

► Chapter 1

Model and Molecule

Phase

These still days after frost have let down
the maple leaves in a straight compression
to the grass, a slight wobble from circular to

the east, as if sometime, probably at night, the
wind's moved that way—surely, nothing else
could have done it, really eliminating the *as*

if, although the *as if* can nearly stay since
the wind may have been a big, slow
one, imperceptible, but still angling

off the perpendicular the leaves' fall:
anyway, there was the green-ribbed, yellow,
flat-open reduction: I just now bagged it up.

A. R. Ammons¹

Proteins perform many functions in living organisms. For example, some proteins regulate the expression of genes. One class of gene-regulating proteins contains structures known as *zinc fingers*, which bind directly to DNA. Figure 1.1*a* shows a complex composed of a double-stranded DNA molecule and three zinc fingers from the mouse protein Zif268 (PDB 1zaa).

The protein backbone is shown as a yellow ribbon. The two DNA strands are red and blue. Zinc atoms, which are complexed to side chains in the protein, are green. The green dotted lines near the top center indicate two hydrogen bonds in which nitrogen atoms of arginine-18 (in the protein) share hydrogen atoms with nitrogen and oxygen atoms of guanine-10 (in the DNA), an interaction that holds the sharing atoms about 2.8 Å apart. Studying this complex with modern graphics software, you could zoom in, as in Fig. 1.1*b*, measure the hydrogen-bond lengths, and find them to be 2.79 and 2.67 Å. From a closer study, you would also learn that all of the protein–DNA interactions are between protein *side chains* and DNA *bases*; the protein backbone does not come in contact with the DNA. You could go on to discover all the specific interactions between side chains of Zif268 and base pairs of DNA. You could enumerate the additional hydrogen bonds and other contacts that stabilize this complex and cause Zif268 to recognize a specific sequence of bases in DNA. You might gain some testable insights into how the protein finds the correct DNA sequence amid the vast amount of DNA in the nucleus of a cell. The structure might also lead you to speculate on how alterations in the sequence of amino acids in the protein might result in affinity for different DNA sequences, and thus start you thinking about how to design other DNA-binding proteins.

Now look again at the preceding paragraph and examine its *language* rather than its content. The language is typical of that in common use to describe molecular structure and interactions as revealed by various experimental methods, including single-crystal X-ray crystallography, the primary subject of this book. In fact, this

¹“Phase,” from *The Selected Poems, Expanded Edition* by A. R. Ammons. Copyright © 1987, 1977, 1975, 1974, 1972, 1971, 1970, 1966, 1965, 1964, 1955 by A. R. Ammons. Reprinted by permission of W. W. Norton & Company, Inc.

