Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models

Second Edition

Gale Rhodes
Department of Chemistry
University of Southern Maine
Portland, Maine
CMCC Home Page: www.usm.maine.edu/~rhodes/CMCC

ACADEMIC PRESS

San Diego San Francisco New York Boston London Sydney Tokyo

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. Ammons[1]

[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.

1 Model and Molecule

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. Plate 1 shows a complex composed of a double-stranded DNA molecule and three zinc fingers from the mouse protein Zif268.

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 and measure the hydrogen-bond lengths, and find them to be 2.79 and 2.67 Å. 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 language is shorthand for more precise but cumbersome statements of what we learn from structural studies. First, Plate 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 be as much as several days.

To draw the conclusions found in the first paragraph requires bringing additional knowledge to bear upon the graphics image, including knowledge of just 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 a hydrogen bond joins 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? A final question may surprise you: 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?

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, 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, 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, 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 reading the crystallographic papers and data files that report the new structure and extracting from them criteria of model quality. In Chapter 8, I discuss these criteria and provide a guided exercise in extracting them. The exercise takes the form of annotated excerpts from a published structure determination and its supporting data. Equipped with the background of previous chapters and experienced with the real-world exercise of a guided tour through a recent publication, you should be able to read new structure publications in the scientific literature and understand how the structures were obtained and be aware of just what is known—and what is still unknown—about the molecules under study.

Chapter 9, "Other Diffraction Methods," builds upon 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 Types of 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, even on desktop computers. This chapter also provides information on how to use the World Wide Web to obtain graphics programs and learn how to use them. It also provides an introduction 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, brief 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 page is devoted to making sure that you can find all the Internet resources mentioned here. Because many Internet resources and addresses change rapidly, I did not include them in these pages; but instead, 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 model 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 hard-won products of technically difficult data collection and powerful but subtle data analysis. This book is concerned with where our models of structure come from and how to use them wisely.

2 An Overview of Protein Crystallography

I. Introduction

The most common experimental means of obtaining a detailed picture of a large molecule, allowing the resolution of individual atoms, is to interpret the diffraction of X rays from many identical molecules in an ordered array like a crystal. This method is called *single-crystal X-ray crystallography.* As of this writing, roughly 8000 protein and nucleic-acid structures have been obtained by this method. In addition, the structures of roughly 1300 macromolecules, mostly proteins of fewer than 150 residues, have been solved by nuclear magnetic resonance (NMR) spectroscopy, which provides a model of the molecule in solution, rather than in the crystalline state. Finally, there are theoretical models, built by analogy with the structures of known proteins having similar sequence, or based on simulations of protein folding. All methods have their strengths and weaknesses, and they will undoubtedly coexist as complementary methods for the foreseeable future. One of the goals of this book is to make users of crystallographic models aware of the strengths and weaknesses of X-ray crystallography, so that users' expectations of the resulting models are in keeping with the limitations of crystallographic methods. Chapter 10 provides, in brief, complementary information about other types of models.

This chapter provides a simplified overview of how researchers use the technique of X-ray crystallography to learn macromolecular structures. Chapters 3–8 are simply expansions of the material in this chapter. In order to keep the language simple, I will speak primarily of proteins, but the concepts I describe apply to all macromolecules and macromolecular assemblies that possess ordered structure, including carbohydrates, nucleic acids, and nucleoprotein complexes like ribosomes and whole viruses.

A. Obtaining an image of a microscopic object

When we see an object, light rays bounce off (are diffracted by) the object and enter the eye through the lens, which reconstructs an image of the object and focuses it on the retina. In a simple microscope, an illuminated object is placed just beyond one focal point of a lens, which is called the *objective* lens. The lens collects light diffracted from the object and reconstructs an image beyond the focal point on the opposite side of the lens, as shown in Fig. 2.1.

For a simple lens, the relationship of object position to image position in Fig. 2.1 is $(OF)(IF') = (FL)(F'L)$. Because the distances FL and $F'L$ are constants (but not necessarily equal) for a fixed lens, the distance OF is inversely proportional to the distance IF'. Placing the object near the focal point

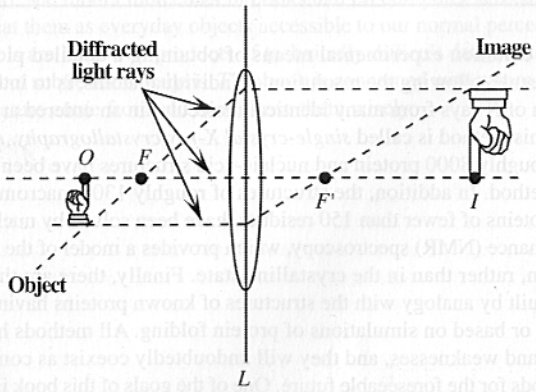


Figure 2.1 Action of a simple lens. Rays parallel to the lens strike the lens and are refracted into paths passing through a focus. Rays passing through a focus strike the lens and are refracted into paths parallel to the lens axis. As a result, the lens produces an image at I of an object at O, such that $(OF)(IF') = (FL)(F'L)$.

F results in a magnified image produced at a considerable distance from F' on the other side of the the lens, which is convenient for viewing. In a compound microscope, the most common type, an additional lens, the *eyepiece*, is added to magnify the image produced by the objective lens.

B. Obtaining images of molecules

In order for the object to diffract light and thus be visible under magnification, the wavelength (λ) of the light must be, roughly speaking, no larger than the object. Visible light, which is electromagnetic radiation with wavelengths of 400–700 nm (nm = 10^{-9} m), cannot produce an image of individual atoms in protein molecules, in which bonded atoms are only about 0.15 nm or 1.5 Å (Å = 10^{-10} m) apart. Electromagnetic radiation of this wavelength falls into the X-ray range, so X rays are diffracted by even the smallest molecules. X-ray analysis of proteins seldom resolves the hydrogen atoms, so the protein models described in this book include elements on only the second and higher rows of the periodic table. The positions of all hydrogen atoms can be deduced on the assumption that bond lengths, bond angles, and conformational angles in proteins are just like those in small organic molecules.

Even though individual atoms diffract X rays, it is still not possible to produce a focused image of a molecule, for two reasons. First, X rays cannot be focused by lenses. Crystallographers sidestep this problem by measuring the directions and strengths (intensities) of the diffracted X rays and then using a computer to simulate an image-reconstructing lens. In short, the computer acts as the lens, computing the image of the object and then displaying it on a screen or drawing it on paper (Fig. 2.2).

