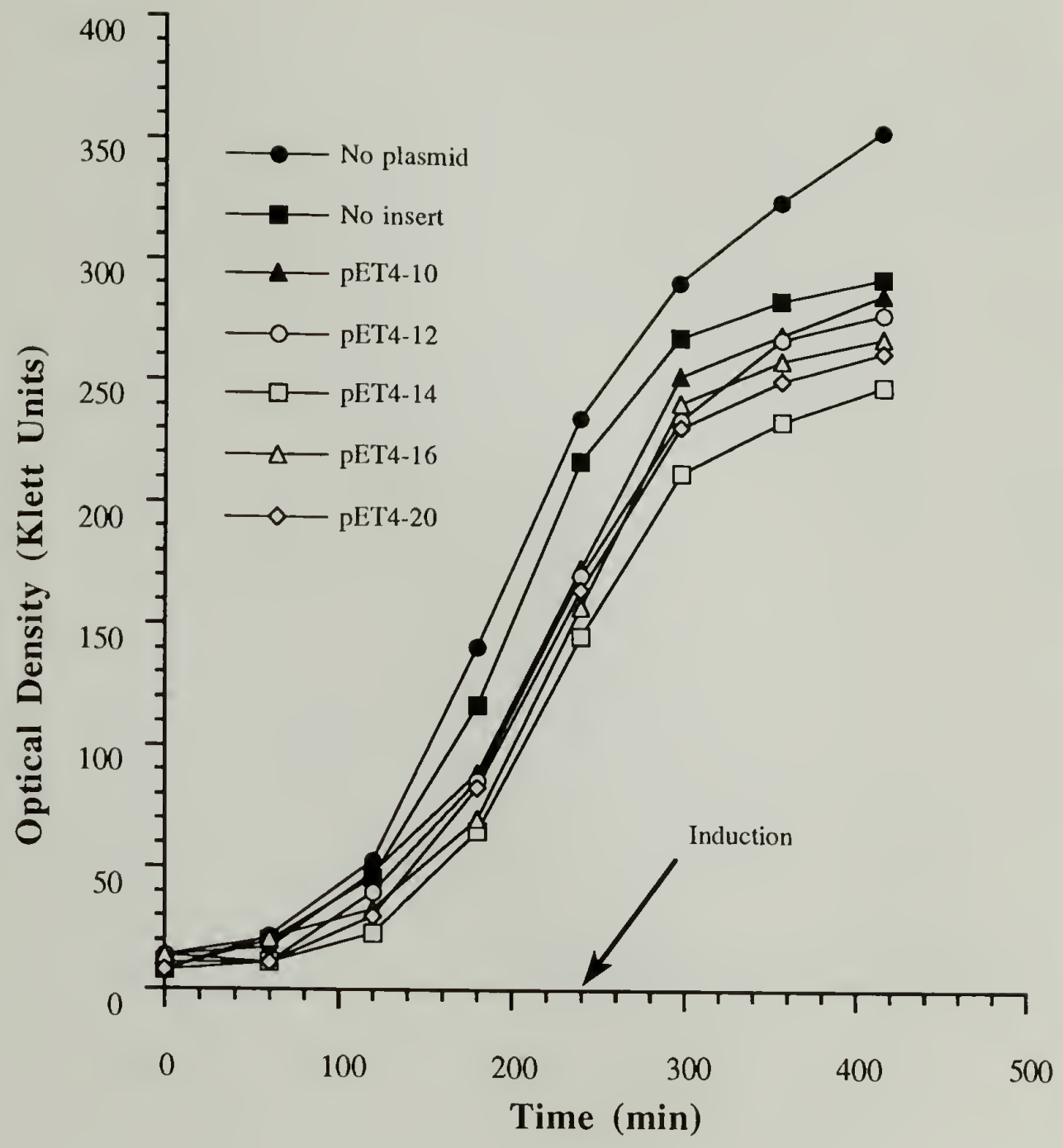


Figure 3.11. Cell growth plot of BL21(DE3)pLysS/pET4-m *E. coli*. Cells with no pET plasmid (no induction control), pET3-b (no insert control), and pET4-m plasmids were grown in LB medium containing antibiotics and the growth monitored using a Klett-Summerson Colorimeter, starting with average optical densitites of 10 KU at inoculation. Cells containing pET plasmids are seen to grow at a slightly slower rate than the no pET plasmid control. Protein production was induced by addition of IPTG 240 minutes after inoculation. Cell growth is observed to continue as long as 180 minutes after induction at a slightly retarded rate in the strains with the pET plasmids (relative to the no pET plasmid control); the rate reduction is observed to start as early as 30 minutes after induction. Expression of the HC4-m proteins is discussed in section 3.3.6.

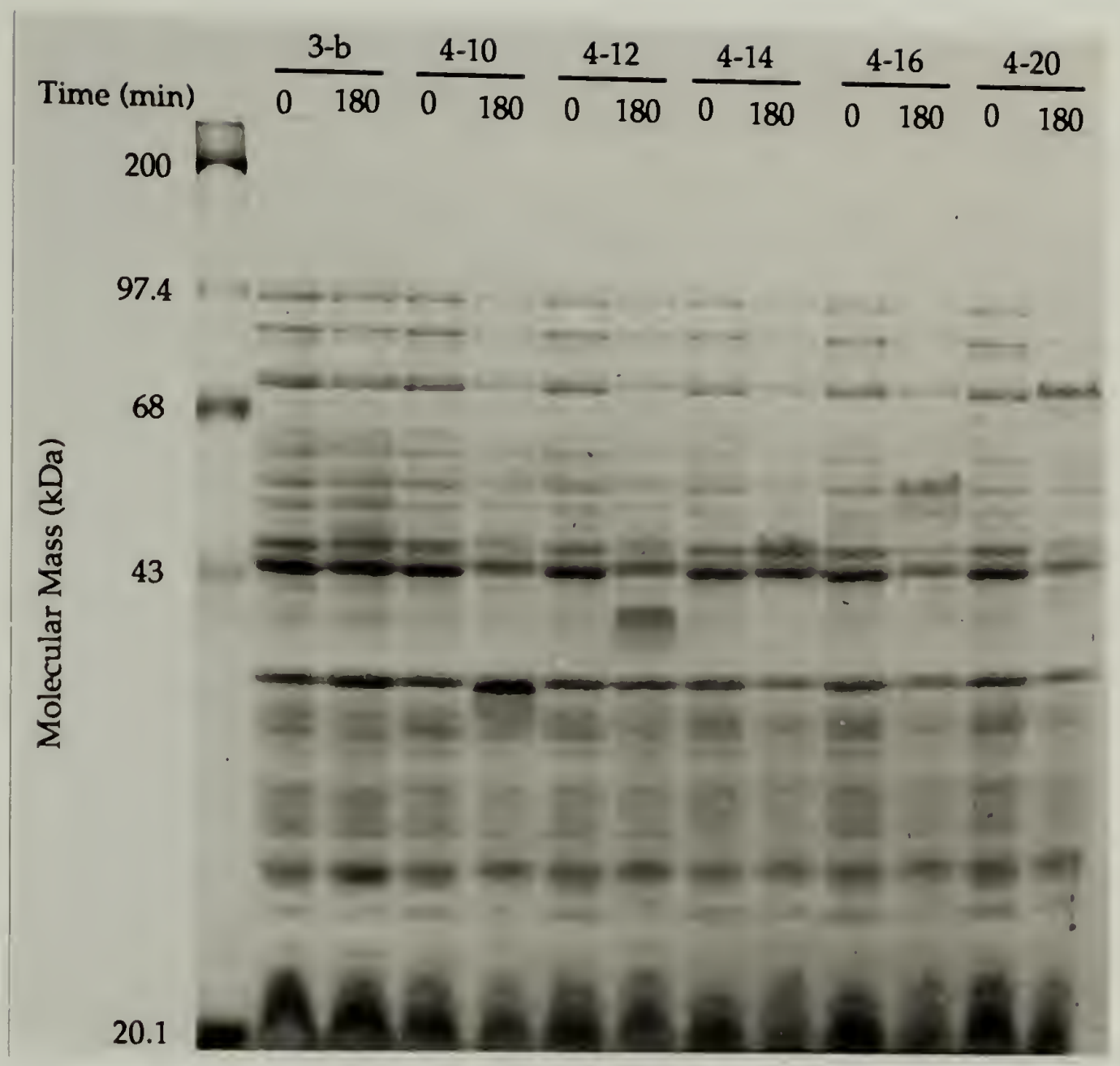


Figure 3.12. SDS-PAGE analysis of HC4-m proteins expressed in BL21(DE3)pLysS *E. coli*. Proteins produced at induction with IPTG (t = 0) and 180 minutes post-induction (t = 180) from approximately equal numbers of cells were separated on a 10% SDS-polyacrylamide gel at 12 mA constant current and stained with Coomassie Brilliant Blue G-250. Lanes 2-3 contain proteins from BL21(DE3)pLysS/pET3-b (No insert control). The background of *E. coli* proteins are visible in these lanes, with no significant new bands appearing after induction. Prominent new bands of significant intensity are shown 180 minutes post-induction for each of the HC4-m samples. The apparent molecular masses of the HC4-m proteins (31,000, 42,000, 49,000, 58,000, and 78,000 Da for m = 10, 12, 14, 16, and 20, respectively) differ significantly from the calculated values (14,024, 15,616, 17,207, 18,799, and 21,982 Da, respectively). The significance of this observation is discussed in section 4.1. Expression of HC4-m proteins is discussed in section 3.3.6.

As these proteins were intended for use strictly as diagnostic samples in the mass analysis discussed in Chapter 4, they were isolated on a small scale from one liter cultures using the purification scheme outlined in section 3.2.6. Typical yields were 50 mg of purified fusion protein. No further chemical analysis was attempted.

3.4. Synthesis of HCn-16 (n = 3, 4, 5, 6)

3.4.1. Synthesis of pUCn-2 (n = 5, 6)

The construction of recombinant pUC plasmids containing inserts encoding two repeats of **3.1** with n = 5, and n = 6 was accomplished using the scheme outlined in section 3.2.1 with a few minor exceptions. Single strand, complementary oligonucleotide pairs **3.3** and **3.4** (n = 5) and **3.5** and **3.6** (n = 6) were synthesized, purified, treated with kinase and annealed to yield 94 and 106 bp duplexes, respectively. These duplexes were inserted in *Bam*H I and *Eco*R I cut and gel purified pUC18 with T4 DNA ligase, and the recombinant vectors used to transform *E. coli* strain HB101. Selected colonies were screened for the presence of the insert, the plasmids were isolated and the sequence of the insert verified. The recombinant plasmids were isolated on a large scale using the Qiagen column procedure described in section 3.3.1.

A sample of pUC18-$[(AG)_3PEG]_2$ (hereafter referred to as pUC3-2) was obtained from Kevin McGrath and used without further manipulation; its design and synthesis are discussed.[7] The three recombinant plasmids pUCn-2 (n = 3, 5, and 6) were digested with *Ban* I to recover the DNA monomers. The digestion products were separated by electrophoresis for 2 hours at 80 V constant voltage on a 4% NuSieve® GTG low melting point agarose gel containing ethidium bromide. The 54, 78, and 90 base pair *Ban* I monomers were identified by comparison to the pBR322/*Msp* I digestion fragments used as a size standard. The DNAs were recovered from excised gel slices as described in section 3.3.2. After purification, the dry DNA pellet was resuspended in 10 µL of TE.

A portion of the monomer solution was analyzed on a 4% NuSieve® minigel and the concentration estimated at 200 ng/µL for each solution by comparison to size standards.

5'-AATTCG TAA GGT GCC GGC GCT GGT GCG GGC GCT GGG GCT GGT CCG GAA GGT GCT GGT GCA GGC GCT GGC GCG GGC GCG GGC CCG GAA GGT GCC G-3' **3.3**

5'-GC ATT CCA CGG CCG CGA CCA CGC CCG CGA CCC CGA CCA GGC CTT CCA CGA CCA CGT CCG CGA CCG CGC CCG CGC CCG GGC CTT CCA CGG CCTAG-3' **3.4**

5'-AATTCG TAA GGT GCC GGC GCA GGA GCT GGT GCG GGC GCT GGG GCT GGT CCG GAA GGT GCT GGT GCA GGC GCT GGC GCG GGC GCTGGC GCG GGC CCG GAA GGT GCC G-3' **3.5**

5'-GC ATT CCA CGG CCG CGT CCT CGA CCA CGC CCG CGA CCC CGA CCA GGC CTT CCA CGA CCA CGT CCG CGA CCG CGC CCG CGA CCG CGC CCG GGC CTT CCA CGG CCTAG-3' **3.6**

3.4.2. Isolation of Target Multimers

The multimer isolation protocols described section 3.3.4 were modified to bias the results to the isolation of DNA multimers containing 8 repeats of the monomer. A 1 µg portion of each of the DNA monomers was suspended in 10 µL of New England Biolabs ligase buffer containing 1 mM ATP. T4 DNA ligase was added and the reaction incubated on ice for one hour. The ligation mix was loaded on a 1.2% LMP agarose gel containing ethidium bromide and the DNA separated by electrophoresis at 80 V for two hours. The separated multimers were visualized under UV light and large ligation products were seen. The target multimers were not resolved, so gel slices containing DNA in the desired size range were excised and the DNA recovered. The purified DNA was ligated into p937.51, the ligation mix used to transform HB101, the plasmids from selected colonies were isolated, digested with *Bam*H I and the digestion products analyzed on a 1.5% low melting point agarose gel. The result of this analysis are outlined in Table 3.11: the desired DNA multimers with n = 3, and 5 were recovered, but no small fragments were seen for n = 6.

Table 3.11. Summary of isolated DNA multimers encoding HCn-16 proteins.

n	Total number of colonies	Number of colonies analyzed	Summary of recovered inserts (# of plasmids*/insert size**)
3	90	10	3/16, 1/11, 2/9, 1/8, 3/<7
5	30	15	2/8, 2/7, 5/6, 6/<4
6	15	15	2/7, 5/6, 5/5, 3/<5

*Multimer containing recombinant p937.51 plasmids isolated from HB101 *E. coli* cells.
**Insert size x = number of DNA monomers in multimer. The target size was 8.

In order to isolate multimers with the correct size for n = 6, another portion of the 90 bp *Ban* I monomer was suspended in 10 µL of ice-cold New England Biolabs ligase buffer containing 1 mM ATP. T4 DNA ligase was added and the mixture immediately loaded on a 1.5% agarose gel and separated by electrophoresis. Surprisingly, despite the short ligation time, DNAs as large as 5000 bp were visible on the gel. The band containing the target DNA (calcd 795 bp) was clearly resolved: a gel slice containing the band was excised and the DNA isolated. A 5 ng portion (of 50 ng purified yield) was mixed with 5 ng of p937.51, the DNAs ligated as usual, and the ligation mix used to transform *E. coli* strain HB101. Recombinant plasmids were isolated, digested with *Bam*H I and the digestion products analyzed on a 1.5% low melting point agarose gel. Of 15 colonies analyzed, 11 contained recombinant plasmids containing multimers of the correct size, the remainder contained multimers with 7 repeats of the monomer. This procedure represents the most well controlled multimer isolation performed in this study.

3.4.3. Construction of Expression Systems

The target DNAs (507, 699, and 795 bp for n = 3, 5, and 6 respectively) were isolated from the low melting point gels used for the size analysis and purified as usual. The multimers were ligated into pET3-b which had been treated with *Bam*H I and

phosphatase as described in section 3.3.3. The ligation mixes were used to transform competent BL21(DE3)pLysS cells prepared as described in section 2.3.4. Transformants containing recombinant pET plasmids with the insert in the correct orientation (verified by separate digestions with *Bam*H I and *Ava* I) were isolated and maintained for protein expression. The *Bam*H I digestion products from all pETn-16 plasmids isolated from the BL21(DE3)pLysS strains were analyzed on a 1% agarose gel (Figure 3.13).

3.4.4. Expression and Analysis of HCn-16 Polypeptides

Expression of the HCn-16 (n = 3, 4, 5, and 6) was demonstrated using the protocol outlined in section 3.3.6. Saturated overnight cultures of the expression strains and controls (no pET plasmid, plasmid with no insert) were used to inoculate 50 mL of LB medium containing antibiotics. Cell growth at 37°C was monitored and protein production induced at an optical density of 180-200 KU by the addition of IPTG to 0.4 mM. The cells were incubated for an additional three hours with cell samples extracted periodically. The growth curves for the expressions are seen in Figure 3.14.

The proteins in the cell lysates were separated on a discontinuous 10% SDS polyacrylamide gel by electrophoresis at 12 mA constant current for 11 hours. The separated proteins were visualized with Coomassie Blue dye; the gel is seen in Figure 3.14. A prominent new band appears in the cell lysates three hours after induction, once again with anomalously high apparent molecular weights when compared to the standards.

3.4.5. Large-scale Synthesis and Purification

Each of the HCn-16 proteins was produced in two batch fermentations of either 24 or 20 L. Two 10 or 12 L fermentations were performed simultaneously (e.g. HC3-16 was isolated from 2 x [2 x (12 L)] = 48 L of total cell culture). Proteins were isolated from the combined cell pellets from all production runs. Typically, 125 mL saturated cultures of BL21(DE3)pLysS/pETn-16 were used to inoculate 10 or 12 L of LB medium containing

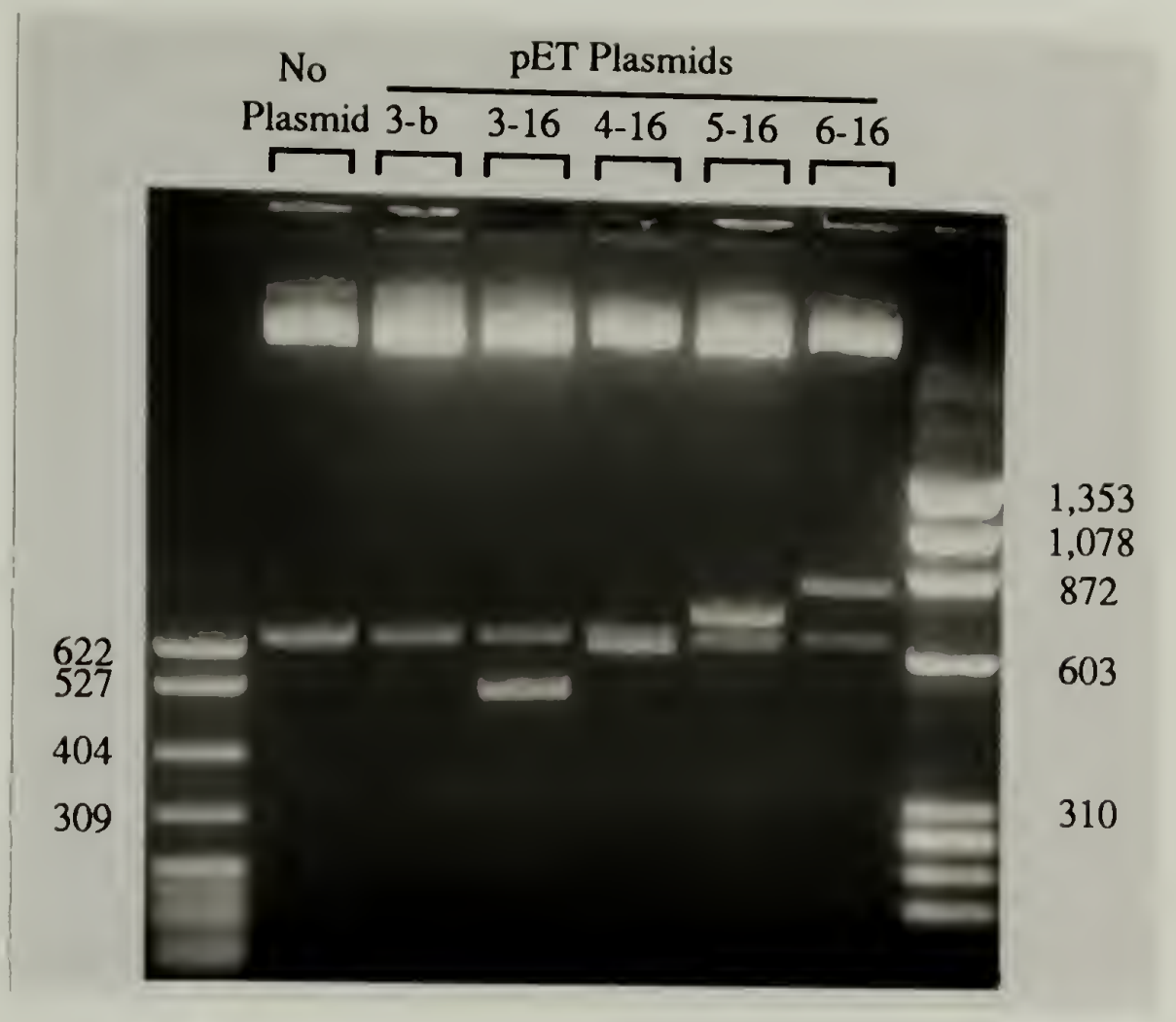


Figure 3.13. Analysis of pETn-16 plasmids isolated from BL21(DE3)pLysS *E. coli*. Samples were analyzed on a 1% agarose gel and separated DNA visualized with ethidium bromide. Lane 2 (BL21(DE3)pLysS: no pET plasmid control) shows the 621 bp *Bam*H I fragment from pLysS as well as the 4244 bp vector band. In lane 3 (BL21(DE3)pLysS/pET3-b: no insert control) the expected *Bam*H I cut pET3-b vector band (4637 bp) is unresolved from the pLysS linear vector band. Lanes 4-7 show bands corresponding to multimer inserts from pETn-16 (n = 3, 4, 5, and 6) at 507, 603, 699 and 795 base pairs, respectively. The *Bam*H I fragment from pLysS (631 bp) is seen in all lanes. All insert sizes were referenced to DNA standards in lanes 1 and 8. Construction of the pET plasmids is discussed in section 3.4.3.

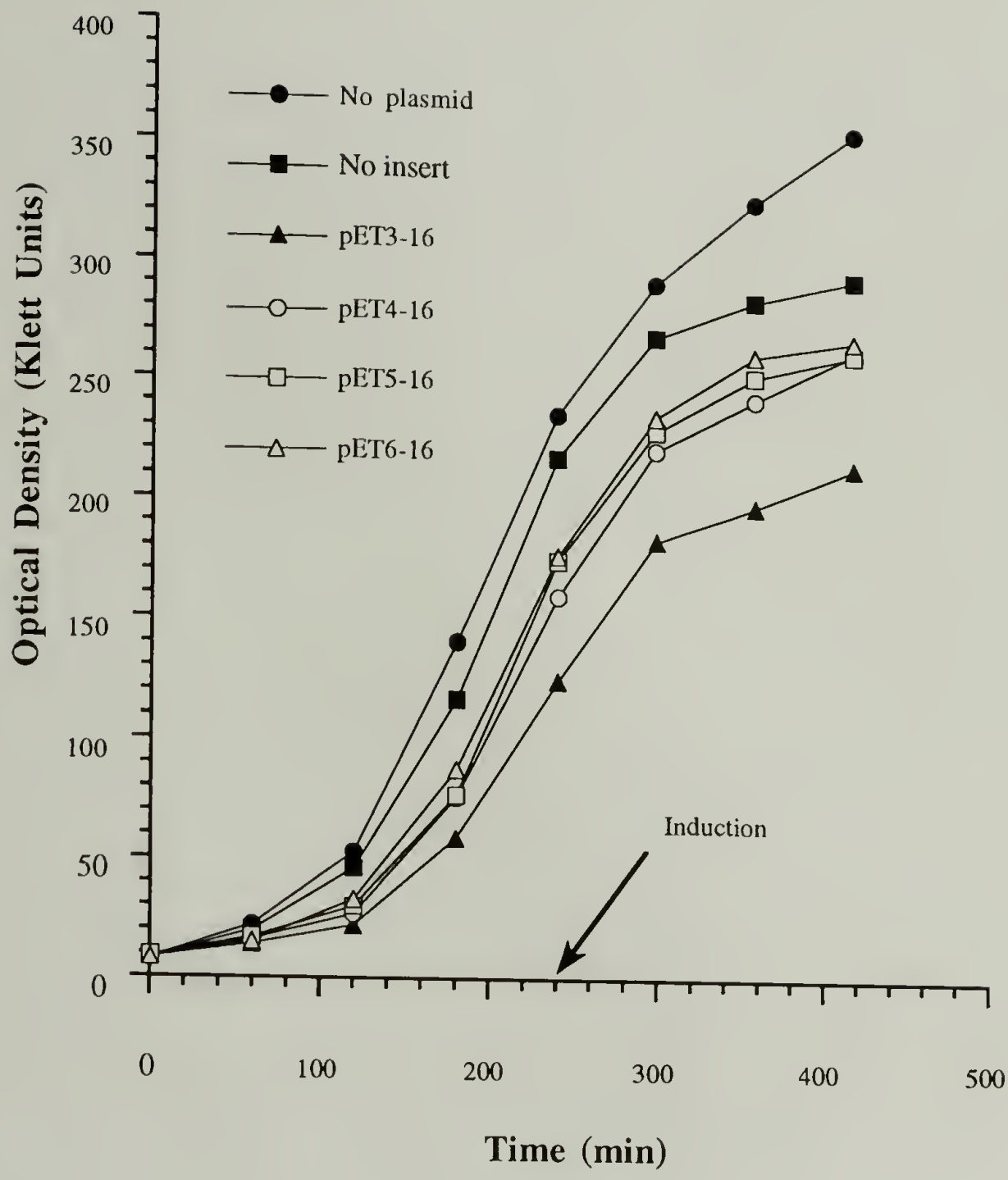


Figure 3.14. Cell growth plot of BL21(DE3)pLysS/HCn-16 *E. coli*. Cells with no pET plasmid (no induction control), pET3-b (no insert control), and pET4-m plasmids were grown in LB medium containing antibiotics and the growth monitored using a Klett-Summerson Colorimeter, starting with average optical densitites of 10 KU at inoculation. Cells containing pET plasmids are seen to grow at a slightly slower rate than the no pET plasmid control. Protein production was induced by addition of IPTG 240 minutes after inoculation. Cell growth is observed to continue as long as 180 minutes after induction at a slightly retarded rate in the strains with the pET plasmids (relative to the no pET plasmid control); the rate reduction is observed to start as early as 30 minutes after induction. Expression of the HC4-m proteins is discussed in section 3.4.4.

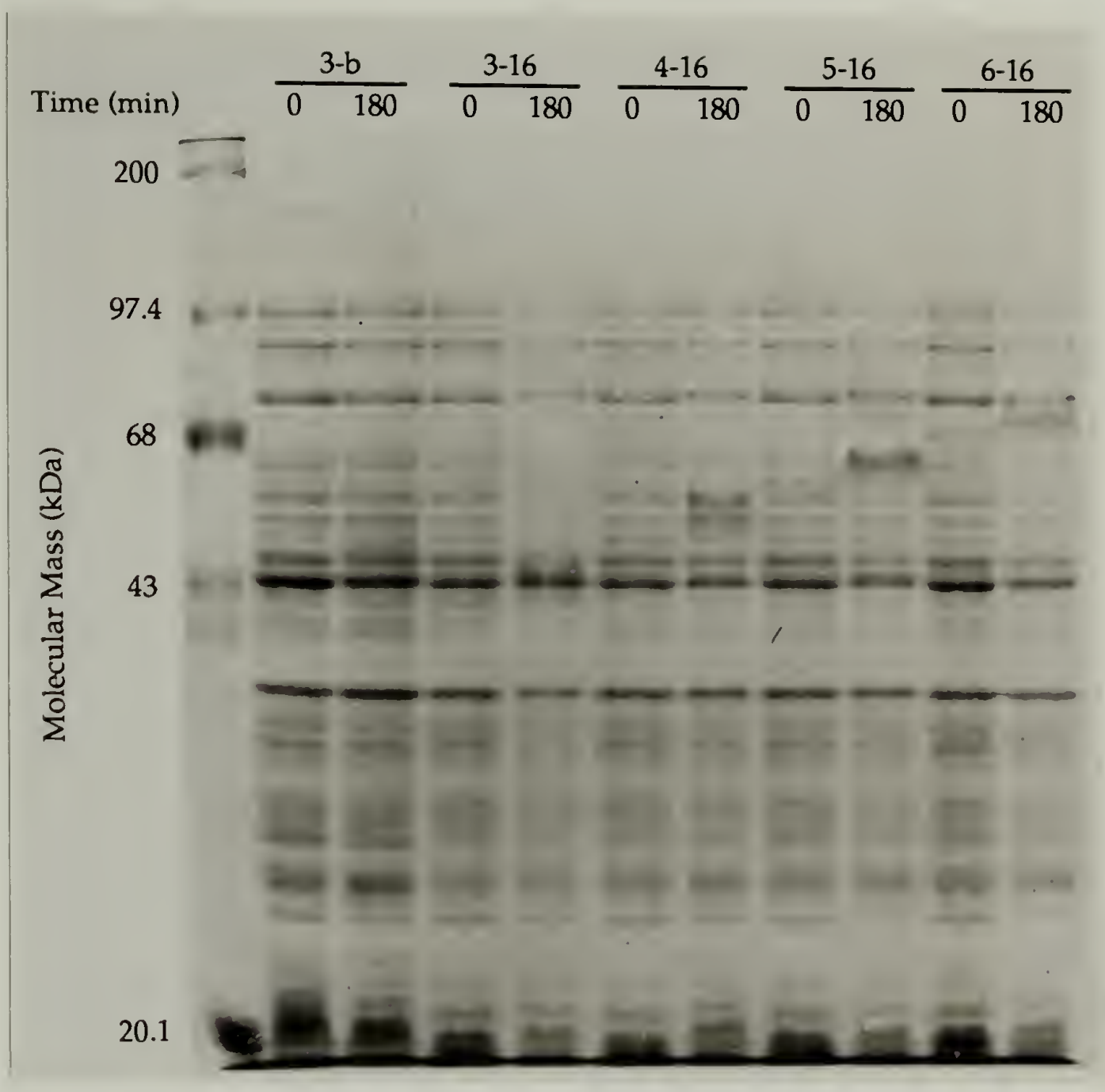


Figure 3.15. SDS-PAGE analysis of HCn-16 proteins expressed in BL21(DE3)pLysS *E. coli*. Proteins produced at induction with IPTG (t = 0) and 180 minutes post-induction (t = 180) from approximately equal numbers of cells were separated on a 10% SDS-polyacrylamide gel at 12 mA constant current and stained with Coomassie Brilliant Blue G-250. Lanes 2-3 contain proteins from BL21(DE3)pLysS/pET3-b (No Insert control). The background of *E. coli* proteins are visible in these lanes, with no significant new bands appearing after induction. Prominent new bands of significant intensity are shown 180 minutes post-induction for each of the HC4-m samples. The apparent molecular masses of the HCn-16 proteins (47,000, 57,000, 65,000, and 75,000 Da for n = 3, 4, 5, and 6, respectively) differ significantly from the calculated values (16,749, 18,799, 20,849, and 22,899 Da, respectively). The significance of this observation is discussed in section 4.1. Expression of HCn-16 proteins is discussed in section 3.4.4.

antibiotics in a New Brunswick Scientific Inc. Microferm Fermenter to an optical density of 10 KU. The temperature was equilibrated to 37°C and 1 mL of Antifoam A concentrate added to suppress foaming. The cells were grown to an optical density of 180-225 KU (~ 3 hours) with vigorous aeration (15 psi house compressed air at a rate of 9,000 cc/min with 400 rpm agitation). Protein synthesis was induced by the addition of IPTG to 0.4 mM (0.95 or 1.14 g in 10 mL of dimethylformamide). After three hours, the cells from both fermenters were harvested in four 1 L bottles by centrifugation at 4000 rpm (Sorvall RC-3 centrifuge, HG-4L rotor) for 20 min at 4°C. The cells were resuspended in 100 mL water per 12 L of cell culture (i.e. 200 mL water per 24 L batch production), placed in a 250 mL centrifuge bottle and frozen at -70°C.

The proteins were recovered from the cells using a modification of the protocol described in section 3.2.6. The suspended cells were thawed on ice, 750 µL of 1 M magnesium chloride and 3 mg of DNAse and RNAse were added and the mixtures incubated at room temperature for 15 minutes. This digestion served to considerably reduce the viscosity of the solution. The lysate was centrifuged at 16,300 x g at 4°C for 30 min to remove insoluble cell debris. The supernatant was decanted (avoiding residual cell debris) and the gelatinous pellet washed with 20 mL of water. The suspension was centrifuged and the supernatants combined. After the initial clarification, the cell extracts were acidified to pH 6 by the addition of glacial acetic acid, then centrifuged at 16,300 x g at 4°C for 30 min to remove precipitated contaminants. The supernatant was removed to a new bottle, acidified to pH 5 with glacial acetic acid and precipitated contaminants removed by centrifugation 16,300 x g at 4°C for 30 min. This treatment was repeated with adjustment of the pH to 4.5 and then to pH 4.0. The pH 4 supernatant (~ 200 mL) of the HCn-16 proteins with n = 3, 4, and 5 was then adjusted to 40% ethanol by the addition of 100% ethanol and incubated for 6 hours at 10°C to precipitate more contaminants. The HC6-16 supernatant was adjusted to 20% ethanol in a similar fashion. The suspensions were centrifuged 16,300 x g at 4°C for 30 min, the supernatant removed to a 1 L centrifuge

bottle, adjusted to 80% ethanol and the recombinant protein allowed to precipitate at room temperature for an hour. The precipitated material was recovered by centrifugation at 16,300 x g at 4°C for 30 min washed with several portions of 80% ethanol and dried *in vacuo*. A summary of the large scale production is shown in Table 3.12.

Table 3.12. Results of large scale production of HCn-16 proteins.

HCn-16	Total cell culture volume (L)	Mass of purified protein (g)	Production scale (mg protein/L cell culture)
3	48	4.24	88
4	40	3.64	91
5	40	4.19	105
6	48	6.21	129

A portion of each protein was further purified using one of two methods. Water soluble HC3-16 and HC4-16 (2 g)were dissolved in 100 mL 0.1 M ammonium bicarbonate, dialyzed against water overnight, recovered by lyophilization and dried *in vacuo*. Two grams of water insoluble HC5-16 and HC6-16 were washed three times with 100 mL portions of water and three times with 100 mL portions of methanol and dried *in vacuo*. All samples were dried in a vacuum oven at 50 °C for twelve hours. The results of the elemental analysis of the proteins are shown in Table 3.13. The values for HC3-16 and HC4-16 were corrected for the presence of 8.5% water. The values for HC5-16 and HC6-16 are uncorrected. All samples showed a low level of inorganic residue (<0.1% ash).

NMR spectra of these materials were recorded by Dr. Janos Borbely. Samples were dissolved in deuterated formic acid and showed the expected resonances of a glycine, alanine, glutamic acid and proline rich protein with no indication of organic contaminants (the spectra are collected in Appendix C). In the absence of suitable mass analysis data for high precision confirmation of the mass of the protein, the amino acid composition analysis

was considered to be the crucial test of the integrity of the amino acid sequence. The results (Table 3.14) confirmed the compositions of the fusion proteins within experimental error. In summary, the chemical analysis of the HCn-16 fusion proteins indicated that the samples were recovered with low levels of organic (NMR analysis) and inorganic (elemental analysis) contaminants and were of the correct amino acid composition.

3.4.6. Cleavage of Fusion Ends

The N- and C-terminal fusions were cleaved from the purified HCn-16 proteins using a scaled-up procedure based on the one described in section 3.2.8. The mass of protein calculated to yield one gram of cleaved material was dissolved in sufficient 70% formic acid (HC6-16 was first dissolved in 90% formic acid and diluted to 70% with water) to make a 5 mg/mL solution. The stirred solution was alternately purged with nitrogen and evacuated several times over the course of an hour. A 50 mol excess of cyanogen bromide was added and the reaction slowly stirred for twenty hours.

For all materials except HC6-16, the solvent was removed using a rotary evaporater and the residue treated with 100 mL of water. Both HC4-16c and HC5-16c were washed five times with 20 mL portions of water, then three times with 20 mL portions of methanol and dried *in vacuo*. HC3-16c was dissolved in water, dialyzed against water, recovered by lyophilization and dried. After drying at 50°C for 20 hours, the proteins were stored desiccated at -10°C. All of these materials were recovered at ~75% yield and used in the analyses described in Chapter 5 without further purification.

HC6-16c consistently formed an insoluble gel during cleavage. This gel was stable to 70% formic acid, water and methanol. Reaction conditions (lower concentration without stirring) could be controlled to prevent gelation during the reaction, but the material would gel during solvent removal. All attempts to recover HC6-16c in a soluble form failed. Samples for solid-state analysis were prepared by washing the recovered HC6-16c

Table 3.13. Elemental analysis of HCn-16 proteins.

Atom	HC3-16		HC4-16		HC5-16		HC6-16	
	Calcd*	Exptl	Calcd*	Exptl	Calcd	Exptl	Calcd	Exptl
C	49.27	48.96	49.01	48.60	48.80	49.01	48.62	48.02
H	6.43	7.21	6.42	7.01	6.41	7.05	6.39	7.41
N	18.99	19.10	19.30	19.70	19.55	20.10	19.76	19.23
Ash	—	<0.1%	—	<0.1%	—	<0.1%	—	<0.1%

*Calculated based on the presence of 8.5% water.

Table 3.14. Amino acid composition analysis of HCn-16 proteins.

	HC3-16		HC4-16		HC5-16		HC6-16	
	Calcd	Exptl	Calcd	Exptl	Calcd	Exptl	Calcd	Exptl
Y	1.01	1.01	0.87	0.89	0.76	0.80	0.68	0.77
S	1.01	1.16	0.87	1.12	0.76	0.85	0.68	1.00
P	10.05	10.26	8.66	8.96	7.60	7.43	6.78	6.47
F	0.50	0.56	0.43	0.55	0.38	0.44	0.34	0.44
M	2.51	2.64	2.16	2.41	1.90	2.36	1.70	2.19
K	1.51	1.49	1.30	1.31	1.14	1.01	1.02	1.02
L	1.01	0.89	0.87	0.67	0.76	0.54	0.68	0.63
I	0.50	0.44	0.43	0.33	0.38	ND*	0.34	ND
H	0.50	ND	0.43	ND	0.38	ND	0.34	ND
G	34.67	35.73	36.80	36.84	38.40	39.72	39.66	40.66
E/Q	11.56	11.57	9.96	10.26	8.75	8.38	7.80	7.70
D/N	2.01	2.23	1.73	2.03	1.52	1.67	1.36	1.68
R	3.02	3.07	2.60	2.63	2.28	2.17	2.03	2.04
A	29.15	27.94	32.04	31.10	34.22	33.86	35.93	34.62
T	1.01	1.00	0.87	0.90	0.76	0.76	0.68	0.77
V	0.00	ND	0.00	ND	0.00	ND	0.00	ND
C	0.00	ND	0.00	ND	0.00	ND	0.00	ND
W	0.00	ND	0.00	ND	0.00	ND	0.00	ND

*ND = not detected

containing gel five times with 40 mL of 70% formic acid, 40 mL water and 40 mL methanol. The sample was stored as a methanol suspension at -10°C.

HC3-16c, HC4-16c, and HC5-16c were analyzed by NMR spectroscopy which revealed the expected resonances and a low level of organic contaminants (the NMR spectra are collected in Appendix C). The amino acid compositions of the materials were confirmed within experimental error; the results are summarized in Table 3.15. Once again, the amino acid data was taken as good evidence of the integrity of the amino acid sequence.

Due to solubility limitations of HC6-16c, NMR and amino acid compostion analysis was impossible. The level of purity of the HC6-16 fusion protein (Tables 3.13 and 3.14), taken together with the success of the cyanogen bromide treatment of the other HCn-16 proteins (witnessed by the amino acid composition data), was taken as a sufficient assessment of the purity of HC6-16c for solid state analysis.

Table 3.15. Amino acid composition analysis of HCn-16c proteins.

	HC3-16c		HC4-16c		HC5-16c	
	Calcd	Exptl	Calcd	Exptl	Calcd	Exptl
P	10.81	10.47	8.89	8.83	7.55	7.39
L	—	0.98	—	—	—	—
G	43.92	40.54	45.00	45.26	45.76	45.89
E/Q	10.81	11.86	8.89	9.46	7.55	7.99
D/N	—	1.30	—	—	—	—
R	0.68	2.07	0.56	ND	0.47	ND
A	33.11	32.79	36.11	36.45	38.21	38.73
HSL	0.68	ND	0.56	ND	0.47	ND

3.5. Discussion of the Synthetic Technique

The synthesis of copolypeptides using the techniques of genetic engineering was outlined in this chapter. Three separate, yet related syntheses of materials with repetitive

sequence **3.1** were accomplished. The success of the techniques is highlighted by recapping the exceptionally well controlled synthesis of HC6-16. Polymerization of the DNA monomer was controlled sufficiently to bias the multimer population in the target range. The desired multimer was specifically lifted from this population and almost exclusively cloned and isolated. After construction of the expression system, large scale batch production and the application of a simple purification scheme resulted in the isolation of over 6 g of protein with a high degree of purity. The entire production and purification process took a total of 4 days. Purified yields should increase with improved knowledge of fermentation technology and optimized expression conditions.

3.6. References

1. E.-L. Winnacker, *From Genes to Clones: Introduction to Gene Technology* (VCH Verlagsgesellschaft, Weinheim, 1987), p. 243.

2. S.-I. Aota, T. Gojobori, F. Ishibashi, T. Maruyama, T. Ikemura, *Nucleic Acids Res.* **16**, r315 (1988).

3. L. J. McBride and M. H. Caruthers, *Tetrahedron Lett.* **24**, 245 (1983).

4. J. Sambrook, E. F. Fritsch, T. Maniatis, *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Press, Cold Spring Harbor, 1989), Appendix E.

5. K. P. McGrath, thesis, University of Massachusetts (1991).

6. K. P. McGrath, personal communication, (1991).

7. K. P. McGrath, M. J. Fournier, T. L. Mason, D. A. Tirrell, *J. Am. Chem. Soc.* **114**, 727 (1992).

8. K. Wüthrich, *NMR of Proteins and Nucleic Acids* (John Wiley and Sons, New York, 1990), Chapter 1.

9. R. D. B. Fraser and T. P. MacRae, *Conformation in Fibrous Proteins and Related Synthetic Polypeptides* (Academic Press, New York, 1973), p. 326.

