PREFACE TO THE FIRST EDITION

This book completes a trilogy, joining *Introduction to protein architecture: the structural biology of proteins*, and *Introduction to bioinformatics*. The other two books are more specialized—*Bioinformatics* to sequences, and *Architecture* to structures. This one is intended as an introduction to the other two. It also complements them, with a focus on protein function, including the integration and regulation of function: this book is as much about the logic of life as about its chemistry. These three components—sequence, structure and function—define the significance of proteins in biology.

Sydney Brenner once stated the questions to ask about any living system:

1. How does it work?
2. How does its structure form?
3. How did it evolve?

Protein science has appealed to many scientific disciplines to address these problems:

- Parts of protein science have developed out of *chemistry*, including physical methods such as spectroscopy, kinetics, and techniques of structure determination; and organic/biochemical methods for working out the mechanism of enzymatic catalysis. Much of classical protein chemistry and enzymology studied purified molecules in isolation. These studies revealed detailed information about the structures and activities of individual molecules.
- Parts of protein science have developed out of *molecular and cellular biology*. The challenge is to coordinate our knowledge of the properties and functions of isolated molecules, into an understanding of the biological context and integration of protein activity.

High-throughput methods of genomics and proteomics provide very large amounts of data about sequences and expression patterns. We know the full sequences of the DNA of many organisms. From these genome sequences we can infer the amino acid sequences of an organism's full complement of proteins. Proteomics tells us how the expression patterns of these proteins vary. We have pieces of a jigsaw puzzle that extends in both space and time. How can we fit the pieces together to understand the complex and delicate instrument that is the living cell?

- Parts of protein science have developed out of *evolutionary biology*. Evolution takes place on the molecular scale, exploring variations in sequences, protein structures and functions, and patterns of expression. Much of what we see in living organisms depends on, and reveals, historical events no longer observable directly.
- Parts of protein science have emerged from *computing*. Computer science is a flourishing field with the goal of making the most effective use of information technology hardware. Computers are essential for storing and analysing the very large amounts of data that have been coming onstream. Databanks archive and organize access to sequences of genomes; to amino acid sequences, structures, expression patterns, and interaction networks of proteins; to metabolic pathways; to the molecular correlates of diseases; and to the scientific literature. Distribution of the information requires facilities of computer networks and the World Wide Web. Computer programs provide access to the data, and tools for their analysis. *Bioinformatics* is a hybrid of biology, chemistry, physics, and computer science. Its goal is to bring the data to bear on biological problems, including applications to clinical medicine, agriculture, and technology.

Some readers may wonder why nucleic acids receive substantial attention in a book on protein science. The answer is that they belong here. It is impossible to make sense of the biology of proteins without discussion of the ribosome, and of control of gene expression, just to pick two examples. Nucleic acids provide the logical framework of the stage on which proteins act.

Whole-genome sequencing projects have transformed molecular biology. Before, we were limited to studying selected available examples. The new results lay bare the choices that Nature has made.

The three-dimensional counterpart of genome-sequence determination, the provision of *structures* of all the proteins in an organism, is known as *structural genomics.*

When we know the sequences and structures of all macromolecules in an *E. coli* cell, we shall have a complete but *static* knowledge of the components of a living object. The next step will be to learn about how the components are assembled and how their individual activities are integrated.

The most profound significance of proteins lies in their collective properties as a class of biomolecules. These properties include all their potential, as well as their actual, characteristics: proteins have an underlying chemical unity, they have the ability to organize themselves in three dimensions; and the system that produces them can create inheritable structural variations, conferring the ability to evolve. Recent developments in genomics, proteomics, and structural genomics permit study of the entire complement of proteins of an organism and their coordinated activities, and give glimpses of the *proteosphere*—the entire spectrum of the proteins in Nature.

Proteins are where the action is

Proteins are fascinating molecular devices. They play a variety of roles: there are structural proteins (molecules of the cytoskeleton, epidermal keratin, viral coat proteins); catalytic proteins (enzymes); transport and storage proteins (haemoglobin, myoglobin, ferritin); regulatory proteins, (including hormones, many kinases and phosphatases, and proteins that control gene expression); and proteins of the immune system and the immunoglobulin superfamily (including antibodies, and proteins involved in cell–cell recognition and signalling).

Amino acid sequences are now available for over 1 million proteins, and detailed atomic structures for over 24 000. These data support a number of fascinating and useful scientific endeavours, including:

*(1) **Interpretations of the mechanisms of function of individual proteins.*** The catalytic activities of enzymes can be explained in terms of physical-organic chemistry, on the basis of the interactions of substrates with residues in the active site of the protein.

