PREFACE

Six years have elapsed since the first edition of this book was published. The field of structural bioinformatics has sustained a high level of excitement in that time, leading to innovative developments and considerable progress throughout the topics covered in the first edition and in the extension to many new domains. Through the efforts of the authors of this new edition, we have tried to capture these developments and to provide an accurate, detailed view of the current field. One way of picturing the advances or defining the "structural" change is relatively straightforward; namely, the number of experimental macromolecular structures has doubled since the first edition of this book was published. The Protein Structure Initiative has also led to an increase in the number of novel structures and folds. Overall, the continued growth in experimental structures has created an even richer data source for much of the work described herein. But numbers do not tell the whole story. The complexity of structures, the methods used, the ways structure is represented, our ability to model structures, our understanding of proteomes and their structural coverage, and so on, have also changed.

Describing the advances in "bioinformatics" per se is more difficult. Change in this case reflects both scientific advances and an increase in recognition within the biological sciences for the importance of computational methods. Due in part to the explosion in high-throughput experimental methods, bioinformatics is certainly more mainstream than it was 6 years ago and most experimental (i.e., non-computational) life scientists would acknowledge the role bioinformatics now plays in furthering our understanding of living systems. Some years from now, whether or not bioinformatics will exist as a separate entity, rather than as a core effort in every biological science department, is a subject for debate. What is important here is that there is an active effort to apply computational methods to a rapidly growing corpus of macromolecular structure data. Our primary goal is to provide a comprehensive description of what this field has accomplished to date and to make the reader aware of what we have gained and could gain in the future toward our understanding of living systems through the study of macromolecular structure and the continued, rigorous application of bioinformatics. As such, this edition should provide a fully current, useful reference to those already in the field, and a suitable text for those educating others. The first edition already encouraged new scholars to enter the field and we believe the case for engaging in structural bioinformatics is stronger than ever.

To meet this goal, the second edition includes not only updated chapters, but also new chapters covering mass spectrometry, genome annotation, immunology, protein dynamics and disorder, membrane proteins, protein design capabilities, and evolutionary biology as they relate to macromolecular structure.

Macromolecular structure is often underappreciated and bypassed in practice during the current era of high-throughput biology, especially since researchers can jump directly from

genomic sequence to phenotype and conduct biochemical studies on large-scale protein–protein interactions that are often involved in pathways associated with disease states. While a great deal can be learned from such studies, ultimately the devil is in the details which do not arise from traditional functional studies; the field of molecular biophysics, now termed structural biology, came into existence to obtain and use structures to provide those details. We believe structure will play an ever increasingly important role as genome studies seek to explore deeper into the mechanisms of life. As such, computational approaches that analyze structure are essential and we hope this book will be there to guide you.

We begin by describing the scope of this book and the history of the field (Chapter 1). The remainder of the introductory Section I is devoted to the understanding of the data itself, namely protein, DNA and RNA structure, respectively (Chapters 2 and 3). Understanding the nuances (scope, accuracy, completeness, etc.) of structural data is prerequisite to any effective use of that data. Effective data use in turn requires an understanding of the experiments or experimental method that produce the data. The most popular methods for deriving macromolecular structure data are, in order, X-ray crystallography (Chapter 4), NMR spectroscopy (Chapter 5), and electron microscopy (Chapter 6). Constructing structural models of molecules can also be guided with hydrogen–deuterium and cross-linking experiments coupled with mass spectrometry (Chapter 7). The raw data from these methods are most often a set of Cartesian coordinates representing the positions of the atoms in these structures, which are well suited for analysis by computer, but alternative representations of this information rich content are sometimes needed to conduct wide-scale bioinformatics analysis (Chapter 8). That is, structural biology and structural bioinformatics are inherently visual sciences—the tabular output of atomic coordinates can be useful as input for computation, but not for human insight. The visualization of structure has evolved along with the science and many useful tools, mostly free, are available (Chapter 9).

In the early days of structural biology (up to the late 1970s), those in the field could name all the structures that had been solved, some of which had Nobel prizes attached to them. As the field grew this was no longer possible, and databases of structure data began to appear. Consistent use of structural data contained within these databases (and indeed the construction of the databases themselves) requires consistent data representation and Section II is devoted to this topic. Chapter 10 introduces the common data representations used by today's software. The field is very fortunate to have scientists who recognize the importance of having a single source of primary data, the worldwide PDB (wwPDB—Chapter 11), from which a variety of secondary resources are derived. Examples of such resources are provided in Chapters 12 and 13.