10. B. J. Smith, *Methods Mol. Biol.* **3**, 70 (1988).

CHAPTER 4

ANALYSIS OF COPOLYPEPTIDES USING MASS SPECTROMETRY

4.1. Overview

A major premise of this work is that the techniques of molecular biology allow the synthesis of monodisperse polymers of defined sequence. The ability to construct such well defined polymer systems demands the development of analytical techniques capable of verifying the accuracy of the synthesis. A crucial variable is the molecular mass of the protein, which, if analyzed with sufficient precision, yields valuable insight into the composition of the chains and the presence of contaminating sequences.

The present study is an example of expanded synthetic ability taxing the capability of an established analytical technique. This is illustrated in Figure 4.1, which shows the anomalous migration of the copolypeptides used in this study relative to molecular mass standards during SDS-PAGE. The observed molecular mass of the proteins is roughly 2.5 times the calculated value for all sequences. It is implicit in the linear relationship between $\log_{10}$(molecular mass) and mobility that the proteins bind a constant mass ratio of SDS—this behavior underlies the ability to compare mobilities of unknown samples to molecular mass standards. However, anomalous migration of acidic proteins has been observed that has been attributed to reduced binding of SDS.[1] This behavior invalidates the principle of the SDS-PAGE technique and renders it useless for analysis of the molecular masses of acidic proteins of the kind used in this study. Moreover, even under rigorous conditions, SDS-PAGE allows an estimation of protein mass to only 5-10%.

In a recent review, Chait and Kent have described techniques of mass spectrometry developed for the purpose of accurately "weighing naked proteins" of high molecular mass.[2] The instrumentation for accurate mass analysis of proteins consists of two

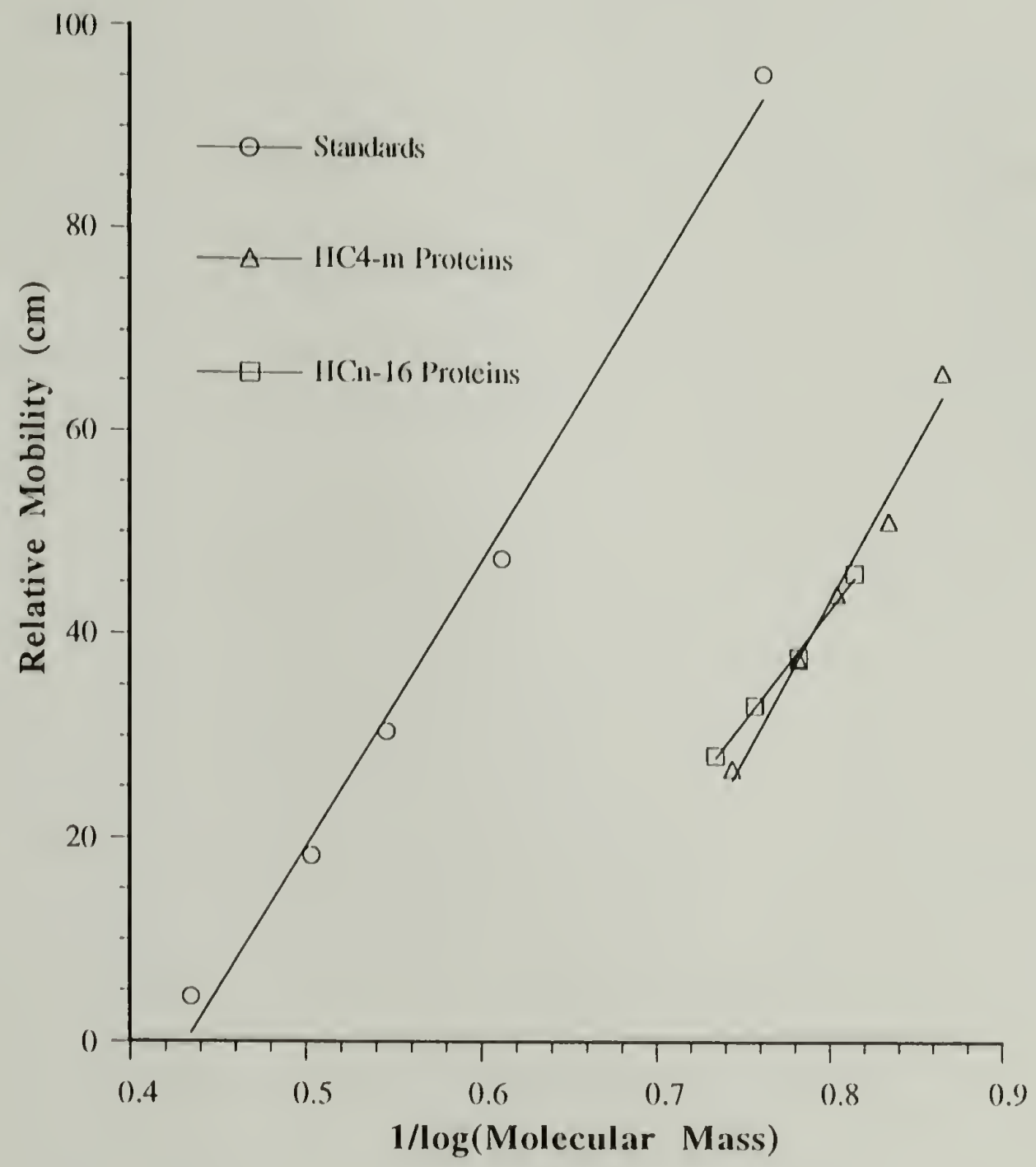


Figure 4.1. Comparison of mobility of HCn-m proteins versus molecular mass standards during SDS-PAGE. The HC4-m (m=10, 12, 14, 16, and 20) and the HCn-16 (n=3, 4, 5, and 6) proteins are plotted along with protein standards: myosin, 200 kDa; phosphorylase B, 97.4 kDa; bovine serum albumin, 68 kDa; ovalbumin, 43 kDa; and soybean trypsin inhibitor, 20.1 kDa. The repetitive polypeptides are seen to migrate with an apparent molecular mass from 2.2 (HC4-10) to 3.8 (HC6-16) times higher than predicted by comparison to the standards. These observations are discussed in section 4.1.

components: an ion source and a mass analyzer. The challenge for accurate mass analysis is the creation and introduction into the gas phase of intact molecular ions from solid protein samples. Some of the methods of ionization that have been used successfully include fast atom bombardment, field and plasma desorption, secondary ion generation and electrospray (for a review see Harrison and Cotter[3]). All the techniques have some inherent disadvantage, whether excessive fragmentation, low mass range or the need for elaborate, expensive instrumentation. However, Beavis and Chait[4], expanding early work of Karas and Hillenkamp[5], have pioneered the application of matrix-assisted laser desorption/ionization, which, when coupled with a relatively inexpensive time-of-flight (TOF) mass analyzer, enables the rapid and highly accurate mass analysis of large proteins (as high as 200 kDa). This analysis, in turn, gives previously inaccessible insight into the chemical structure of biopolymers (for reviews see references 2, 6-8).

The work presented in this chapter was motivated by routine analysis of HC4-14 using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (hereafter referred to as mass spectrometry). The analysis emphasized the inaccuracy and low sensitivity of protein mass analysis by SDS-PAGE. In particular, accurate mass analysis prompted the discovery of two sequence errors in the DNA code for the protein, and detailed analysis of low molecular mass fragments observed in the mass spectrum enabled the determination of a portion of the protein sequence, highlighting the utility of mass spectrometry in the determination of protein molecular mass, purity and sequence.

4.2. Description of the Technique

The experimental set-up for the mass spectrometer is shown in Figure 4.2.[2, 9] In the technique, intact ions are generated by pulsing a focused laser of appropriate wavelength on the protein sample imbedded in a large excess of a matrix material. The laser is tuned to the absorption frequency of the matrix, so the majority of the energy is absorbed by the matrix, not the analyte. The laser "shot" simultaneously triggers one desorption

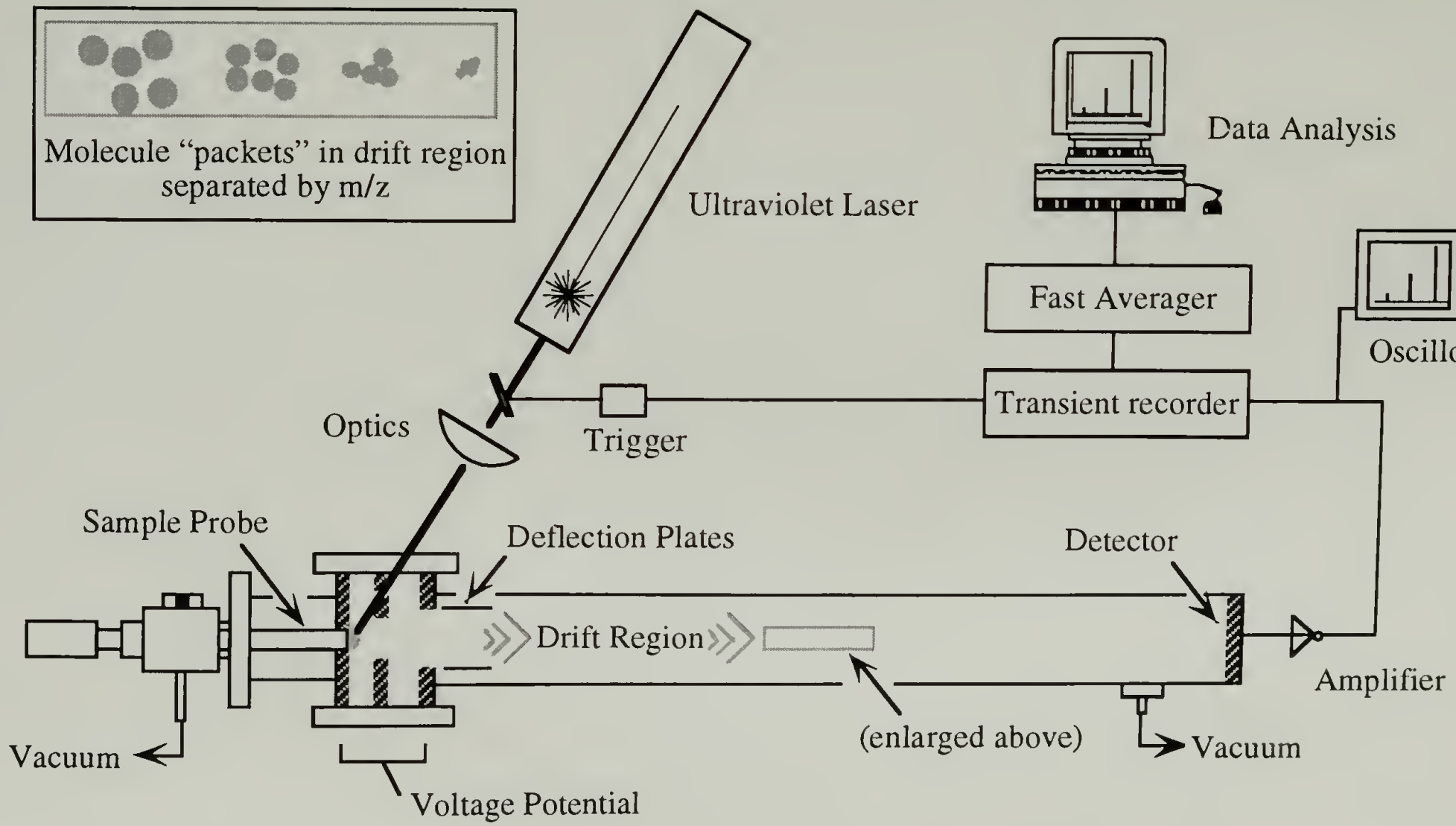


Figure 4.2. Schematic of the matrix-assisted laser-desorption/ionization mass spectrometer developed by Beavis and Chait.[4] The details of the method are described in section 4.2. Inset is a representation of the molecular ion fractions separated by *m/z* in the field-free drift region of the spectrometer.

event and signals the recorder. The matrix material vaporizes, carrying the intact protein into the gas phase, and a charge is transferred to the intact chains. The gaseous ions are all accelerated to a velocity proportional to $(m/z)^{-1/2}$ by a fixed potential difference and as they drift down an evacuated tube, separate into discrete mass fractions. The arrival of the separated ion fractions at the end of the tube is detected by a transient recorder which produces a signal upon impact of each ion group. The records of the drift times of ions created during a number of desorption events are summed and the resulting histogram displayed as a TOF spectrum. The difference between the common start time for all ions and the arrival times for each individual ion group is proportional to $(m/z)^{+1/2}$, this relation is used to convert the TOF spectrum to a mass spectrum with simple data manipulation.

This technique allows the rapid analysis of protein samples as small as 1 pmol containing relatively high concentrations of non-proteinacious material. More importantly, the matrix induced ionization results in the production of only intact ions, so the masses are readily accessible with no requirement for elaborate interpretation of the mass spectrum.

The matrix material is the important parameter in the analysis. While several have been employed, the most successful analyses have been conducted using 3,5,-dimethoxy-4-hydroxy-*trans*-cinnamic acid (sinapinic acid). The absorption frequency of sinapinic acid is in the range of the 355 nm light of neodymium-yttrium aluminum garnet (Nd-YAG) lasers and that of relatively inexpensive nitrogen lasers (337 nm). In theory, a solution containing the protein sample mixed with the matrix in large excess can be deposited on a substrate to form (upon solvent evaporation) a glass with well separated protein chains homogenously dispersed in the matrix. In practice, however, insolubility or aggregation of the protein in the matrix often prevents vaporization and thus sample analysis.

All spectra were recorded at the Laboratory for Mass Spectrometry and Gas Phase Ion Chemistry by Dr. Ronald Beavis, Dr. Rong Wong or Dr. Brian Chait. In a typical experiment, a small amount (1-10 pmol) of the polypeptide analyte was mixed with a ~10^4-fold molar excess of sinapinic acid in an aqueous 30% acetonitrile solution

containing 0.1% trifluoroacetic acid. The mixture was deposited on a metal substrate and placed inside the evacuated sample chamber of the mass spectrometer. The frequency tripled output of a Lumonics HY400 Nd-YAG laser was focused in a 10 nsec pulse onto the sample through a fused silica window. The product ions formed by each laser shot were accelerated by a 30 kV static electric potential into a 2 m evacuated tube and detected using a Lecroy TR8828D transient recorder. Data manipulation were performed at a VAX workstation using software developed by Dr. Ronald Beavis.[10]

All mass values were calculated using an algorithm that first calculated the molecular formula of the protein using the formulae of neutral residues, added two hydrogens and an oxygen, then calculated the mass using isotopically averaged atomic mass values. This was done for ease of calculation and is equivalent to summing the isotopically averaged values for each residue (Appendix A) and adding the mass of water.

4.3. Mass Spectral Analysis of HC4-14

4.3.1. Overview of Full Spectrum

The mass spectrum for HC4-14 is shown in Figure 4.3. Ribonuclease A was added as an internal calibrant. A molecular ion $(M+H)^+$ peak is observed with a *m/z* value of 17,264±2 (all experimental *m/z* values are for protonated chains and are assumed to be measured with an error of ±0.01%). The doubly ionized peak $(M+2H)^{2+}$ = 8,633 is also visible. All of the prominent peaks have an additional satellite peak corresponding to a routinely observed, photo-chemically generated matrix-protein adduct $(M+206)^+$.

A large number of additional low molecular mass fragments can be seen in the spectrum that can be divided into two groups: a series of low intensity peaks in the 8,000-15,000 *m/z* range apparently related in some periodic fashion, and a cluster of higher intensity peaks with *m/z* of 5,000-8,000, roughly divided into three groups, also with an apparent periodic nature.

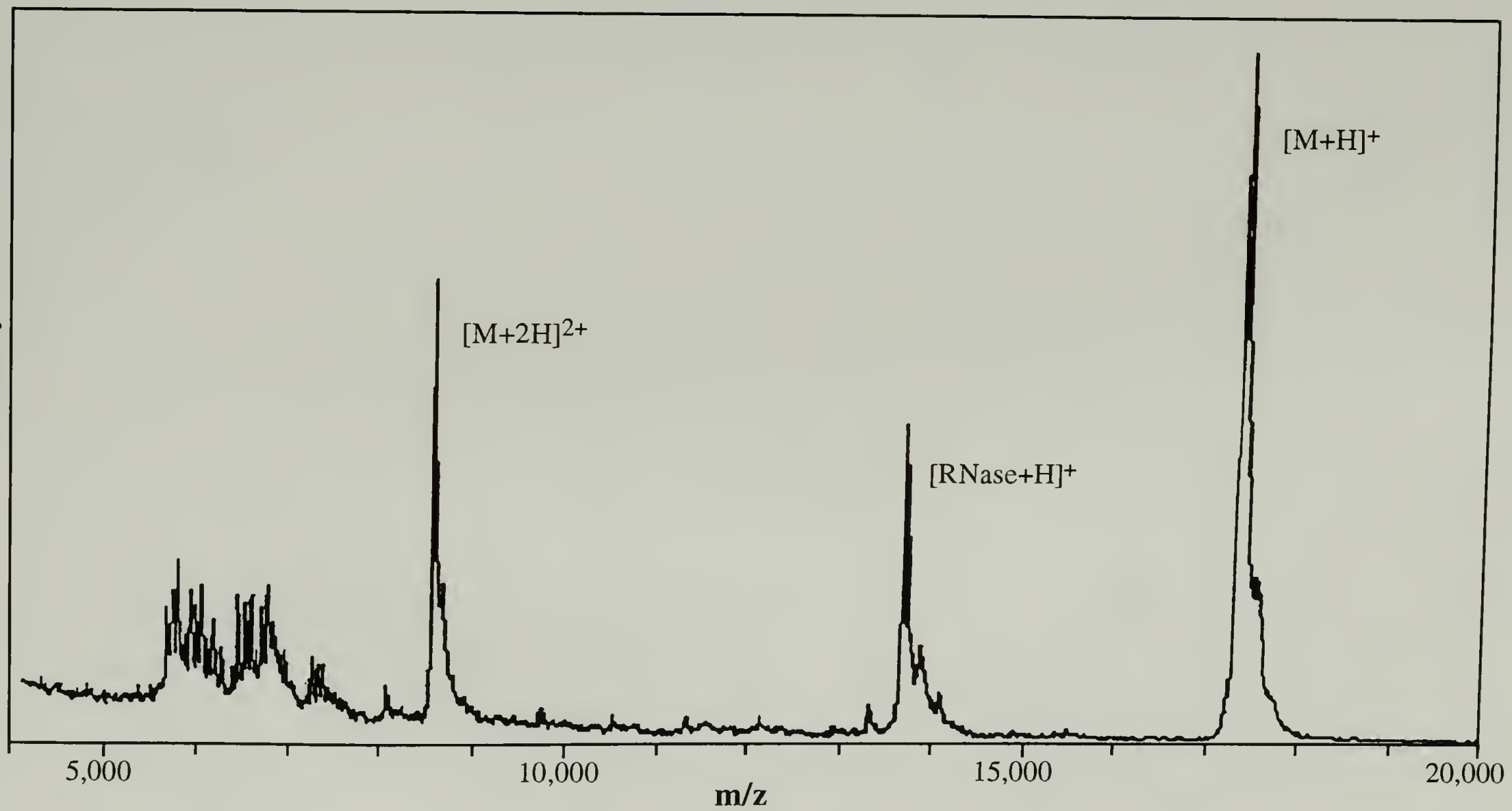


Figure 4.3. Full mass spectrum of HC4-14. RNase was added as an internal calibrant. The molecular ion $[M+H]^+$ with an *m/z* value of 17,264±2 (calcd 17,207) and the doubly ionized molecular ion $[M+2H]^{2+}$ are prominent. A cluster of ions are seen in the *m/z* 5,000-8,000 region, as well as a series of small peaks from *m/z* 8,000-14,000. The analysis of these peaks are presented in section 4.3. Reprinted from Beavis and coworkers[11], with the permission of the American Chemical Society.

Since this technique results in the production of intact molecular ions, with few laser induced chain cleavage fragments, the abundance of other peaks observed in the spectrum were considered discrete polypeptide fragments in the sample, not artifacts of the analytical technique. In the following analysis, it was assumed that the peaks were the molecular ions of fragments that were present due to some biological process, perhaps the action of proteolytic enzymes. The calculated *m/z* values of the fragments were taken as those consistent with polypeptides comprised of unmodified, neutral residues and intact N- and C- termini. These assumptions were used in an attempt to resolve the total mass discrepancy and identify fragments corresponding to the observed signals.

4.3.2. Molecular Mass Analysis

4.3.2.1. Implications of Molecular Mass Discrepancy

The measured *m/z* of 17,264±2 for the polypeptide was significantly larger than the expected value of 17,208 for **4.1**. Although this 0.3% error is far less than the error associated with mass analysis using SDS-PAGE, given that the demonstrated accuracy of the technique is in fact 0.01%[2], this discrepancy suggested the presence of a sequence defect or other modification of the protein chain.

H-ASMTGGQQMGRDPMFKYSRDPMG$[(AG)_4PEG]_{14}$ **4.1**

ARMHIRPGRYQLDPAANKARKEAELAAATAEQ-OH

One possible source of the discrepancy (and perhaps the most straightforward to investigate) was an alteration of the DNA code leading to an amino acid substitution in the expressed protein. For instance, a single C to A mutation in the second position of a GCA or GCG codon for alanine would cause a substitution of a glutamic acid residue, the molecular mass difference of these two residues accounting for the apparent discrepancy.

A single nucleotide substitution represents the simplest case—multiple substitutions could also account for the mass difference. In order to determine if DNA alterations give rise to the mass discrepancy, the sequencing of the coding fragment in pET3-b was undertaken.

4.3.2.2. Sequencing of the pET4-14 DNA Insert

The entire DNA sequence as expressed in pET4-14, including that encoding the N- and C-terminal fusion sequences, was isolated by digestion with *Xba* I and *Eco*R V. This yields a fragment beginning at the *Xba* I site (T/CTAGA) 39 bp upstream of the *Nde* I site in the gene 10 promoter region (see Figure 2.2), spanning the start and stop signals and ending at the *Eco*R V site (GAT/ATC) at the end of the ϕ10 termination sequence. The length of the DNA fragment required the use of single stranded sequencing protocols, so the fragment was cloned into the *Sma* I (CCC/GGG) and *Xba* I sites of the replicative form (double stranded) of M13mp18 and M13mp19 filamentous phage and single stranded phage particles generated using standard protocols.[12, 13]

pET4-14 (2 µg) was digested with restriction enzymes *Eco*R V (N. E. Biolabs, 40 U) and *Xba* I (N. E. Biolabs, 40 U) in 50 µL of 1X NEBuffer #2 with bovine serum albumin added to 0.1 mg/ml. After incubation at 37°C for six hours, the digest was loaded onto a 0.8% low melting point agarose gel and the DNA fragments separated by electrophoresis at 40V constant voltage for 4 hours. The gel slice containing the 732 bp *Xba* I/*Eco*R V fragment (XE insert) was excised and incubated at 65°C for 10 minutes. The melted solution (~150 µL) was extracted sequentially with equal volumes of phenol, 1:1 phenol/chloroform and chloroform. The purified DNA was precipitated by addition of 15 µL of 3 M sodium acetate (pH 4.9) and 300 µL 100% ethanol. The precipitated DNA was washed, dried *in vacuo* and resuspended in 20 µL TE. Gel analysis indicated a solution concentration of the XE insert of ~50 ng/µL.

M13mp18 RF and M13mp19 RF DNA (4 µg, New England Biolabs) were digested with *Sma* I (N. E. Biolabs, 40 U) in 50 µL 1X NEBuffer #4 at 25°C for four

hours. The digestion mix was diluted to 100 µL, extracted with an equal volume of 1:1 phenol/chloroform and chloroform, precipitated as described above and resuspended in 20 µL water. Agarose gel analysis showed that digestion was complete. The DNA was then digested with *Xba* I under the conditions described above. The linear plasmids were purified from LMP agarose and recovered in approximate concentrations of 100 ng/µL.

Approximately 5 ng of the 732 bp XE insert was mixed with 5 ng of each of the purified plasmids in 10 µL of 80 mM Tris-HCl pH 7.5, 10 mM magnesium chloride, 1 mM ATP and 20 mM dithiothreitol. This represented an insert/vector molar ratio of ~10:1. T4 DNA ligase (400 U) was added and the mixture incubated at 15°C overnight.

The standard protocol[13] for the isolation of recombinant single stranded templates from M13 was followed explicitly and is briefly outlined here. Each ligation mix (5 µL) was used to transform fresh competent *E. coli* strain JM101 cells. The transformation mix was mixed with an aliquot of saturated JM101 culture, plated in top agar containing X-gal and IPTG and incubated at 37°C. Infected cells containing recombinant phage plasmids were identified as white plaques against a lawn of host cells. Phage picked from selected plaques were used to infect a fresh culture of JM101 cells and grown at 37°C for 4-6 hours. The cells were recovered by centrifugation and the double stranded replicative form of the vector isolated by a standard plasmid mini-prep procedure. The presence of the insert was verified by restriction digestion of the isolated RF vector with *Hin*d III and *Eco*R I and subsequent agarose gel analysis of the digestion fragments. The single stranded phage from selected samples were precipitated from the supernatent by treatment with poly(ethylene glycol) and sodium chloride. The phage pellet was resuspended and extracted with phenol to remove associated proteins. The single stranded template was precipitated, recovered and resuspended in TE. Recovery was checked on an agarose gel by comparison of the migration of the recombinant DNA with a standard template containing no insert.

The ^{T7}Sequencing™ protocol was used to sequence the templates. The 17 bp M13 sequencing primer (5'-GTAAAACGACGGCCAGT-3') and [α–^{35}S] dATP were used.

Due to the high G/C content of the sequence, band compressions in the sequencing gel were anticipated and dGTP was replaced with the 7-deaza-dGTP and with 7-deaza-dITP in sequencing reactions. The Deaza ^{T7}Sequencing™ Mixes were used and the supplied protocol followed exactly. Sequencing fragments were separated on a 0.2 mm thick, 6% polyacrylamide gel containing 50% urea by electrophoresis at 60 W constant power. The fragments were analyzed in two loadings: a second aliquot of sequence reactions was loaded after the xylene cyanol marker of the first loading had migrated 3/4 the length of the gel. The gel was recovered when the bromophenol blue marker of the second loading reached the bottom of the gel. The gel was dried at 80°C *in vacuo* and the sequence examined by autoradiography.

4.3.2.3. Altered Sequence of HC4-14

The sequence was confirmed by comparison of the sequence of 3 samples of the DNA fragment sequenced in both directions (a total of 6 templates); for varying reasons, no one reaction gave the entire sequence unambiguously. However, examination of the signals on the autoradiograph revealed the presence of two 2 C to T mutations in codons 96 (GCC to GTC) and 101 (GCA to GTA) relative to the start codon in pET4-14. The sequence alterations were prominent in all sequences on the autoradiograph. The experimentally determined sequence is given in Figure 4.4.

These DNA mutations result in the substitution of valines (encoded by GTN) for two alanines (GCN) in the seventh and eighth repeats of the expressed protein. The protein sequence **4.2**, predicted from these data, has a calculated molecular mass of 17,263. *This is exactly the value found experimentally —17,263 ± 2.*

H-ASMTGGQQMGRDPMFKYSRDPM **4.2**

$[(GA)_4GPE]_6[(GA)_3(GV)GPE][(GV)(GA)_3GPE][(GA)_4GPE]_6$

GARMHIRPGRYQLDPAANKARKEAELAAATAEQ-OH

Start Ala Ser Met Thr Gly Gly Gln Gln Met Gly Arg Asp Pro Met Phe Lys Tyr Ser Arg Asp Pro Met Gly

ATG GCT AGC ATG ACT GGT GGA CAG CAA ATG GGT CGG GAT CCT ATG TTT AAA TAT TCT CGC GAT CCG ATG GGT

Ala Gly Ala Gly Ala Gly Ala Gly Pro Glu Gly

1 GCC GGC GCT GGT GCT GGG GCC GGT CCG GAA GGT 2 GCC GGC GCT GGT GCT GGG GCC GGT CCG GAA GGT ①

3 GCC GGC GCT GGT GCT GGG GCC GGT CCG GAA GGT 4 GCC GGC GCT GGT GCT GGG GCC GGT CCG GAA GGT ②

5 GCC GGC GCT GGT GCT GGG GCC GGT CCG GAA GGT 6 GCC GGC GCT GGT GCT GGG GCC GGT CCG GAA GGT ③

Ala Gly Ala Gly Ala Gly Val Gly Pro Glu Gly

7 GCC GGC GCT GGT GCT GGG GTC GGT CCG GAA GGT 8 GTC GGC GCT GGT GCT GGG GCC GGT CCG GAA GGT ④

9 GCC GGC GCT GGT GCT GGG GCC GGT CCG GAA GGT 10 GCC GGC GCT GGT GCT GGG GCC GGT CCG GAA GGT ⑤

11 GCC GGC GCT GGT GCT GGG GCC GGT CCG GAA GGT 12 GCC GGC GCT GGT GCT GGG GCC GGT CCG GAA GGT ⑥

13 GCC GGC GCT GGT GCT GGG GCC GGT CCG GAA GGT 14 GCC GGC GCT GGT GCT GGG GCC GGT CCG GAA GGT ⑦

Ala Arg Met His Ile Arg Pro Gly Arg Tyr Gln Leu Asp Pro Ala Ala Gln Lys Ala Arg Lys Glu Ala Glu

GCC CGC ATG CAC ATC CGC CCG GGT CGA TAT CAG CTG GAT CCG GCT GCT AAC AAA GCC CGA AAG GAA GCT GAG

Leu Ala Ala Ala Thr Ala Glu Gln Stop

TTG GCT GCT GCC ACC GCT GAG CAA TAA

① DNA Monomer Number ——— DNA Sequence From pET3-b

1 Protein Repeat Sequence Number ——— DNA Sequence From p937.51

Figure 4.4. Experimentally determined sequence of the polypeptide coding region of pET4-14. The sequence encoding the fusion protein is presented. The sequence contributed by the p937.51 and pET3-b plasmids (encoding the *N*- and *C*- terminal fusions) is indicated, as is each of the individual DNA monomers comprising the repetitive sequence and the two altered codons (resulting in alanine to valine substitutions, as indicated).

4.3.3. Analysis of Protein Fragments

4.3.3.1. Analysis of Peaks in 8,000-15,000 *m/z* Range

A series of peaks were observed in the 8,000-15,000 *m/z* range of the spectrum shown in Figure 4.3. Initially it was observed that the *m/z* values for some of these peaks (Table 4.1) differed from each other on average by the mass of the undecapeptide repeat unit of the protein [$(GlyAla)_4GlyProGlu$, *m/z* = 797], but the measured *m/z* values of the peaks could not be assigned to known fragments. However, it remained possible to interpret this periodic peak pattern as arising from fragments generated by some endoproteinase that cleaves specifically at a residue sequence in the repeating portion of the copolypeptide.

Table 4.1. Summary of fragments in the 8,000-15,000 *m/z* range

Exptl *m/z*	8,145	8,969	9,767	10,557	11,357	12,148	12,942
Difference*	824	798	790	800	791	794	
Repeat #	8	9	10	11	12	13	
Calcd *m/z*	8,148	8,972	9,768	10,564	11,360	12,156	12,952

*The calculated difference between the observed signals on the spectrum with the indicated exptl *m/z*. The calculated *m/z* value of a $(GlyAla)_4GlyProGlu$ repeat unit is 797.

After the DNA sequence analysis, all of the peaks could be rationalized on the basis of the valine for alanine substitutions represented in sequence **4.2**. The peak at *m/z* 8,145 is thought to arise from fragment **4.3** which has a calculated mass of 8,148 Da. The fragment seen at *m/z* 8,146 arises from the larger sequence at *m/z* 8,969 by removal of the eighth repeat unit [$(GV)(GA)_3GPE$] with a calculated mass of 824. The remaining peaks are assigned by virtue of the 796 mass of the [$(GA)_4GPE$] repeat unit. The assignments are summarized in Table 4.1 and shown in Figure 4.5.

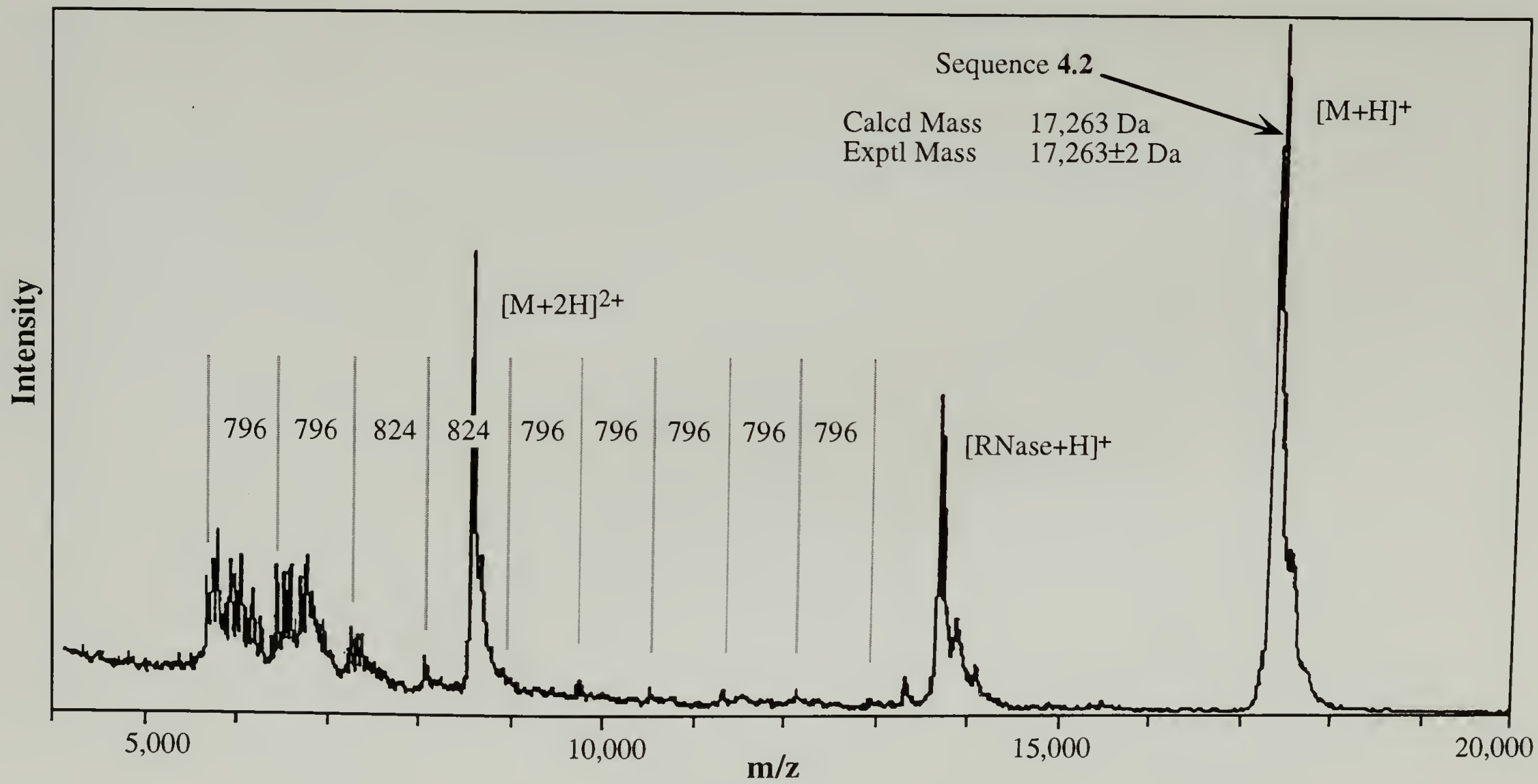


Figure 4.5. Reinterpreted full mass spectrum of HC4-14. The discovery of two sequence errors in the DNA code for HC4-14 allows the interpretation of the spectrum in terms of altered sequence **4.2** with a calculated molecular mass of 17,263, exactly the value found experimentally (17,263±2). Protein fragments arising from sequential cleavage between the glutamic acid and glycine residues in the undecapeptide repeat units are represented on the spectrum as differences between mass values of the observed fragments, progressing from left to right starting with the peak at 5,730, correponding to sequence **4.4** (calcd 5,733), through each of the remaining repeats in the sequence. The altered seventh and eighth repeat units [(GV)$(GA)_3$GPE], with a calculated mass of 824, are visible. The remaining peaks are assigned by virtue of the 796 mass of the unaltered [$(GA)_4$GPE] repeat unit. The assignments are summarized in Table 4.1. Adapted from Beavis and coworkers[11], with the permission of the American Chemical Society.

H-ASMTGGQQMGRDPMFKYSRDPM
[(GA)$_4$GPE]$_6$[(GA)$_3$(GV)GPE]G-OH **4.3**

4.3.3.2. Analysis of Peaks in 5,000-8,000 *m/z* Range

A cluster of peaks is visible in the 5,000-8,000 *m/z* region of the spectrum shown in Figure 4.3. Analysis of these signals, summarized in Table 4.2 and displayed in expanded form in Figure 4.6 was based on the clearly visible peak at *m/z* 5,730 corresponding to fragment **4.4** which is composed of the intact N-terminal sequence of the fusion protein followed by four copies of the periodic undecapeptide. The calculated mass of the molecular ion derived from **4.4** is 5,733 Da, a value in good agreement with the observed *m/z* value.

H–ASMTGGQQMGRDPMFKYSRDPMG[AGAGAGAGPEG]$_4$–OH **4.4**

More striking, however, is the fact that each succeeding signal can be rationalized by addition of a single amino acid residue, proceeding in the N–to–C–terminal direction along sequence **4.2**, *and one can read the periodic sequence directly from the mass spectrum, including one of the substituted valines.* From this analysis, it is apparent that the analytical method for rapid, direct protein sequencing is in place (however the controlled generation of protein fragments suitable for sequencing remains a major obstacle). This analysis also highlights the level of detail on the complex mass state of the polypeptide provided by mass spectrometry; such details are unattainable with other analytical methods.

4.3.3.3. Removal of Low Molecular Fragments

In order to confirm that the low molecular mass protein fragments did not arise during the mass analysis, 50 mg of the HC4-14 sample S1 with the spectrum presented in Figure 4.3 was dissolved in 50 mL of 0.1% ammonium bicarbonate, the solution was

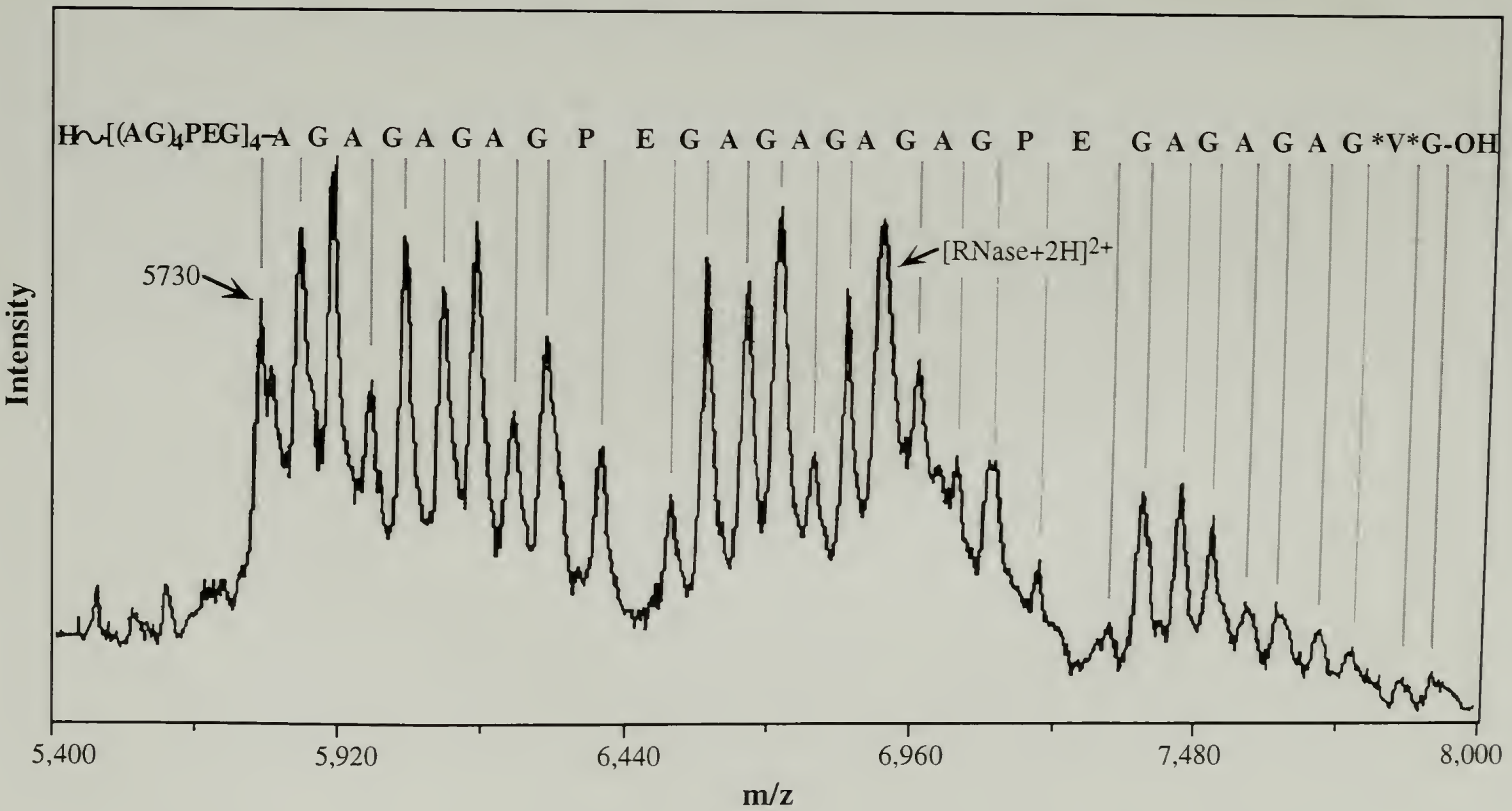


Figure 4.6. Mass spectrum of HC4-14 expanded in the 5,000-8,000 *m/z* region. The peaks progress starting with indicated N-terminal fragment **4.4** (calcd *m/z* 5,733, found 5,730) with the sequential addition of amino acid residues through three repeats of the target repetitive sequence. The substituted valine in the seventh repeat is visible. The analysis of the peaks is discussed in section 4.3.3.2. Reprinted from Beavis and coworkers[11], with the permission of the American Chemical Society.