*(2) **Patterns of molecular evolution.*** For many families of proteins, we know dozens or even hundreds of amino acid sequences, and many structures. We can observe the structural and functional roles of the sets of residues that are strongly conserved. We can describe the consequences of mutations, insertions and deletions in the amino acid sequence that perturb protein conformation, and identify constraints that selection imposes on the structure to preserve function. When proteins evolve with changes in function, these constraints on the structure are relaxed—or rather, replaced by alternative constraints—and the sequences and structures change more radically.

On a broader scale, *comparative genomics* traces the changes and distribution of proteins and regulatory sequences in different species, and similarities and differences in genome organization.

*(3) **The amino acid sequences of proteins dictate their three-dimensional structures, and their folding pathways.***

(a) Under physiological conditions of solvent and temperature, proteins fold spontaneously to an active native state. Although we cannot yet confidently predict the conformation of a novel protein structure from its amino acid sequence in all cases, many principles of protein structure and folding are now understood.

(b) Relationships between the evolutionary divergence of amino acid sequence and structure in protein families do make possible the reliable prediction of protein structure from amino acid sequence, whenever we know the sequence and structure of one or more sufficiently closely related proteins. As we come closer to having a structure of every type that appears in Nature, then the nearest relative of known structure will provide a reasonable quantitative model for almost all unknown proteins. In the future, this will give us access to structures of the proteins encoded in any genome.

(c) The amino acid sequence of a protein must not only preferentially stabilize the native state, it must contain a road map telling the protein how to get there, starting from the many diverse

conformations that comprise the unfolded state. A combination of physicochemical measurements and studies of the effects of mutations on the kinetics of folding and unfolding are illuminating the process of folding.

(4) Mapping the logic and the mechanisms of integration and control of life processes.

(a) We can measure and compare patterns of biochemical activity in different tissues, or under different growth conditions. Comparisons of protein expression patterns, or genotypes embedded in the DNA, using microarrays, have direct clinical applications. Many diseases involve failures in control mechanisms. Therefore differences in expression patterns in normal and disease states permit precise diagnosis of disease, producing a more accurate prognosis and choice of treatment. Prediction, from details of genotypes, of enhanced susceptibility to disease may suggest preventative measures, including alteration of lifestyle.

(b) We can trace pathways of information flow in regulatory networks. Proteins receive and transmit signals, often using conformational changes to control activity. Proteins 'talk' to one another, and also to DNA. DNA–protein interactions control gene expression.

(c) We can study patterns in time as well in space. The cell cycle, and the unfolding of developmental programmes, are coordinated with changes in protein expression. Development of the nervous systems of higher organisms presents great challenges to diversity plus accuracy in molecular signalling and recognition.

Advances in protein science have spawned the biotechnology industry:

(5) Protein engineering. It is now possible to design and test modifications of known proteins, and to design novel ones. Applications include:

(a) Attempts to enhance thermostability, for example by introducing disulphide or salt bridges, or by optimizing the choice of amino acids that pack the protein interior.

(b) Clinical applications, for instance therapeutic antibodies. Rats can raise antibodies against human tumours. Transfer of the active site from a rat antibody to a human antibody framework can produce a molecule that retains therapeutic activity in humans but reduces the side effects arising from patients' immunological response against the antigenic rat protein.

(c) Modifying antibodies to give them catalytic activity. Two features of enzymes are: the ability to bind substrate specifically, and the juxtaposition of bound substrate with residues that catalyse a chemical change. Antibodies can provide the binding and discrimination; the challenge to the chemist is to introduce the catalytic function.

(d) Modifications to probe mechanisms of function or folding, such as the identification of properties of folding intermediates by the effects of mutations on stability and folding kinetics.

(6) Drug discovery. There are many proteins specific to pathogens that we want to deactivate. Knowing the structure of the AIDS proteinase, or the neuraminidase of influenza virus, it should be possible to design molecules that will bind tightly and specifically to an essential site on these molecules, to interfere with their function. Knowing the structures of receptors it should be possible to design agonists and antagonists of the signals to which they respond.

These and other topics comprise a new scientific speciality, *protein science*. By a combination of approaches and methods, gathered from many other fields of science, we are gaining insight into the principles governing this fascinating class of molecules. Let us apply this knowledge in medicine, agriculture, and technology to improve the lot of mankind.