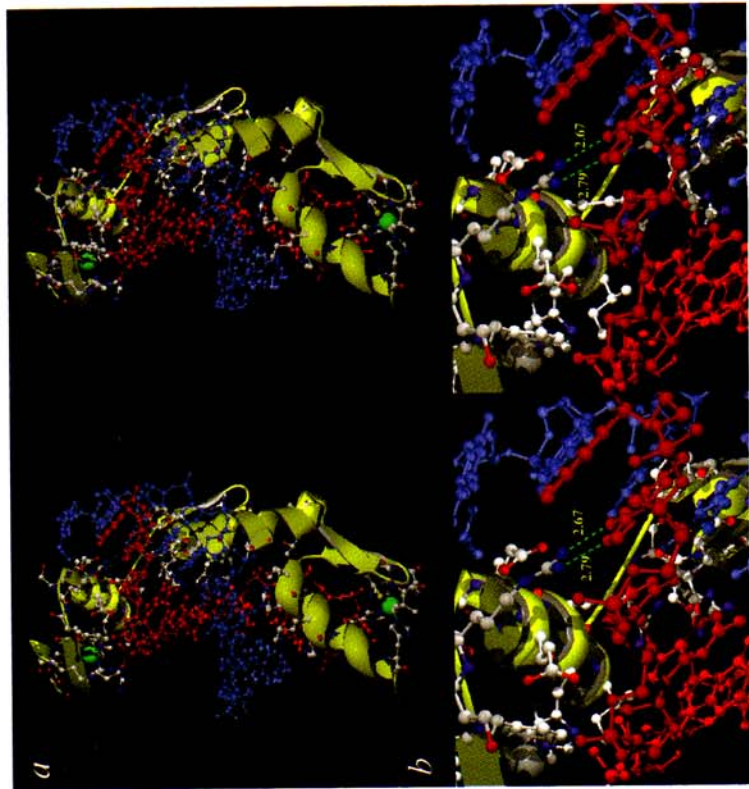


Figure 1.1 ▶ (a) Divergent stereo image of Zif268/DNA complex (N. P. Pavletich and C. O. Pabo, *Science* **252**, 809, 1991). (b) Detail showing hydrogen bonding between arginine-18 of the protein and guanidine-10 of the DNA. Atomic coordinates for preparing this display were obtained from the Protein Data Bank (PDB), which is described in Chapter 7. The PDB file code is 1zaa. To allow easy access to all models shown in this book, I provide file codes in this format: PDB 1zaa. Image created by DeepView (formerly called Swiss-PdbViewer), rendered by POV-Ray. To obtain these programs, see the CMCC home page at <http://www.usm.maine.edu/~rhodes/CMCC/index.html>. For help with viewing stereo images, see Appendix, page 293.

language is shorthand for more precise but cumbersome statements of what we learn from structural studies.

First, Fig. 1.1, of course, shows not molecules, but *models* of molecules, in which structures and interactions are *depicted*, not shown. Second, in this specific case, the models are of molecules not in solution, but in the crystalline state, because the models are derived from analysis of X-ray diffraction by crystals of the Zif268/DNA complex. As such, these models depict the average structure of somewhere between 10^{13} and 10^{15} complexes throughout the crystals that

were studied. In addition, the structures are averaged over the time of the X-ray experiment, which may range from minutes to days.

To draw the conclusions found in the first paragraph requires bringing additional knowledge to bear upon the graphics image, including a more precise knowledge of exactly what we learn from X-ray analysis. The same could be said for structural models derived from spectroscopic data or any other method. In short, the graphics image itself is incomplete. It does not reveal things we may know about the complex from other types of experiments, and it *does not even reveal all that we learn from X-ray crystallography*.

For example, how accurately are the relative positions of atoms known? Are the hydrogen bonds precisely 2.79 and 2.67 Å long, or is there some tolerance in those figures? Is the tolerance large enough to jeopardize the conclusion that hydrogen bonds join these atoms? Further, do we know anything about how rigid this complex is? Do parts of these molecules vibrate, or do they move with respect to each other? Still further, in the aqueous medium of the cell, does this complex have the same structure as in the crystal, which is a solid? As we examine this model, are we really gaining insight into cellular processes? Two final questions may surprise you: First, does the model fully account for the chemical composition of the crystal? In other words, are any of the known contents of the crystal missing from the model? Second, does the crystallographic data suggest additional crystal contents that have not been identified, and thus are not shown in the model?

The answers to these questions are not revealed in the graphics image, which is more akin to a cartoon than to a molecule. Actually, the answers vary from one model to the next, and from one region of a model to another region, but they are usually available to the user of crystallographic models. Some of the answers come from X-ray crystallography itself, so the crystallographer does not miss or overlook them. They are simply less accessible to the noncrystallographer than is the graphics image.