Table 4.2. Summary of fragments in the 5,000-8,000 *m/z* range.

m/z Exptl	Assigned Residue[a]	m/z Calcd[b]	m/z Exptl	Assigned Residue	m/z Calcd
5730		5732	6527		6528
	70 *Ala*			69 *Ala*	
5800		5803	6596		6599
	57 *Gly*			58 *Gly*	
5857		5860	6654		6656
	71 *Ala*			71 *Ala*	
5928		5931	6725		6727
	58 *Gly*			57 *Gly*	
5986		5988	6782		6784
	71 *Ala*			69 *Ala*	
6057		6059	6851		6855
	59 *Gly*			57 *Gly*	
6116		6116	6908		6912
	68 *Ala*			75 *Ala*	
6184		6187	6978		6983
	59 *Gly*			60 *Gly*	
6243		6244	7043		7040
	97 *Pro*			95 *Pro*	
6340		6342	7138		7137
	129 *Glu*			128 *Glu*	
6469[c]		6471	7266		7267
	57 *Gly*			56 *Gly*	
6527		6528	7322		7323
				69 *Ala*	
			7391		7395

[a]Residue mass values were calculated as differences between masses of adjacent peaks: indicated by shaded lines. All residues were assumed to be neutral. The calculated residue mass values used: glycine, 57; alanine, 71; proline, 97; glutamic acid, 129.

[b]Calculated m/z values were obtained by sequentially adding residue masses commencing with the N-terminal fragment (*m/z* = 5732):
H-ASMTGGQQMGRDPMFKYSRDPM(GAGAGAGAGPE)$_4$G-OH

[c]Underlined values are interpolated values of unlabeled *m/z* peaks with appropriate residue masses determined by best fit to differences between labeled peaks.

[d]The peak corresponding to the fragment with interpolated *m/z* = 6851 is believed to be obscured by the observed ribonuclease $(M+2H)^+$ fragment at *m/z* = 6841.

placed in a sealed dialysis bag (Spectrum Spectra/Por® 4, MWCO 12-14,000) and dialyzed against water for two days. The protein (32 mg) was recovered by lyophilization and the mass spectrum recorded.

The spectrum (Figure 4.7) reveals only the singly, doubly and triply ionized peaks of intact chains of **4.2**. The absence of low molecular mass signals indicates that the fragments are inherent in the sample and do not arise during ionization. This analysis also demonstrates that with a simple modification of the purification protocol, completely monodisperse protein samples can be produced.

4.3.3.4. SDS-PAGE Analysis of HC4-14 samples

In order to demonstrate the quality of the mass analysis, the samples were reanalyzed by SDS-PAGE. Cell lysates collected at various times during the expression of the pET4-14 (as described in section 3.3.6), along with the undialyzed HC4-14 sample S1 and dialyzed sample S2 were analyzed on a 25 cm discontinuous SDS 15% polyacrylamide gel using the methods described in section 2.4. The electrophoresis conditions were controlled to allow the resolution of the 3,000 Da molecular mass marker. The gel was stained with Coomassie Brilliant Blue G-250 solution to visualize the proteins and destained in 30% methanol containing 10% acetic acid.

It is evident from Figure 4.8 that the degradation products prominent in the mass spectrum shown in Figure 4.3 are not visible by SDS-PAGE—the presence or absence of the low molecular mass fragments observed in the mass analysis is not discernible in the migration patterns of either of the purified samples.

4.3.3.5. HPLC Fractionation of Low Mass Fragments

Due to the nature of the sample preparation and data collection in the mass spectral analysis, no quantitative information is available as to the amount of the low mass fragments present relative to the parent protein. It is possible that the fragments are present

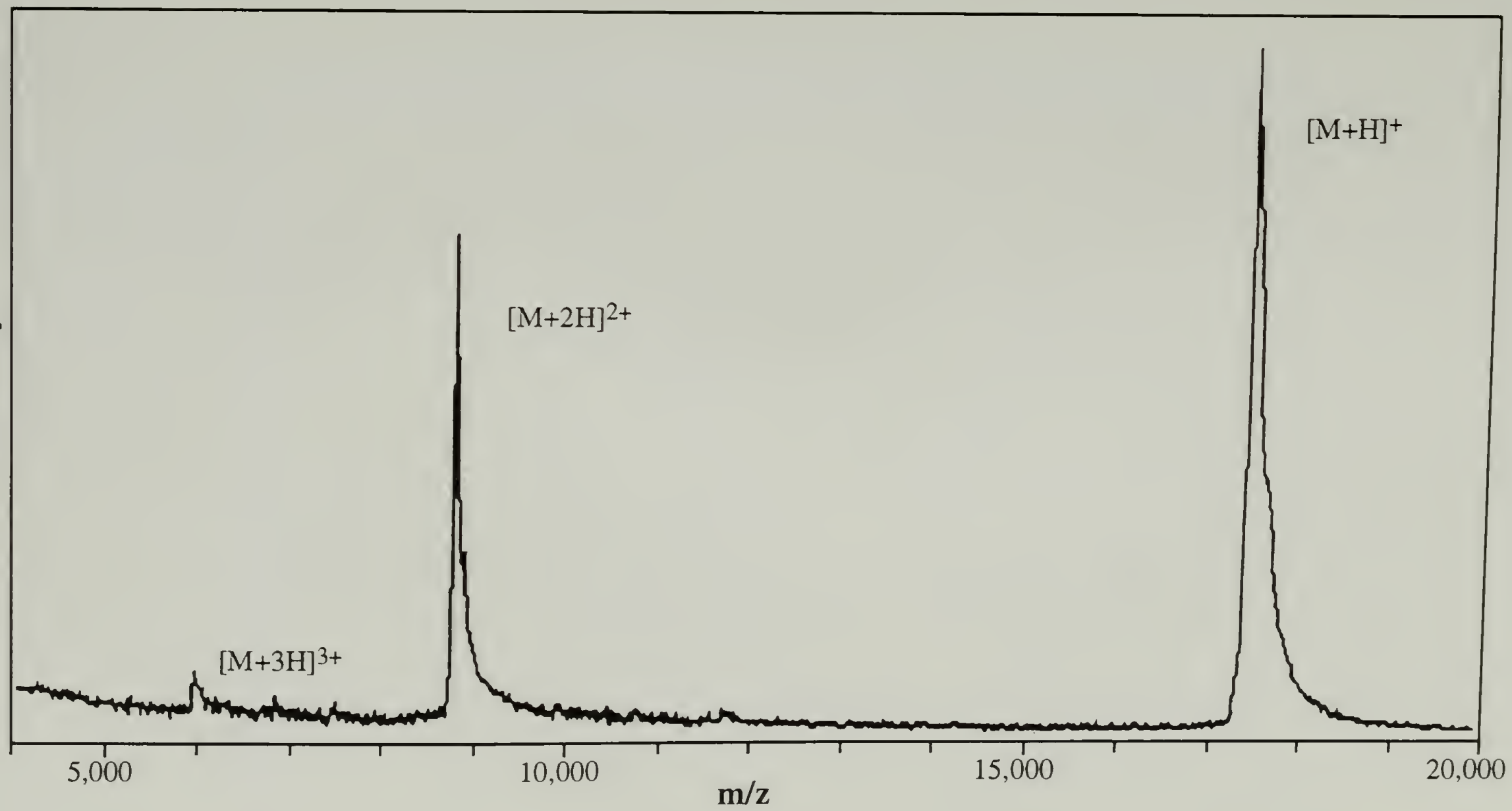


Figure 4.7. Mass spectrum of HC4-14 after removal low mass fragments. The HC4-14 protein sample (S1) with the mass spectrum in Figure 4.3 was dialyzed against water in a MWCO 12-14,000 dialysis bag. Mass analysis of this sample (S2) shows only the singly, doubly, and triply ionized molecular ions. The analysis is discussed in section 4.3.3.3. Reprinted from Beavis and coworkers[11], with the permission of the American Chemical Society.

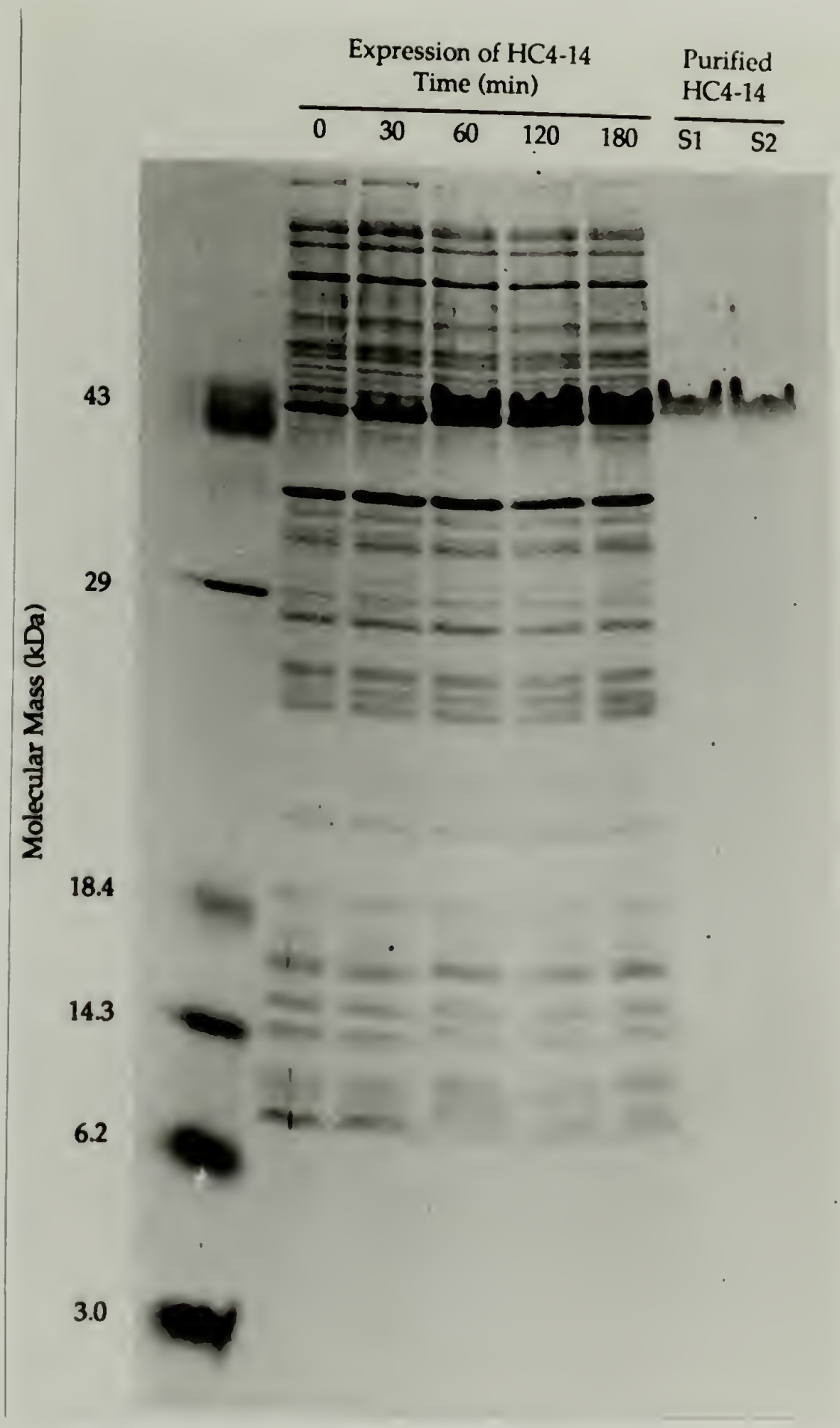


Figure 4.8. SDS-PAGE analysis of HC4-14 samples used in the mass analysis. BL21(DE3)pLysS/pET4-14 *E. coli* cell lysates were collected at various indicated times after addition of IPTG and analyzed, along with the undialyzed HC4-14 sample S1 and dialyzed sample S2, on a 25 cm 15% discontinuous SDS polyacrylamide gel. The electrophoresis conditions were controlled to allow the resolution of the 3 kDa molecular mass marker. The proteins were stained with Coomassie Brilliant Blue G-250. The low-mass fragments seen in Figure 4.3 are not apparent on the gel: both dialyzed and undialyzed HC4-14 samples show a single protein band migrating with an apparent mass 2.5 times higher than calculated. The analysis highlights the inadequacies of SDS-PAGE in the analysis of the repetitive polypeptides. The analysis is discussed in section 4.3.3.3.

in trace amount and by some mechanism involving selective solubility or ionization are artificially enhanced in the collection of the TOF data.

In order to assess the relative amount of the contaminating fragments in the HC4-14 sample S1, the protein was analyzed by reversed-phase high performance liquid chromatography (HPLC) on a system consisting of two Beckman Model 114M pumps controlled by a Beckman Model 421 gradient controller. A Polymer Laboratories PLRP-S 150x4.6 mm column with 8 μm polystyrene particles of a pore size of 300 Å was used for the separations and eluted compounds were detected at 214 nm using a Kratos Spectroflow Model 783 UV detector. The analog output was converted to a digital signal using an Analog Devices 6B A/D board and the data recorded and manipulated using Labtech Acquire software.

The solvent system consisted of a 95% acetonitrile/5% aqueous solution (eluent A) and an aqueous solution containing 5% acetonitrile (eluent B). Both eluent solutions contained 0.1% trifluoroacetic acid (TFA). Solutions of the proteins dissolved in eluent B were injected in 100 μL sample volumes into an 80% eluent B mobile phase. Two minutes after injection, the solvent was ramped linearly with time to 70% eluent B over the course of 15 min and the absorption of the eluted material recorded.

HPLC analysis of the protein is shown in Figure 4.9. The shape of the peaks in the chromatograph suggested that the low intensity peaks on the HPLC trace correspond to the fragments in the low mass region of the mass spectrum in Figure 4.3. In order to confirm this premise, a series of six eluent fractions (indicated in Figure 4.9) were collected, the solvent evaporated and the residue analyzed by mass spectrometry. The mass spectra of the first four fractions are shown in Figures 4.10 and 4.11. The spectra of fractions 5 and 6 were identical to that of fraction 4. It is clear from analysis of these spectra that the low mass fragments are partially resolved sequentially by reversed-phase HPLC, with fraction 1 containing the majority of the fragments seen in the expanded region in Figure 4.6 and fractions 4, 5, and 6 containing solely the parent protein. The *m/z* values of the

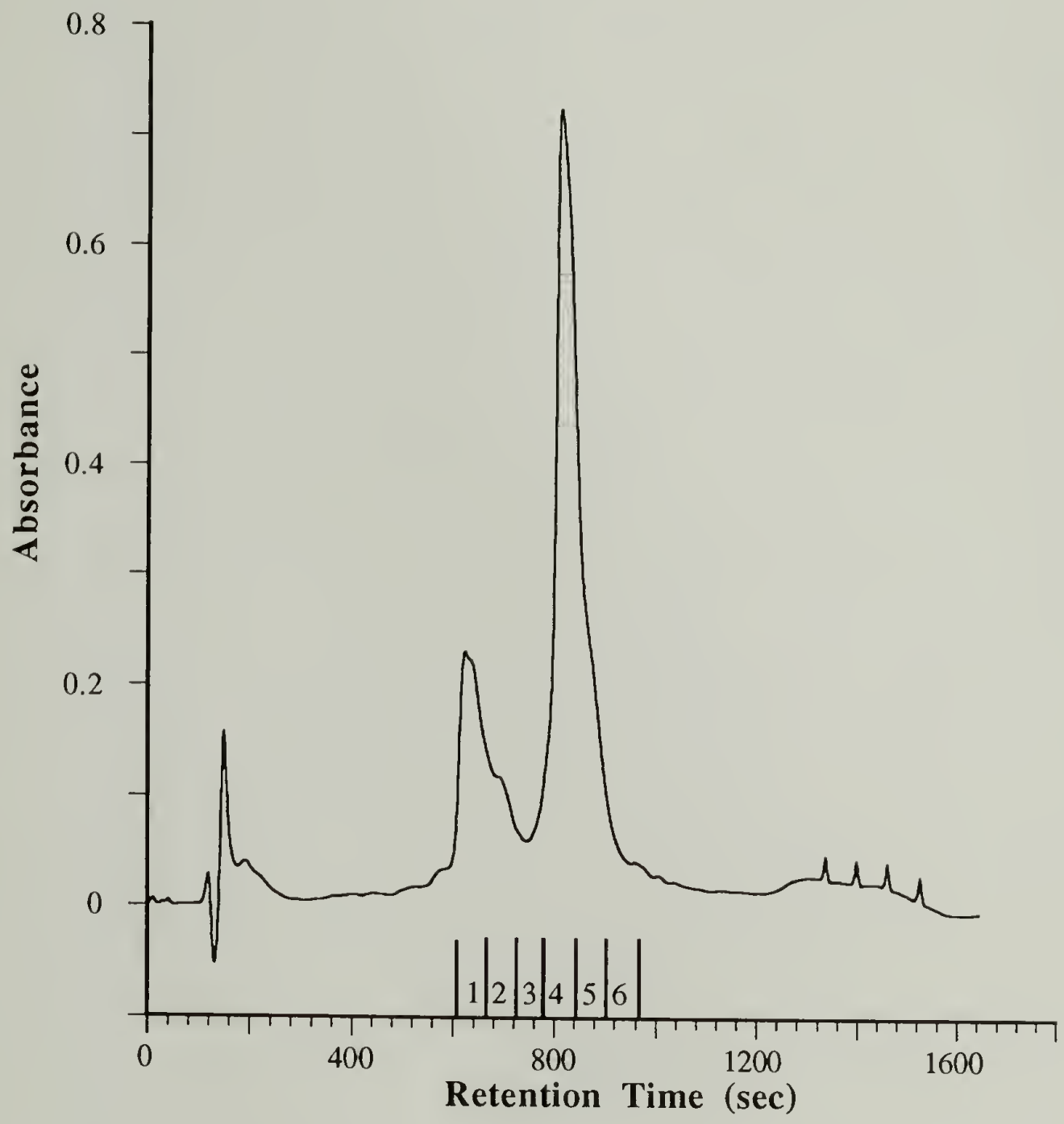


Figure 4.9. Analysis of HC4-14 by high-performance liquid chromatography. The details of the analysis are given in section 4.3.3.5. The shape of the chromatograph was reminiscent of the mass spectrum of HC4-14 (Figure 4.3), with the lower intensity peaks at lower retention time resembling the low mass protein fragments visible in the 5,000-8,000 *m/z* range of the mass spectrum. To test this premise, six eluent fractions were collected at the indicated retention times and prepared for mass analysis (Figures 4.10 and 4.11).

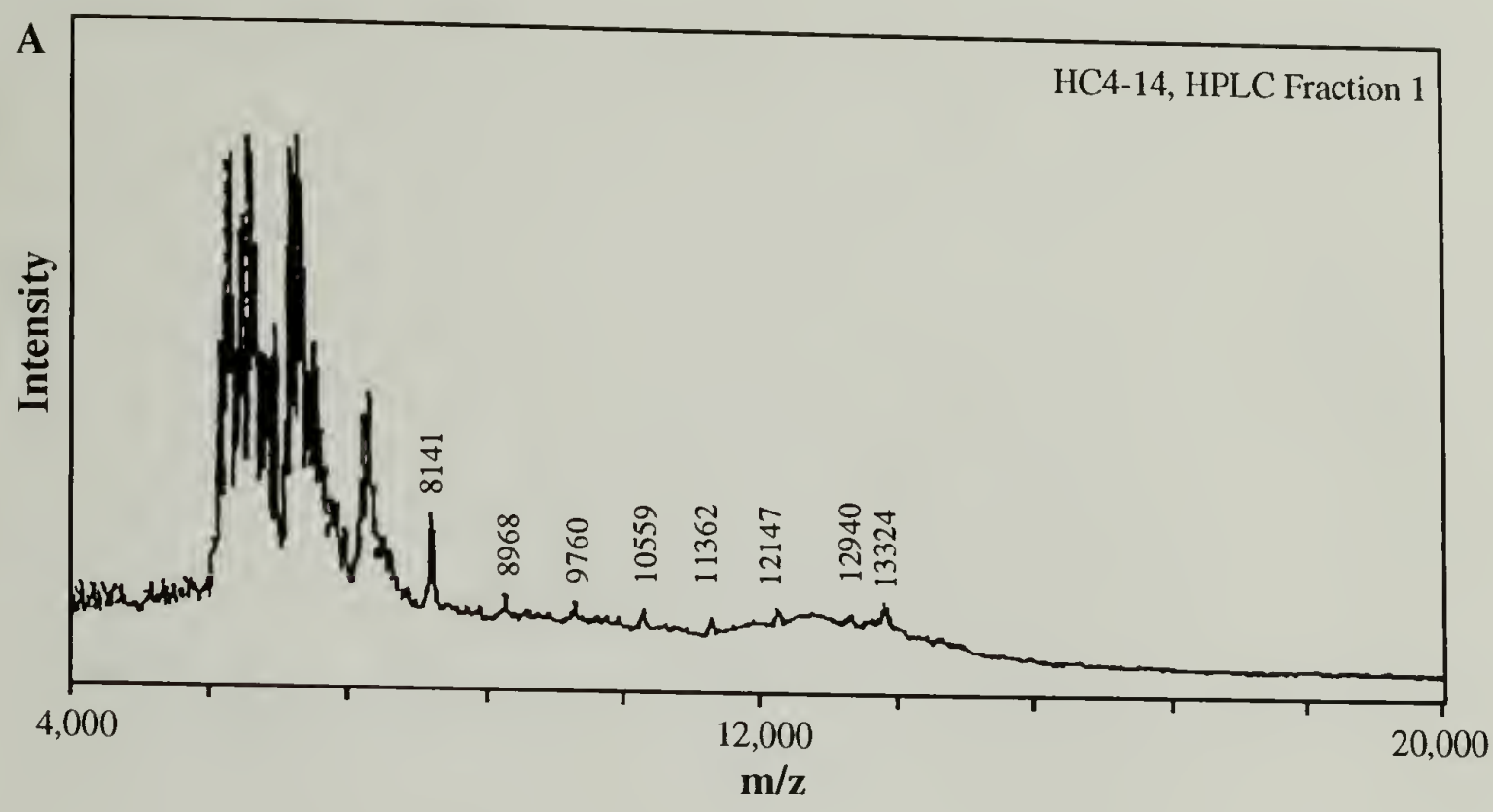
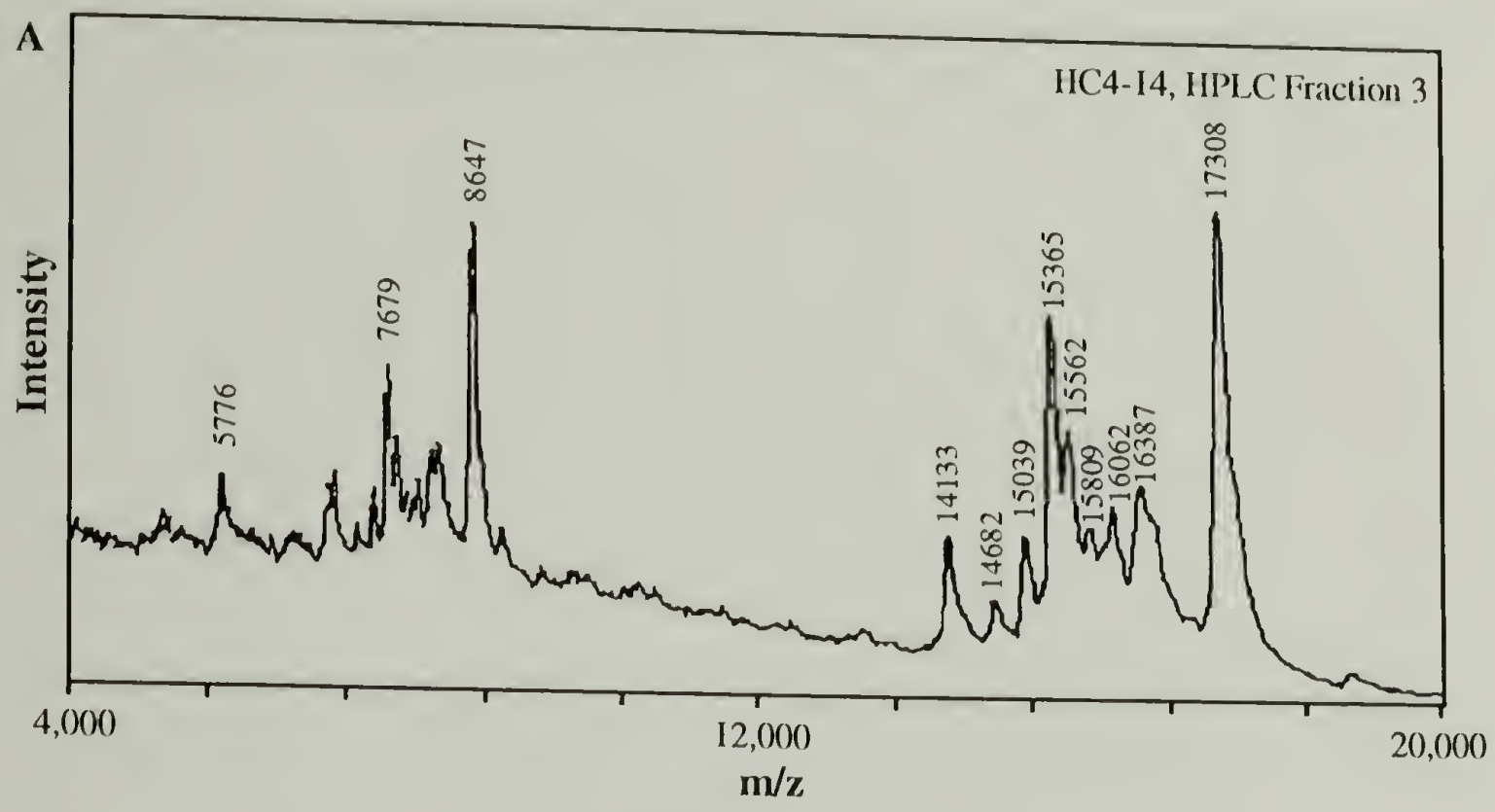


Figure 4.10. Mass analysis of HPLC fractions 1-2. **A**: Fraction 1 was comprised of the low mass HC4-14 proteins seen in the 5,000-8,000 *m/z* region of the mass spectrum (Figure 4.3) and low-intensity signals corresponding to fragments arising from cleavage of the repeat unit (discussed in section 4.3.3.1 and summarized in Table 4.1; **B**: Fraction 2 also contained the low mass fragments. Each of the peak clusters defined by the prominent repetitive signals indicated in Table 4.1 is seen to have the same general shape, suggesting that, like the more intense signals in the 5,000-8,000 *m/z* region (Figure 4.5), the peak clusters arise from fragments with masses rationalized in terms of the amino acid sequence of HC4-14. The higher intensity peaks in the 13,600-16,800 region are unassigned. The analysis is discussed in section 4.3.3.5.



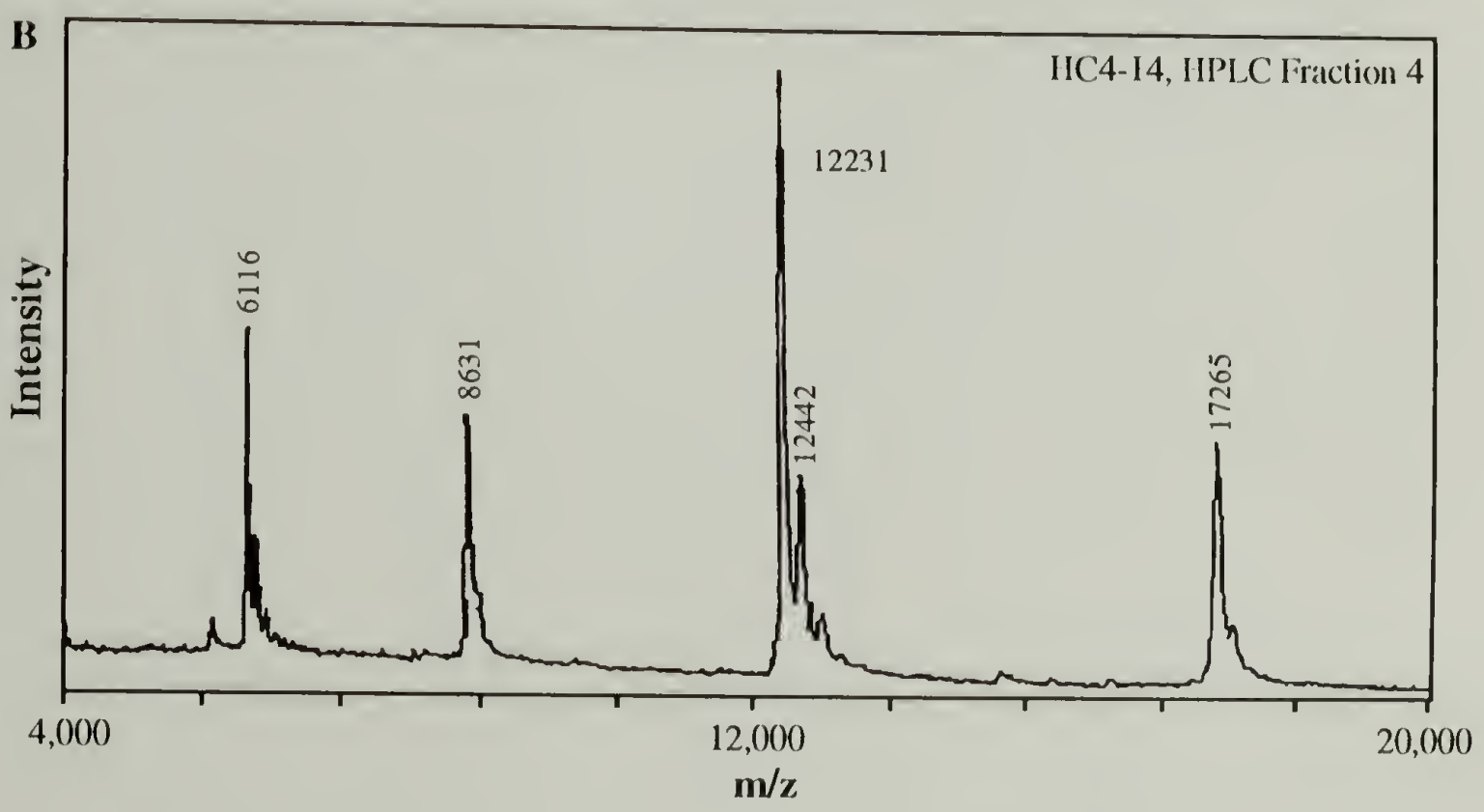


Figure 4.11. Mass analysis of HPLC fractions 3-4. **A**: Fraction 3 is comprised of a number of unidentifiable peaks. The signals at 17,308 and 8,647 are thought to be the singly and doubly ionized molecular ions of HC4-14, the mass error could be a result of the lack of calibration of the spectrum. **B**: Fraction 4 is comprised solely of the singly, doubly and triply ionized HC4-14 molecular ions and the molecular ion of the internal calibrant. The peaks at 6,116, 12,331, and 12,442 are associated with the internal calibrant. Fraction 4, 5, and 6 had identical mass spectra. The analysis is discussed in section 4.3.3.5.

large mass fragments observed in fraction 3 (Figure 4.11A) have not been rigorously assigned to peptide fragments derived from sequence **4.1**.

The relative level of low mass fragment contamination in the sample was estimated to be 33% by integration of the areas under the peaks in the HPLC trace. This value is in good agreement with the 36% weight loss upon removal of the fragments by dialysis.

4.4. Analysis of Other HC4-m Proteins

In order to assess the generality of the biological processes giving rise to the fragments observed in the mass spectrum of HC4-14, chain length variants of HC4-m (m=10, 12, 16, and 20) were expressed in *E. coli*, recovered and subjected to analysis using mass spectrometry and HPLC.

The mass spectra are shown in Figures 4.12 and 4.13. All of the spectra are qualitatively similar to that of HC4-14 discussed in section 4.3.1: each spectrum features multiply charged molecular ion peaks, the periodic fragments resulting from the cleavage of the repeat unit and the smaller *m/z* peaks seen in the expanded spectrum of HC4-14 (Figure 4.6).

The results of the molecular mass analysis for the HC4-m proteins are tabulated in Table 4.3. For three of the proteins, m = 12, 16, and 20, the calculated mass was verified exactly within experimental error. However, the experimental *m/z* value for HC4-10 indicates the same +57 *m/z* discrepancy found for HC4-14. Attempts to sequence the DNA encoding this protein failed: the quality of the sequencing data did not allow unambiguous determination of the sequence. It is interesting to note that the two proteins with molecular mass discrepancies are encoded by DNA multimers comprised of an *odd* number of DNA monomers.

The periodic peak pattern seen in the 5,000-8,000 *m/z* range is also visible in all of the spectra to varying degrees. Closer examination of the expanded region indicates that in all cases the initial peak of the series corresponds to sequence **4.4** and the succeeding

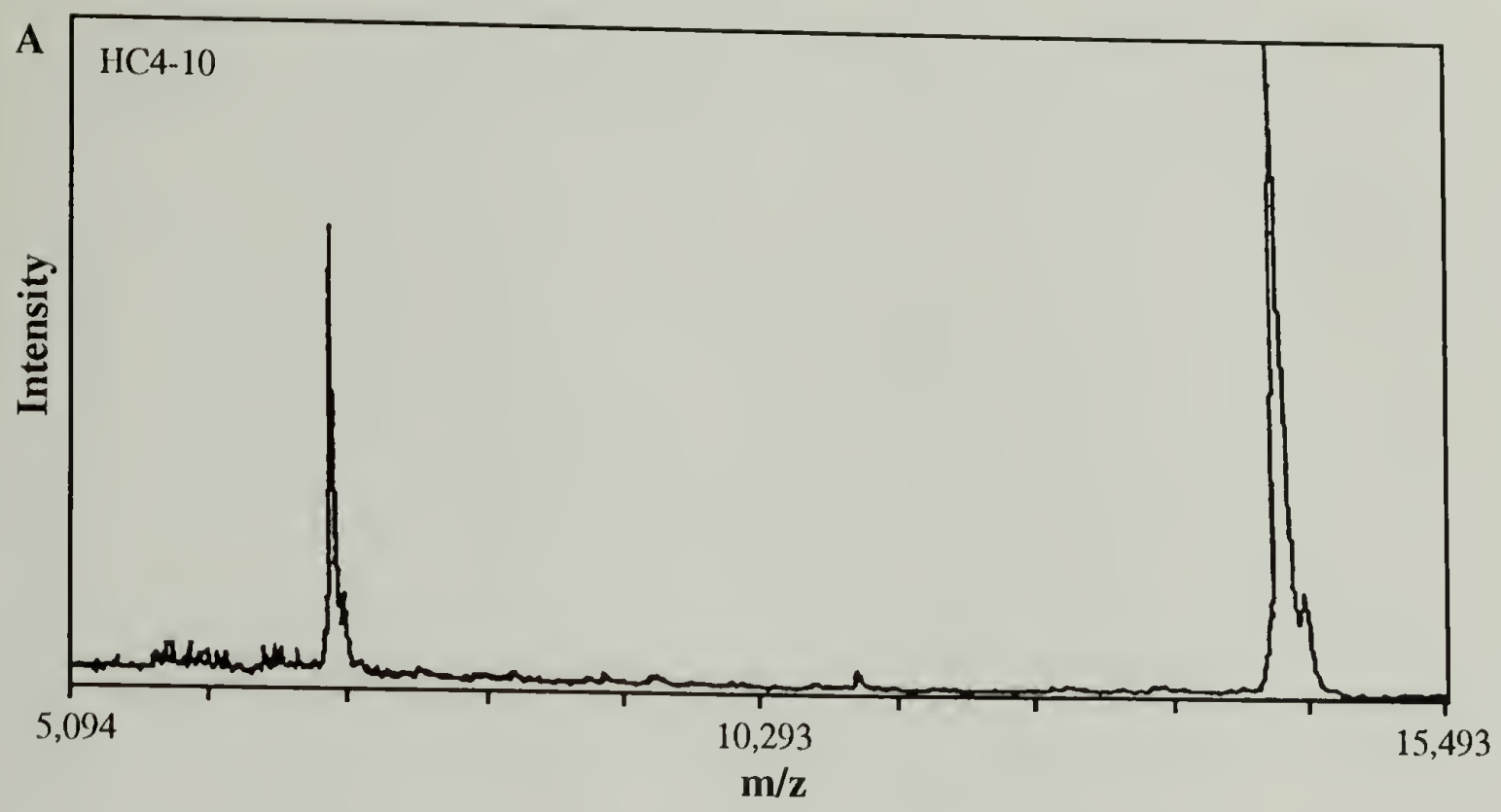
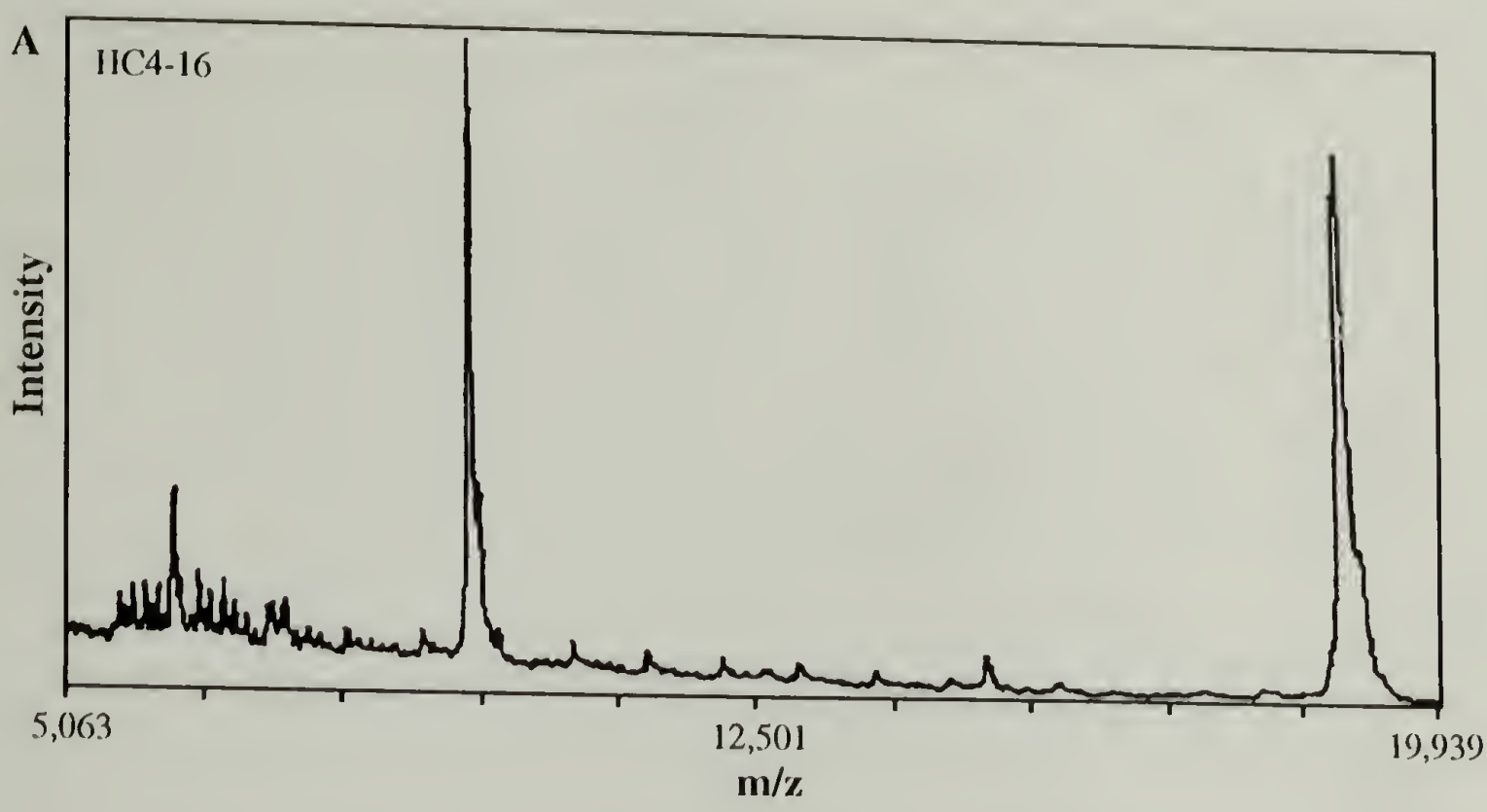


Figure 4.12. Mass analysis of HC4-10 and HC4-12. **A**: Mass spectrum of HC4-10 **B**: Mass spectrum of HC4-12. Both spectra show singly and doubly ionized molecular ion peaks, a fragment cluster in the 5,000-8,000 *m/z* region and various low-intensity signals in the intermediate *m/z* region. The triply ionized HC4-12 molecular ion is visible at *m/z* 5,206. The details of the peak assignments are discussed in section 4.4.