I thank F. Arnold, M. Bashton, C. Chothia, J. Clarke, T.J. Dafforn, A. Doherty, D.S. Eisenberg, I. Fearnley, J.R. Fresco, A. Friday, M. Gait, E. Gherardi, W.B. Gratzer, R. Henderson, O. Herzberg, T.J.P. Hubbard, J. Janin, T. Kiefhaber, G. Kleywegt, P.A. Lawrence, E.L. Lesk, V.E. Lesk, V.I. Lesk, L. Liljas, L. Lo Conte, D.J. Lomas, J. Löwe, P.A. Lyons, A.J. McCoy, J.D. Mollon, S. Moran, J. Moult, A.G. Murzin, D. Neuhaus, M. Oliveberg, V. Ramakrishnan, R. Read, G.D. Rose, B. Rost, K. Scott, L. Serpell, R. Staden, R. Tregear, M. Vendruscolo, C. Vogel, G. Vriend, A.G. Weeds, J.C. Whisstock, A. Yonath, and L. You.

PREFACE TO THE SECOND EDITION

This new edition of *Introduction to protein science: structure, function, and genomics* seeks to retain the successful features of the first edition, and to expand the coverage of the subject. This includes a fuller treatment of methods of structure determination, of enzymatic catalysis, and of protein evolution. There have been significant advances in protein science since the first edition, both experimental and computational. In presenting these advances, I emphasize how, for many topics, the combination of theory and experiment has proved to be a very powerful and effective partnership.

In the first edition, *Introduction to protein science* was described as a companion volume to *Introduction to protein architecture: the structural biology of proteins*, and *Introduction to bioinformatics*. Some of *Architecture* has been integrated into this volume. A more recent book, *Introduction to genomics*, is another companion volume.

Proteins claim the attention of many readers, spanning a range of interests in the fields of biology and medicine, chemistry and physics. In using this book as a text for a course, instructors will find ample opportunity to tailor the material to their students' interests and needs, especially through choice of exercises, problems, and weblems.

Plan of the book

Chapter 1: A general introduction to the subject, setting out the topics developed in the rest of the book, emphasizing that amino acid sequences of proteins determine their three-dimensional structures and functions.

Chapter 2: Detailed description of the structural chemistry of amino acids and proteins; survey of varieties of proteins.

Chapter 3: Protein structure determination, integrating experimental and computational methods.

Chapter 4: Protein bioinformatics: databases of sequences, structures and other experimental results, and tools for information retrieval and analysis.

Chapter 5: Enzyme catalysis and mechanisms, and the use of classical and modern tools to illuminate mechanisms of enzyme activity.

Chapter 6: Protein interactions, including functional aspects of complex formation.

Chapter 7: Protein evolution: by diverging in sequence and structure, proteins can retain, modify, or profoundly alter their function.

Chapter 8: Mechanism of the folding process by which proteins achieve their native states.

Chapter 9: Systems biology: the regulatory networks by which cells control and integrate the activities of their constituent molecules.

Learning aids throughout the book include boxes with a purple background containing summaries of key points in context and boxes with a green background providing additional guidance and notes to help readers grasp key concepts. Boxes headed Proteins in health and disease describe clinical applications.

The list of people to thank has grown. In addition to those mentioned in the first edition, I acknowledge with gratitude Drs P. Aloy, H. and F. Bernstein, J. Finer-Moore, K. Henrick, J. Lecomte, C. Praul, N. Seeman, L. Serpell, J. Sussman, V.E. Womble, and E.B. Ziff for reading specific sections and answering questions, and many other colleagues and publishers for courtesy in allowing me to reproduce illustrations.

Viewing stereo images

Readers may find a stereo viewer useful. The most common type comes with a stand appropriate for material lying flat on a table. Lorgnettes, held directly in front of the eyes, are more comfortable for reading a book. Many readers will become adept at seeing stereo unaided. The trick is to develop an unusual control of the eye muscles. In ordinary vision, when an object moves away from us we diverge our eyes *and* refocus for the greater distance. We are accustomed to coupling these muscular adjustments automatically. To view stereo pictures, we must diverge our eyes, but keep the focus appropriate for an object that is relatively near. A useful initial goal in learning to do this is to diverge the eyes without refocusing—imagine looking through the page at a distant object behind it—to produce a view in which each eye is receiving only its proper image, but these images are out of focus. Then try to focus them. (You will see *three* images—two from each eye, with the two in the centre overlapping; it is the centre one that will appear in stereo.)