Second, a single molecule is a very weak scatterer of X rays. Most of the X rays will pass through a single molecule without being diffracted, so the diffracted beams are too weak to be detected. Analyzing diffraction from crystals, rather than individual molecules, solves this problem. A crystal of a protein contains many ordered molecules in identical orientations, so each molecule diffracts identically, and the diffracted beams for all molecules augment each other to produce strong, detectable X-ray beams.

C. A thumbnail sketch of protein crystallography

In brief, determining the structure of a protein by X-ray crystallography entails growing high-quality crystals of the purified protein, measuring the directions and intensities of X-ray beams diffracted from the crystals, and using a computer to simulate the effects of an objective lens and thus produce an

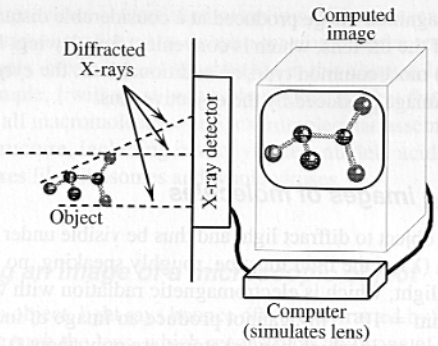


Figure 2.2 Crystallographic analogy of lens action. X-rays diffracted from the object are received and measured by a detector. The measurements are fed to a computer, which simulates the action of a lens to produce a graphics image of the object.

image of the crystal's contents, like the small section of a molecular image shown in Plate 2*a*. Finally, that image must be interpreted, which entails displaying it by computer graphics and building a molecular model that is consistent with the image (Plate 2*b*).

The resulting model is often the only product of crystallography that the user sees. It is therefore easy to think of the model as a real entity that has been directly observed. In fact, our "view" of the molecule is quite indirect. Understanding just how the crystallographer obtains models of protein molecules from diffraction measurements is essential to fully understanding how to use models properly.

II. Crystals

A. The nature of crystals

Under certain circumstances, many molecular substances, including proteins, solidify to form crystals. In entering the crystalline state from solution, individual molecules of the substance adopt one or a few identical orientations. The resulting crystal is an orderly three-dimensional array of molecules, held together by noncovalent interactions. Figure 2.3 shows such a crystalline array of molecules.

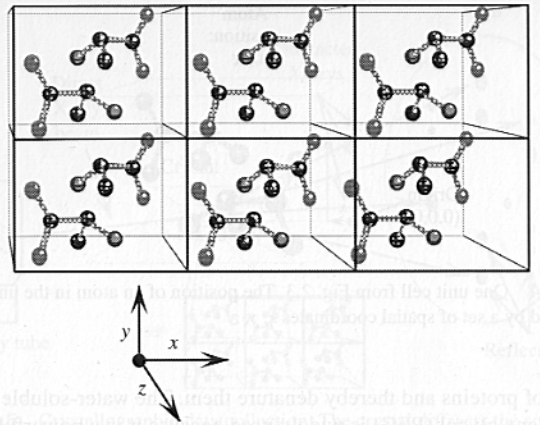


Figure 2.3 Six unit cells in a crystalline lattice. Each unit cell contains two molecules of alanine (hydrogen atoms not shown) in different orientations.

The lines in the figure divide the crystal into identical *unit cells*. The array of points at the corners or vertices of unit cells is called the *lattice*. The unit cell is the smallest and simplest volume element that is completely representative of the whole crystal. If we know the exact contents of the unit cell, we can imagine the whole crystal as an efficiently packed array of many unit cells stacked beside and on top of each other, more or less like identical boxes in a warehouse.

From crystallography, we obtain an image of the electron clouds that surround the molecules in the average unit cell in the crystal. We hope this image will allow us to locate all atoms in the unit cell. The location of an atom is usually given by a set of three-dimensional Cartesian coordinates, x, y, and z. One of the vertices (a lattice point or any other convenient point) is used as the origin of the unit cell's coordinate system and is assigned the coordinates $x = 0$, $y = 0$, and $z = 0$, usually written (0,0,0). See Fig. 2.4.

B. Growing crystals

Crystallographers grow crystals of proteins by slow, controlled precipitation from aqueous solution under conditions that do not denature the protein. A number of substances cause proteins to precipitate. Ionic compounds (salts) precipitate proteins by a process called "salting out." Organic solvents also cause precipitation, but they often interact with hydrophobic

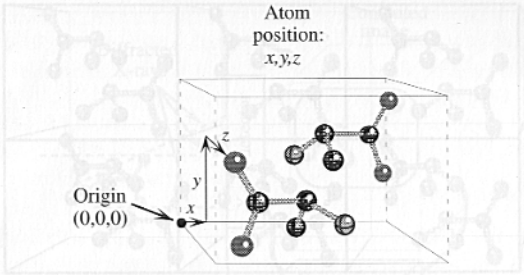


Figure 2.4 One unit cell from Fig. 2.3. The position of an atom in the unit cell can be specified by a set of spatial coordinates *x, y, z*.

portions of proteins and thereby denature them. The water-soluble polymer polyethylene glycol (PEG) is widely used because it is a powerful precipitant and a weak denaturant. It is available in preparations of different average molecular masses, such as PEG 400, with average molecular mass of 400 daltons.

One simple means of causing slow precipitation is to add denaturant to an aqueous solution of protein until the denaturant concentration is just below that required to precipitate the protein. Then water is allowed to evaporate slowly, which gently raises the concentration of both protein and denaturant until precipitation occurs. Whether the protein forms crystals or instead forms a useless amorphous solid depends on many properties of the solution, including protein concentration, temperature, pH, and ionic strength. Finding the exact conditions to produce good crystals of a specific protein often requires many careful trials and is perhaps more art than science. I will examine crystallization methods in Chapter 3.

III. Collecting X-ray data

Figure 2.5 depicts the collection of X-ray diffraction data. A crystal is mounted between an X-ray source and an X-ray detector. The crystal lies in the path of a narrow beam of X rays coming from the source. A simple detector is X-ray film, which when developed exhibits dark spots where X-ray beams have impinged. These spots are called *reflections* because they emerge from the crystal as if reflected from planes of atoms.