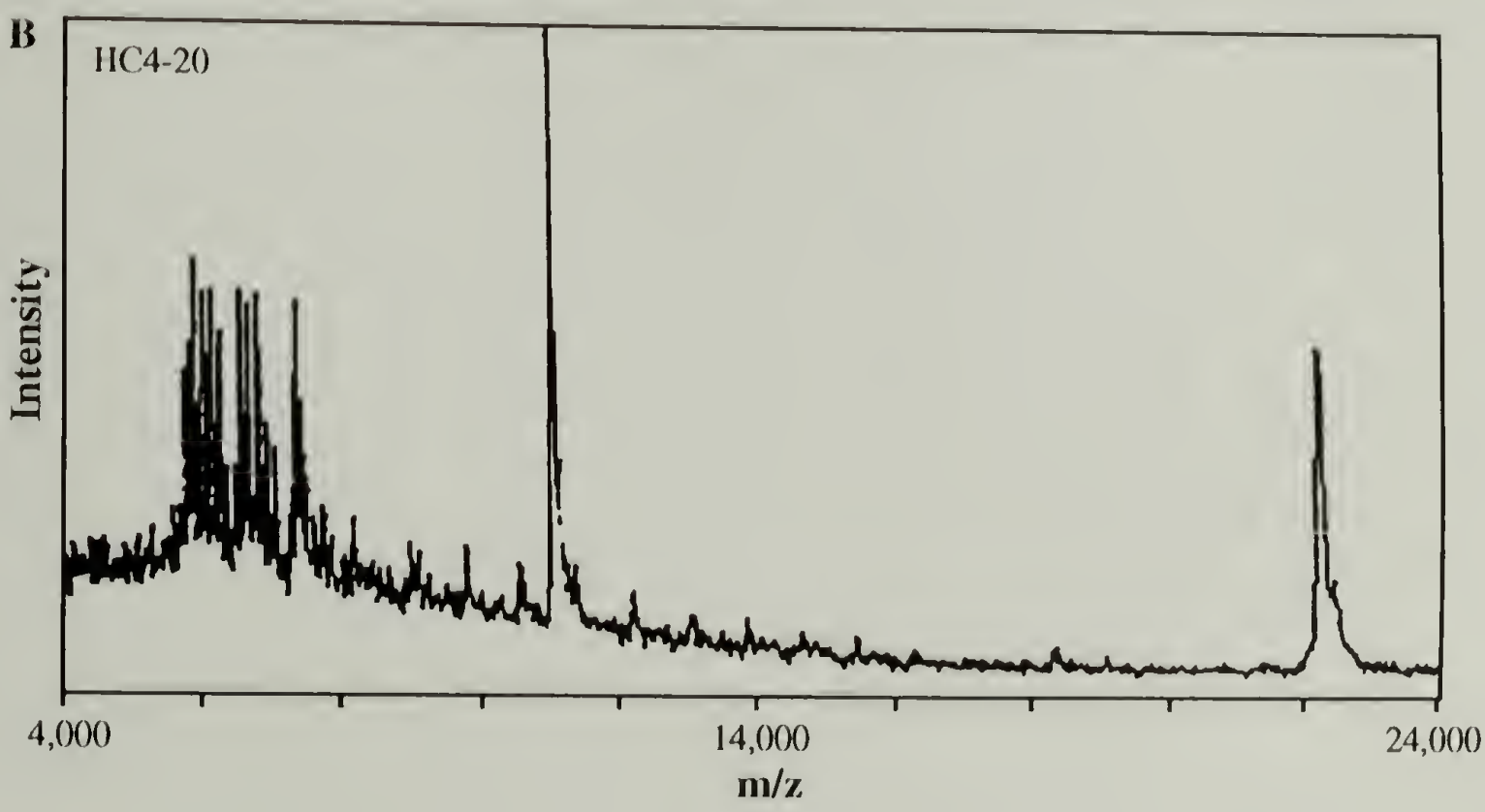


Figure 4.13. Mass analysis of HC4-16 and HC4-20. **A**: Mass spectrum of HC4-16 **B**: Mass spectrum of HC4-20. Both spectra show singly and doubly ionized molecular ion peaks, a fragment cluster in the 5,000-8,000 *m/z* region and various low-intensity signals in the intermediate *m/z* region. The triply ionized HC4-16 molecular ion is visible at *m/z* 6,267. Note the lower intensity of the molecular ion compared to the fragment peaks. The details of the peak assignments are discussed in section 4.4.

Table 4.3. Summary of molecular mass analysis of intact HC4-m proteins.

# Protein repeats (m)	*m/z* (calcd)	*m/z* (exptl)	Δ *m/z*
10	14,024	14,081	+57
12	15,616	15,614	-2
14	17,207	17,264	+57
16	18,799	18,802	+2
20	21,982	21,981	-1

peaks can be rationalized in a similar fashion as discussed in section 4.3.3.2. HPLC analysis of these samples (Figure 4.14) verify the presence of peaks of similar shape and relative mobility to those assigned by fractionation and mass analysis to the low mass fragments of HC4-14. The variation of the mobility of the intact protein with chain length is unexplained as is the additional peak seen in the chromatograph of HC4-20.

In addition to the above described features of the mass spectra for the HC4-m proteins, an additional phenomenon can be observed by comparison of prominent large fragments observed in the spectra (Table 4.4). In three of the five proteins a signal is seen corresponding to a N-terminal fragment of sequence **4.5**. This fragment arises from the removal of the C-terminal fragment **4.6**, (calcd *m/z* 3927).

No prominent peak of this kind appears in the spectrum of HC4-10 and the value (*m/z* 3908) indicated for HC4-20 cannot be assigned to any fragment from the intact chain. These fragments probably arise from some discrete proteolysis event, but the specifics of that process are unclear.

H-ASMTGGQQMGRDPMFKYSRDPMG[$(AG)_4PEG]_{(m-1)}(AG)_3$-OH **4.5**

m = 12, 14, and 16

H-AGPEGARMHIRPGRYQLDPAANKARKEAELAAATAEQ-OH **4.6**

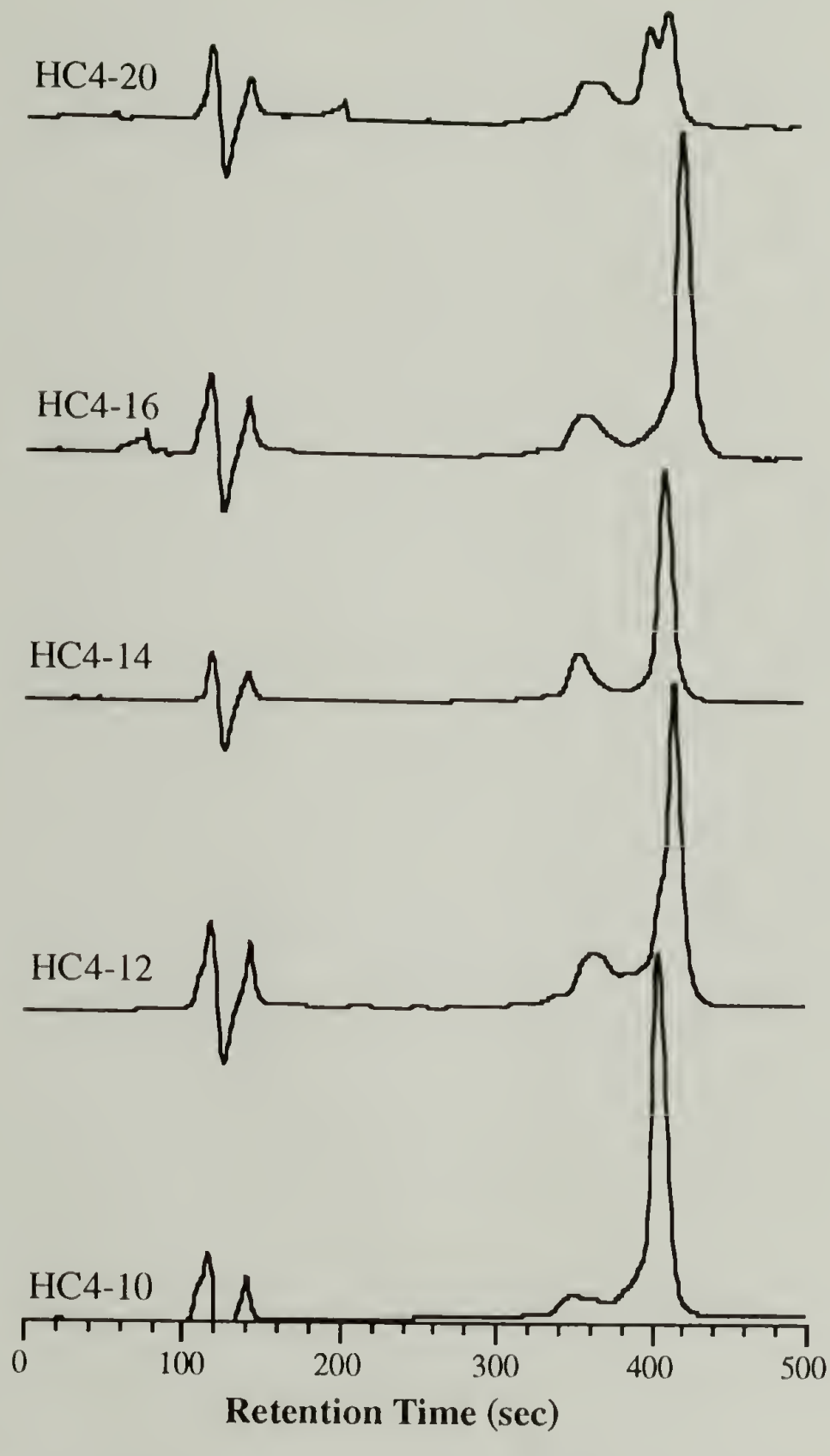


Figure 4.14. HPLC analysis of HC4-m proteins. In all samples, a prominent peak is seen in addition to a less intense peak at shorter retention times. In a previous analysis of HC4-14 (Figure 4.9), the larger peak was shown to correspond to the intact molecular ion and the smaller peak to be comprised of the same HC4-14 protein fragments that give rise to the peak cluster in the 5,000-8,000 region of the mass spectra (see section 4.3.3). The smaller peak appears at approximately the same retention time in all traces, consistent with appearance similar low-mass fragments in the mass spectra for each chain-length variant (Figures 4.12 and 4.13). The pattern of the retention times for the intact proteins is not readily apparent. The chromatograph for HC4-20 has a slightly different shape than the other chain-length variants, an observation consistent with the difference in peak intensities in the mass spectrum of HC4-20 compared to those of the other proteins (Figure 4.13).

Table 4.4. Summary of large fragments in the mass spectra of HC4-m proteins.

# Protein repeats (m)	*m/z* Intact protein	*m/z* Large fragment	Δ *m/z**
10	14,081	—	—
12	15,614	11,687	3927
14	17,264	13,335	3929
16	18,801	14,873	3929
20	21,983	21,983	3908

*The mass difference is thought to correspond to C-terminal fragment **4.6**.

4.5. Summary of Results

The above analysis shows that a single mass spectrum taken using matrix-assisted laser desorption mass spectrometry revealed a wealth of information inaccessible by conventional methods. The significant highlights of the mass analysis of the HC4-m proteins can be summarized as follows:

•Mass analysis exactly confirmed the molecular mass of three of the five proteins of sequence **4.1**. (m=12, 14, and 16). Sequencing of the DNA encoding HC4-14 revealed two amino acid substitutions, exactly accounting for the 57 mass unit discrepancy. A similar 57 mass unit discrepancy was found for HC4-10, which is as yet unresolved. It is noted that the two proteins showing mass discrepancies are expressed from DNA comprised of an odd number of DNA monomers.

•All preparations contained a set of peaks arising from cleavage at sites within the repetitive portion of the protein resulting in the sequential removal of a fragment with the sequence $[(AG)_4PEG]$.

•All preparations contained a series of fragments based on the intact *N*-terminal sequence of the fusion protein (see sequence **4.1**) followed by four copies of the periodic undecapeptide. The relationship of each succeeding signal can be rationalized by addition of a single amino acid residue, proceeding in the *N*–to–*C*–terminal direction along sequence **4.1**: the sequence of a portion of the protein can be read directly from the mass spectrum.

•Three of the five protein preparations (m=12, 14, and 16) contain fragments arising from the removal of a C-terminal fragment with a *m/z* value of 3929. It is noted that the HC4-10 protein shows no such prominent peak and the value of the peak for HC4-20 could not be assigned to any known fragment.

It is not clear whether mutations of the kind found in this analysis are a general phenomenon in the production of synthetic proteins of repetitive structure. The sequence analysis presented here is the first direct evidence of mutations in synthetic DNA produced in our laboratory. Potential sources of the DNA alterations can be considered; the earliest instance a mutation could have occurrred is in the large scale production of pUC4-2. As described in section 3.2.1.5, four DH5α/pUC4-2 *E. coli* colonies were selected from the original transformation plate and restreaked on a second plate. Cells from the selected colonies on this second plate were cultured and stored, and the DNA sequence of the monomer inserts verified. Subsequent to the discovery of the sequence errors described in section 4.3.2.3, the sequence of the monomer DNA in the recombinant pUC4-2 recovered in large scale (section 3.2.1.7) was verified. Despite the verification of the pUC4-2 insert sequence, a DNA monomer with an altered sequence could have been present at some level in this preparation, and subsequently incorporated into the multimers upon enzymatic polymerization: the fact that the two sequence errors occur in the same DNA monomer (Figure 4.4) supports this contention. Subsequent manipulations in the p937.51 vector

could have resulted in the preferential selection of the altered multimer sequence cloned into pET3-b. However, it is also possible that the mutations occurred in p937.51 or in pET3-b; as the intermediate p937.51 containing strains have been lost, pinpointing the exact cause of the sequence errors may be impossible.

The source of the protein degradation fragments is also unclear. The fact that the same fragments are isolated in all preparations suggests a common biological origin in the expression strain. However, it is not clear whether the protein fragments are generated in the cell during cell growth or subsequent lysate processing step. The selective cleavage of the repeat units and the precise "nibbling" of individual amino acids could be the result of concerted proteolysis events or arise through separate processes. Only detailed analysis of the action of the expression strain will reveal the source of the fragmentation.

Whatever the nature of the biological processes underlying the mutations and protein degradation, the analysis of the results of those processes was virtually impossible before the refinement of the present mass spectrometry technique. The analysis of a single mass spectrum allowed characterization of the molecular mass, chemical sequence and purity level of the proteins with unprecedented detail. These features provide a persuasive argument for the widespread application of matrix-assisted laser-desorption/ionization mass spectrometry in protein analysis.

4.6. References

1. A. K. Dunker and R. R. Ruebert, *J. Biol. Chem* **227**, 680 (1969).
2. B. T. Chait and S. B. H. Kent, *Science* **257**, 1885 (1992).
3. A. G. Harrison and R. J. Cotter, *Methods Enzymol.* **193**, 3 (1990).
4. R. C. Beavis and B. T. Chait, *Proc. Natl. Acad. Sci. USA* **87**, 6873 (1990).
5. M. Karas and F. Hillenkamp, *Anal. Chem.* **60**, 2299 (1988).
6. F. Hillenkamp and M. Karas, *Meth. Enzymol.* **193**, 281 (1990).
7. K. Biemann, *Annu. Rev. Biochem.* **61**, 977 (1992).

8. K. O. Börnsen, M. Schär, H. M. Widmer, *Chimia* **44**, 412 (1990).

9. R. D. Cox, M. A. Park, S. Della-Negra, E. Schweikert, *J. Trace Microprobe Tech.* **10**, 55 (1992).

10. R. Beavis, personal communication, (1991).

11. R. C. Beavis, et al., *J. Am. Chem. Soc.* **114**, 7584 (1992).

12. J. Sambrook, E. F. Fritsch, T. Maniatis, *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Press, Cold Spring Harbor, 1989).

13. M13 Cloning/Dideoxy Sequencing Instruction Manual, Bethesda Research Laboratories, (1986).

CHAPTER 5

SOLID-STATE CHARACTERIZATION OF COPOLYPEPTIDES

5.1. Overview

5.1.1. Preliminary Exploration of the Stem-length Premise

A preliminary X-ray scattering analysis of HC4-14c was performed to demonstrate the validity of the early premise of this work: that the insertion of a single alanyl-glycine dyad in the repeat sequence of KM3-27 (HC3-54 in the present nomenclature), a material shown to be essentially amorphous, will result in an increase in solid-state order. A sample of HC4-14c as recovered after cyanogen bromide cleavage (section 3.2.7) was dried extensively *in vacuo* at 80°C, the dry powder packed in a 1.5 mm glass capillary and the X-ray scattering pattern recorded. All X-ray scattering patterns discussed in this chapter were recorded on a Statton camera using a Rigaku Ni filtered Cu K_α X-ray source with a wavelength of 1.5418 Å. Wide angle patterns were generated at a sample-to-film distance (SFD) of 50.33 mm, calibrated periodically with calcium carbonate. Smaller angle patterns were taken at a SFD of 170 mm.

A direct comparison of the scattering pattern with the essentially featureless pattern of KM3-27[1] is shown in Figure 5.1. While there were no evident wide angle reflections associated with the formation of a ß-sheet, the general sharpening of the strong 4.7 Å reflection from KM3-27 to HC4-14c seemed to validate the stem-length premise. Additional support was derived from the observation of 1527 cm^{-1} amide II and 1703 and 1630 cm^{-1} amide I peaks in the infrared spectrum of a sample prepared as a pressed KBr pellet; these data were consistent with the presence of some anti-parallel ß-sheet.

Based on the results of this preliminary analysis, the synthesis of longer stem-length variants was accomplished. In order to provide a direct comparison of the effect of

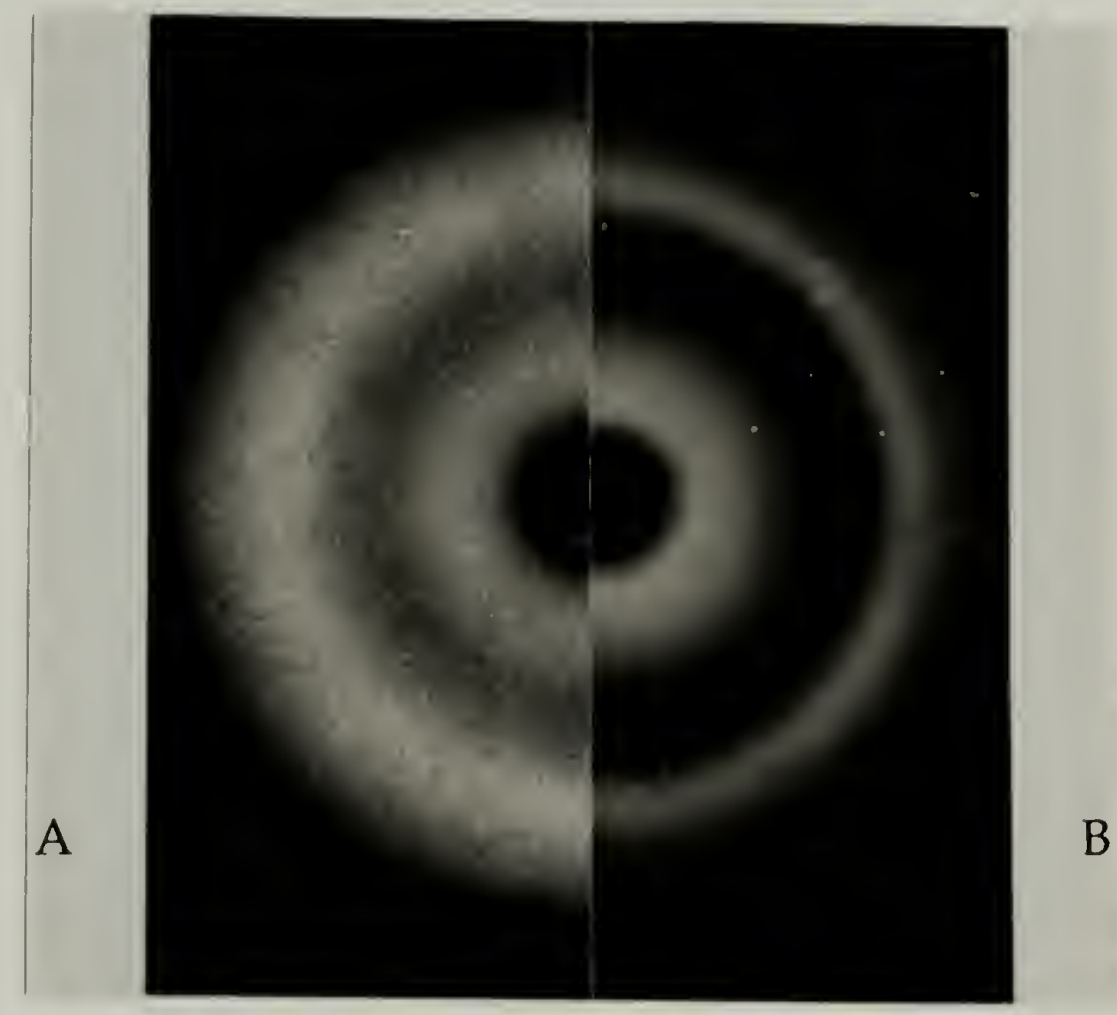


Figure 5.1. X-ray powder patterns of KM3-27 and HC4-14. **A**: KM3-27 and **B**: HC4-14. The relative increase in the sharpness of the X-ray reflections from **A** to **B**, indicating the adoption of a more ordered structure with the insertion of an alanylglycine dyad in the repeat sequence of KM3-27, was taken as support for the stem-length premise, as discussed in section 5.1.1.

the increased stem-length, the chain length variable *m* of sequence **5.1** was held constant. The choice of chain length of m=16 was made based on the results of the mass analysis presented in Chapter 4, which indicated the correct molecular mass for HC4-16. The discussion below is concerned with the details of the solid-state structure of the series of materials described by sequence **5.1**.

$$[(AG)_nPEG]_m \qquad m=16,\ n=3,\ 4,\ 5,\ 6 \qquad \mathbf{5.1}$$

5.1.2. Approach to the Analysis of HCn-16c Proteins

The analysis of HCn-16c polymers presented below was conducted as part of a general program aimed at accumulating a broad and consistent base of structural information on the requirements for chain folding of polypeptides. In particular, the procedures used in this analysis were designed to correlate with the work of Krejchi[2], who demonstrated that the repetitive polypeptide **5.2** adopted a chain-folded lamellar architecture of defined thickness upon crystallization from solution. Since the sequence of **5.2** is closely related to that of **5.1**, a defined set of procedures were adopted based on those already established in the analysis of **5.2**.

Although this work was performed in a larger context, the major thrust of the analysis was a *comparative* study of the four stem length variants. The following discussion of the solid-state behavior of these materials therefore focuses on relative trends in data collected and not on details of the structure of the individual samples.

$$[(AG)_3EG]_{36} \qquad \mathbf{5.2}$$

5.2. Analysis of HCn-16c Powders

In the following discussion, the HCn-16c proteins will be referred to as the appropriate stem-length variable **n**. For instance, the cleaved polypeptide HC3-16c will be

referred to as **3**, HC4-16c as **4**, etc. Since the analysis is focused on a limited number of sample preparations, suffices pertinent to the discussion are included to define individual sample preparations where appropriate—no detailed sample nomenclature was employed.

5.2.1. X-ray Powder Diffraction Analysis of HCn-16c Proteins

The HCn-16 proteins as recovered after cyanogen bromide cleavage were analyzed by X-ray diffraction. Samples of **3**, **4**, and **5** were recovered after extensive washing in methanol as described in section 3.4.6, and dried *in vacuo* at 50°C for 15 hours. A sample of gelled **6** (*vide infra*) was recovered from a methanol suspension by centrifugation, and the sedimented gel dried *in vacuo* at 50°C for 15 hours. X-ray patterns were taken of the powders packed in a 1.5 mm glass X-ray capillaries. The diameters of the observed reflections were measured in millimeters and *d* spacings calculated using Bragg's law.

The powder patterns are shown in Figure 5.2. A comparative analysis of the patterns reveals a consistent, progressive sharpening and increase in intensity of reflections from **3** to **6**. The *d* spacings and intensities of the reflections are summarized in Table 5.1.

Overall, the level of disorder implied by the patterns is high: all the reflections are broad and diffuse with the exception of the weak lower angle spacing at 17.3 Å. The sharp, strong 4.3 Å reflection is not well resolved from the intense 4.7 Å spacing, and is only evident upon close examination of the X-ray films. The prominent reflections at 9.7 and 4.7 Å, consistent with intersheet and interchain spacing of ß-sheets, are the only features common to all patterns. The diffuse maxima in **3** are sharpened with the addition of a single alanyl-glycine dyad in **4**, and the prominent reflections are seen to increase in intensity and sharpness in the progression to longer stem-lengths. The scattering pattern associated with the antiparallel ß-sheet arrangement of PLAG I and silk II is well established in **6**. The gradual sharpening of the 3.7 Å spacing and appearance of the weak wide-angle spacings at 2.25 and 2.07 Å reflections are the best indicators of this continuum trend to a more defined ß-structure.

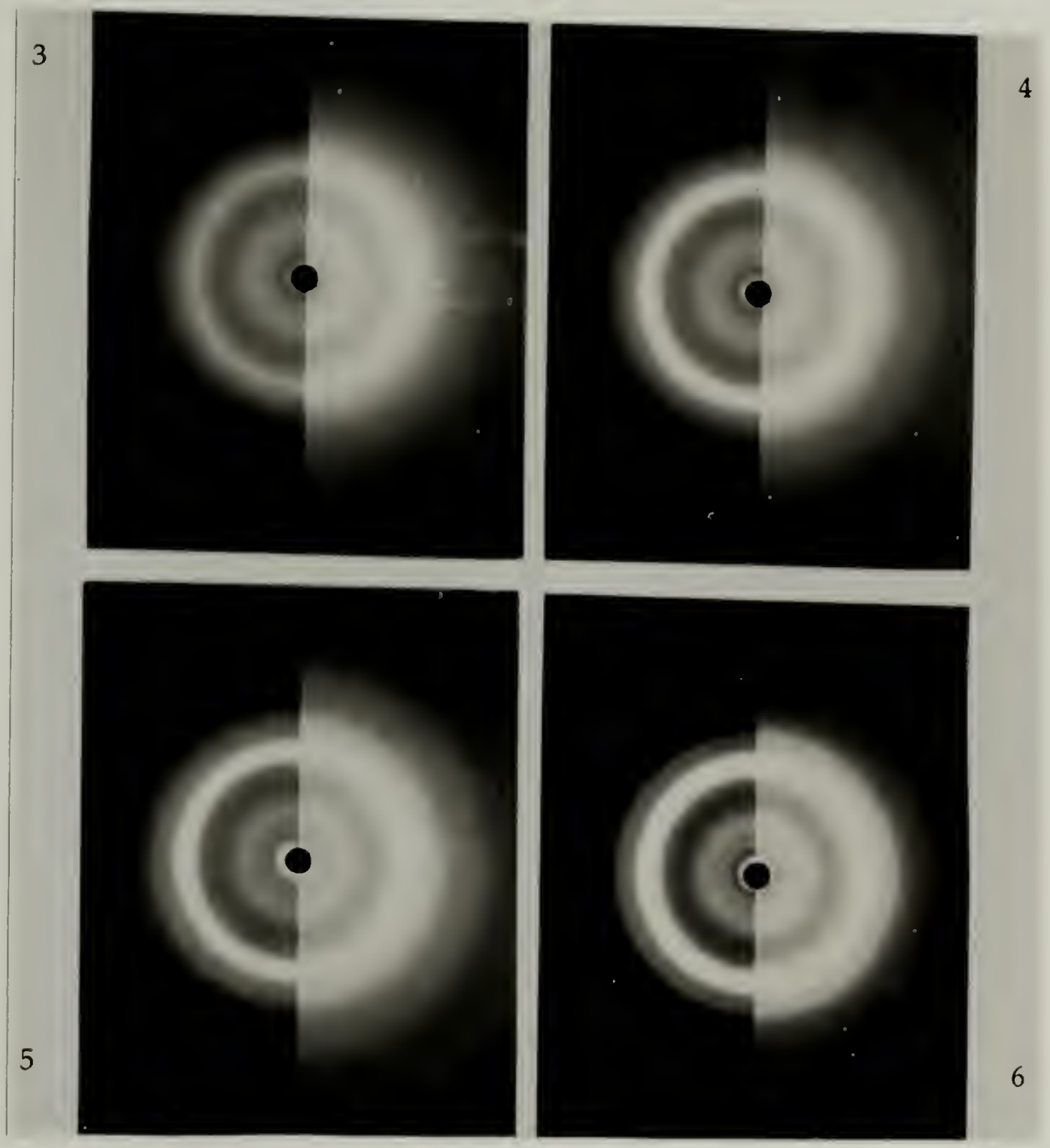


Figure 5.2. X-ray powder patterns of HCn-16c proteins. Each pattern is presented at two different exposures for clarity. Each successive pattern, starting with **3**, shows a general increase in the sharpness of the reflections, a trend best illustrated by examination of the prominent 4.5 Å reflection, which is present as a diffuse halo in **3** and increases in sharpness in the patterns for **4** and **5**. The resolution of this signal into two reflections at 4.7 and 4.3 Å in **6** is faintly discernible in the reproduction (these signals are resolvable for **4**, **5**, and **6** by direct examination of X-ray negatives). The assignments of the peaks are given in Table 5.1 and the patterns are discussed in section 5.2.1.

Table 5.1. Summary of X-ray powder diffraction analysis of HCn-16c proteins.

d Spacing (Å)					
HC3-16c	HC4-16c	HC5-16c	HC6-16c	Models	
				Silk[3]	PLAG I[4]
	17.3 w, sh	17.3 w, sh	17.3 w, sh		
~9.7 (df halo)	9.7±1.2 m, df	9.7±1.2 m, df	9.7±1.2 m, df	9.2	8.87
		5.8 vw, sh	5.8 vw, sh		
	4.7±0.1 vs, br	4.7±0.1 vs, br	4.7±0.1 vs, br	4.60	4.68
4.5±0.3 vs, df					4.45
	4.3 s, sh	4.3 vs, sh	4.3 vs, sh	4.19	4.22
	3.7 m, df	3.7 m, sh	3.7 s, sh		3.76
					3.61
	2.9 vw, df	2.9 w, df	2.9 w, df	3.07	2.97
					2.74
	2.29 vw, br	2.29 vw, br	2.29 w, br	2.35	2.25
		2.07 vw, br	2.07 w, br	2.09	2.07

Intensity: very weak (vw), weak (w), med (m), strong (s), very strong (vs).
Appearance: sharp (sh), broad (br), diffuse (df). Broad reflections are given with limits.

In addition to the ß-pattern, a relatively sharp reflection of low intensity is seen at lower angles in **4**, **5**, and **6** (and may also appear in **3**). The consistent intensity and interval of the spacing (17.3 Å) suggests that it is an indication of some longer range order in the samples which is independent of stem-length. The origin of this reflection is unclear, but its consistent appearance in all patterns suggests that the spacing is not indicative of a stem-length dependent lamellar repeat distance, but rather arises in some way from a chain packing arrangement unique to these materials. The 17.3 Å spacing also consistently appeared in over-exposed patterns recorded at a SFD of 170 mm. However, no longer period spacings were observed in any of these patterns (this camera position allows the resolution of sharp reflections at *d* spacings of ca. 50 Å).

5.2.2. Analysis of Amorphous HC3-16c

The amorphous halos in the scattering pattern of **3** shown in Figure 5.2A were sharper than the patterns generated for KM3-27 (Figure 5.1A). Since this difference could have arisen from the large disparity in the length of the chains between the two samples, it was considered important to correlate this work to that of McGrath to address any molecular weight effects on structure. To that end a sample of **3** was prepared in an attempt to match the amorphous halos of KM3-27 shown in Figure 5.1A.

A sample of **3** (100 mg) was dissolved in 100 mL water and the pH adjusted to 7 with sodium hydroxide to make a neutral 0.1% aqueous solution. The solution was quenched at -80°C and the protein recovered by lyophilization. The powder was dried *in vacuo* as before and the X-ray scattering pattern recorded as shown in Figure 5.3.

The pattern is essentially identical to that of KM3-27 in Figure 5.1A, showing only two diffuse amorphous halos centered at 4.5 and 9.3 Å. Thus, the sharper pattern shown for the sample of **3** (as recovered after cyanogen bromide treatment) in Figure 5.2 results from the sample preparation technique and is not a consequence of the shorter chain length. The diffuse pattern seen in Figure 5.3 was established as a baseline sample with a completely disordered conformation (100% amorphous content) for comparison with subsequent sample preparations (the sample is designated HC3-16cA or **3A**).

5.2.3. Densitometry of HCn-16c Powder Diffraction Patterns

The radial scattering profiles of the scattering patterns seen in Figure 5.2 were determined by densitometry on an Optronics International System C-4500 scanning densitometer at a raster size of 50 μm. The scanned intensity was radially averaged over the image and plotted as 1/d (measured in reciprocal ångstroms) versus X-ray intensity. The collected scans of the X-ray patterns are shown in Figure 5.4.

The traces confirm the sharpening of the prominent reflections from **3** to **6** as well as the overall disorder observed in the patterns. The reflection at 17.3 Å is apparent

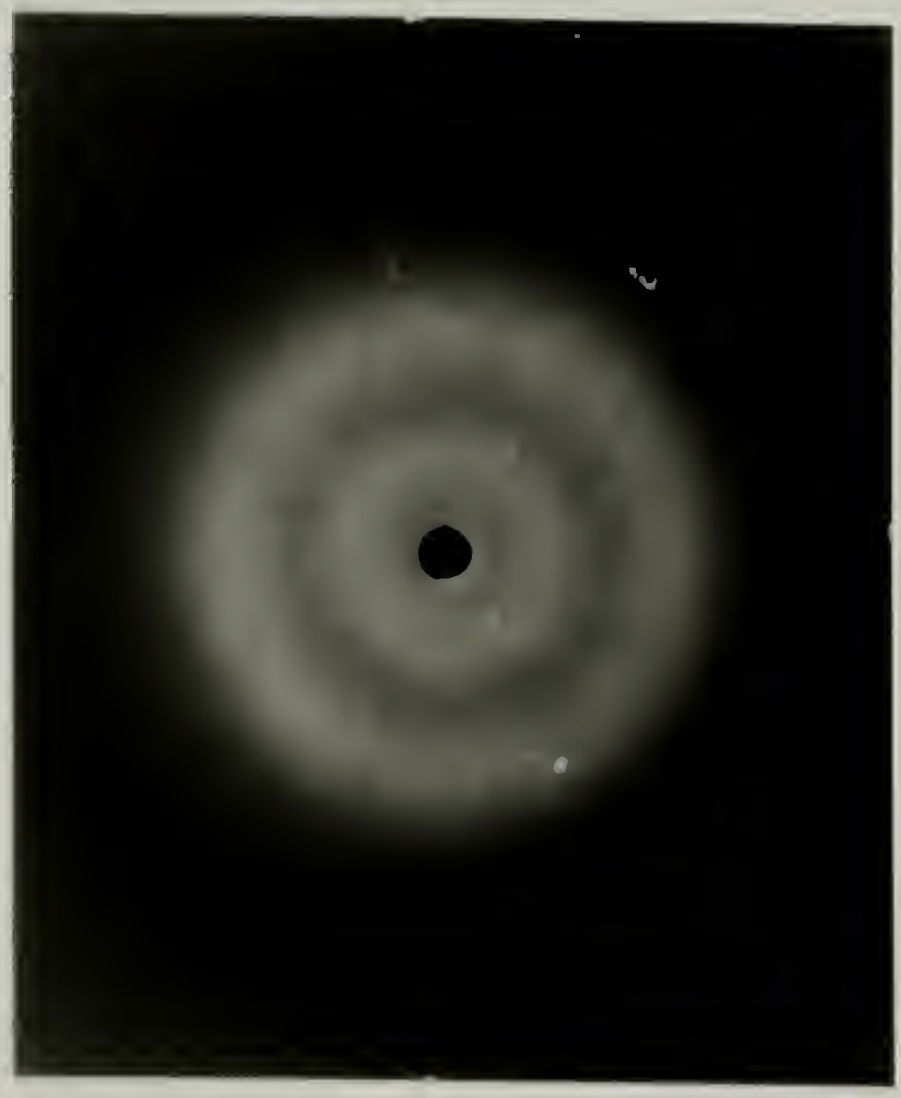


Figure 5.3. X-ray powder patterns of amorphous sample HC3-16cA. The pattern shows only broad, diffuse halos centered at ~9.7 Å and ~4.5 Å. This preparation was established as a baseline sample representing a completely disordered conformation (100% amorphous content) for subsequent sample preparations. The preparation and analysis of this sample is discussed in section 5.2.2.

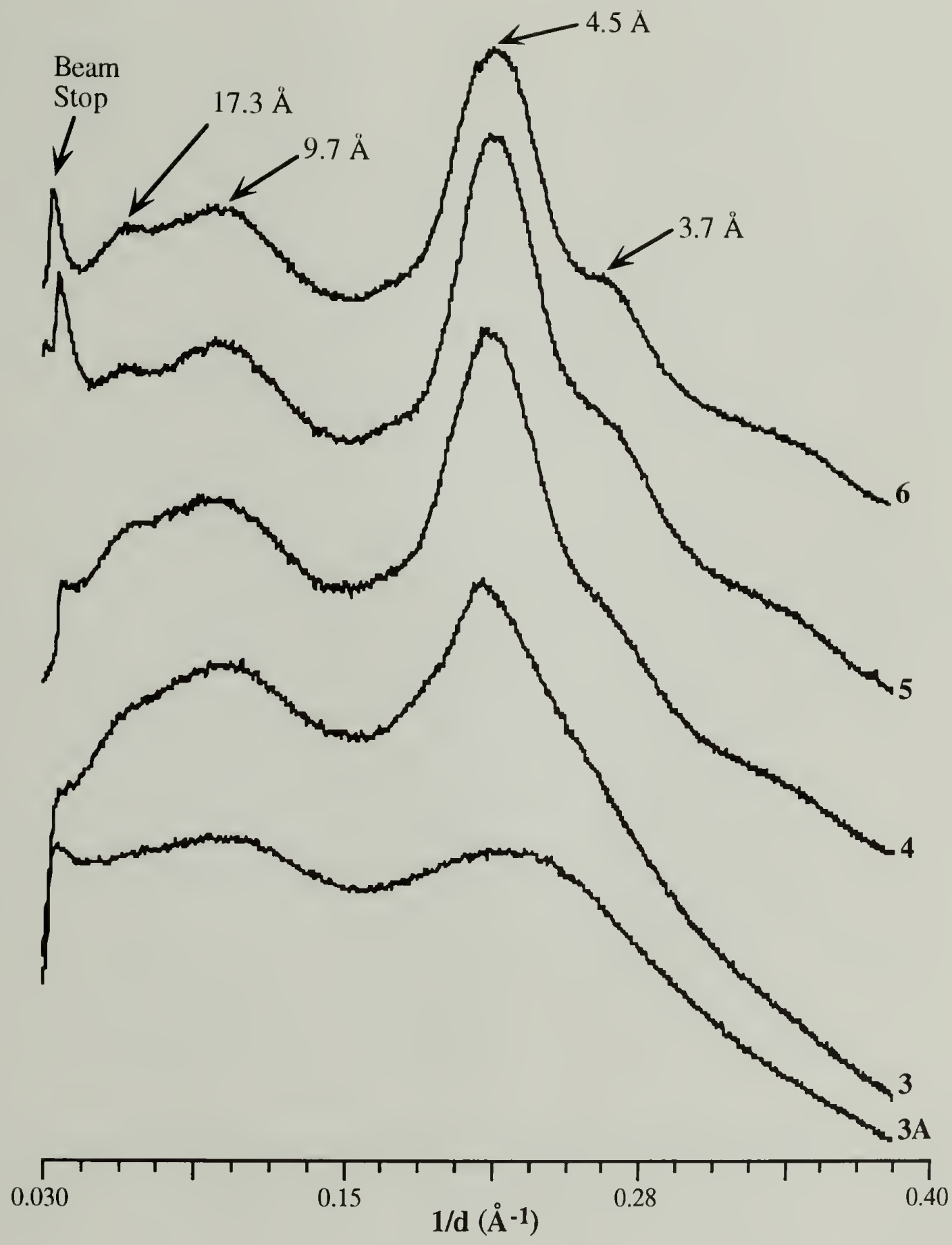


Figure 5.4. Radial scattering profile of HCn-16c X-ray powder patterns. The scattering patterns reproduced in Figure 5.2 were scanned on an Optronics International System C-4500 scanning densitometer at a raster size of 50 μm. The scanned intensity was radially averaged over the image and plotted against *1/d* (Å). The scanned image of amorphous sample **3A** reflects the broad, diffuse halos of the X-ray pattern seen in Figure 5.3. There is an increase in intensity and sharpness of the peaks at 17.3, 9.7, 4.5, and 3.7 Å from **3A** to **6** (the corresponding peaks for **6** are labeled). The presence of separate 4.7 and 4.3 Å reflections is not apparent in **4**, **5**, or **6**; the densitometer reveals only a broad maximum at 4.5 Å. The analysis is discussed in section 5.2.3.

in **4**, **5**, and **6** and the sharpening of the 4.7 Å reflection of **3** in comparison to that of amorphous sample **3A** is evident. In addition, the observed sharpening of the 3.7 Å and (to a lesser extent) the 9.7 Å reflections with increasing stem-length is also apparent.

It is often possible to determine the degree of crystallinity of a sample from the calculated area under the diffraction curve, based on the notion that the scattered intensity is proportional to the number of electrons, and hence the mass of the scattering phase. Assuming the presence of only two phases (crystalline and amorphous) in an isotropic sample, the ratio of the integrated intensity of the reflections associated with the crystalline phase with the total integrated intensity is a direct measure of the crystalline content in the sample.

In the present case it was impossible to make any sort of definitive statement about the relative amount of amorphous and crystalline components in the diffraction curves, even using the trace of the amorphous sample **3A** as a basis for determining the peak shape of the amorphous component. The difficulty lay in assigning the appropriate baseline. Due to differences in the overall intensities arising from sample-dependent exposure conditions, there was no way to determine the appropriate arrangement of the traces and assign a baseline with any rigor. Attempts to compare the traces "by eye" failed to yield any consistent or meaningful results.

Despite the impediments to the assessment of the absolute amount of crystallinity in each of the samples, the relative sharpening of the reflections with increasing stem length inferred from the X-ray patterns is clearly demonstrated in the diffraction curves.

5.2.4. FTIR Analysis of HCn-16c Powders

All infrared spectra were taken on a FTIR spectrometer comprised of a Brüker optics bench and laser controlled by Nicolet software. Samples were prepared as pressed KBr pellets and analyzed under a nitrogen purge.

5.2.4.1. Amide A Region

The 3800-2800 cm^{-1} regions of the spectra of all HCn-16c proteins (including amorphous **3A**) are shown in Figure 5.5. The region is dominated by the amide A absorption, with two broad, overlapping components at 3490 and 3270 cm^{-1} arising from N-H stretch vibrations. The amide B vibration at 3080 cm^{-1} is also apparent.