As the number of structures has increased, much can be learnt from comparative analysis (Section III), where similarities and differences provide new insights. Chapters 14 and 15 describe structure validation, which is important in understanding the accuracy of the data you are dealing with before a 3D comparison and alignment of structures can be made (Chapter 16). When structure comparisons are made and similarities found, reductionism can be applied to make sense of the vast amount of data. Such reductionism leads to classification in various ways, such as by fold, domain, family, and superfamily (Chapters 17 and 18).

The more we know from comparing structures the more we can learn about structure and functional assignment (Section IV). Secondary structure assignment can now be made consistently and reliably for the majority of structures (Chapter 19). Proteins exist as one or more domains or compact structural and functional units. Hence, automated assignment of domains is important (Chapter 20). Through the structural genomics projects and the NIH

Protein Structure Initiative, structure determination is moving from a functional to a genomic initiative. That is, structures were traditionally determined in an effort to elucidate further details about a known function, and these structural efforts were established based on very extensive prior biological, biochemical and often genetic research, and were done in parallel with continuing biological research on functional properties. In contrast, high-resolution structures with no elucidated or known function are being determined at an accelerated pace, thus making functional assignment critical (Chapter 21). The use of structural information to identify distantly related proteins also serves in annotating genomes (Chapter 22) and clarifying evolutionary relationships (Chapter 23).

Proteins do not act in isolation, that is, most proteins do not function by themselves but act as the result of complex protein–protein, protein–ligand and protein–solvent interactions and are often part of larger macromolecular assemblies. Section V describes these interactions beginning with an introduction to electrostatic forces that have a fundamental impact on recognition between molecules (Chapter 24). The majority of these interactions are not captured in the experimental structure of a complex, but as an apo form of the structure with a signature that can be teased out to predict that interaction. Understanding these signatures when found in protein–DNA and protein–RNA interactions (Chapter 25) and in protein–protein interactions (Chapter 26) aids, for example, in the identification of new transcription sites and reconstruction of protein signaling networks. After the sites of interactions are identified, docking of the molecules is simplified, which is important in drug design (Chapter 27).

While the number of structures is increasing rapidly, the number of protein sequences is increasing much more rapidly; thus, the idea of predicting protein structure from its sequence remains an "obsessive" goal (Section VI). Spurred by an unusual biannual competition, referred to as CASP—the Critical Assessment of Protein Structure Prediction (Chapter 28), progress is being made in subcategories of structure prediction efforts within CASP and the field in general. Structure prediction categories include homology modeling (Chapter 30), fold recognition (Chapter 31) and *ab initio* structure prediction (Chapter 32). Other forms of prediction include secondary structure and membrane components for proteins (Chapter 29). Advances in understanding and predicting RNA structures have also been made and are discussed in Chapter 33.

Structural bioinformatics is playing an increasingly important role in the development of new pharmaceuticals (Section VII). The identification of drug targets, understanding the action of drug binding, and the design of promising leads all involve structural bioinformatics (Chapter 34). In addition to the development of small molecule and peptide-based drugs, contributions are also being made in identifying antigen recognition sites that aid in antibody-based therapeutics (Chapter 35).

Finally, Section VIII identifies challenges at the frontiers of structural bioinformatics. Membrane associated proteins, whose structures are difficult to characterize *in vitro* and thus are underrepresented experimentally, are one example (Chapter 36). Proteins are not static under physiological conditions, yet understanding the dynamics (Chapter 37) and the impact of disorder and conformational variants (Chapter 38), while important to protein function, are all still poorly understood. As our understanding of protein structure improves, so do our design rules and capacity for engineering new proteins for their functions that improve upon nature or provide the potential for novel processes (Chapter 39). The best way to push back these frontiers is with more structures and enhanced generalizations about their roles; structure genomics is doing just that (Chapter 40) and is thus a fitting place to end our tour of structural bioinformatics.

The words that follow are written by many of the leaders in the field and we thank them for their time and energy in sharing what motivates them to unravel the mysteries of nature, which are so beautifully displayed before us in an ever increasing number of macromolecular structures.

Jenny Gu
Philip E. Bourne

Color files of all figures from this book are available for download from the following web address: ftp://ftp.wiley.com/public/sci_tech_med/structural_bioinformatics