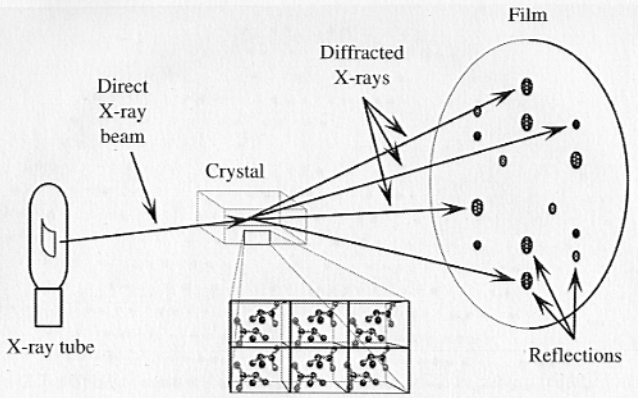


Figure 2.5 Crystallographic data collection. The crystal diffracts the source beam into many discrete beams, each of which produces a distinct spot (reflection) on the film. The positions and intensities of these reflections contain the information needed to determine molecular structures.

Figure 2.6 shows the complex diffraction pattern of X rays produced on film by a protein crystal. Notice that the crystal diffracts the source beam into many discrete beams, each of which produces a distinct reflection on the film. The greater the intensity of the X-ray beam that reaches a particular position, the darker the reflection.

An optical scanner precisely measures the position and the intensity of each reflection and transmits this information in digital form to a computer for analysis. The position of a reflection can be used to obtain the direction in which that particular beam was diffracted by the crystal. The intensity of a reflection is obtained by measuring the optical absorbance of the spot on the film, giving a measure of the strength of the diffracted beam that produced the spot. The computer program that reconstructs an image of the molecules in the unit cell requires these two parameters, the beam intensity and direction, for each diffracted beam.

Although film for data collection has largely been replaced by devices that feed diffraction data (positions and intensities of each reflection) directly into computers, I will continue to speak of the data as if collected on film because of the simplicity of that format, and because diffraction patterns are usually published in a form identical to their appearance on film. I will discuss other methods of collecting data in Chapter 4.

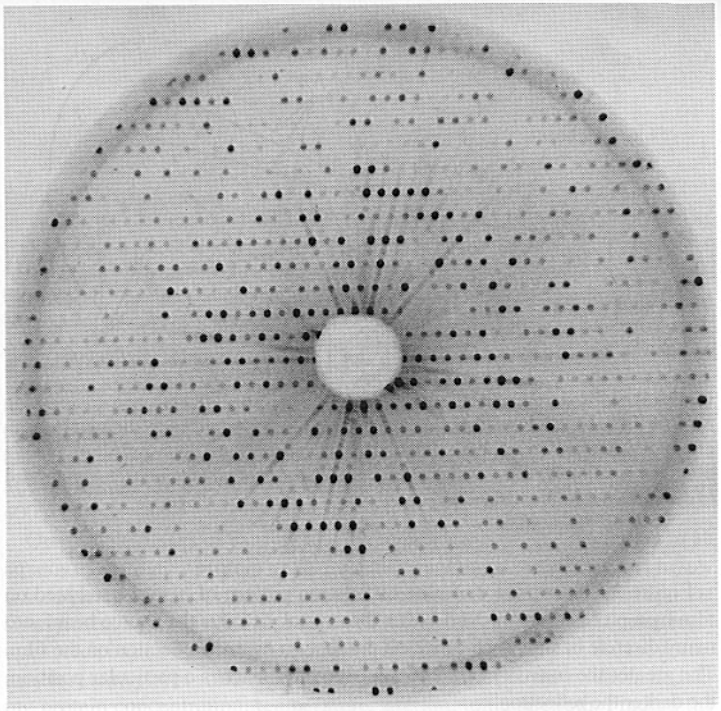


Figure 2.6 Diffraction pattern from a crystal of the MoFe (molybdenum–iron) protein of the enzyme nitrogenase from *Clostridium pasteurianum*. Notice that the reflections lie in a regular pattern, but their intensities (darkness of spots) are highly variable. [The hole in the middle of the pattern results from a small metal disk (beam stop) used to prevent the direct X-ray beam, most of which passes straight through the crystal, from destroying the center of the film.] Photo courtesy of Professor Jeffery Bolin.

IV. Diffraction

A. Simple objects

You can develop some visual intuition for the information available from X-ray diffraction by examining the diffraction patterns of simple objects like spheres or arrays of spheres (Figs. 2.7–2.10). Figure 2.7 depicts diffraction by a single sphere, shown in cross section on the left. The diffraction pattern, on

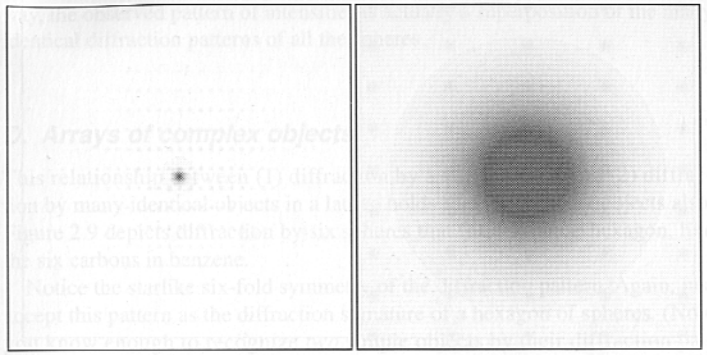


Figure 2.7 Sphere (cross-section, on left) and its diffraction pattern (right). Images for Figures 2.7–2.10 were generously provided by Dr. Kevin Cowtan.

the right, exhibits high intensity at the center, and smoothly decreasing intensity as the diffraction angle increases.[1]

For now, just accept the fact that diffraction by a sphere produces this pattern, and think of it as the diffraction signature of a sphere. In a sense, you are already equipped to do very simple structure determination; that is, you can now recognize a simple sphere by its diffraction pattern.

B. Arrays of simple objects: Real and reciprocal lattices

Figure 2.8 depicts diffraction by a crystalline array of spheres, with a cross section of the crystal on the left, and its diffraction pattern on the right.

The diffraction pattern, like that produced by crystalline nitrogenase (Fig. 2.6), consists of reflections (spots) in an orderly array on the film. The spacing of the reflections varies with the spacing of the spheres in their array. Specifically, observe that although the lattice spacing of the crystal is smaller vertically, the diffraction spacing is smaller horizontally. In fact, there is a simple inverse relationship between the spacing of unit cells in the crystalline lattice, called the *real lattice*, and the spacing of reflections in the lattice on the film, which, because of its inverse relationship to the real lattice, is called the *reciprocal lattice*.

[1]The images shown in Figures 2.7–2.10 are computed, rather than experimental, diffraction patterns. Computation of these patterns involves use of the Fourier transform (Section V.E).