The higher frequency vibration at 3490 cm^{-1}, prominent in the amorphous **3A** sample, is seen to lessen in intensity in sample **3** relative to the peak at 3270 cm^{-1}. This shift of intensity to lower frequency is generally associated with an increase in hydrogen bonding.[3] Thus the prominent high frequency peak in the spectrum of the amorphous sample **3A** is thought to arise from a high percentage of amide groups not participating in hydrogen bonding. The peak shapes of samples **4**, **5**, and **6** are qualitatively similar, suggesting that some plateau in hydrogen bonding content is reached.

The amide A vibration is often difficult to differentiate from the vibrations associated with water because the higher frequency peaks overlap. Although the samples were scrupulously dried and handled in the presence of desiccants, some water could have been trapped in the sample, giving rise to the higher frequency vibration. To test the effect of hydration on the infrared spectra of the materials, portions of **3**, **4**, **5**, and **6** were allowed to equilibrate with the ambient water content (roughly 60% relative humidity) for 36 hours and the infrared spectra of pressed KBr pellets measured as before.

This treatment caused no change in the spectra of **4**, **5**, and **6**, but as shown in Figure 5.6A the higher frequency amide A vibration of **3** disappeared, with the spectrum showing only the lower frequency vibration associated with hydrogen bonded amide groups. This result indicates that a significant portion of **3** is accessible to diffusion of water molecules, which associate with the "free" amide hydrogens (those not participating in hydrogen bonds) and lower the vibrational frequency. This is not the case for **4**, **5**, and **6** which show no such change with exposure to water; the spectra of wet and dry **4** are seen in Figure 5.6B.

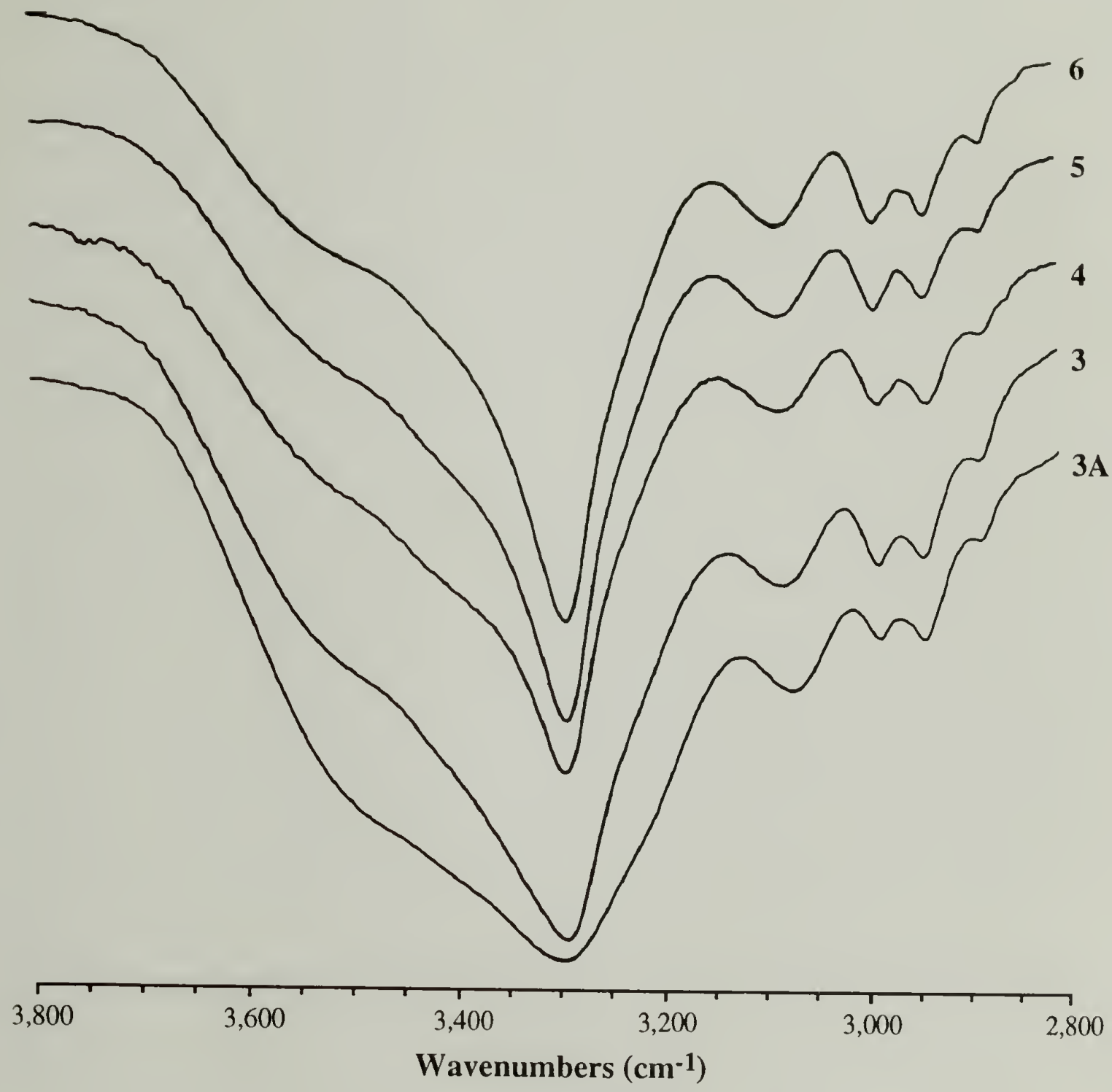


Figure 5.5. FTIR spectra of HCn-16c (n=3-6) proteins: Amide A region. The region is dominated by the amide A absorption, with two broad, overlapping components at 3490 and 3270 cm^{-1} arising from N-H stretch vibrations. The amide B vibration at 3080 cm^{-1} is also apparent. The prominent shoulder at 3490 cm^{-1} is seen to diminish from **3A** to **3** and (to a lesser extent) from **3** to **4**, with the peak patterns of **4**, **5**, and **6** virtually identical. This observation can be rationalized in terms of a general increase in the number of amide groups participating in hydrogen bonds with increasing stem-length: the analysis is discussed in section 5.2.3.1.

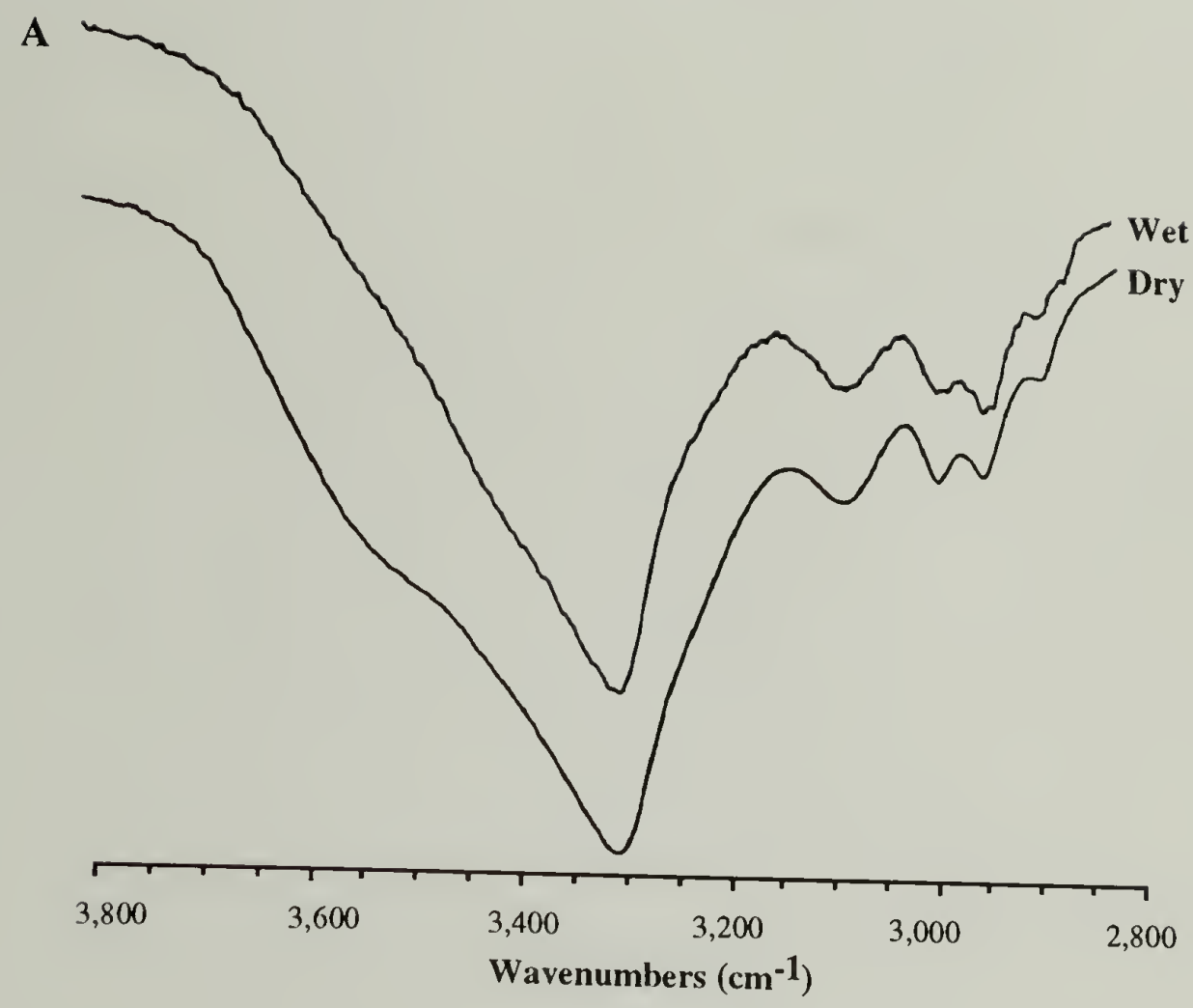


Figure 5.6. FTIR spectra of HCn-16c (n=3, 4) proteins: Effect of hydration. **A**: Hydration of HC3-16c. Upon exposure of dried **3** to ambient moisture (~60% relative humidity) ("Wet" spectrum) the higher frequency amide A vibration, prominent in the "Dry" spectrum, disappears, and shows only the lower frequency vibration associated with hydrogen bonded amide groups. **B**: Hydration of HC4-16c.The FTIR spectra of dry **4, 5**, and **6** show no significant change upon hydration with ambient moisture: the spectra of "Wet" and "Dry" samples of **4** are representative. These data were interpreted in terms of the ease of diffusion of water into the material as discussed in section 5.2.3.1.

This hydration experiment indicates that the lower frequency amide A vibration is a qualitative diagnostic for the level of inherent hydrogen bonding in the samples, assuming a consistent water content in the dry samples. Thus, in the samples as recovered from cyanogen bromide cleavage, there seems to be a significant increase in the number of amide groups participating in hydrogen bonding in **4** compared to **3**; the level then seems to plateau with increasing stem-length.

The only other stem-length dependent feature of the spectrum in this region is a general sharpening of the CH_3 asymmetric stretch vibration at 2982 cm^{-1} and the lower frequency CH_3 symmetric stretch vibration at 2930 cm^{-1}, indicative of the increased alanine content with longer stem length.[5]

5.2.4.2. Amide I and II Region

The region of the spectrum characteristic of amide I (carbonyl stretch) vibrations was used to explore the conformational state of the polypeptide samples. As discussed in section 1.6.2.1, this is possible due to frequency splitting of the carbonyl stretch bands from discrete vibrational interactions among peptide groups participating in particular structural units. This structure dependent frequency shift is also evident for the amide II (N-H bending) vibrations.

The extensive study of the infrared spectra polypeptides and proteins has defined the vibrational frequencies associated with the dominant structural motifs of these materials.[3, 6] The average frequencies associated with the α-helix, antiparallel ß-sheet and random structures are shown in Table 5.2.

These average frequencies must be used with care when examining spectra of new materials, because individual peptide vibrations can become coupled and thus the associated peaks shifted and split. In addition, because of peak overlap, the use of relative peak intensities in assigning relative structural content can be misleading as shown in Figure 5.7.[3] The amide I peaks associated with the antiparallel ß-sheet of silk fibroin are

Table 5.2. Average values of diagnostic infrared frequencies for protein conformations. Adapted from Fraser and MacRae[3]

Structural Motif	Amide I (cm^{-1})	Amide II (cm^{-1})
α-helix	1652	1516,1546
Antiparallel ß-sheet	1702, 1630	1530
Random structure (turns)	1658	1535

deconvoluted to show the relative contributions of the split vibrations [$\upsilon_{\perp}(\pi,0)$, 1630 cm^{-1} and $\upsilon_{\parallel}(0,\pi)$, 1700 cm^{-1}] associated with the antiparallel ß-sheet and the υ_a (1657 cm^{-1}) vibration of the random structure. The figure demonstrates the need for careful interpretation of relative peak intensities. Despite these caveats, the spectral characteristics of poly(L-alanylglycine) and the silks, materials similar to those examined here, can be used as a basis for a comparative analysis of the spectra, with the understanding that there may be differences in structural details among the materials.

A comparison of the expanded spectra for the HCn-16c protein powders is shown in Figure 5.8. The spectrum of amorphous **3A**, included as a baseline, shows strong absorbances at 1657 and 1540 cm^{-1} indicative of random structure, with no other peaks from more complex structures visible, even as shoulders. On the other hand, the presence of a strong amide II vibration at 1525 cm^{-1} and shoulders at 1703 and 1627 cm^{-1} indicates the presence of some antiparallel ß-sheet structure in **3**. The fact that these peaks appear as shoulders on the very strong 1657 cm^{-1} band suggests that the ß-sheet is not the dominant conformation. It should be noted however, that well-defined turn structures show absorbances in this region, and since **3**, if folded, has the lowest ratio of stem to turn content of the series, this large absorbance may not necessarily be indicative of inherent disorder.

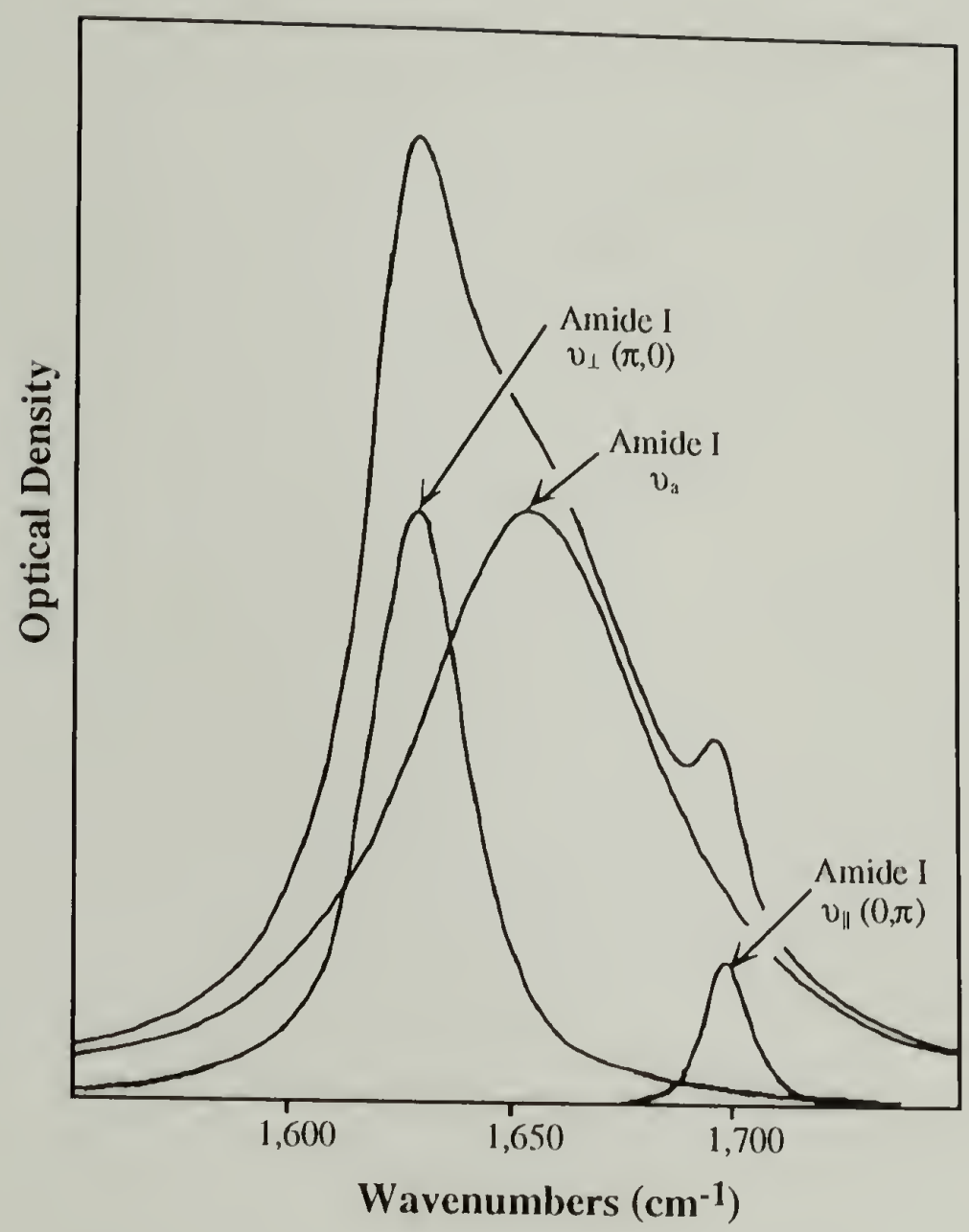


Figure 5.7. IR spectrum of isotropic silk fibroin showing amide I vibrations.The amide I peaks associated with the antiparallel ß-sheet of silk fibroin are deconvoluted to show the relative contributions of the split vibrations [$\upsilon_{\perp}(\pi,0)$, 1630 cm^{-1} and $\upsilon_{\parallel}(0,\pi)$, 1700 cm^{-1}] associated with the antiparallel ß-sheet and the υ_a (1657 cm^{-1}) vibration of the random structure. This analysis demonstrates the need for careful interpretation of relative intensities of overlapping peaks. Adapted from Fraser and MacRae[7]; reprinted with permission of the Academic Press.

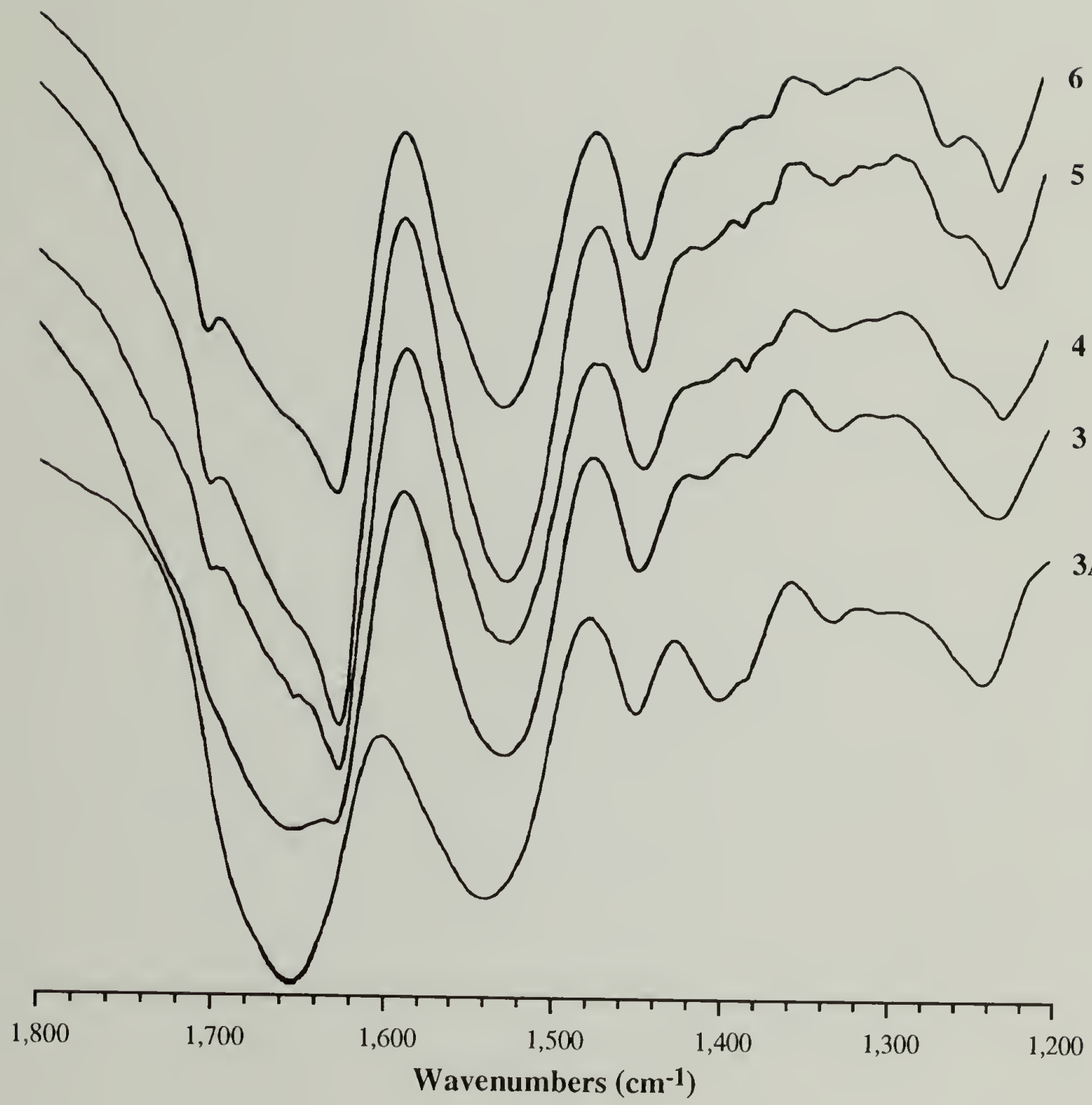


Figure 5.8. FTIR spectra of HCn-16c (n=3-6) proteins: Amide I, II and III region. The broad bands at 1657 cm^{-1} (amide I) and 1540 cm^{-1} (amide II) in the spectrum of **3A** indicates that the material consists primarily of a disordered phase. The presence of the 1703 cm^{-1}, 1627 cm^{-1} and 1657 cm^{-1} amide I peaks, and the amide II signal at 1525 cm^{-1} in the spectrum of **3**, indicate the presence of some ß-sheet content for this material. The split amide I vibrations (1627 and 1703 cm^{-1}) become more prominent with increasing stem length (from **3** to **6**), indicating the increasing dominance of the ß-sheet structure in solid-state structure of these materials. The significance of the other vibrations, namely the amide III absorbances at ~1260 cm^{-1} and the methyl bend peak at 1450 cm^{-1}, are discussd in section 5.2.3.2.

Continuing with the spectra of the polymers with increasingly longer stem length, a sharpening of the vibrational frequencies (1703, 1627, and 1525 cm^{-1}) associated with an antiparallel ß-sheet is seen in the spectra of **4**, **5**, and **6**, indicating the predominance of the sheet motif in these materials. Comparison of all four spectra of the HCn-16c proteins is once again indicative of an increase in solid-state order with stem-length.

Two other spectral features of Figure 5.7 are of note. The materials show a sharpening and resolution of the 1260 and 1230 cm^{-1} amide III (N-H in-plane bending) vibrations observed in spectra of antiparallel ß-sheets of poly(L-alanylglycine) and silk fibroin.[4] This suggests a relative increase in the number of peptide units participating in the sheet in correlation with the increasing numbers of alanyl-glycine dyads in the stem. In addition, all spectra show a strong peak at 1450 cm^{-1} from the alanine CH_3 asymmetric bending vibration.[5]

Another spectral feature in this region is the observed peaks at 1405 and 1375 cm^{-1} in the amorphous sample **3A**. These frequencies are associated with asymmetric CH_2 wagging and the CH_3 symmetric bending, respectively and are seen as very weak signals in the spectrum of antiparallel ß-sheets of poly(L-alanylglycine).[5] The increased intensity of these peaks relative to those in the spectra of the other, more ordered materials shown in Figure 5.7 is intriguing, as it is another prominent signal of the development of solid-state order from **3A** to **3**. The exact explanation for this observation is not readily apparent.

All of the spectral features of the 1800-1200 cm^{-1} infrared spectral region discussed in this section are consistent with the establishment of an antiparallel ß-sheet structure upon increasing the stem length of **3** by one alanyl-glycine unit. The structure of **3** itself is not featureless, with an indication of peaks associated with a sheet structure: the amorphous sample **3A** shows the spectral features associated with polypeptides in a mostly disordered state. Thus the increase in conformational order with increasing stem-length, especially from **3** to **4**, is verified by this analysis.

5.2.5. DSC Analysis of HCn-16c Powders

Differential scanning calorimetric traces for **3** through **6** are shown in Figure 5.9. The samples were analyzed using a Dupont DSC 10 Differential Scanning Calorimeter referenced against an indium standard. All materials were first ramped from 50 to 220°C at 20°C/min. On the first run a large endotherm representing the loss of water was recorded. Subsequent ramps were reproducible; Figure 5.9 is a plot of the results of the fourth cycle. The materials were then cooled to -40°C and ramped again to 220°C; no further features were visible on any thermogram. In the final scan, the temperature was raised until the onset of degradation, as witnessed as the start of an endotherm (generally at 240°C) and the temperature quickly lowered to prevent loss of material into the DSC head. Once again, no additional transitions were visible.

Figure 5.9 shows a reproducible glass transition temperature (T_g) at ~180°C for **3** and **4**, and no visible transitions for **5** and **6** (up to the onset of degradation). Since the crystalline melting transitions of proteins are generally above their decomposition points, the lack of a first order melting transitions in any of the materials is not unusual. The T_g shown by **3** and **4** is indicative of the amorphous content of the material, and the absence of such a transition in the traces of **5** and **6** may indicate either a low amorphous content or that the transition, if present, is above the decomposition temperature.

5.3. Analysis of Oriented HCn-16c Mats

5.3.1. Preparation of Gelled of HCn-16c Protein Samples

In an attempt to obtain oriented crystalline X-ray scattering patterns, a series of trial sample preparation conditions of **3**, **4** and **5** were attempted based on the work of Krejchi[2] and Deguchi[8] that involved the crystallization of polypeptides in sheared (stirred) formic acid solutions. In general, they observed a physical gelation of the solution at some point during the preparation and found that the onset of gelation was an indicator of the formation

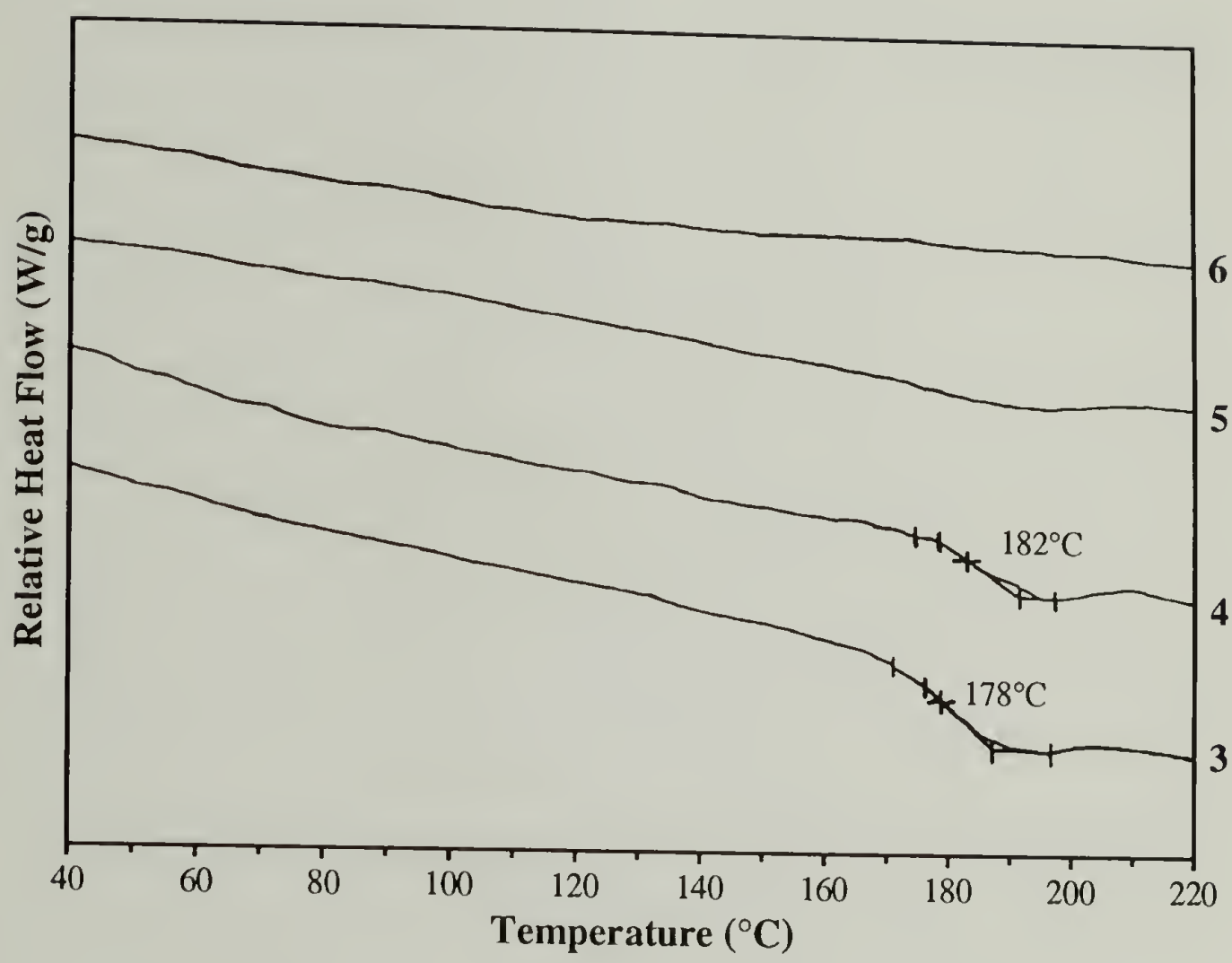


Figure 5.9. Differential scanning calorimetry traces of HCn-16c (n=3-6) proteins. Samples were ramped several times from 50 to 220°C. After the first scan (which indicated the evolution of water) the scans were reproducible (the fourth scan of each protein is shown). Polypeptides **3** and **4** indicate a reproducible glass transition at ~180°C, with no other transitions visible in the range of -40 to 240°C (other regions not shown; see section 5.2.4. for details). There were no transitions evident for **5** and **6** up to the decomposition point (~240°C).

of crystallites. This type of behavior is not unprecedented; Stamhuis and Pennings observed the formation of dense gels upon crystallization of Nylon 6 from dilute solutions of benzyl alcohol.[9].

The attempted sample preparations are summarized in Table 5.2. In general, proteins were dissolved at 20 mg/mL in 1-4 mL of formic acid and the solutions were filtered and stirred with a teflon micro stir-bar in sealed glass vials. All attempts to gel the samples in this fashion failed; the solutions remained clear with no increase in viscosity indicative of gelation. Quenching these solutions by the addition of a ten-fold excess of methanol resulted in the instant formation of a gel that was stable in methanol. The degree of swelling of the gel was observed to decrease from **3** to **5**. These gels were washed repeatedly with methanol and recovered each time by centrifugation at 27,000x g for 20 min. X-ray scattering analysis of these samples showed no significant orientation. No further analyses were performed on these samples.

This method of sample preparation was abandoned with the observation that formic acid solutions of the proteins consistently gelled when stirred in an atmosphere saturated with methanol vapor. Various conditions were tried, varying both formic acid and protein concentration. The best results were obtained for 3 mL of a 20 mg/mL protein solution in 70% formic acid (note that **5** is not soluble directly in 70% formic acid: solutions were prepared by diluting solutions made in 90% formic acid). The solutions were filtered and placed in an open glass vial, and the vial placed in a larger, sealed vial containing a few mL of methanol. The protein solutions were stirred with a teflon stir bar in saturated methanol vapor until 24 hours after the onset of gelation. The results of these trials are outlined in Table 5.3. After gelation, the gel was suspended in an excess of methanol, washed repeatedly with 20 mL methanol portions and stored as a methanol suspension at -10°C for at least 24 hours before further use.

The preparation of gelled samples of **6** was dictated by the behavior of the material upon cleavage of the fusion ends. No matter the conditions attempted, **6** gelled during

Table 5.3. Sample preparation conditions for HCn-16 (n =3, 4, and 5) proteins.

Sample Preparation	Gelation during processing		
	HC3-16c	HC4-16c	HC5-16c
90% Formic acid Stir 7 days	No	No	No
70% Formic acid Stir 7 days	No	No	No
90% Formic acid Stir 7 days w/water vapor	No	No	No
70% Formic acid Stir 7 days w/water vapor	No	No	No
70% Formic acid Stir 7 days w/MeOHvapor	Yes 7 days	Yes 3 days	Yes 2 days
30% Formic acid Stir 4 days w/MeOH vapor	Yes 3 days	—	—

cyanogen bromide cleavage or during solvent evaporation in the work-up. In all cases, the gel was stable to formic acid, water and methanol, and when isolated from methanol solution and dried, was insoluble in virtually all solvents. More importantly, X-ray analysis of oriented samples from all these different preparations were consistent. It was apparent that the material adopted a preferred, insoluble solid-state structure upon removal of the fusion ends. The analyses below were perfomed on a gelled sample of **6** prepared as discussed in section 3.4.6. The majority of the solvent was removed *in vacuo* until a swollen gelatinous mass was obtained. The gel was washed repeatedly with 20 mL aliquots of formic acid, water and methanol and stored as a methanol suspension at -10°C for 24 hours before use in the orientation protocol.

5.3.2. Electron Microscopy of Gelled HCn-16c Samples

The gels were examined using transmission electron microscopy on a Jeol 100CX STEM. The samples were prepared by placing a drop of diluted gel suspension on a carbon coated copper grid and allowing the solvent to evaporate. The micrographs were taken at

20,000X magnification. In the reproductions shown in Figure 5.10 accounting for enlargement of the image, the scale is 1 mm = 18.5 nm.

At lower magnification large, 500 μm size particles were seen to be spread uniformly over the grid in all samples. The images shown in Figure 5.10 are magnified views of the periphery of these large particles. The large aggregates of **3** are seen to be relatively featureless, with a seemingly interlaced network of ill-defined nature. No isolated components of this network are resolved in the image, so it is difficult to assess their relative dimensions. The make-up of the aggregates of gelled **4** is more evident: a network of needle-like particles of average size 0.2 x 0.01 μm are clearly visible on the periphery.

These needle-like particles are better resolved in the images of **5** and **6**, with isolated needles clearly visible. The structure of these fibrils is seemingly well defined, with an average length of 1-2 μm and a constant lateral dimension in the range of 5-10 nm over their entire length. These types of structures, of similar dimension, were seen as the structural component of the nylon 6 gels.[9] In those structures, several of the fibrillar crystals were aggregated into sheaves; the random orientation of the fibrils displayed Figure 5.10 may be a result of their formation with stirring. The nylon 6 crystals could be examined by selected area diffraction, and the chain axis and hydrogen bond directions assigned to the lateral and long axes of the crystallite, respectively. No electron diffraction was possible with these samples because the isolated single fibrils were degraded during irradiation.

It is apparent from observation of the dependence of the gelation behavior on stem-length (Table 5.3 and gelation of **6** during cleavage) and an examination of the structure of the gels, that the notion that gelation occurs upon crystallization is a sound one and the onset of gelation is a reasonable indication of the formation of crystallites in the solution. The gelation of **3** is encouraging in the light of the demonstrated structural organization of gelled samples of **6**. The gelled samples were used in an attempt to obtain oriented X-ray scattering patterns.

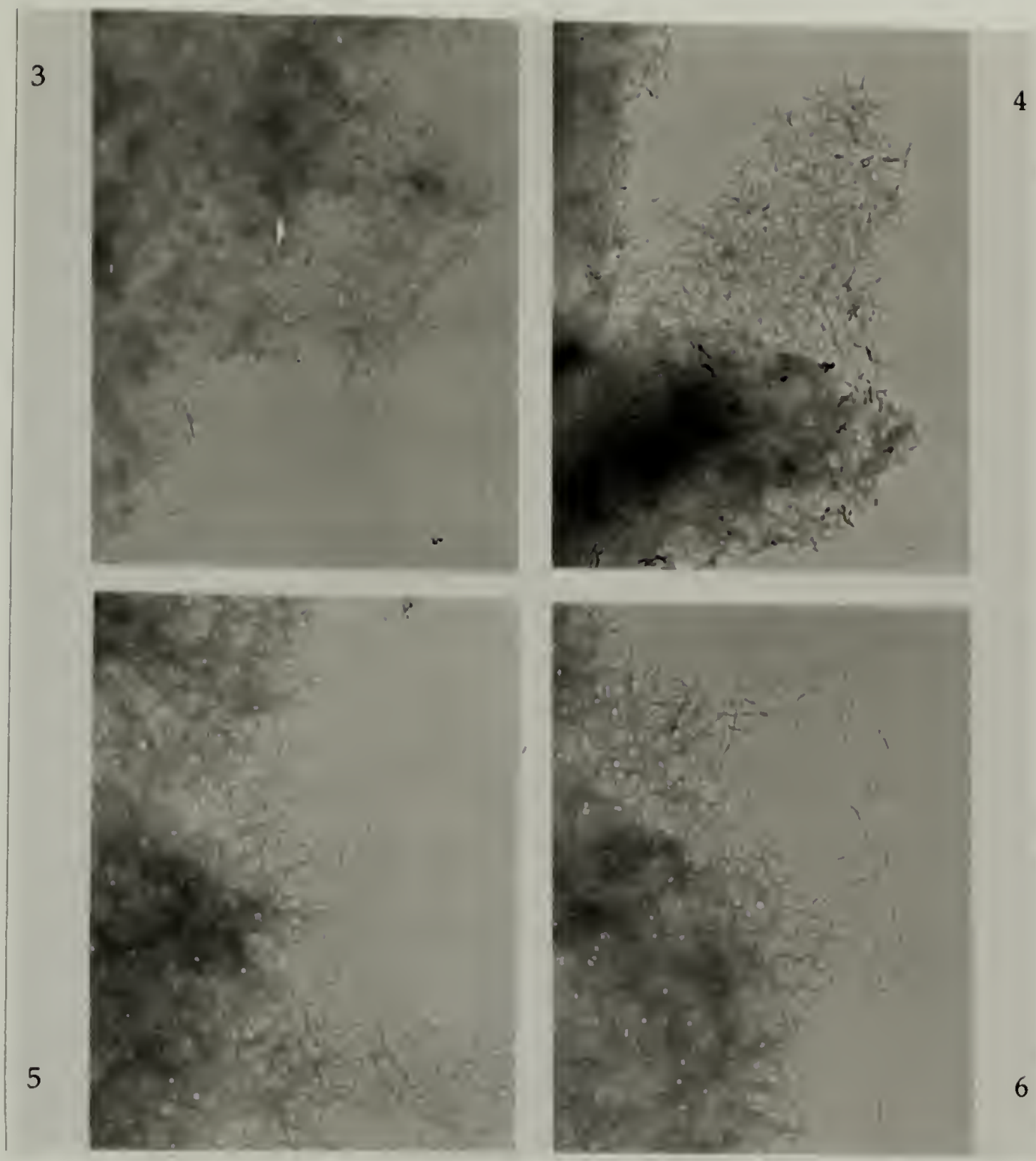


Figure 5.10. Transmission electron micrographs of HCn-16c (n=3-6) protein gel samples. The gelled proteins samples were prepared as described in section 5.3.1. A drop of the gel suspension was placed on a carbon coated copper grid and the solvent allowed to evaporate (see section 5.3.2). At low magnification, large (500 μm) size particles were visible; magnified views of the periphery of these particles are presented (scale: 1mm = 170 Å). In each sample the large aggregates are seen to be comprised of small needle-like particles. In each succeeding sample, beginning with the relatively featureless aggregates of **3**, the network of interlaced needles is better defined, with gelled sample **6** seemingly comprised of well-defined fibrils of average length 1-2 μm and showing a constant lateral dimension in the range of 5-10 nm over their entire length. The significance of these observations, in terms of the relative solid-state order of the samples, is discussed in sections 5.3.2-5.3.5.

5.3.3. Preparation of Oriented HCn-16c Protein Crystal Mats

The technique of obtaining oriented samples of polymer crystallites by sedimentation from solution is well established in studies of the structure of the nylons.[10,11] In general, the growth of the crystallites in the chain axis (*c*) dimension is expected to be limited; the stacking of individual lamellae is slow in the time scale of crystal growth. For lamellar single crystals, growth in the *a* and *b* dimensions is much more rapid and these lengths dominate the crystallite dimensions. Upon sedimentation, such thin, flat crystalites would be expected to settle on the *ab* face, with the chain axis perpendicular to the substrate (parallel with the direction of sedimentation). The backbone separation of packed polypeptide chains in a non-helical arrangement is generally 4.6-4.8 Å, so these strong reflections can be used to assign the hydrogen bonding direction. X-ray diffraction patterns of sedimented crystallites taken with the beam perpendicular to the direction of sedimentation should show this strong interchain signal direction biased on the equator. With this as a basic assignment, the dimensions in other directions can be explored.

The process is outlined in Figure 5.11. A suspension of the polymer gel in methanol was placed in a 1x15 cm glass tube and the gel particles allowed to settle on Whatman #40 filter paper placed on a coarse frit. A constant flow of methanol was maintained by wicking of the solvent from the filter paper. This was continued until the solvent was observed to be clear, then the excess solvent was removed by aspiration using a drawn out glass pipette until the top of the packed gel tablet was exposed. A teflon piston was used to compress the tablet with gentle pressure as a vacuum was applied from below to remove more solvent. The compressed, moist tablet was removed from the tube, placed between teflon filters and pressed between glass plates until dry. The dried tablet was recovered as brittle flakes, several of which were selected from the middle of the compressed tablet and stacked vertically to make a sample approximately 1 mm thick (the direction of orientation).