Molecular models obtained from crystallography are in wide use as tools for revealing molecular details of life processes. Scientists use models to learn how molecules “work”: how enzymes catalyze metabolic reactions, how transport proteins load and unload their molecular cargo, how antibodies bind and destroy foreign substances, and how proteins bind to DNA, perhaps turning genes on and off. It is easy for the user of crystallographic models, being anxious to turn otherwise puzzling information into a mechanism of action, to treat models as everyday objects seen as we see clouds, birds, and trees. But the informed user of models sees more than the graphics image, recognizing it as a static depiction of dynamic objects, as the average of many similar structures, as perhaps lacking parts that are present in the crystal but not revealed by the X-ray analysis, as perhaps failing to show as-yet unidentified crystal contents, and finally, as a fallible interpretation of data. The informed user knows that the crystallographic model is richer than the cartoon.

In the following chapters, I offer you the opportunity to become an informed user of crystallographic models. Knowing the richness and limitations of models requires an understanding of the relationship between data and structure.

In Chapter 2, I give an overview of this relationship. In Chapters 3 through 7, the heart of the crystallography in this book, I simply expand Chapter 2 in enough detail to produce an intact chain of logic stretching from diffraction data to final model. Topics come in roughly the same order as the tasks that face a crystallographer pursuing an important structure.

As a practical matter, informed use of a model requires evaluating its quality, which may entail using online *model validation tools* to assess model quality, as well as reading the crystallographic papers and data files that report the new structure, in order to extract from them criteria of model quality. In Chapter 8, I discuss these criteria and provide guided exercises in extracting them from model files themselves and from the literature. Chapter 8 includes an annotated version of a published structure determination and its supporting data, as well as an introduction to online validation tools. Equipped with the background of previous chapters and experienced with the real-world exercises of using validation tools and taking a guided tour through a recent publication, you should be able to read new structure publications in the scientific literature, understand how the structures were obtained, and be aware of just what is known—and what is still unknown—about the molecules under study. Then you should be better equipped to use models wisely.

Chapter 9, “Other Kinds of Macromolecular Methods,” builds on your understanding of X-ray crystallography to help you understand other methods in which diffraction provides insights into the structure of large molecules. These methods include fiber diffraction, neutron diffraction, electron diffraction, and various forms of X-ray spectroscopy. These methods often seem very obscure, but their underlying principles are similar to those of X-ray crystallography.

In Chapter 10, “Other Kinds of Macromolecular Models,” I discuss alternative methods of structure determination: NMR spectroscopy and various forms of theoretical modeling. Just like crystallographic models, NMR and theoretical models are sometimes more, sometimes less, than meets the eye. A brief description of how these models are obtained, along with some analogies among criteria of quality for various types of models, can help make you a wiser user of all types of models.

For new or would-be users of models, I present in Chapter 11 an introduction to molecular modeling, demonstrating how modern graphics programs allow users to display and manipulate models and to perform powerful structure analysis, as well as model validation, on desktop computers. I also provide information on how to use the World Wide Web to obtain graphics programs and learn how to use them. Finally, I introduce you to the Protein Data Bank (PDB), a World Wide Web resource from which you can obtain most of the available macromolecular models.

There is an additional chapter that does not lie between the covers of this book. It is the Crystallography Made Crystal Clear (CMCC) home page on the World Wide Web at www.usm.maine.edu/~rhodes/CMCC. This web site is devoted to making sure that you can find all the Internet resources mentioned here. Because even major Internet resources and addresses may change (the Protein Data Bank

moved while I was writing the second edition of this book), I include only one web address in this book. For all web resources that I describe, I refer you to the CMCC home page. At that web address, I maintain links to all resources mentioned here or, if they disappear or change markedly, to new ones that serve the same or similar functions. For easy reference, the address of the CMCC home page is shown on the cover and title page of this book.

Today’s scientific textbooks and journals are filled with stories about the molecular processes of life. The central character in these stories is often a protein or nucleic acid molecule, a thing never seen in action, never perceived directly. We see models of molecules in books and on computer screens, and we tend to treat them as everyday objects accessible to our normal perceptions. In fact, models are hand-won products of technically difficult data collection and powerful but subtle data analysis. And they are richer and more informative than any single image, or even a rotating computer image, can convey. This book is concerned with where our models of structure come from and how to use them wisely.