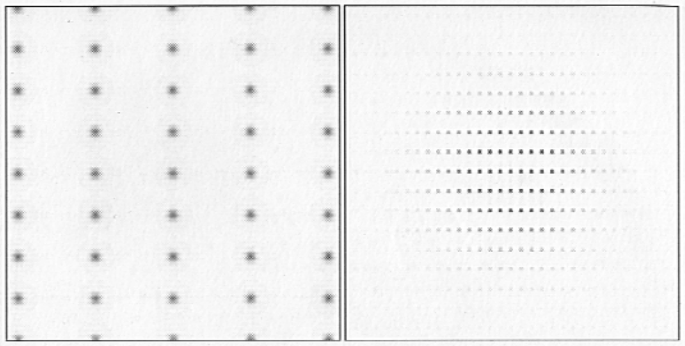


Figure 2.8 Lattice of spheres (left) and its diffraction pattern (right). If you look at the pattern and blur your eyes, you will see the diffraction pattern of a sphere. The pattern is that of the average sphere in the real lattice, but it is sampled at the reciprocal lattice points.

Because the real lattice spacing is inversely proportional to the spacing of reflections, crystallographers can calculate the dimensions, in angstroms, of the unit cell of the crystalline material from the spacings of the reciprocal lattice on the X-ray film (Chapter 4). The simplicity of this relationship is a dramatic example of how the macroscopic dimensions of the diffraction pattern are connected to the submicroscopic dimensions of the crystal.

C. Intensities of reflections

Now look at the intensities of the reflections in Fig. 2.8. Some are intense ("bright"), whereas others are weak or perhaps missing from the otherwise evenly spaced pattern. These variations in intensity contain important information. If you blur your eyes slightly while looking at the diffraction pattern, so that you cannot see individual spots, you will see the intensity pattern characteristic of diffraction by a sphere, with lower intensities farther from the center, as in Fig. 2.7. (You just determined your first crystallographic structure.) The diffraction pattern of spheres in a lattice is simply the diffraction pattern of the average sphere in the lattice, but this pattern is incomplete. The pattern is *sampled* at points whose spacings vary inversely with real-lattice spacings. The pattern of varied intensities is that of the *average* sphere because all the spheres contribute to the observed pattern. To put it another

way, the observed pattern of intensities is actually a superposition of the many identical diffraction patterns of all the spheres.

D. Arrays of complex objects

This relationship between (1) diffraction by a single object and (2) diffraction by many identical objects in a lattice holds true for complex objects also. Figure 2.9 depicts diffraction by six spheres that form a planar hexagon, like the six carbons in benzene.

Notice the starlike six-fold symmetry of the diffraction pattern. Again, just accept this pattern as the diffraction signature of a hexagon of spheres. (Now you know enough to recognize *two* simple objects by their diffraction patterns.) Figure 2.10 depicts diffraction by these hexagonal objects in a lattice of the same dimensions as that in Fig. 2.8.

As before, the spacing of reflections varies reciprocally with lattice spacing, but if you blur your eyes slightly, or compare Figs. 2.9 and 2.10 carefully, you will see that the starlike signature of a single hexagonal cluster is present in Fig. 2.10. From these simple examples, you can see that the reciprocal-lattice spacing (the spacing of reflections in the diffraction pattern) is characteristic of (inversely related to) the spacing of identical objects in the crystal, whereas the reflection intensities are characteristic of the shape of the individual objects. From the reciprocal-lattice spacing in a diffraction pattern, we can compute the dimensions of the unit cell. From the intensities of the reflections,

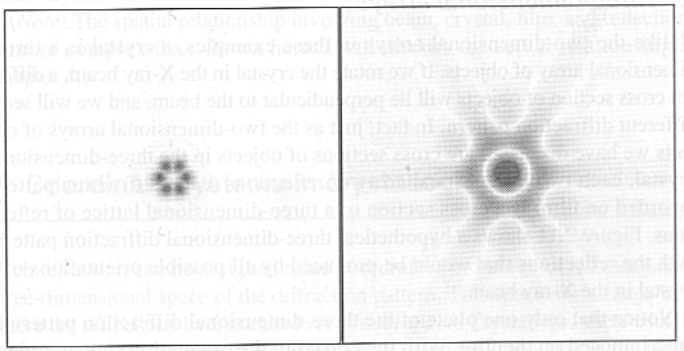


Figure 2.9 A planar hexagon of spheres (left) and its diffraction pattern (right).

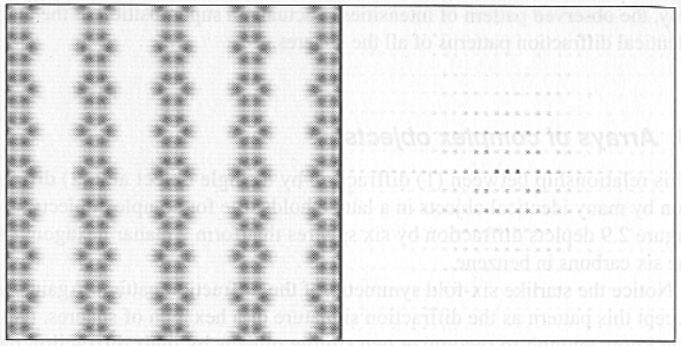


Figure 2.10 Lattices of hexagons (left) and its diffraction pattern (right). If you look at the pattern and blur your eyes, you will see the diffraction pattern of a hexagon. The pattern is that of the average hexagon in the real lattice, but it is sampled at the reciprocal lattice points.

we can learn the shape of the individual molecules that compose the crystal. It is actually advantageous that the object's diffraction pattern is sampled at reciprocal-lattice positions. This sampling reduces the number of intensity measurements we must take from the film and makes it easier to program a computer to locate and measure the intensities.

E. Three-dimensional arrays

Unlike the two-dimensional arrays in these examples, a crystal is a three-dimensional array of objects. If we rotate the crystal in the X-ray beam, a different cross section of objects will lie perpendicular to the beam, and we will see a different diffraction pattern. In fact, just as the two-dimensional arrays of objects we have discussed are cross sections of objects in the three-dimensional crystal, each two-dimensional array of reflections (each diffraction pattern recorded on film) is a cross section of a three-dimensional lattice of reflections. Figure 2.11 shows a hypothetical three-dimensional diffraction pattern, with the reflections that would be produced by all possible orientations of a crystal in the X-ray beam.