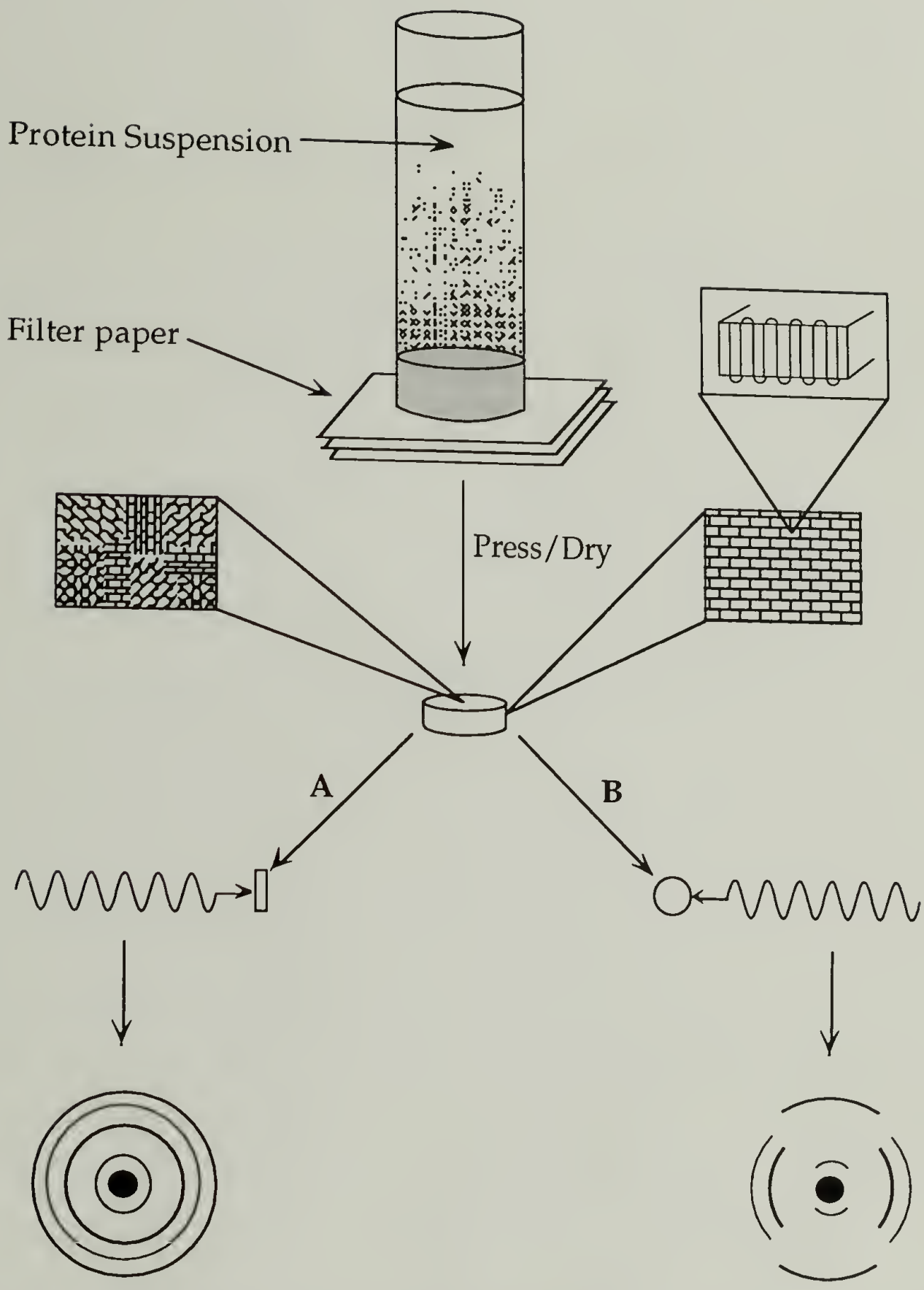


Figure 5.11. Idealized process for preparation and analysis of oriented protein crystal mats. In general for lamellar single crystals, growth in the *a* and *b* dimensions is faster than in the *c* direction. Thus, upon sedimentation from a suspension in a confined geometry (as illustrated), the resulting thin, flat crystalites would be expected to preferentially orient on the *ab* face, with the chain axis parallel to the direction of the orientation. Upon drying, the sedimented tablet shows two distinct orientation axes: *parallel* and *perpendicular* with the sedimentation direction. **A**: X-ray analysis parallel with the sedimentation axis yields a powder diffraction pattern indicative of the random orientation of the *ab* faces of the crystals. **B**: X-ray analysis perpendicular with the sedimentation axis yields a diffraction pattern showing strong interchain signals (hydrogen bond direction) biased on the equater and reflections arising from chain axis spacings on the meridian. This separation of signals facilitates assignment of the chain dimensions, as discussed in section 5.3.3. Adapted from Dougherty.[12]

The sedimentation times for all the samples were on the order of 2-4 hours. Upon charging the apparatus with the gel suspension, the particles were seen to aggregate and begin to settle quickly. In the belief that this fast sedimentation was an impediment to obtaining well oriented samples, a sample of **6** was diluted in methanol and sonicated briefly (2X 5 sec pulse). The result of this treatment was a homogeneous suspension of fine particles that did not settle. Varying the sonication conditions resulted in one sample where a small sedimented mat was observed after 4 days, even though the solvent remained turbid. Analysis of this sample showed no appreciable improvement in the orientation. Other sample treatments were tried, including dilution and the application of a vacuum to the stable suspensions, again with no observation of increased orientation. The samples analyzed below were prepared *via* the protocol outlined in the beginning of this section.

5.3.4. X-ray Analysis of Oriented HCn-16c Protein Mats

X-ray scattering patterns were taken of pressed mat samples, with the beam oriented either perpendicular or parallel to the direction of sedimentation. The patterns are seen in Figures 5.12 (the mat normal is vertical) and 5.13, respectively and the reflections are summarized in Tables 5.4 and 5.5, respectively.

Upon examination of the scattering patterns in Figure 5.12, the level of disorder again appears high, but there is definite bias of intensity both on the equator and the meridian. A comparison of the patterns taken in both direction shows that no new reflections appear relative to the powder patterns in Figure 5.2; all the reflections can be accounted for in one or the other pattern. The spacings in the parallel patterns in Figure 5.13 are exceptionally sharp: the separate 4.7, 4.3 and 3.7 Å spacings are discernible in all patterns, even the characteristically more disordered **3**. Absent from these patterns is the sharp 17.3 Å reflection which is seen with meridional bias in the perpendicular patterns. This low angle spacing is the only completely oriented reflection.

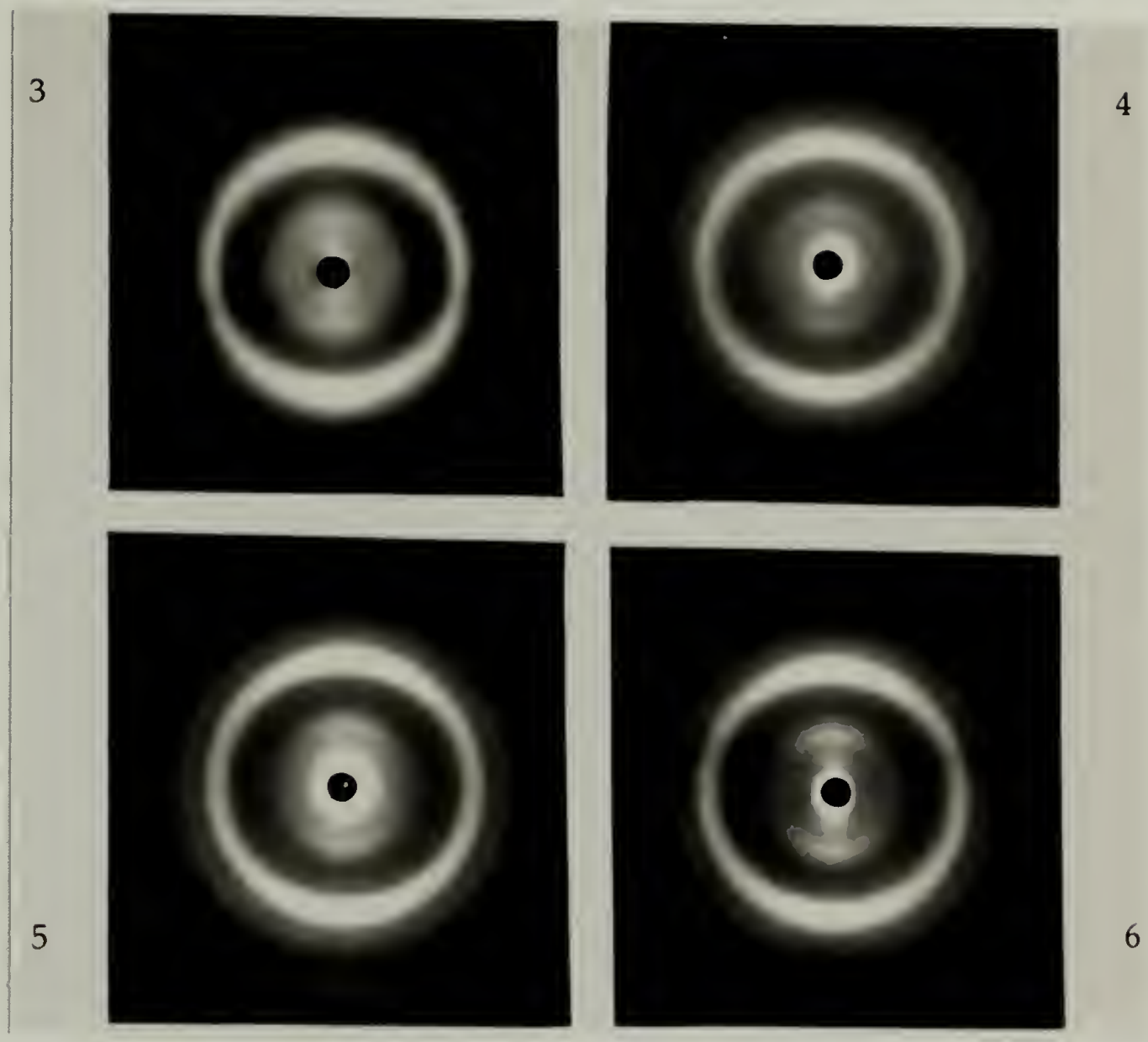


Figure 5.12. Sedimented HCn-16c protein mats: X-ray analysis perpendicular to mat normal. Patterns are presented with mat normal direction vertical. As in the powder patterns (Figure 5.2) there is a general increase in the number and sharpness of the reflections with each succeeding sample. This is best illustrated by examination of the prominent meridional reflection (centered at 4.6 Å) which seen to increase in sharpness from **3** to **6**. The reflections are summarized in Table 5.4 and discussed in section 5.3.3.

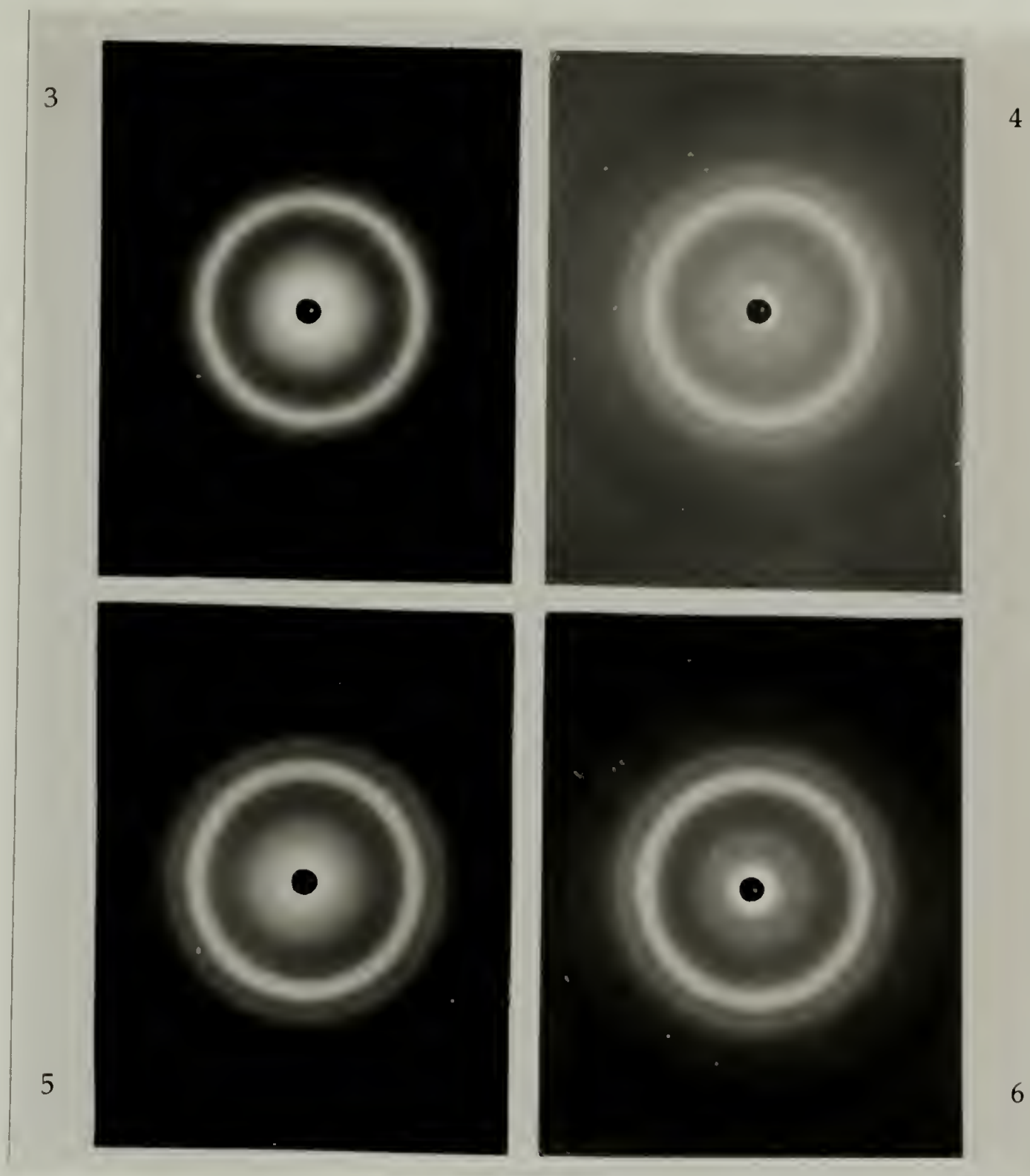


Figure 5.13. Sedimented HCn-16c protein mats: X-ray analysis parallel with mat normal. In general, the signals are sharper than the powder patterns shown in Figure 5.2. The prominent 4.7 and 4.3 Å reflections are sharp and well resolved (though not discernible in the reproduction, these signals are apparent upon examination of the X-ray negatives of **3**). The intensity and sharpness of these signals remains fairly constant in **4**, **5**, and **6**; however, more wide angle signals are discernible in each succeeding pattern. The absence of the 17.3 Å signal, prominent in Figure 5.2 for **5**, and **6**, is striking. The reflections are summarized in Table 5.5 and discussed in section 5.3.3.

Table 5.4. Sedimented HCn-16c protein mats: X-ray analysis perpendicular to mat normal.

d Spacing (Å)			
HC3-16c	HC4-16c	HC5-16c	HC6-16c
Meridional Reflections			
9.2±1.2 m, br, df 4.5±.4 vs, br	Beam Stop* 9.3±1.1 m, br, df 4.6±0.3 vs, br 3.0 w, df	Beam Stop* 9.2±1.1 s, df 4.6±0.3 vs, br 3.0 m, df	17.3±0.2 s, br 9.2±1.2 s, df 4.6±0.3 vs, br 3.0 m, df
Equatorial Reflections			
4.5±0.2 vs, br	4.5±0.2 vs, br 3.7 w, df	4.7 s, sh 4.3 vs, sh 3.7 m, sh 3.0 vw, sh	4.7 s, sh 4.3 vs, sh 3.7 m, sh 3.0 vw, sh
Other Reflections†			
		2.3 w, df 2.1 w, df	2.3 w, df 2.1 w, df

Intensity: very weak (vw), weak (w), med (m), strong (s), very strong (vs).
Appearance: sharp (sh), broad (br), diffuse (df). Broad reflections are given with limits.
*The presence of 17.3 Å reflections is suggested, but obscured by beam stop scattering.
†Reflections of indeterminate orientation.

Table 5.5. Sedimented HCn-16c protein mats: X-ray analysis parallel with mat normal.

d Spacing (Å)			
HC3-16c	HC4-16c	HC5-16c	HC6-16c
~9.2 w, df 4.5±0.2 vs, br* 3.7±0.2 w, br, df	~9.2 w, df 4.7 s, sh 4.3 vs, sh 3.7±0.1 m, br, df 2.9 vw, df	~9.2 w, df 4.7 s, sh 4.3 vs, sh 3.7 m, sh 2.9 w, br 2.3 vw, br 2.1 vw, br	~9.2 w, df 4.7 s, sh 4.3 vs, sh 3.7 m, sh 2.9 w, br 2.3 w, br 2.1 w, br

*Sharp 4.7 and 4.3 Å reflections are suggested but not well resolved on the reproduction.

As in the powder patterns, the general appearance of the X-ray scattering intensity is consistent in all samples, with a significant increase in sharpness of the reflections from **3** to **4**, and a gradually sharpening and increase in intensity of the patterns of the longer stem-length samples. The reflections seen in the oriented patterns can be summarized as follows:

•The low angle spacing at 17.3 Å appears only in the analysis perpendicular to the mat normal, arrayed on the meridian as a diffuse reflection of medium intensity. There is a component of the 9.2 Å reflection in both directions, with the intensity biased on the meridian in the perpendicular patterns and appearing as a diffuse ring of weak intensity when analyzed parallel with the mat normal.

•The most prominent reflection in the patterns in either orientation appears at 4.6 Å on the meridian of the pattern taken perpendicular to the mat normal. Two sharp, medium intensity spacings appear on the equator at 4.7 and 4.3 Å; rings with the same spacings and relative intensities appear in the parallel patterns.

•A weak, equatorial, perhaps off-diagonal reflection at 3.7Å is seen in the patterns taken perpendicular to the mat normal. This spacing is seen as a ring of similar appearance and intensity in the parallel orientation.

•A diffuse reflection appears on the meridian in the mat-perpendicular pattern at 3.0 Å: the corresponding ring in the pattern taken in the opposite direction is extremely weak. The wider angle reflections at 2.3 and 2.1 Å appear as very weak signals in both directions, with no appreciable orientation apparent.

This analysis is similar to the X-ray and FTIR analysis of the powder samples, in that there is a consistent change in the observed patterns with increasing stem-length; all of

the patterns are related in some systematic fashion, witnessed by a gradual increase in the intensity of the observable reflections.

5.3.5. X-ray Analysis of Annealed Oriented HCn-16c Protein Mats

In an attempt to improve the quality of the orientation, the dry, matted flakes were annealed between teflon sheets for 30 min in an autoclave at 110°C and 110 kPa in the presence of water vapor.

This treatment was not feasible with **3** as the matted flakes were water soluble and the flakes dissolved. The X-ray scattering patterns perpendicular to the sedimentation direction were recorded and are shown in Figure 5.14. The appearance of the patterns of **5** and **6** is essentially unchanged from those of the unannealed samples in Figure 5.12. No new reflections or changes in orientation or intensity are evident.

The interesting and most significant result of the annealing experiment is the general sharpening of the pattern of **4**. The intense 4.6 Å meridional reflection is much sharper, and the resolution of the 4.7 and 4.3 Å equatorial reflections is increased. The pattern reveals no new reflections, only a general sharpening of the reflections is seen. It is thought that this sharpening is a result of the heat, pressure and humidity in the autoclave causing a glass transition in the amorphous phase of the material. This increase in conformational freedom allowed some rearrangement of the chains participating in the ß-sheet, giving rise to the sharpened X-ray pattern. The absence of an accessible glass transition for **5** and **6** made such rearrangement impossible, witnessed by the similar patterns of the annealed and unannealed samples.

5.3.6. FTIR Analysis of HCn-16c Protein Mats

In order to confirm the evidence for the presence of an antiparallel ß-sheet structure in the oriented samples, the dried mat flakes were ground, prepared as pressed KBr pellets and analyzed by FTIR spectroscopy.

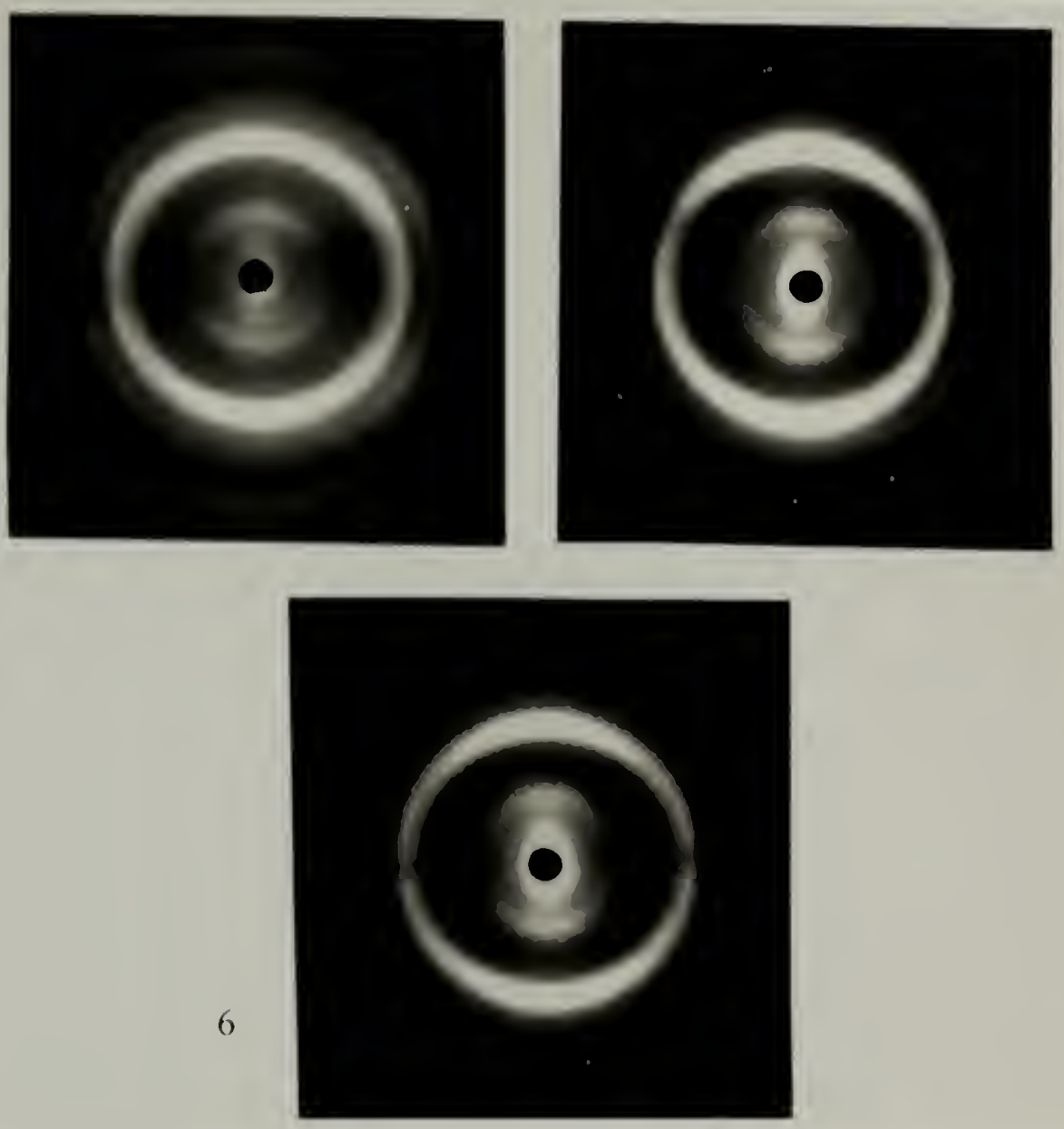


Figure 5.14. X-ray analysis of annealed, sedimented HCn-16c protein mats. Samples were annealed in an autoclave in an attempt to improve the orientation of the chains. Patterns were taken with the X-ray beam oriented perpendicular to sedimentation direction and are presented with mat normal direction vertical. Analysis of **3** was not possible, as it dissolved during annealing. There is no substantial change in the patterns of **5** and **6** upon annealing: the patterns seen here are essentially identical with those of the unannealed samples shown in Figure 5.12. The pattern of **4** does show considerable improvement in the sharpness and resolution of the X-ray reflections with annealing. The prominent 4.6 Ä meridional signal is much sharper and the 4.7 and 4.3 Ä equatorial reflections are sharper and more well resolved than the corresponding reflections in the unannealed sample. The significance of these results is discussed in section 5.3.4.

In general, the spectra for **4**, **5**, and **6** were unchanged from the spectra discussed in Section 5.2.4. To illustrate this, a comparison of the carbonyl region from the powder and mat spectra of **4** is shown in Figure 5.15. This comparison is representative of the appearance of the same spectra of **5** and **6**. In fact, this is a general result from this analysis: the adoption of a well defined ß-structure is well established in the infrared spectra of all preparations of **4**, **5**, and **6**. Controlled crystallization, matting and annealing may affect the relative orientation and intensities of the X-ray reflections, but there is no significant change in the pattern of absorbances associated with the ß-sheet structure.

There is, however, significant sample preparation dependence in the sharpness of the absorbances in the infrared spectrum of **3**. To illustrate this, a comparison of the carbonyl regions of the spectra of amorphous sample **3A** ("Amorphous" spectrum), the dry powder sample as recovered after CNBr treatment ("Dry" spectrum), this sample after hydration ("Wet" spectrum) and the dried, oriented mat of **3** ("Mat" spectrum) is shown in Figure 5.16. The only sample not indicating a ß-sheet structure is amorphous **3A**, which was intentionally isolated in a disordered state. The adoption of a ß-sheet structure in the powder samples is supported by the amide II vibration at 1524 cm^{-1} and the amide I shoulders at 1703 and 1627 cm^{-1}. Upon gelation, the antiparallel ß-sheet structure is seen to dominate the amide I region. This general sharpening of the absorbances with processing is reflected in the appearance of some orientation of the intense reflections seen in Figure 5.12 and the (relatively) sharp X-ray powder rings for this material seen in Figure 5.13. A comparison of the carbonyl region for matted samples **3** and **4** is seen in Figure 5.17.

5.4. Conclusion

All of the data presented in this chapter supports the premise that the stem-length *n* of copolypeptides of sequence **5.1** controls the solid-state architecture adopted by the chains upon crystallization. Two different sets of samples were examined in this study that can be loosely designated unprocessed (powder samples) and processed (oriented mats).

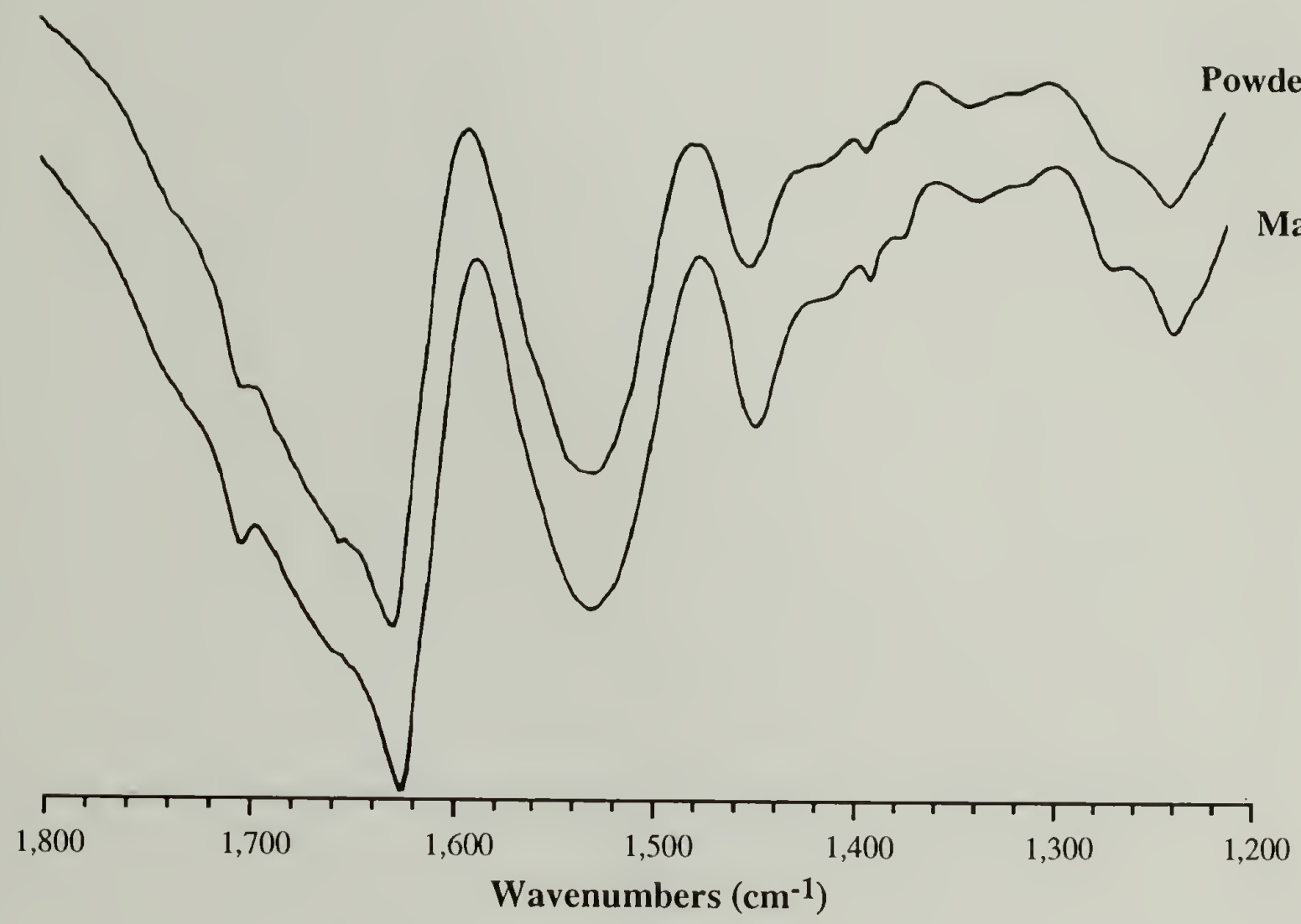


Figure 5.15. FTIR analysis of different preparations of HC4-16c. The infrared spectrum of powder samples used in the X-ray analysis shown in Figure 5.2 ("Powder" spectrum) and the sedimented mats ("Mat" spectrum) (the X-ray patterns of these samples are shown in Figures 5.12 and 5.13) are presented to illustrate that the signals associated with the presence of the ß-sheet are present in all the preparations analyzed: the peak patterns for the two samples are virtually identical. This observation is true for the corresponding preparations of **5** and **6**. The spectra are discussed in section 5.3.5.

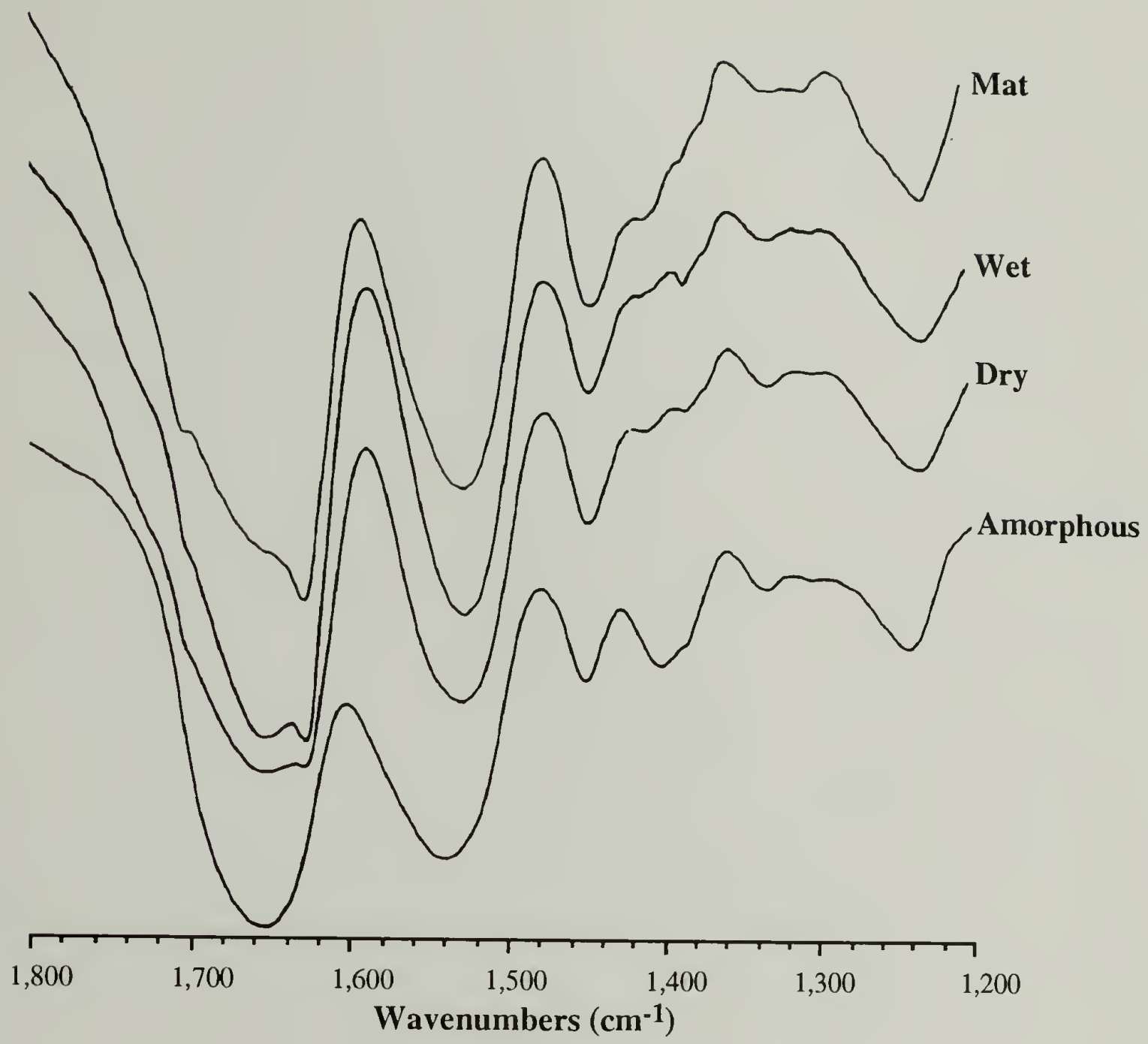


Figure 5.16. FTIR analysis of different preparations of HC3-16c. The infrared spectrum of **3** is intimately dependent on sample preparation. The bottom spectrum ("Amorphous" spectrum) is that of amorphous sample **3A** discussed in section 5.2.2: this sample shows only the 1657 and 1540 cm^{-1} absorbances associated with a polypeptide chain in a random structure. [Note the increased intensity (relative to the other spectra) of the peaks at 1405 and 1375 cm^{-1}, associated with asymmetric CH_2 wagging and the CH_3 symmetric bending]. The analysis of **3** as recovered after cyanogen bromide cleavage ("Dry" spectrum) shows clear indication of the 1703 and 1627 cm^{-1} amide I, and 1525 cm^{-1} amide II signals indicative of the presence of an antiparallel ß-sheet. Upon hydration ("Wet" spectrum), this pattern remains, with the 1627 cm^{-1} peak sharpened somewhat. Upon gelation from formic acid solution, sedimentation and drying ("Mat" spectrum), the characteristic pattern of the ß-sheet structure is firmly established.

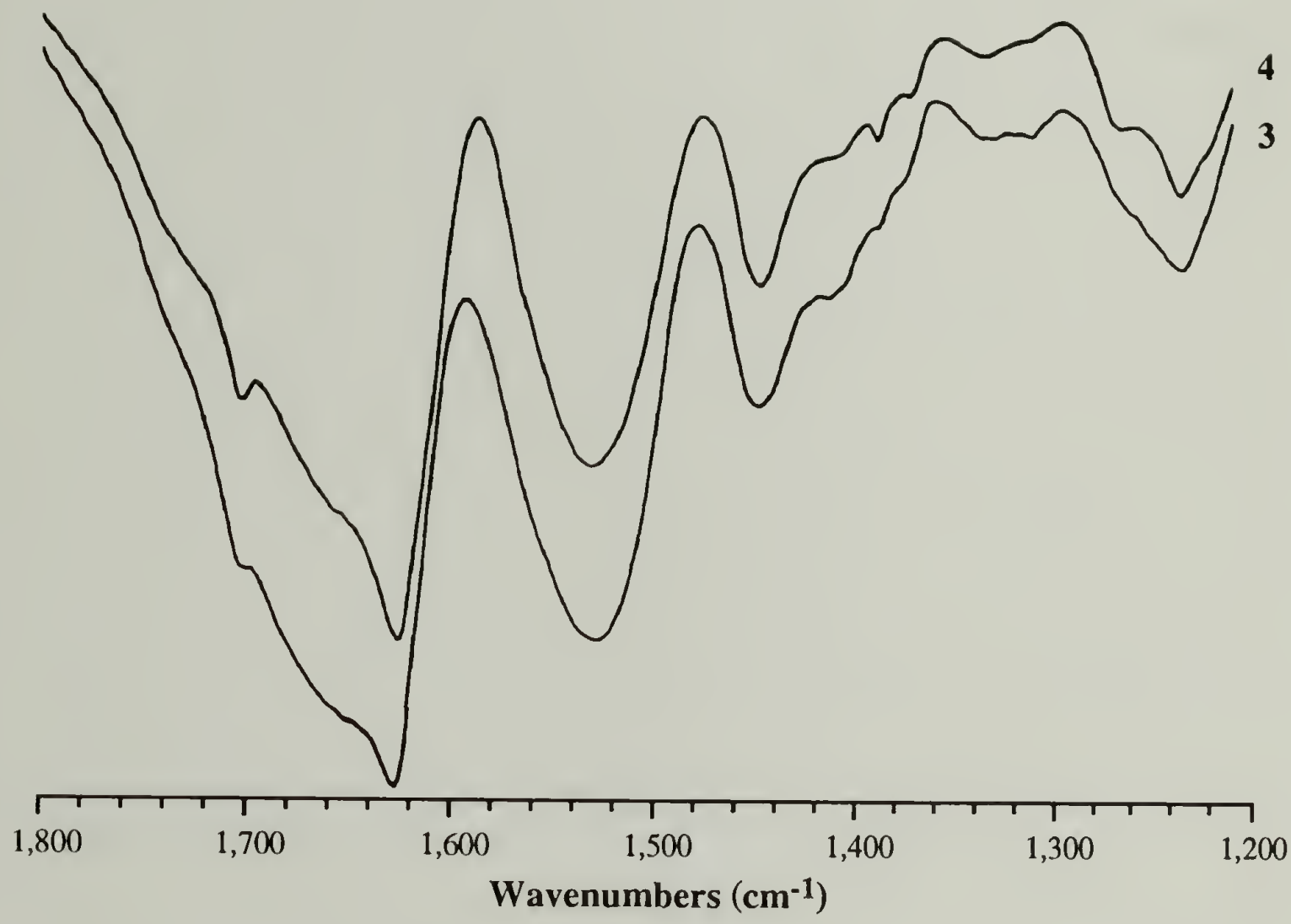


Figure 5.17. FTIR analysis of dried mats of HC3-16c and HC4-16c. The spectra of the two samples are similar, with the antiparallel ß-sheet signature firmly established in each. The spectra of mats of **5** and **6** are similar to **4**. From these data it is asserted that upon crystallization from formic acid, the ß-sheet structure is the preferred conformation of all of the HCn-16 proteins. FTIR is a more sensitive probe of the presence of ß-sheet structure than X-ray analysis: the appropriate ß-sheet reflections are absent in the X-ray scattering patterns of **3**. Thus the systematic increase in the sharpness and intensity of the X-ray reflections of the HCn-16c protein mat samples, from **3** to **6** (section 5.3.3), arises from an increase in the dominance of the ß-sheet in the overall solid-state structure.

The powders examined in Section 5.2 were prepared with no special attention to achieving a particular solid-state order; the structure was that obtained from the work-up after removal of the fusion ends, not from particular processing. The FTIR spectra show the presence of antiparallel ß-sheet in all samples. The only prominent difference in the spectra is the relatively larger 1657 cm^{-1} absorbance associated with random structure indicated by **3**. The X-ray patterns corroborate this, revealing reflections consistent with those of materials known to form ß-sheets. There is a large increase in the sharpness and intensity of the reflections from **3** to **4**, which continues in a systematic fashion in the patterns of the longer stem-length materials. Thus related, the patterns of each can be interpreted in terms of the reflections indicative of the ß-sheet seen in the pattern of HC6-16c. The thermal analysis shows that **3** and **4** are related by the indication of a glass transition temperature; **5** and **6** show no such transition. This analysis serves to underscore the relationship between the materials, with the observed crystalline order in the powder patterns of the longer stem-length polymers consistent with the absence of a discernible T_g in **5** and **6** . An assignment of the relative percent crystallinity from the diffraction curves may support this contention. This awaits a more rigorous treatment.

In the processed samples, those crystallized and oriented, the relative trend is similar, except the solid-state order of **3** has been improved relative to the other samples. The FTIR spectra for the dried protein mats of **4**, **5**, and **6** were essentially unchanged from those of the unprocessed samples. In the dried mat samples of **3**, however there was a significant sharpening and resolution of the split amide I vibrations associated with the adoption of an antiparallel ß-sheet. In addition, the X-ray analysis correlates with the observed increase in solid-state order for processed **3**: taken parallel with the mat normal, the patterns show relatively sharp rings and when analyzed perpendicular to the mat normal, there is a slight meridional bias of the intense reflections unobserved in the unprocessed samples.

Processing, therefore, brings the solid-state order of **3** in-line with the systematic increase observed for the longer stem-length materials. In other words, the relatively larger "order gap" between unprocessed **3** and **4**, compared to the relative increase in order observed for unprocessed **4**, **5**, and **6**, is narrowed by processing. Thus the relative increase in order of the processed samples, witnessed by the increase in sharpness and intensity of the associated X-ray reflections, is now consistent with the increase in stem-length over all four materials. In this light, the results can be summarized as follows:

•With the exception of the deliberately prepared amorphous samples **3A**, the infrared spectra of all preparations indicate the presence of an antiparallel ß-sheet.

•Infrared and X-ray analysis of the unprocessed powders support the existence of two relative structural "transitions": a large jump in order from **3** to **4**, then a gradual increase from **4** to **6**. This notion is supported by comparison of the shape of the amide I vibrations in the infrared spectrum (Figure 5.8) and the general appearance and intensity of the X-ray reflections (Figure 5.2).

•The thermal analysis (Figure 5.9), the hydration experiment (Figure 5.6) and the mat annealing experiment (Figure 5.14) support the presence of another, separate relative "transition": a change from amorphous phase dominated material properties in **3** and **4** to behavior dominated by a crystalline phase in **5** and **6**.

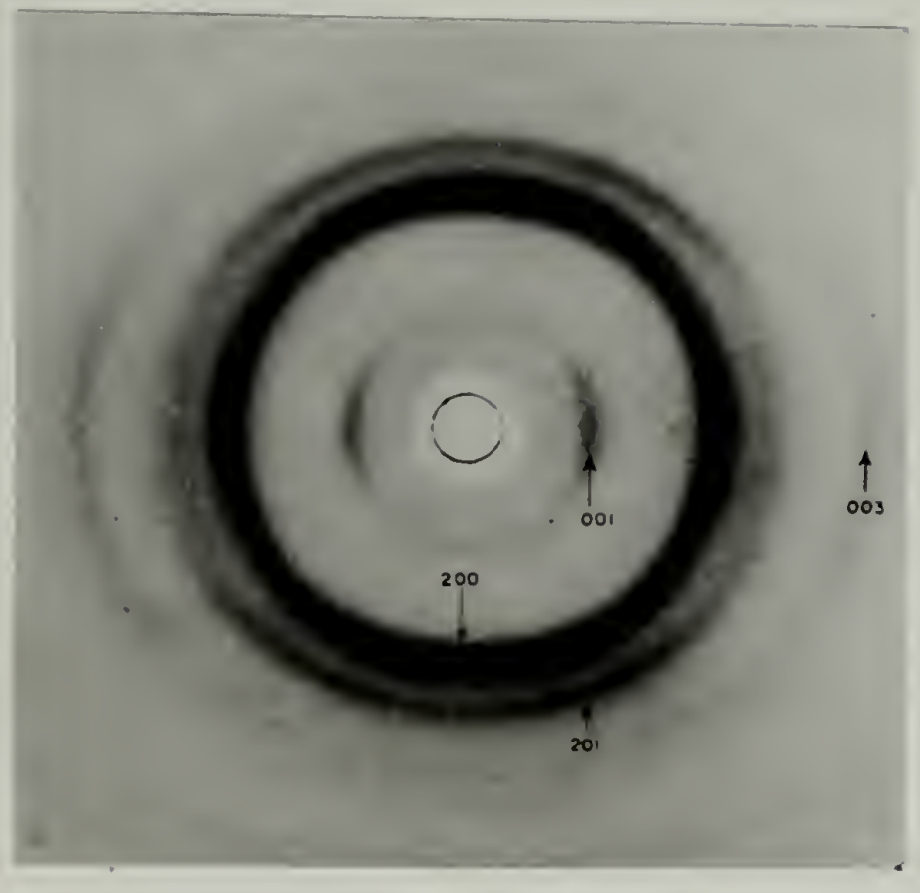
•The structural analysis of the processed HCn-16 polypeptides is consistent with a monotonic increase in the relative content of ß-sheet structure with increasing stem-length: the presence of the ß-sheet in *all* materials is supported by appropriate signals in the infrared spectrum (Figure 5.8). The relative increase in the content of ß-sheet *between* materials with increasing stem-length is supported by the

consistent sharpening and increase in intensity of the diagnostic X-ray reflections in the correponding patterns (Figure 5.2 versus Figures 5.12 and 5.13).

•There is no indication of long range order from the assembly of any structures in a superlattice: the smallest angle reflection observed at 17.3 Å seems to be stem-length independent.

In a general sense then, the stem-length premise is validated: there is a consistent dependence of the solid-state structure adopted by polypeptides of sequence **5.1** on the number of alanylglycine dyads in the sequence. However, there is no concrete evidence for chain folding in the form of low angle spacings suggestive of stacked lamellae.

Nevertheless, it is difficult to dismiss the possibility that the structures are, in fact, chain-folded. A comparison of orientation of the X-ray spacings of representative sample HC6-16c with those of poly(L-alanylglycine)[4] are found in Table 5.6. For each major reflection in the pattern of PLAG I there is a correponding signal in the pattern of HC6-16c. The only exception is the the low angle spacing at 17.3 Å seen in the pattern of HC6-16c and not observed for poly(L-alanylglycine). Direct comparison of the patterns, seen in Figure 5.18, reveals that the PLAG I pattern is exactly orthogonal to that of HC6-16c; i.e. a reflection found on the meridian in the PLAG I pattern is seen on the equator in the HC6-16c pattern. Since the oriented samples were obtained *via* different methods—PLAG I by stroking and HCn-16c by sedimentation—it is not unlikely that the chains align differently upon orientation. Since the chains of PLAG I are known to be in a cross-ß conformation[13] (aligned normal to the direction of orientation), the fact that the PLAG I pattern appears to be exactly orthogonal to that of HC6-16c is consistent with the observation that chain-folded lamella preferentially orient with the chains parallel with the direction of sedimentation. The observed differences in signal intensities (Table 5.6) between the two patterns may be an artifact of this difference in orientation.



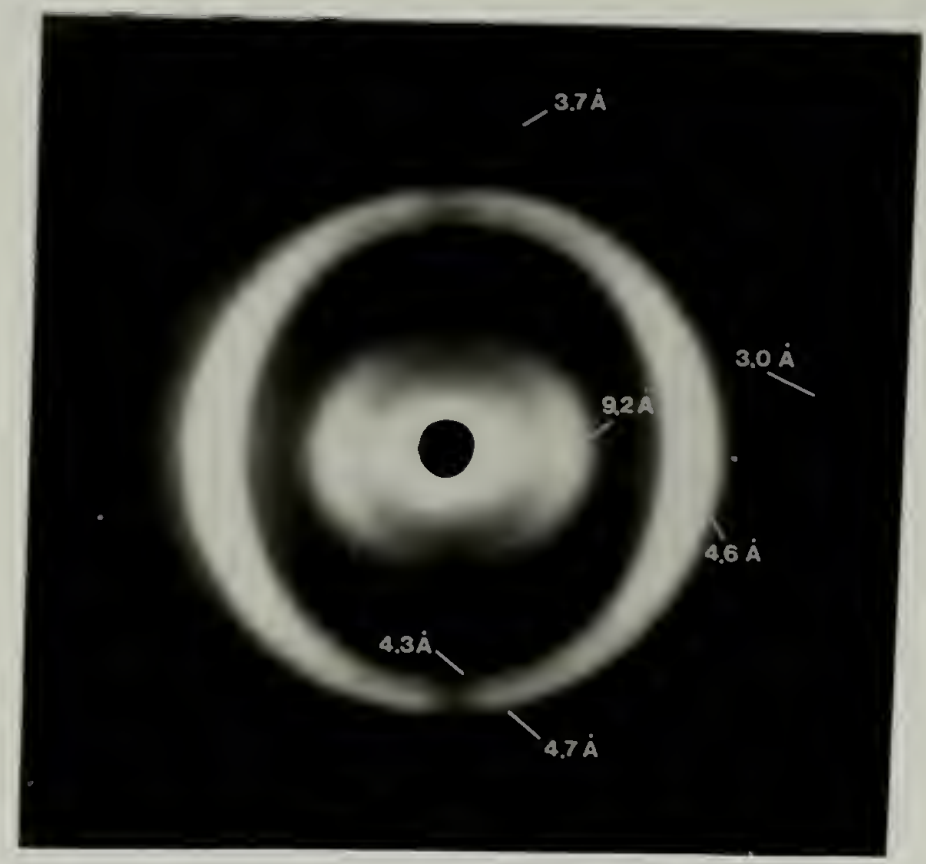


Figure 5.18. X-ray scattering patterns of oriented PLAG I and HC6-16c. **A**: The wide-angle scattering pattern of poly(L-alanylglycine) in Form I, reprinted from Fraser and coworkers[4], with permission from Academic Press. The assignment of prominent lattice spacings is indicated and summarized in Table 5.6. [Note that the reflections listed in the reproduction of the PLAG I X-ray pattern are those originally presented by Fraser and coworkers, who assigned *b* as the chain axis for PLAG I. However, the reflections listed in the text and Table 5.6 are consistent with the assignment of the *c* as the chain axis: ie, the 001 reflection indicated on the reproduction of the PLAG I pattern is referred to as the 010 in the text, etc.] **B**: The wide-angle scattering pattern of dried sedimented mats of HC6-16c. The reflections corresponding to those seen in the pattern of PLAG I (as summarized in Table 5.6) are indicated. The pattern is presented with the mat normal horizontal, to correspond with orientation of PLAG I. The observed differences in the intensities of corresponding reflections between the two materials may be a consequence of the difference in orientation direction. The significance of the similarity of the patterns is discussed in section 5.5.

Table 5.6. Comparison of X-ray data of oriented PLAG I and HC6-16c.

PLAG I[4]		HC6-16c
Equatorial		Meridional
8.87, m,	(010)*	9.2±1.2 m, df
4.45, vs	(020)	4.6±0.3 vs, br
2.97, m	(030)	3.0 w, df
Meridional		Equatorial
4.68, m	(200)	4.7 s, sh
4.22, vs	(210)	4.3 s, sh
3.76, m	(120)	3.7 m, sh

* Note that Fraser and coworkers indexed the unit cell for PLAG I with *b* as the chain direction. These reflections are reported with the now more common assignment of *c* to the chain direction. In addition, the values of the initially reported reflections have been slightly modified, as reflected in the description of the unit cell reported by Fraser and MacRae and listed below.

As a review, the structure of PLAG I is indexed on an orthogonal unit cell 9.44 Å in the hydrogen bond (*a*) direction, 8.96 Å in the intersheet (*b*) direction and 6.94 Å along the chain axis (*c* direction). The chains form extended sheets decorated on only one side with alanyl methyls; the sheets pack with alanyl-alanyl face and glycine-glycine face contacts. This determines the observed relative intensities of the 010 and 030 reflections. The intersheet spacing is partitioned into a b_{Gly} distance of 3.79 Å and a b_{Ala} distance of 5.17 Å. The 210 and 200 reflections are interpreted in terms of a *a*/4 stagger between sheets.[4]

Is there a structure that is consistent with the exact correlation of the X-ray scattering data of HC6-16c with the PLAG I structure and can be rationalized in terms of chain folding? In fact, such a structure can be envisioned—comprised of extended sheets of the kind shown in Figure 5.19, which are decorated on only one side with alanyl methyls. Similar sheets pack with glycine-glycine face and alanine-alanine face contacts: the constraints of packing the relatively congested fold surfaces giving rise to the slight

increase in the observed 0*k*0 spacings of HC6-16c over those of PLAG I. The *a*/4 sheet stagger indicated by the (200) and (210) reflections is also reasonable.

The trends seen in the solid-state analysis of the HCn-16c proteins can also be rationalized with such a structure. The systematic increase in the ß-sheet content with stem-length is consistent with the increasing predominance of the bulk sheet structure over the relatively disordered turn structures: these disordered phases undergo a thermal transition in the materials with the lowest ratios of turn to sheet structure (**3** and **4**). The ß-sheet content of these structures is not sufficient to "lock" the disordered phases into a particular conformation. Conversely, the increased disorder allows sufficient freedom in **4** that sharpening of the bulk structure occurs upon annealing.

The structure shown in Figure 5.19 is the same figure used in Chapter 1 to illustrate the effect of an odd number of amino acids on the trajectory of a chain, postulating equivalent turn structures in each stem-turn repeat. The equivalent ß-turn structures used in that model directed the chains on a spiral trajectory. However, the bulk sheet structure may stabilize other equivalent turn structures in each repeat with a conformation that allows the correct chain trajectory and distributes the mismatched hydrogen bond pair equally in each stem-turn. The inherent disorder in the fold surface may prevent the observation of the low angle spacings from lamellar stacking; the observed 17.3 Å reflection may be from some long repeat across the turn structures.

What if the turns are not equivalent? The repetitive structure could be envisioned as encompassing *two* repeats of sequence **5.1** in the order [stem-turnA-stem-turnB]. Turn A could be a ß-turn with normal dihedrals and turn B could be some forced turn that allows the correct reentry of the chain with the strategic placement of the odd hydrogen bonding pair. The differences in the fold surfaces could explain the lack of low angle superlattice repeats—the random alignment of the disparate faces precluding any periodic low angle scattering in the *c* direction. Once again, the sharp 17.3 Å reflection could be indicative of some repeat within the fold surfaces.

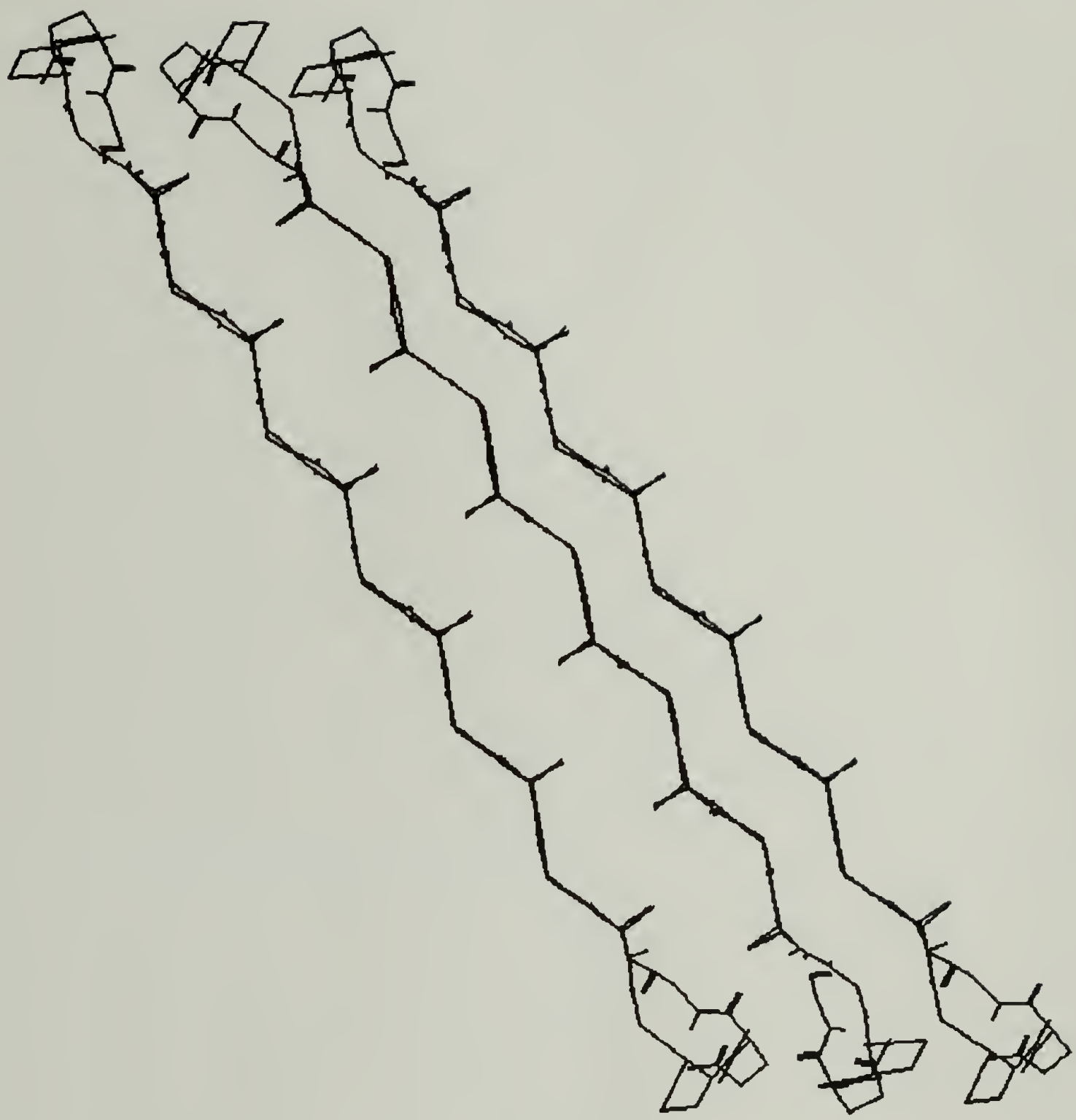


Figure 5.19. Proposed structure of sheets of folded HC6-16c chains. The chain-folded sheets have alanyl methyls decorating one face, and adjacent sheets align with alternating alanine-alanine and glycine-glycine face contacts. The structure of the bulk ß-sheet (excluding the fold surfaces) is that of PLAG I[4]: the demonstrated similarity of the X-ray scattering patterns of HCn-16c and PLAG I reinforces the proposed structure. The nature of the turn structure is unspecified; however, the disorder of the fold surface relative to the bulk ß-sheet structure is emphasized. The proposed structure is discussed in section 5.5. The structure was generated on a Silicon Graphics work station using Polygraf software. The sheets are presented tilted merely for aesthetics; no actual chain tilt is implied.

It is difficult to rationalize the close correlation of the intersheet spacing of the HCn-16 proteins with that of PLAG I in terms of placement of two bulky groups, proline and glutamic acid, in the extended sheets. The 010 and 030 reflections have been interpreted for the structure of PLAG I in terms of the spacings of sheets with alternating Gly-Gly and Ala-Ala contacts.[4] The presence of those reflections supports a similar sheet arrangement adopted by the HCn-16c proteins. The presence of bulky residues glutamic acid or proline should affect the spacing of the sheets, thus changing the spacings of the X-ray reflections. This notion is consistent with the demonstrated exclusion of proline from the bulk structure of ß-sheets (a ß-sheet breaker).[14]

The refinement of the exact structure adopted by the HCn-16c polypeptides awaits the preparation of better oriented samples and more extensive and quantitative modeling of the crystallographic data. However, the data stongly suggests the adoption of a ß-sheet stucture identical with that exhibited by poly(L-alanylglycine). The presence of such a structure can be rationalized in terms of chain folding as discussed above. Rigorous analysis of the absolute structure should only serve to reinforce the stem-length premise developed in this study.

5.5. References

1. K. P. McGrath, personal communication, (1991).
2. M. T. Krejchi, personal communication, (1992).
3. R. D. B. Fraser and T. P. MacRae, *Conformation in Fibrous Proteins and Related Synthetic Polypeptides* (Academic Press, New York, 1973),p. 326.
4. R. D. B. Fraser, T. P. MacRae, F. H. C. Stewart, E. Suzuki, *J. Mol. Biol.* **11**, 706 (1965).
5. W. H. Moore and S. Krimm, *Biopolymers* **15**, 2465 (1976).
6. T. Miyazawa and E. R. Blout, *J. Am. Chem. Soc.* **83**, 712 (1961).
7. R. D. B. Fraser and T. P. MacRae, p. 120.
8. Y. Deguchi, personal communication, (1992).

9. J. E. Stamhuis and A. J. Pennings, *Br. Polym. J.* **10**, 221 (1978).

10. P. Dreyfuss and A. Keller, *J. Macromol. Sci., Phys.* **B4**, 811 (1970).

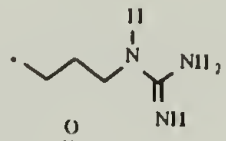
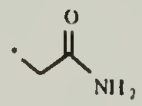
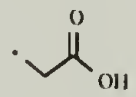
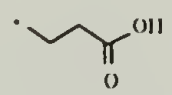
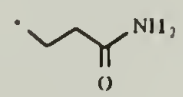
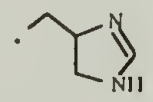
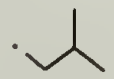
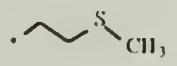
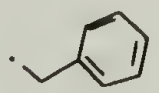
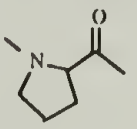
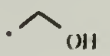
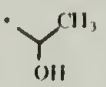
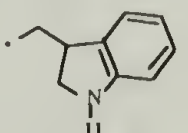
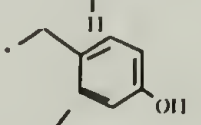
11. E. D. T. Atkins, A. Keller, D. M. Sadler, *J. Polym. Sci., Polym. Phys. Ed.* **10**, 863 (1972).

12. M. Dougherty, personal communication, (1992).

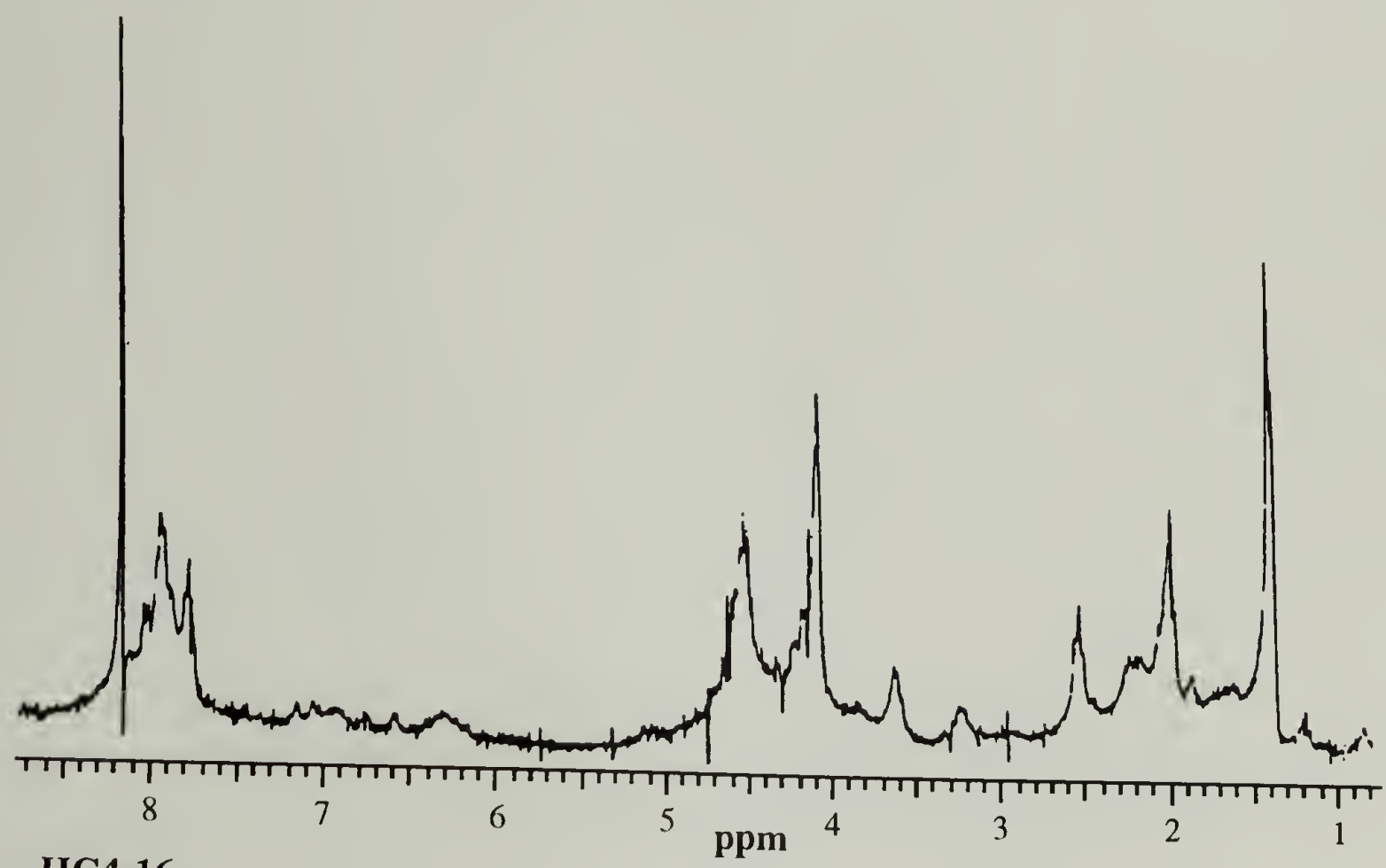
13. R. D. B. Fraser and T. P. MacRae, p. 239.

14. J. S. Richardson and D. C. Richardson, in *Prediction of Protein Structure and the Principles of Protein Conformation* G. D. Fasman, Ed. (Plenum Press, New York, 1989).

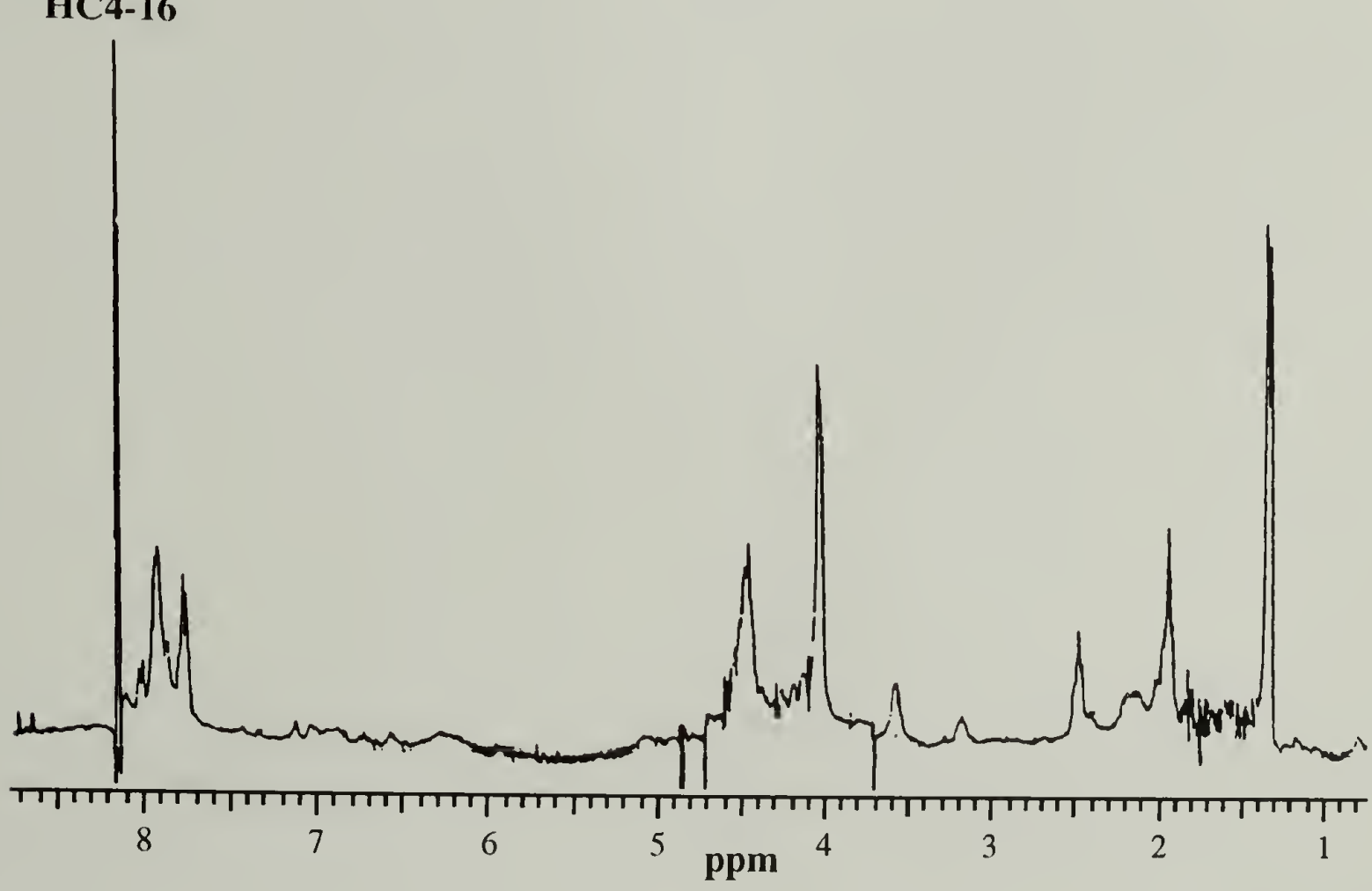
APPENDIX A: AMINO ACID DATA

Amino Acid	Code		Residue Mass	Side-chain Structure
Alanine	Ala	A	71.079	
Arginine	Arg	R	156.188	
Asparagine	Asn	N	114.104	
Aspartic Acid	Asp	D	115.089	
Cysteine	Cys	C	103.144	
Glutamic acid	Glu	E	129.116	
Glutamine	Gln	Q	128.131	
Glycine	Gly	G	57.052	
Histidine	His	H	137.142	
Isoleucine	Ile	I	113.160	
Leucine	Leu	L	113.160	
Lysine	Lys	K	128.174	
Methionine	Met	M	131.198	
Phenylalanine	Phe	F	147.177	
Proline	Pro	P	97.117	
Serine	Ser	S	87.078	
Threonine	Thr	T	101.105	
Tryptophan	Trp	W	186.213	
Tyrosine	Tyr	Y	163.17	
Valine	Val	V	99.133	