Notice that only one plane of the three-dimensional diffraction pattern is superimposed on the film. With the crystal in the orientation shown, reflections shown in the plane of the film (solid spots) are the only reflections that produce spots on the film. In order to measure the directions and intensities of

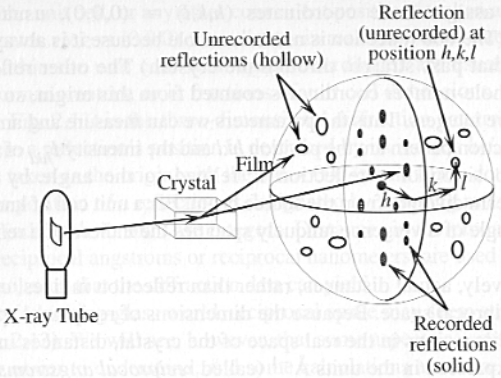


Figure 2.11 Crystallographic data collection, showing reflections measured at one particular crystal orientation (solid, on film) and those that could be measured at other orientations (hollow, within the sphere but not on the film). The relationship between measured and unmeasured reflections is more complex than shown here (see Chapter 4).

all additional reflections (shown as hollow spots), the crystallographer must collect diffraction patterns from all unique orientations of the crystal with respect to the X-ray beam. The direct result of crystallographic data collection is a list of intensities for each point in the three-dimensional reciprocal lattice. This set of data is the raw material for determining the structures of molecules in the crystal.

(*Note*: The spatial relationship involving beam, crystal, film, and reflections is more complex than shown here. I will discuss the actual relationship in Chapter 4.)

V. Coordinate systems in crystallography

Each reflection can be assigned three coordinates or *indices* in the imaginary three-dimensional space of the diffraction pattern. This space, the strange land where the reflections live, is called *reciprocal space*. Crystallographers usually use h, k, and l to designate the position of an individual reflection in the reciprocal space of the diffraction pattern. The central reflection (the round solid spot at the center of the film in Fig. 2.11) is taken as the origin in reciprocal

space and assigned the coordinates $(h,k,l) = (0,0,0)$, usually written $hkl = 000$. (The 000 reflection is not measurable because it is always obscured by X rays that pass straight through the crystal.) The other reflections are assigned whole-number coordinates counted from this origin, so the indices *h, k,* and *l* are integers. Thus the parameters we can measure and analyze in the X-ray diffraction pattern are the position *hkl* and the intensity I_{hkl} of each reflection. The position of a reflection is related to the angle by which the diffracted beam diverges from the source beam. For a unit cell of known dimensions, the angle of divergence uniquely specifies the indices of a reflection (see Chapter 4).

Alternatively, actual distances, rather than reflection indices, can be measured in reciprocal space. Because the dimensions of reciprocal space are the inverse of dimensions in the real space of the crystal, distances in reciprocal space are expressed in the units $Å^{-1}$ (called *reciprocal angstroms*). Roughly speaking, the inverse of the reciprocal-space distance from the origin out to the most distant measurable reflections gives the potential resolution of the model that we can obtain from the data. So a crystal that gives measurable reflections out to a distance of 1/(3 Å) from the origin should yield a model with a resolution of 3 Å.

The crystallographer works back and forth between two different coordinate systems. I will review them briefly. The first system (see Fig. 2.4) is the unit cell (real space), where an atom's position is described by its coordinates *x,y,z*.

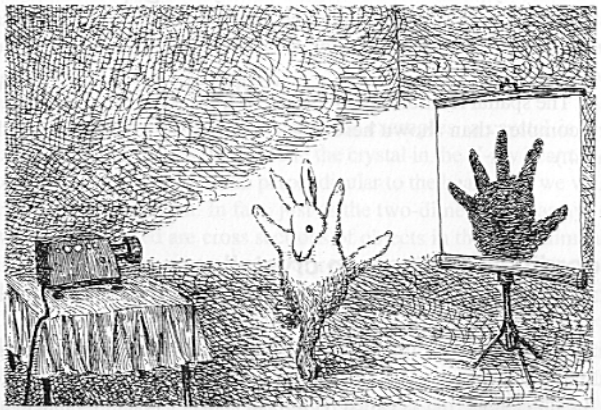


Figure 2.12 Fun in reciprocal space. © The New Yorker Collection, 1991. John O'Brien, from cartoonbank.com. All rights reserved.

A vertex of the unit cell, or any other convenient position, is taken as the origin, with coordinates $x,y,z = (0,0,0)$. Coordinates in real space designate real spatial positions within the unit cell. Real-space coordinates are usually given in angstroms or nanometers, or in fractions of unit cell dimensions. The second system (see Fig. 2.11) is the three-dimensional diffraction pattern (reciprocal space), where a reflection's position is described by its indices *hkl*. The central reflection is taken as the origin with the index 000 (round black dot at center of sphere). The position of a reflection is designated by counting reflections from 000, so the indices *h, k,* and *l* are integers. Distances in reciprocal space, expressed in reciprocal angstroms or reciprocal nanometers, are used to judge the potential resolution that the diffraction data can yield.

Like Alice's looking-glass world, reciprocal space may seem strange to you at first (Fig. 2.12). We will see, however, that some aspects of crystallography are actually easier to understand, and some calculations are more convenient, in reciprocal space than in real space (Chapter 4).

VI. The mathematics of crystallography: A brief description

The problem of determining the structure of objects in a crystalline array from their diffraction pattern is, in essence, a matter of converting the experimentally accessible information in the reciprocal space of the diffraction pattern to otherwise inaccessible information about the real space inside the unit cell. Remember that a computer program that makes this conversion is acting as a simulated lens to reconstruct an image from diffracted radiation. Each reflection is produced by a beam of electromagnetic radiation (X rays), so the computations entail treating the reflections as waves and recombining these waves to produce an image of the molecules in the unit cell.

A. Wave equations: Periodic functions

Each reflection is the result of diffraction from complicated objects, the molecules in the unit cell, so the resulting wave is complicated also. Before considering how the computer represents such an intricate wave, let us consider mathematical descriptions of the simplest waves.

A simple wave, like that of visible light or X rays, can be described by a periodic function, for instance, an equation of the form

$$f(x) = F\cos 2\pi(hx + \alpha) \tag{2.1}$$

or

$$f(x) = F\sin 2\pi(hx + \alpha). \tag{2.2}$$

In these functions, $f(x)$ specifies the vertical height of the wave at any horizontal position x along the wave. The variable x and the constant α are angles expressed in fractions of the wavelength; that is, $x = 1$ implies a position of one full wavelength (2π radians or 360°) from the origin. The constant F specifies the amplitude (the height of the crests and troughs) of the wave. For example, the crests of the wave $f(x) = 3\cos 2\pi x$ are three times as high and the troughs are three times as deep as those of the wave $f(x) = \cos 2\pi x$ (compare *b* with *a* in Fig. 2.13).

The constant h in a simple wave equation specifies the frequency or wavelength of the wave. For example, the wave $f(x) = \cos 2\pi(5x)$ has five times the frequency (or one-fifth the wavelength) of the wave $f(x) = \cos 2\pi x$ (compare *c* with *a* in Fig. 2.13). (In the wave equations used in this book, h takes on integral values only.)

Finally, the constant α specifies the phase of the wave, that is, the position of the wave with respect to the origin of the coordinate system on which the wave is plotted. For example, the position of the wave $f(x) = \cos 2\pi(x + 1/4)$ is shifted by one-quarter of 2π radians (or one-quarter of a wavelength, or 90°) from the position of the wave $f(x) = \cos 2\pi x$ (compare Fig. 2.13*d* with Fig. 2.13*a*). Because the wave is repetitive, with a repeat distance of one wavelength or 2π radians, a phase of 1/4 is the same as a phase of 1 1/4, or 2 1/4, or 3 1/4, and so on. In radians, a phase of 0 is the same as a phase of 2π, or 4π, or 6π, and so on.

These equations describe one-dimensional waves, in which a property (in this case, the height of the wave) varies is one direction. Visualizing a one-dimensional function $f(x)$ requires a two-dimensional graph, with the second dimension used to represent the numerical value of $f(x)$. For example, if $f(x)$ describes the electrical part of an electromagnetic wave, the x-axis is the direction the wave is moving, and the height of the wave at any position on the x-axis represents the momentary strength of the electrical field at a distance x from the origin. The field strength is in no real sense perpendicular to x, but it is convenient to use the perpendicular direction to show the numerical value of the field strength. In general, visualizing a function in n dimensions requires $n + 1$ dimensions.

B. Complicated periodic functions: Fourier series

As discussed in Section VI.A, any simple sine or cosine wave can therefore be described by three constants—the amplitude F, the frequency h, and the

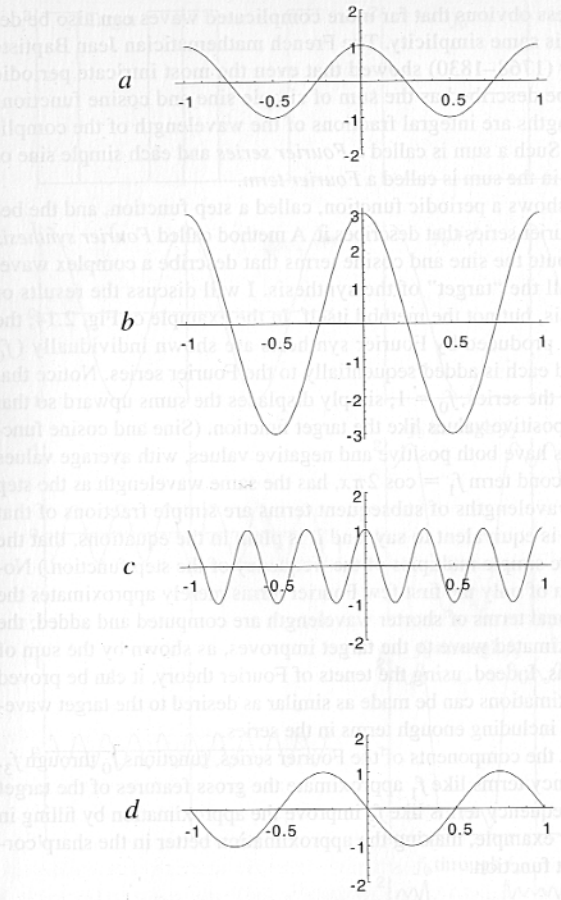


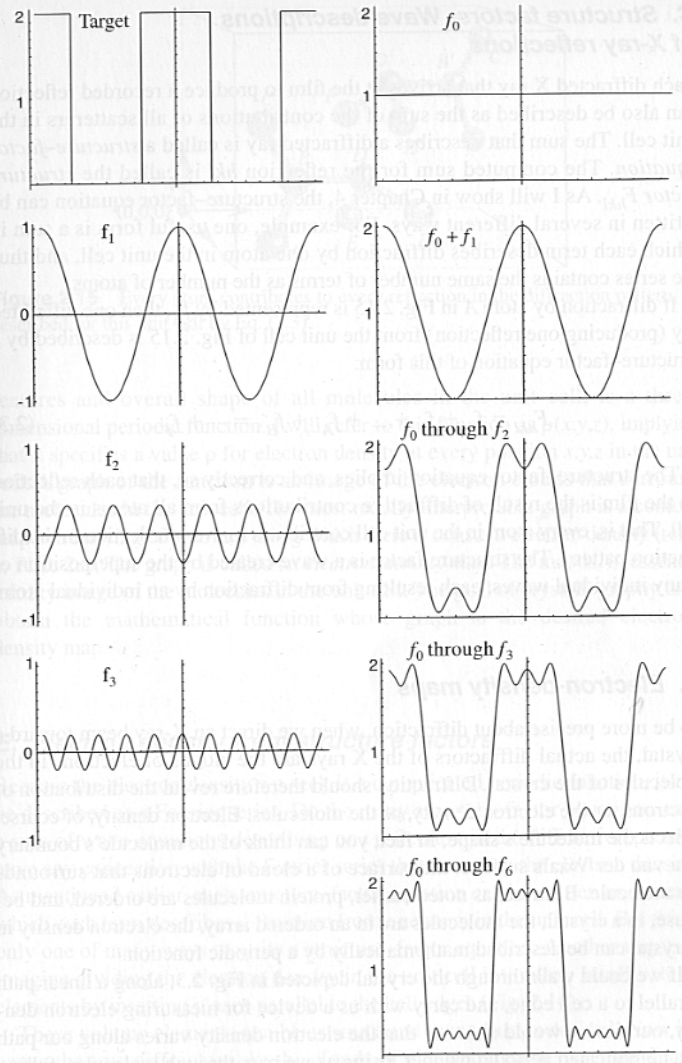
Figure 2.13 Graphs of four simple wave equations $f(x) = F\cos 2\pi(hx + \alpha)$. (*a*) $F = 1$, $h = 1$, $\alpha = 0$: $f(x) = \cos 2\pi(x)$. (*b*) $F = 3$, $h = 1$, $\alpha = 0$: $f(x) = 3\cos 2\pi(x)$. Increasing F increases the amplitude of the wave. (*c*) $F = 1$, $h = 3$, $\alpha = 0$: $f(x) = \cos 2\pi(3x)$. Increasing h increases the frequency (or decreases the wavelength λ) of the wave. (*d*) $F = 1$, $h = 1$, $\alpha = 1/4$: $f(x) = \cos 2\pi(x + 1/4)$. Changing α changes the phase (position) of the wave.