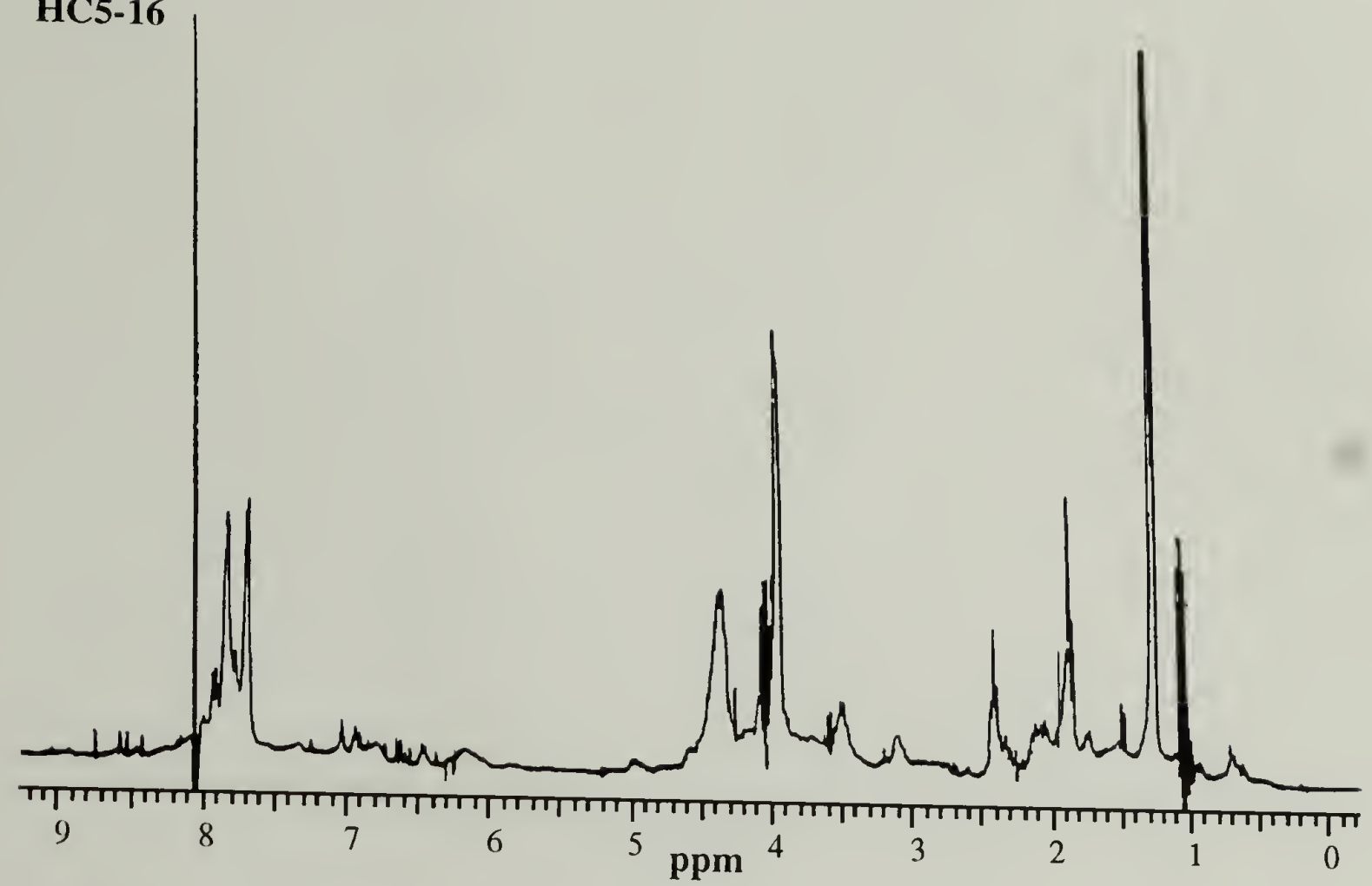
HC3-16



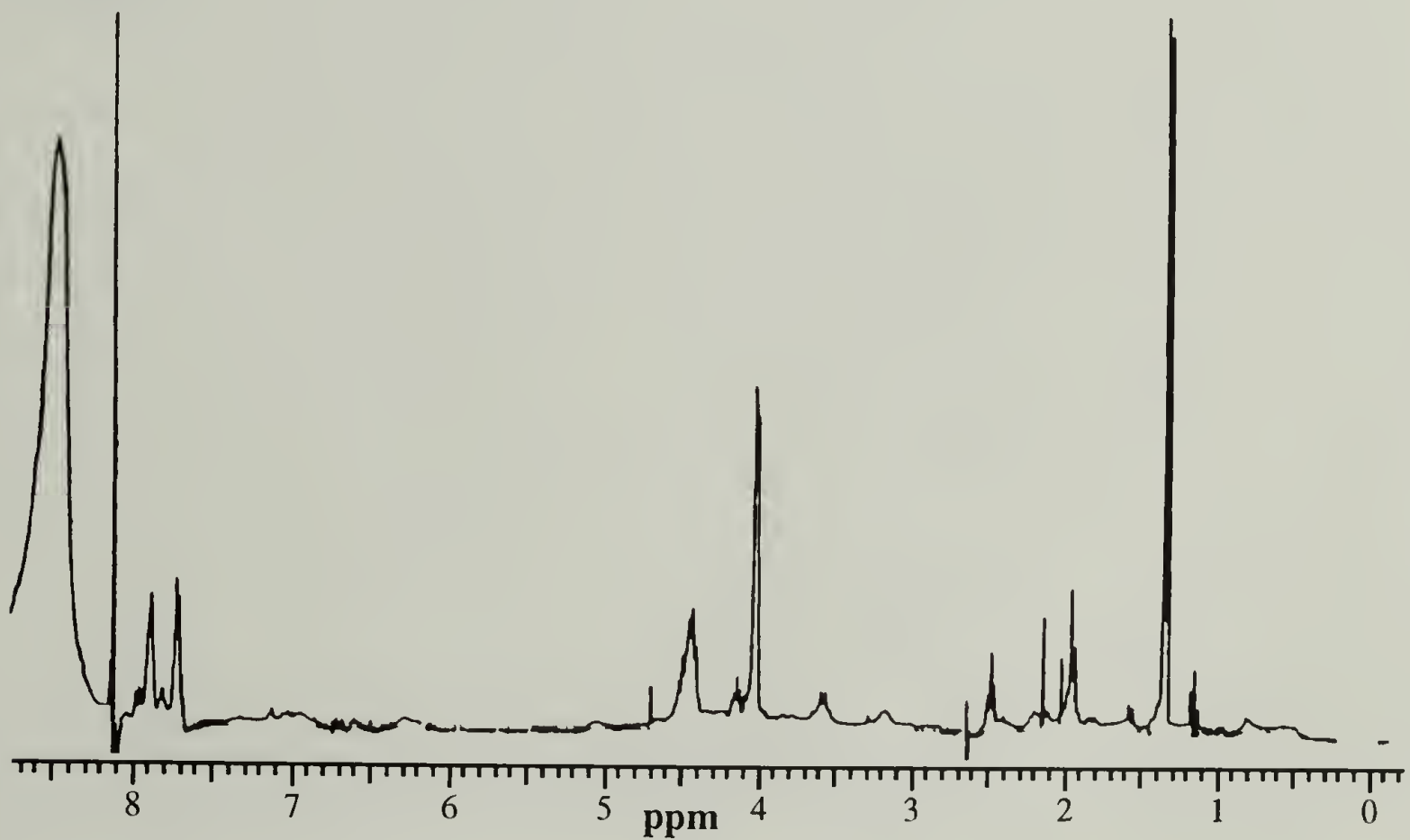
HC4-16



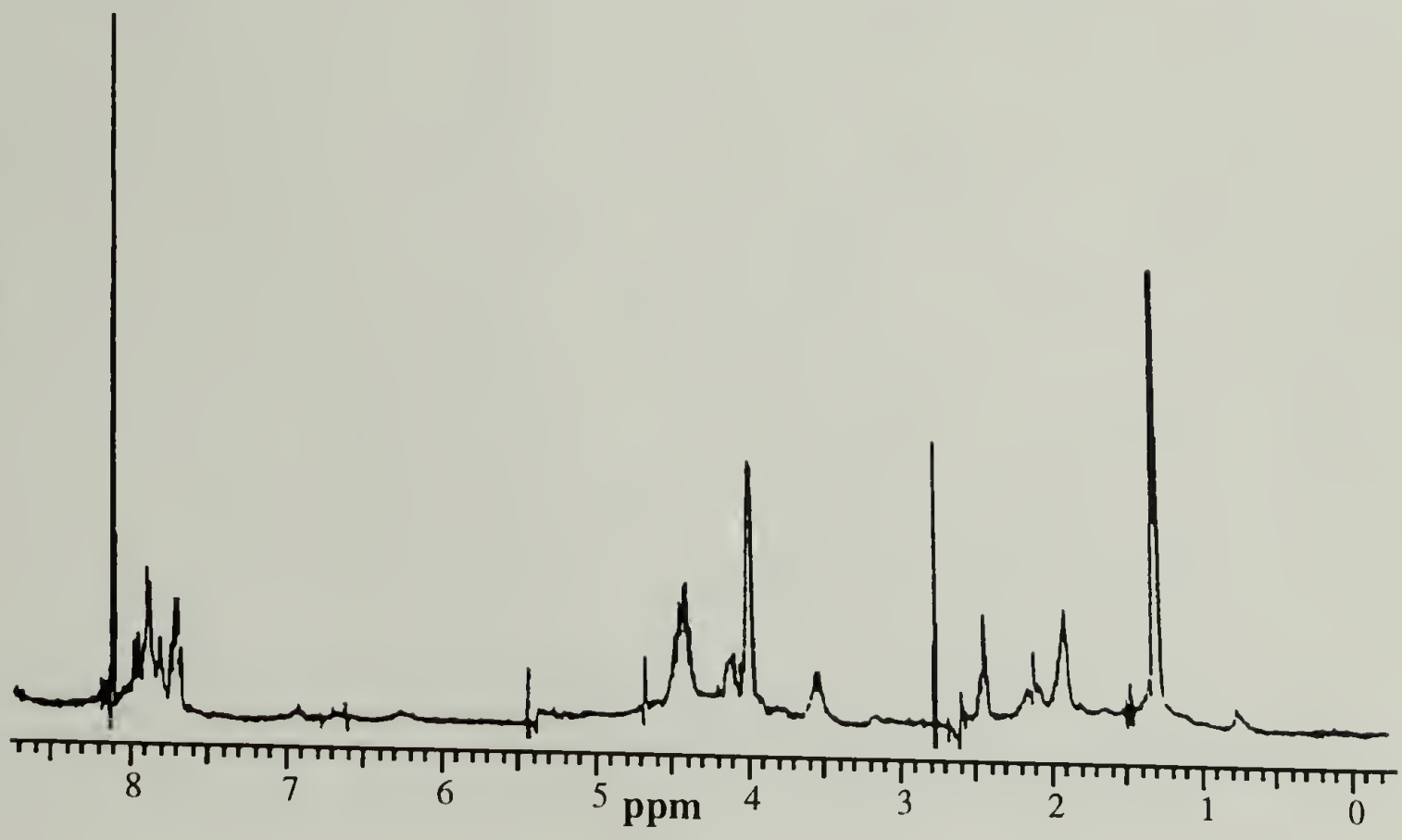
HC5-16



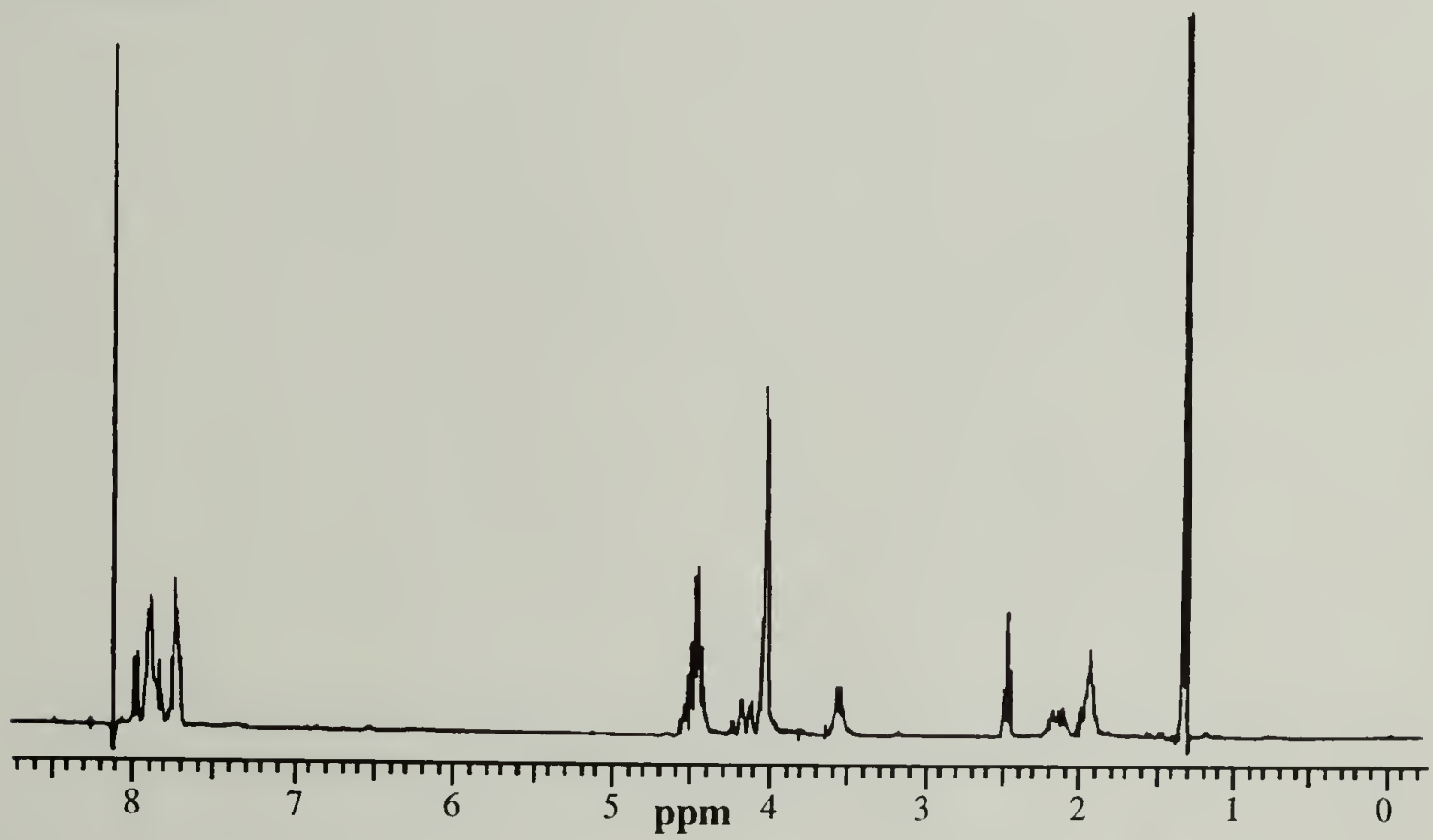
HC6-16

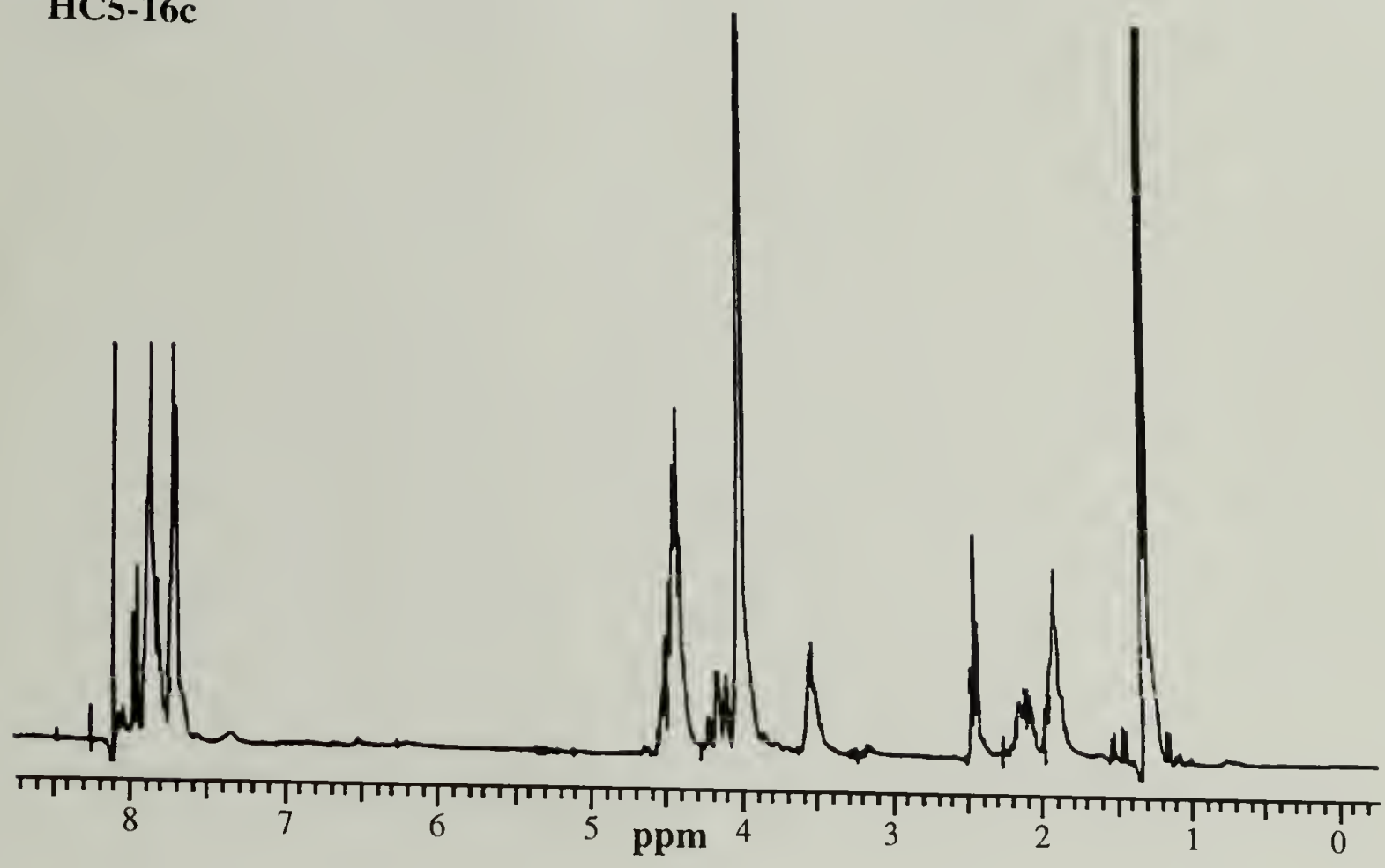


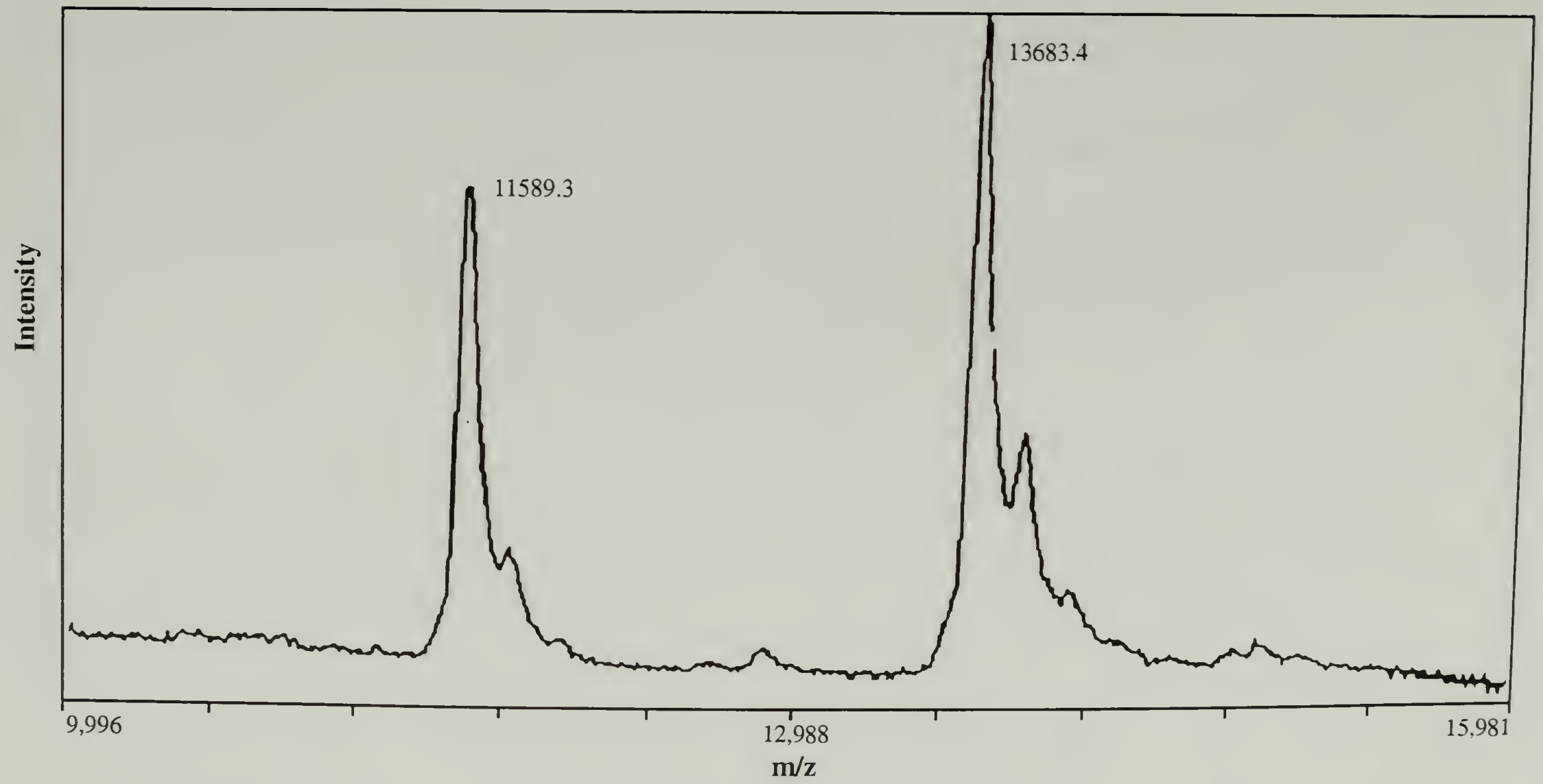
HC3-16c



HC4-16c




HC5-16c
8
7
6
5
ppm
4
3
2
1
0



APPENDIX C: MALDI MASS SPECTRUM OF HC4-14C

BIBLIOGRAPHY

CHAIN FOLDING MODELS

Alexander, L. E., "Chain Folding Conformations in the Polyamides: Models From the Structures of Cyclic Oligomers" *Polym Lett.* **10**, 759-767 (1972).

Atkins, E. D. T., Hill, M., Hong, S. K., Keller, A., Organ, S., "Lamellar Structure and Morphology of Nylon 46 Crystals. A New Chain Folding Mechanism of Nylons" *Macromolecules* **25**, 917-924 (1992).

Atkins, E. D. T., Keller, A., Sadler, D. M., "Structure Analysis of Chain-Folding Lamellar Polyamide Crystals From X-Ray Diffraction" *J. Polym. Sci., Polym. Phys. Ed.* **10**, 863-875 (1972).

Bassett, D. C., *Principles of Polymer Morphology* (Cambridge University Press, Cambridge, 1981).

Bunn, C. W. and Garner, E. V., "The Crystal Structures of Two Polyamides ('Nylons')" *Proc. Royal Soc. A* **189**, 39-68 (1947).

Burmester, A. F., Dreyfuss, P., Geil, P., Keller, A., "On the Annealing of Polyamide Crystal Mats" *Polym. Lett.* **10**, 769-775 (1972).

Dreyfuss, P. and Keller, A., "Chain Folding in Polyamides: A Study on Nylons 66, 610, and 612 as Crystallized From Solution" *J. Macromol. Sci., Phys.* **4**, 811-836 (1970).

Geil, P. H., "Nylon Single Crystals" *J. Polym. Sci.* **64**, 449-458 (1960).

Hallos, R. S., "A 'Chain-Fold Band' in the Infrared Spectra of Nylon 6" *J. Appl. Polym. Sci.* **29**, 3907-3914 (1983).

Holmes, D. R., Bunn, C. W., Smith, D. J., "The Crystal Structure of Polycaproamide: Nylon 6" *J. Polym. Sci.* **17**, 159-177 (1955).

Keller, A., "A Note on Single Crystals in Polymers: Evidence for a Folded Chain Configuration" *Philos. Mag.* **2**, 1171-1175 (1957).

Keller, A., "Morphology of Polymers" *Pure and Appl. Chem.* **64**, 193-204 (1992).

Koenig, J. L. and Itoga, M., "Infrared Studies of Chain Folding in Uniaxially Oriented Polyhexamethylene Adipamide" *J. Macromol. Sci., Phys* **6**, 309-326 (1972).

Koenig, J. L. and Itoga, M., "Mechanisms for Regularization of Chain Folds During Annealing of Polyhexamethylene Adipamide" *J. Macromol. Sci., Phys* **6**, 327-342 (1972).

Lotz, B., "Structural and Morphological Features of Synthetic and Natural Polymers" *Mater. Res. Soc. Symp. Proc.* **255**, 95-104 (1992).

Sperling, L. H., *Introduction to Physical Polymer Science* (Wiley-Interscience, New York, 1986).

Stamhuis, J. E. and Pennings, A. J., "Gelation and Crystallisation of Nylon-6 From Dilute Solution" *Brit. Polym. J.* **10**, 221-225 (1978).

Wunderlich, B., *Macromolecular Physics* (Academic Press, New York, 1973).

CROSS-BETA STRUCTURE

Bradbury, E. M., Brown, L., Downie, A. R., Elliott, A., Fraser, R. D. B., Hanby, W. E., McDonald, T. R. R., "The 'Cross-Beta' Structure in Polypeptides of Low Molecular Weight" *J. Mol. Biol.* **2**, 276-287 (1960).

Carbone, F. R. and Leach, S. J., "Studies of Repeating Synthetic Peptides Designed to Adopt a Cross-Beta Conformation. II. Conformational Analysis of the Repeating Polymers" *Int. J. Peptide Protein Res.* **26**, 225-235 (1985).

Geddes, A. J., Parker, K. D., Atkins, E. D. T., Beighton, E., "'Cross-Beta' Conformation in Proteins" *J. Mol. Biol.* **32**, 343-358 (1968).

Hepburn, H. R., Chandler, H. D., Davidoff, M. R., "Extensometric Properties of Insect Fibroins: The Green Lacewing Cross-Beta, Honeybee Alpha-Helical and Greater Waxmoth Parallel-Beta Conformations" *Insect Biochem.* **9**, 69-77 (1979).

Parker, K. D. and Rudall, K. M., "The Silk of the Egg-Stalk of the Green Lace-Wing Fly" *Nature (London)* **179**, 905-906 (1957).

Takeda, Y., "X-Ray Diffraction Studies on the Cross-Beta Form of Poly (O-Benzyl-L-Serine)" *Biopolymers* **29**, 1125-1128 (1990).

PROTEIN STRUCTURE

Bamford, C. H., Brown, L., Elliott, A., Hanby, W. E., Trotter, I. F., "Some New Investigations on the Structure of Synthetic Polypeptides" *Proc. Royal Soc. B* **141**, 49-60 (1953).

Bowie, J. U., Luthy, R., Eisenberg, D., "A Method to Identify Protein Sequences That Fold Into a Known Three-Dimensional Structure" *Science* **253**, 164-170 (1991).

Cantor, C. R. and Schimmel, P. R., *Biophysical Chemistry Part I: The Conformation of Biological Macromolecules* (W. H. Freeman, New York, 1980).

Chou, P. Y. and Fasman, G. D., "Conformational Parameters for Amino Acids in Helical, Beta-Sheet, and Random Coil Regions Calculated From Proteins" *Biochemistry* **13**, 211-221 (1974).

Chou, P. Y. and Fasman, G. D., "Prediction of Protein Conformation" *Biochemistry* **13**, 222-245 (1974).

Corey, R. B. and Pauling, L., "Fundamental Dimensions of Polypeptide Chains" *Proc. Royal Soc. B* **141**, 10-20 (1953).

Degrado, W. F., Wasserman, Z. R., Lear, J. D., "Protein Design, a Minimalist Approach" *Science* **243**, 622-628 (1989).

Fossey, S. A., Nemethy, G., Gibson, K. D., Scheraga, H. A., "Conformational Energy Studies of Beta-Sheets" *Biopolymers* **31**, 1529-1536 (1991).

Gething, M. and Sambrook, J., "Protein Folding in the Cell" *Nature (London)* **355**, 33-45 (1992).

Halverson, K. J., Sucholeiki, I., Ashburn, T. T., Lansbury, P. T., Jr., "Location of Beta-Sheet-Forming Sequences in Amyloid Proteins by FTIR" *J. Am. Chem. Soc.* **113**, 6701-6703 (1991).

Jaenicke, R., "Protein Folding: Local Structures, Domains, Subunits, and Assemblies" *Biochemistry* **30**, 3147-3161 (1991).

MacArthur, I., "The Pauling-Corey Models and Fibrous Proteins" *Proc. Royal Soc. B* **141**, 33-39 (1953).

MacArthur, M. W. and Thornton, J. M., "Influence of Proline Residues on Protein Conformation" *J. Mol. Biol.* **218**, 397-412 (1991).

Miyazawa, T. and Blout, E. R., "The Infrared Spectra of Polypeptides in Various Conformations: Amide I and II Bands" *J. Am. Chem. Soc.* **83**, 712-719 (1961).

Pauling, L. and Corey, R. B., "Stable Configurations of Polypeptide Chains" *Proc. Royal Soc. B* **141**, 21-33 (1953).

Pauling, L., Corey, R. B., Branson, H. R., "The Structure of Protein: Two Hydrogen-Bonded Helical Configurations of the Polypeptic Chain" *Proc. Nat. Acad. Sci. USA* **37**, 205-221 (1951).

Privalov, P. L., "Thermodynamic Problems of Protein Structure" *Annu. Rev. Biophys. Biophys. Chem.* **18**, 47-69 (1989).

Ramachandran, G. N. and Sasisekharan, V., "Conformation of Polypeptides and Proteins" *Adv. Protein Chem.* **23**, 239-297 (1968).

Regan, L. and DeGrado, W. F., "Characterization of a Helical Protein Designed From First Principles" **241**, 976-978 (1988).

Richards, F. M., "Folded and Unfolded Proteins: An Introduction" in *Protein Folding* T. E. Creighton, Ed. (W. H. Freeman and Company, New York, 1992), pp 37-51.

Richardson, J. S. and Richardson, D. C., "Principles and Patterns of Protein Conformation" in *Prediction of Protein Structure and the Principles of Protein Conformation* G. D. Fasman, Ed. (Plenum, New York, 1989), pp 1-48.

Richardson, J. S. and Richardson, D. C., "The *De Novo* Design of Protein Structures" *Trends Biochem. Sci.* **14**, 304-309 (1989).

Schultz, G. E. and Schirmer, R. H., *Principles of Protein Structure*. C. R. Cantor, Eds., Springer Advanced Texts in Chemistry (Springer Verlag, New York, 1979).

Traub, W., "Polypeptides as Structural Models for Proteins", in *Proceedings of the First Cleveland Symposium on Macromolecules* A. G. Waltons, Ed. (Elsevier, Cleveland, 1976), pp 23-28.

CHAIN REVERSALS IN PROTEINS

Chou, P. Y. and Fasman, G. D., "Beta-Turns in Proteins" *J. Mol. Biol.* **115**, 135-175 (1977).

Dyson, H. J., Rance, M., Houghten, R. A., Lerner, R. A., Wright, P. E., "Folding of Immunogenic Peptide Fragments of Proteins in Water Solution. I. Sequence Requirements for the Formation of a Reverse Turn" *J. Mol. Biol.* **201**, 161-200 (1988).

Isogai, Y., Nemethy, G., Rackovsky, S., Leach, S. J., Scheraga, H. A.,"Characterization of Multiple Bends in Proteins" *Biopolymers* **19**, 1183-1210 (1980).

Lewis, P. N., Momany, F. A., Scheraga, H. A., "Chain Reversals in Proteins" *Biochim. Biophys. Acta* **303**, 211-229 (1973).

Sibanda, B. L., Blundell, T. L., Thornton, J. M., "Conformation of Beta-Hairpins in Protein Structures: A Systematic Classification With Applications to Modeling by Homology, Electron Density Fitting and Protein Engineering" *J. Mol. Biol.* **206**, 759-777 (1989).

Sibanda, B. L. and Thornton, J. M., "Beta-Hairpin Families in Globular Proteins" *Nature (London)* **316**, 170-174 (1985).

Venkatachalam, C. M., "Stereochemical Criteria for Polypeptides and Proteins. V. Conformation of a System of Three Linked Peptide Units" *Biopolymers* **6**, 1425-1436 (1968).

Wilmot, C. M. and Thornton, J. M., "Analysis and Prediction of the Different Types of Beta-Turns in Proteins" *J. Mol. Biol.* **203**, 221-232 (1988).

SILKS AND REPETITIVE POLYPEPTIDES

Anderson, J. M., Chen, H. H., Rippon, W. B., Walton, A. G., "Characterization of Three Sequential Polydipeptides Containing Glycine" *J. Mol. Biol.* **67**, 459-471 (1972).

Arnott, S., Dover, S. D., Elliott, A., "Structure of Beta-Poly-L-Alanine: Refined Atomic Coordinates for an Anti-Parallel Beta-Pleated Sheet" *J. Mol. Biol* **30**, 201-208 (1966).

Asakura, T., Suzuki, H., Watanabe, Y., "Conformational Characterization of Silk Fibroin in Intact *Bombyx mori* and *Philosamia cynthia ricini* Silkworms by ^{13}C NMR Spectroscopy" *Macromolecules* **16**, 1024-1026 (1983).

Brack, A. and Spach, G., "Synthesis and Conformations of Periodic Copolypeptides of L-Alanine and Glycine" *Biopolymers* **11**, 563-586 (1972).

Colonna-Cesari, F., Premilat, S., Lotz, B., "Conformational Analysis of the Beta Sheet Structure of Poly-L-Alanine and Poly(L-Alanine-Glycine)" *J. Mol. Biol.* **95**, 71-82 (1975).

Colonna-Cesari, F., Premilat, S., Lotz, B., "Structure of Polyglycine I: A Comparison of the Antiparallel Pleated and Antiparallel Rippled Sheets" *J. Mol. Biol.* **87**, 181-191 (1974).

Fossey, S. A., Nemethy, G., Gibson, K. D., Scheraga, H. A., "Conformational Energy Studies of Beta-Sheets of Model Silk Fibroin Peptides. I. Sheets of Poly(Ala-Gly) Chains" *Biopolymers* **31**, 1529-1541 (1991).

Fraser, R. D. B. and MacRae, T. P., *Conformation in Fibrous Proteins and Related Synthetic Polypeptides* (Academic Press, New York, 1973).

Fraser, R. D. B. and Suzuki, E., "Resolution of Overlapping Absorption Bands by Least Squares Procedures" *Anal. Chem.* **38**, 1770-1773 (1966).

Fraser, R. D. B., MacRae, T. P., Stewart, F. H. C., "Poly-L-Alanylglycyl-L-Alanylglycyl-L-Serylglycine: A Model for the Crystalline Regions of Silk Fibroin" *J. Mol. Biol.* **18**, 580-582 (1966).

Fraser, R. D. B., MacRae, T. P., Stewart, F. H. C., Suzuki, E., "Poly-L-Alanylglycine" *J. Mol. Biol.* **11**, 706-712 (1965).

Go, Y., Noguchi, J., Asai, M., Hayakawa, T., "A (Glycyl-L-Alanyl)-Polymer Resembling the Crystalline Part of Silk Fibroin. Confirmation of the Meyer-Mark Model" *J. Polym. Sci.* **21**, 147-150 (1956).

Kaplan, D. L., et al., "Biosynthesis and Processing of Silk Proteins" *MRS Bull.* **17**, 41-47 (October 1992).

Kaplan, D. L., Fossey, S., Viney, C., Muller, W., "Self-Organization (Assembly) in Biosynthesis of Silk Fibers—A Hierarchical Problem" *Mat. Res. Soc. Symp. Proc.* **255**, 19-29 (1992).

Kaplan, D. L., Lombardi, S. J., Muller, W. S., Fossey, S. A., "Silks" in *Biomaterials: Novel Materials From Biological Sources* D. Byron, Eds. (Stockton Press, New York, 1991), pp 3-53.

Lotz, B. and Colonna Cesari, F., "The Chemical Structure and the Crystalline Structures of *Bombyx mori* Silk Fibroin" *Biochimie* **61**, 205-214 (1979).

Lotz, B. and Keith, H. D., "Crystal Structure of Poly(L-Ala-Gly)II. A Model for Silk I" *J. Mol. Biol.* **61**, 201-215 (1971).

Lotz, B., Brack, A., Spach, G., "Beta Structure of Periodic Copolypeptides of L-Alanine and Glycine: Their Relevance to the Structure of Silks" *J. Mol. Biol.* **87**, 193-203 (1974).

Lotz, B., "Crystal Structure of Polyglycine I" *J. Mol. Biol.* **87**, 169-180 (1974).

Lotz, B., Gonthier-Vassal, A., Brack, A., Magoshi, J., "Twisted Single Crystals of *Bombyx mori* Silk Fibroin and Related Model Polypeptides With Beta Structure. A Correlation With the Twist of the Beta Sheets in Globular Proteins." *J. Mol. Biol.* **156**, 345-357 (1982).

Lucas, F., Shaw, J. T. B., Smith, S. G., "Comparative Studies of Fibroins: I. The Amino Acid Composition of Various Fibroins and its Significance in Relation to their Crystal Structure and Taxonomy" *J. Mol. Biol.* **2**, 339-349 (1960).

Marsh, R. E., Corey, R. B., Pauling, L., "An Investigation of the Structure of Silk Fibroin" *Biochim. Biophys. Acta* **16**, 1-34 (1955).

Marsh, R. E., Corey, R. B., Pauling, L., "The Structure of Tussah Silk Fibroin (With a Note on the Structure of Beta-Poly-L-Alanine)" *Acta Cryst.* **8** 710-715 (1955).

Moore, W. H. and Krimm, S., "Vibrational Analysis of Peptides, Polypeptides, and Proteins. I. Polyglycine I" *Biopolymers* **15**, 2439-2464 (1976).

Moore, W. H. and Krimm, S., "Vibrational Analysis of Peptides, Polypeptides, and Proteins. II. Beta-Poly(L-Alanine) and Beta-Poly(L-Alanylglycine)" *Biopolymers* **15**, 2465-2483 (1976).

Padden, J., F. J. and Keith, H. D., "Crystalline Morphology of Synthetic Polypeptides" *J. Appl. Phys.* **36**, 2987-2995 (1965).

Stewart, F. H. C., "Poly-L-Alanylglycyl-L-Alanylglycyl-L-Serylglycine: A Synthetic Model of *Bombyx mori* Silk Fibroin" *Aust. J. Chem.* **19**, 489-501 (1966).

Takahashi, Y., Gehoh, M., Yuzuriha, K., "Crystal Structure of Silk (*Bombyx mori*)" *J. Polym. Sci., Polym. Phys.* **29**, 889-891 (1991).

BIOMATERIALS

Dong, Z., Lewis, R. V., Middaugh, C. R., "Molecular Mechanism of Spider Silk Elasticity" *Arch. of Biochem. Biophys.* **284**, 53-57 (1991).

Dong, Z., Xu, M., Middaugh, C. R., Lewis, R., "Spider Silk Proteins" *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **31**, 197-198 (1990).

Hinman, M. B. and Lewis, R. V., "Isolation of a Clone Encoding a Second Dragline Silk Fibroin. *Nephila clavipes* Dragline Silk is a Two-Protein Fiber" *J. Biol. Chem* **267**, 19320-19324 (1992).

Laursen, R. A., Ou, J.-J., Shen, X.-T., Connors, M. J., "Characterization and Structure of Mussel Adhesive Proteins" *Mat. Res. Soc. Symp. Proc.* **174**, 237-242 (1990).

Lewis, R. V., "Spider Silk: Unraveling of a Mystery" *Acc. Chem. Res.* **25**, 392-398 (1992).

Lombardi, S. J. and Kaplan, D. L., "Isolation, Cloning, and Physiochemical Characterization of Spider Silk From the Golden Orb-Weaver, *Nephila clavipes*" *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **31**, 195-196 (1990).

Rich, A. and Crick, F. H. C., "The Molecular Structure of Collagen" *J. Mol. Biol.* **3**, 483-506 (1961).

Sandberg, L. B., Leslie, J. G., Leach, C. T., Torres, V. L., Smith, A. R., Smith, D. W., "Elastin Covalent Structure as Determined by Solid-Phase Amino Acid Sequencing" *Pathol. Biol.* **33**, 26-34 (1985).

Sandberg, L. B., Soskel, N. T., Leslie, J. B., "Elastin Structure, Biosynthesis, and Relation to Disease States" *N. Engl. J. Med.* **304**, 566-579 (1981).

Sanger, F., Niklen, S., Coulson, A. R., "Elastin Covalent Structure as Determined by Solid-Phase Amino Acid Sequencing" *Proc. Natl. Acad. Sci. USA* **74**, 5463-5479 (1977).

Urry, D. W., "Elastomeric Polypeptide Biomaterials: Structure and Free Energy Transduction" *Mat. Res. Soc. Symp. Proc.* **174**, 243-250 (1990).

Urry, D. W., Haynes, B., Zhang, H., Harris, R. D., Prasad, K. U., "Mechanochemical Coupling in Synthetic Polypeptides by Modulation of an Inverse Temperature Transition" *Proc. Nat. Acad. Sci. USA* **85**, 3407-3411 (1988).

Urry, D. W., Parker, T. M., Minehan, D. S., Nicol, A., Pattanaik, A., Peng, S. Q., "The Capacity to Vary the Bioactive Role of Elastic Protein-Based Polymers" *Polym. Prep. (Am. Chem. Soc., Div. Polym. Mat. Sci. Engr.)* **66**, 399-402 (1992).

Waite, J. H., "Evidence for a Repeating 3,4-Dihydroxyphenylalanine- and Hydroxyproline-Containing Decapeptide in the Adhesive Protein of the Mussel, *Mytilus edulis* L." *J. Biol. Chem.* **258**, 2911-2915 (1983).

Waite, J. H., "Marine Adhesive Proteins: Natural Composite Thermosets" *Int. J. Biol. Macromol.* **12** 139-143 (1990).

Waite, J. H., "Natural Thermoset Proteins" *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **31(1)**, 181-182 (1990).

Xu, M. and Lewis, R. V., "Structure of a Protein Superfiber: Spider Dragline Silk" *Proc. Natl. Acad. Sci. USA* **87**, 7120-7124 (1990).

CHEMICAL SYNTHESIS OF PEPTIDES

Bodanszky, M., *Peptide Chemistry* (Springer Verlag, Berlin, 1988).

Katchalski, E. and Sela, M., "Synthesis and Chemical Properties of Poly-Alpha-Amino Acids" *Adv. Protein Chem.* **13**, 249-289 (1958).

Kent, S. B. H., "Chemical Synthesis of Peptides and Proteins" *Annu. Rev. Biochem.* **57**, 957-989 (1988).

BIOSYNTHESIS OF REPETITIVE POLYPEPTIDES

Biernat, J. and Koester, H., "Expressions of Synthetic Genes Coding for Completely New Nutritionally Rich, Artificial Proteins" *Protein Eng.* **1**, 353-361 (1987).

Cappello, J., "Genetic Production of Synthetic Protein Polymers" *MRS Bull.* **17**, 48-53 (October 1992).

Cappello, J., Crissman, J., Dorman, M., Mikolajczak, M., Textor, G., Marquet, M., Ferrari, F. A., "The Genetic Production of Synthetic Crystalline Protein Polymers" *Mater. Res. Soc. Symp. Proc.* **174**, 267-277 (1990).

Cappello, J., Crissman, J., Dorman, M., Mikolajczak, M., Textor, G., Marquet, M., Ferrari, F., "Genetic Engineering of Structural Protein Polymers" *Biotechnol. Prog.* **6**, 198-202 (1990).

Creel, H. S., Fournier, M. J., Mason, T. L., Tirrell, D. A., "Genetically Directed Syntheses of New Polymeric Materials: Efficient Expression of a Monodisperse Copolypeptide Containing Fourteen Tandemly Repeated -$(AlaGly)_4$ProGluGly- Elements" *Macromolecules* **24**, 1213-1214 (1991).

Creel, H. S., Fournier, M. J., Mason, T. L., Tirrell, D. A., "Synthesis and Expression of Artificial Genes Encoding Proteins With Repeating -$(AlaGly)_4$ProGluGly- Elements" *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **31(2)**, 165-166 (1990).

Doel, M. T., Eaton, M., Cook, E. A., Lewis, H., Patel, T., Carey, N. H., "The Expression in *Escherichia coli* of Synthetic Repeating Polymeric Genes Coding for Poly(L-Aspartyl-L-Phenylalanine)" *Nucleic Acids Res.* **8**, 4575-4592 (1980).

Dougherty, M.J., Kothakota, S., Krejchi, M.T., Mason, T. L., Tirrell, D. A., McGrath, K. P., Atkins, E. D. T., "Biosynthesis of New Polymers of Controlled Molecular Structure" *Makromol. Chem., Macromol. Symp.* **62**, 225-229 (1992).

Fournier, M. J., Creel, H. S., Krejchi, M. T., Mason, T. L., Tirrell, D. A., McGrath, K. P., Atkins, E. D. T., "Genetic Synthesis of Periodic Protein Materials" *J. Bioact. Compat. Polym.* **6**, 327-338 (1991).

Fournier, M. J., Mason, T. L., Tirrell, D. A., "Comments on the Role of Molecular Genetics in Polymer Materials Science" *Mater. Res. Soc. Symp. Proc.* **218**, 45-49 (1991).

Gmachl, M., Berger, H., Thalhammer, J., Kreil, G., "Dermal Glands of *Xenopus laevis* Contain a Polypeptide With a Highly Repetitive Amino Acid Sequence" *FEBS Letters* **260**, 145-148 (1990).

Goldberg, I., Salerno, A. J., Patterson, T., Williams, J. I., "Cloning and Expression of a Collagen-Analog-Encoding Synthetic Gene in *Escherichia coli*" *Gene* **80**, 305-309 (1989).

Goraj, K., Renard, A., Martial, J. A., "Synthesis, Purification and Initial Structural Characterization of Octarellin, a *De Novo* Polypeptide Modelled on the Alpha/Beta Barrel" *Protein Eng.* **3**, 259-266 (1990).

Gupta, S. C., Weith, H. L., Somerville, R. L., "Biological Limitations on the Length of Highly Repetitive DNA Sequences That May Be Stably Maintained Within Plasmid Replicons in *Escherichia Coli*" *Bio/Tech.* **2**, 602-608 (1983).

Hartley, J. L. and Gregori, T. J., "Cloning Multiple Copies of a DNA Segment" *Gene* **13**, 347-353 (1981).

Hecht, M. H., Richardson, J. S., Richardson, D. C., Ogden, R. C., "*De Novo* Design, Expression and Characterization of Felix: A Four-Helix Bundle Protein of Native-Like Sequence" *Science* **249**, 884-891 (1990).

Krejchi, M. T., Fournier, M. J., Mason, T. L., Tirrell, D. A., "Synthesis and Expression of Artificial Genes Encoding Periodic Proteins of the Sequence $[(AlaGly)_3GluGly]_{36}$" *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **32(1)**, 411 (1991).

McGrath, K. P., Fournier, M. J., Mason, T. L., Tirrell, D. A., "Genetically Directed Syntheses of New Polymeric Materials Expression of Artificial Genes Encoding Proteins With Repeating -$(AlaGly)_3$ProGluGly- Elements" *J. Am. Chem. Soc.* **114**, 727-733 (1992).

McGrath, K. P., Fournier, M. J., Mason, T. L., Tirrell, D. A., "Synthesis and Expression of an Artificial Gene Encoding a Novel Sequential Polypeptide" *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **31(1)**, 190-191 (1990).

Salerno, A. J. and Goldberg, I., "Expression of a Synthetic Mussel Adhesive Protein in *E. Coli*" *Polym. Prepr. (Am. Chem. Soc., Div. Mat. Sci. Engr.)* **66**, 398 (1992).

Tirrell, D. A., Fournier, M. J., Mason, T. L., "New Polymers From Artificial Genes: A Progress Report" *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **32(3)**, 704-705 (1991).

Tirrell, D. A., Fournier, M. J., Mason, T. L., "Protein Engineering for Materials Applications" *Curr. Opin. Struct. Biol.* **1**, 638-641 (1991).

Tirrell, D. A., Fournier, M. J., Mason, T. L., "Genetic Engineering of Polymeric Materials" *MRS Bulletin* **17**, 23-28 (July 1991).

Willson, T. A., Oakley, K. M., Nagley, P., "A Simple Method for Constructing Directly Repeated Multimeric DNA Sequences" *Gene Anal. Techn.* **2**, 77-82 (1985).

Zhang, G., Fournier, M. J., Mason, T. L., Tirrell, D. A., "Biological Synthesis of Monodisperse Derivatives of Poly(Alpha L-Glutamic Acid): Model Rodlike Polymers" *Macromolecules* **25**, 3601-3603 (1992).

Zhang, G., Fournier, M. J., Mason, T. L., Tirrell, D. A., "Toward Monodisperse Poly(Gamma-Benzyl Alpha L-Glutamate): Uniform, Polar, Molecular Rods" *Mat. Res. Soc. Symp. Proc.* **255**, 405-409 (1992).

MOLECULAR BIOLOGY TECHNIQUES

Birnboim, H. C. and Doly, J., "Rapid Alkaline Extraction Procedure for Screening Recombinant Plasmid DNA" *Nucleic Acids Res.* **7**, 1513 -1515 (1979).

Crowl, R., Seamans, C., Lomedico, P., McAndrew, S., "Versatile Expression Vectors for High-Level Synthesis of Cloned Gene Products in *Escherichia coli*" *Gene* **38**, 31-38 (1985).

Davis, B. J., "Disc Electrophoresis. II. Method and Application to Human Serum Proteins" *Ann. N. Y. Acad. Sci.* **121**, 404-411 (1964).

Dunker, A. K. and Rueckert, R. R., "Observations on Molecular Weight Determinations on Polyacrylamide Gel" *J. Biol. Chem* **244**, 5074-5080 (1969).

Dunn, J. J. and Studier, F. W., "Complete Nucleotide Sequence of Bacteriophage T7 DNA and the Locations of the T7 Genetic Elements" *J. Mol. Biol.* **166**, 477-535 (1983).

Laemmli, U. K., "Cleavage of Structural Proteins During the Assembly of the Head of Bacteriophage T4" *Nature (London)* **227**, 680-682 (1970).

Ornstein, L., *Ann. N. Y. Acad. Sci.* "Disc Electorophoresis. I. Background and Theory" **121**, 321-336 (1964).

Rosenberg, A. H., Lade, B. N., Chui, D.-S., Lin, W.-W., Dunn, J. J., Studier, F. W., "Vectors for Selective Expression of Cloned DNAs by T7 RNA Polymerase" *Gene* **56**, 125-135 (1987).

Salerno, A. J., "Strategy for Stabilizing Genes Coding for Repetitive Sequence Proteins" *Polym. Prepr. (Am. Chemical Soc., Div. Polym. Chem.)* **31(1)**, 183-184 (1990).

Sambrook, J., Fritsch, E. F., Maniatis, T., *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Press, Cold Spring Harbor, 1989).

Studier, F. W. and Moffatt, B. A., "Use of Bacteriophage T7 RNA Polymerase to Direct Selective High-Level Expression of Cloned Genes" *J. Mol. Biol.* **189**, 113-130 (1986).

Studier, F., Rosenberg, A. H., Dunn, J. J., Dubendorff, J. W., "Use of T7 RNA Polymerase to Direct Expression of Cloned Genes" *Methods Enzymol.* **185**, 60-115 (1990).

Yanisch-Perron, C., Vieira, J., Messing, J., "Improved M-13 Phage Cloning Vectors and Host Strains: Nucleotide Sequence of the M13MP18 and pUC19 Vectors" *Gene* **33**, 103-147 (1985).

MASS SPECTROMETRY OF PROTEINS AND POLYMERS

Bahr, U., Deppe, A., Karas, M., Hillenkamp, F., Giessmann, U., "Mass Spectrometry of Synthetic Polymers by UV-Matrix-Assisted Laser Desorption/Ionization" *Anal. Chem.* **64**, 2866-2869 (1992).

Beavis, R. C., "Matrix-Assisted Ultraviolet Laser Desorption: Evolution and Principles" *Org. Mass Spectrom.* **27**, 653-659 (1992).

Beavis, R. C. and Chait, B. T., "Factors Affecting the Ultraviolet Laser Desorption of Proteins" *Rapid. Commun. Mass Spectrom.* **3**, 233-237 (1989).

Beavis, R. C. and Chait, B. T., "Cinnamic Acid Derivatives as Matrices for Ultraviolet Laser Desorption Mass Spectrometry of Proteins" *Rapid Commun. Mass Spectrom.* **3**, 432-435 (1989).

Beavis, R. C. and Chait, B. T., "Matrix-Assisted Laser-Desorption Mass Spectrometry Using 355 nm Radiation" *Rapid. Commun. Mass Spectrom.* **3**, 436-439 (1989).

Beavis, R. C. and Chait, B. T., "Rapid, Sensitive Analysis of Protein Mixtures by Mass Spectrometry" *Proc. Natl. Acad. Sci.* **87**, 6873-6877 (1990).

Beavis, R. C., Chait, B. T., Creel, H. S., Fournier, M. J., Mason, T. L., Tirrell, D. A., "Analysis of Artificial Proteins by Matrix-Assisted Laser Desorption Mass Spectrometry" *J. Am. Chem. Soc.* **114**, 7584-7585 (1992).

Burlingame, A. L., Baillie, T. A., Russell, D. H., "Mass Spectrometry" *Anal. Chem.* **64**, 467A-502A (1992).

Chait, B. T. and Kent, S. B. H., "Weighing Naked Proteins: Practical, High-Accuracy Mass Measurement of Peptides and Proteins" *Science* **257**, 1885-1894 (1992).

Hillenkamp, F. and Karas, M., "Mass Spectrometry of Peptides and Proteins by Matrix-Assisted Ultraviolet Laser Desorption/Ionization" *Methods Enzymol.* **193**, 280-295 (1990).

Hodgson, J., "Electrophoresis in Thin Air" *Bio/Tech.* **10**, 399-401 (1992).

Karas, M. and Bahr, U., "Laser Desorption Ionization Mass Spectrometry of Large Biomolecules" *Trends Anal. Chem.* **9**, 321-325 (1990).

Karas, M. and Hillenkamp, F., "Laser Desorption Ionization of Proteins With Molecular Masses Exceeding 10000 Daltons" *Anal. Chem.* **60**, 2299-2301 (1988).

Karas, M., Bahr, U., Ingendoh, A., Nordhoff, E., Stahl, B., Strupat, K., Hillenkamp, F., "Principles and Application of Matrix-Assisted UV-Laser Desorption/Ionization Mass Spectrometry" *Anal. Chim. Acta* **241**, 175-185 (1990).