phase α. It is less obvious that far more complicated waves can also be described with this same simplicity. The French mathematician Jean Baptiste Joseph Fourier (1768–1830) showed that even the most intricate periodic functions can be described as the sum of simple sine and cosine functions whose wavelengths are integral fractions of the wavelength of the complicated function. Such a sum is called a *Fourier series* and each simple sine or cosine function in the sum is called a *Fourier term.*

Figure 2.14 shows a periodic function, called a step function, and the beginning of a Fourier series that describes it. A method called *Fourier synthesis* is used to compute the sine and cosine terms that describe a complex wave, which I will call the "target" of the synthesis. I will discuss the results of Fourier synthesis, but not the method itself. In the example of Fig. 2.14, the first four terms produced by Fourier synthesis are shown individually (f_0 through f_3), and each is added sequentially to the Fourier series. Notice that the first term in the series, $f_0 = 1$, simply displaces the sums upward so that they have only positive values like the target function. (Sine and cosine functions themselves have both positive and negative values, with average values of zero.) The second term $f_1 = \cos 2\pi x$, has the same wavelength as the step function, and wavelengths of subsequent terms are simple fractions of that wavelength. (It is equivalent to say, and it is plain in the equations, that the frequencies h are simple multiples of the frequency of the step function.) Notice that the sum of only the first few Fourier terms merely approximates the target. If additional terms of shorter wavelength are computed and added, the fit of the approximated wave to the target improves, as shown by the sum of the first six terms. Indeed, using the tenets of Fourier theory, it can be proved that such approximations can be made as similar as desired to the target waveform, simply by including enough terms in the series.

Look again at the components of the Fourier series, functions f_0 through f_3. The low-frequency terms like f_1 approximate the gross features of the target wave. Higher-frequency terms like f_3 improve the approximation by filling in finer details, for example, making the approximation better in the sharp corners of the target function.

Figure 2.14 Beginning of a Fourier series to approximate a target function, in this case, a step function or square wave. $f_0 = 1; f_1 = \cos 2\pi(x); f_2 = (-1/3)\cos 2\pi(3x)$; $f_3 = 1/5)\cos 2\pi(5x)$. In the left column are the target and terms f_1 through f_3. In the right column are f_0 and the succeeding sums as each term is added to f_0. Notice that the approximaton improves (i.e. each successive sum looks more like the target) as the number of Fourier terms in the sum increase. In the last graph, terms f_4, f_5 and f_6 are added (but not shown separately) to show further improvement in the approximation.



C. Structure factors: Wave descriptions of X-ray reflections

Each diffracted X ray that arrives at the film to produce a recorded reflection can also be described as the sum of the contributions of all scatterers in the unit cell. The sum that describes a diffracted ray is called a *structure–factor equation*. The computed sum for the reflection *hkl* is called the *structure factor* F_{hkl}. As I will show in Chapter 4, the structure–factor equation can be written in several different ways. For example, one useful form is a sum in which each term describes diffraction by one atom in the unit cell, and thus the series contains the same number of terms as the number of atoms.

If diffraction by atom *A* in Fig. 2.15 is represented by f_A, then one diffracted ray (producing one reflection) from the unit cell of Fig. 2.15 is described by a structure–factor equation of this form:

$$F_{hkl} = f_A + f_B + \ldots + f_{A'} + f_{B'} = \cdots + f_{F'}. \qquad (2.3)$$

The structure–factor equation implies, and correctly so, that each reflection on the film is the result of diffractive contributions from all atoms in the unit cell. That is, every atom in the unit cell contributes to every reflection in the diffraction pattern. The structure factor is a wave created by the superposition of many individual waves, each resulting from diffraction by an individual atom.

D. Electron-density maps

To be more precise about diffraction, when we direct an X-ray beam toward a crystal, the actual diffractors of the X rays are the clouds of electrons in the molecules of the crystal. Diffraction should therefore reveal the distribution of electrons, or the electron density, of the molecules. Electron density, of course, reflects the molecule's shape; in fact, you can think of the molecule's boundary as a van der Waals surface, the surface of a cloud of electrons that surrounds the molecule. Because, as noted earlier, protein molecules are ordered, and because, in a crystal, the molecules are in an ordered array, the electron density in a crystal can be described mathematically by a periodic function.

If we could walk through the crystal depicted in Fig. 2.3, along a linear path parallel to a cell edge, and carry with us a device for measuring electron density, our device would show us that the electron density varies along our path in a complicated periodic manner, rising as we pass through molecules, falling in the space between molecules, and repeating its variation identically as we pass through each unit cell. Because this statement is true for linear paths parallel to all three cell edges, the electron density, which describes the surface

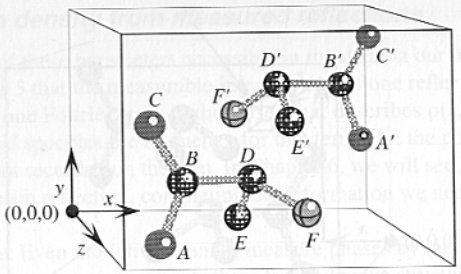


Figure 2.15 Every atom contributes to every reflection in the diffraction pattern, as described for this unit cell by Eq. (2.3).

features and overall shape of all molecules in the unit cell, is a three-dimensional periodic function. I will refer to this function as $\rho(x,y,z)$, implying that it specifies a value ρ for electron density at every position x,y,z in the unit cell. A graph of the function is an image of the electron clouds that surround the molecules in the unit cell. The most readily interpretable graph is a contour map—a drawing of a surface along which there is constant electron density (refer to Plate 2*a*). The graph is called an *electron-density map*. The map is, in essence, a fuzzy image of the molecules in the unit cell. The goal of crystallography is to obtain the mathematical function whose graph is the desired electron-density map.

E. Electron density from structure factors

Because the electron density we seek is a complicated periodic function, it can be described as a Fourier series. Do the many structure–factor equations, each a sum of wave equations describing one reflection in the diffraction pattern, have any connection with the Fourier series that describes the electron density? As mentioned earlier, each structure–factor equation can be written as a sum in which each term describes diffraction from one atom in the unit cell. But this is only one of many ways to write a structure–factor equation. Another way is to imagine dividing the electron density in the unit cell into many small volume elements by inserting planes parallel to the cell edges (Fig. 2.16).

These volume elements can be as small and numerous as desired. Now because the true diffractors are the clouds of electrons, each structure–factor equation can be written as a sum in which each term describes diffraction by the electrons in one volume element. In this sum, each term contains the average numerical value of the desired electron density function $\rho(x,y,z)$ within

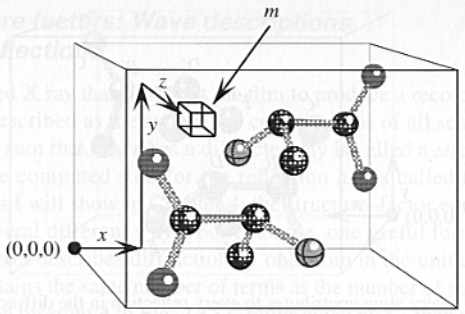


Figure 2.16 Small volume element m within the unit cell, one of many elements formed by subdividing the unit cell with planes parallel to the cell edges. The average electron density within m is $\rho_m(x,y,z)$. Every volume element contributes to every reflection in the diffraction pattern, as described by Eq. (2.4).

one volume element. If the cell is divided into n elements, and the average electron density in volume element m is ρ_m, then one diffracted ray from the unit cell of Fig. 2.16 is described by a structure–factor equation of this form:

$$F_{hkl} = f(\rho_1) + f(\rho_2) + \cdots + f(\rho_m) + \cdots + f(\rho_n). \qquad (2.4)$$

So each reflection is described by an equation like this, giving us a large number of equations describing reflections in terms of the electron density. Is there any way to solve these equations for the function $\rho(x,y,z)$ in terms of the measured reflections? After all, structure factors like Eq. (2.4) describe the reflections in terms of $\rho(x,y,z)$, which is precisely the function the crystallographer is trying to learn. I will show in Chapter 5 that a mathematical operation called the Fourier transform solves the structure–factor equations for the desired function $\rho(x,y,z)$, just as if they were a set of simultaneous equations describing $\rho(x,y,z)$ in terms of the amplitudes, frequencies, and phases of the reflections.

The Fourier transform describes precisely the mathematical relationship between an object and its diffraction pattern. In Figs. 2.7–2.10, the diffraction patterns are the Fourier transforms of the corresponding objects or arrays of objects. To put it another way, the Fourier transform is the lens-simulating operation that a computer performs to produce an image of molecules (or more precisely, of electron clouds) in the crystal. This view of $\rho(x,y,z)$ as the Fourier transform of the structure factors implies that if we can measure three parameters—amplitude, frequency, and phase—of *each* reflection, then we can obtain the function $\rho(x,y,z)$, graph the function, and "see" a fuzzy image of the molecules in the unit cell.

F. Electron density from measured reflections

Are all three of these parameters accessible in the data on our films? We will see in Chapter 5 that the measurable intensity I_{hkl} of one reflection gives the amplitude of one Fourier term in the series that describes $\rho(x,y,z)$, and that the position hkl specifies the frequency for that term. But the phase α of each reflection is not recorded on the film. In Chapter 6, we will see how to obtain the phase of each reflection, completing the information we need to calculate $\rho(x,y,z)$.

A final note: Even though we cannot measure phases by simply collecting diffraction patterns, we can compute them from a known structure, and we can depict them by adding color to images like those of Figures 2.7–2.10. In his innovative World Wide Web *Book of Fourier*[2], Kevin Cowtan illustrates phases in diffraction patterns in this clever manner. For example, Plate 3*a* shows a lattice of simple objects, each one like the carbon atoms in ethylbenzene. Plate 3*b* is the computed Fourier transform of (*a*). Image (*c*) depicts a lattice of the objects in (*a*), and *d*) depicts the corresponding diffraction pattern.

Because patterns (*b*) and (*d*) were *computed* from objects of known structure, rather than measured experimentally from real objects, the phases are known. The phase of each reflection is depicted by its color, according to the color wheel (*f*). The phase can be expressed as an angle between 0° and 360° [this is the angle α in Eqs. (2.1) or (2.2)]. In Plate 3, the phase angle of each region (in *b*) or reflection (in *d*) is the angle that corresponds to the angle of its color on the color wheel (*f*). For example, red corresponds to a phase angle of 0°, and green to an angle of about 135°. So a dark red reflection has a high intensity and a phase angle of 0°. A pale green reflection has a low intensity and a phase angle of about 135°. This depiction, then, gives a full description of each reflection, including the phase angle that we do not learn from diffraction experiments, which would give us only the intensities, as shown in (*e*). In a sense then, Figs. 2.7 through 2.10 and Plate 3*e* show *diffraction* patterns, whereas Plates 3*b* and 3*d* show *structure–factor* patterns, which depict the structure factors fully. Note again that (*d*) is a sampling of (*b*) at points corresponding to the reciprocal lattice of the lattice in (*c*). In other words, the diffraction pattern (*d*) still contains the diffraction signature, including both intensities and phases, of the object in (*a*).

In these terms, I will restate a central problem of crystallography: In order to determine a structure, we need a full-color version of the diffraction pattern—that is, a full description of the structure factors. But diffraction experiments give us only the black-and-white version, the intensities of the

[2]Access to Kevin Cowtan's *Book of Fourier* is provided at the CMCC Home Page, www.usm.maine.edu/~rhodes/CMCC.

reflections, but no information about their phases. We must learn the phase angles from further experimentation, as described fully in Chapter 6.

G. Obtaining a model

Having obtained $\rho(x,y,z)$, we graph the function to produce an electron-density map, an image of the molecules in the unit cell. Finally, we interpret the map by building a model that fits it (refer to Plate 2*b*). In interpreting the molecular image and building the model, a crystallographer takes advantage of all current knowledge about the protein under investigation, as well as knowledge about protein structure in general. Probably the most important information required is the sequence of amino acids in the protein. In a few rare instances, the amino-acid sequence has been learned from the crystallographic structure. But in almost all cases, crystallographers know the sequence to start with, from the work of chemists or molecular biologists, and use it to help them interpret the image obtained from crystallography. In effect, the crystallographer starts with knowledge of the chemical structure, but without knowledge of the conformation. Interpreting the image amounts to finding a chemically realistic conformation that fits the image precisely.

A crystallographer interprets a map by displaying it on a graphics computer and building a graphics model within it. The final model must be (1) consistent with the image and (2) chemically realistic; that is, it must possess bond lengths, bond angles, conformational angles, and distances between neighboring groups that are all in keeping with established principles of molecular structure and stereochemistry. With such a model in hand, the crystallographer can begin to explore the model for clues about its function.

In Chapters 3–7, I will take up in more detail the principles introduced in this chapter.