Chapter 1

Introduction: Preparing for your study

Crystallography borders, naturally, on pure physics, chemistry, biology, mineralogy, technology and also on mathematics, but is distinguished by being concerned with the methods and results of investigating the arrangement of atoms in matter, particularly when that arrangement has regular features.

Paul Ewald, Acta Crystallographica (1948), 1, 2

Chapter 1 provides a brief introduction emphasizing the relation between form and function, followed by an overview of the principles and challenges of macromolecular crystallography. The relevance of protein crystal structures compared with the solution state is discussed. A brief interlude about molecular cryptanalysis leads into an overview of the method, including general guidance on planning and executing a structure determination in order to answer a biological question. A brief overview of coordinate files and the role of the Protein Data Bank concludes the introduction. The main purpose of this chapter is to provide a quick overview of the method, also suitable for entry level courses or users of crystallographic models, while preparing the ground for the serious study of the remaining chapters.

1.1 Molecular structure defines function

The molecular structure of matter defines its properties and function. This simple but far-reaching statement is true for all matter. The properties of gases, liquids, rocks, semiconductors, and small organic molecules are defined by their molecular structure, as are the functions of proteins and their complex macromolecular assemblies.¹ The motivation to use X-ray crystallography as the primary means of macromolecular structure determination is founded on the fact that accurate and precise molecular structure models—often revealing details at the atomic resolution level—can be obtained rapidly and reliably by means of X-ray crystallography. About 90% of all structure models deposited in the Protein Data Bank (PDB) are determined by X-ray crystallography, while the remaining 10% are determined by solution nuclear magnetic resonance (NMR) spectroscopy, largely in the 5 to 25 kDa range of molecular weight.

X-ray crystallography can provide highly detailed molecular structure models of large molecular assemblies. Elucidating the atomic details of molecular interactions is particularly important, for example, in the clarification of enzymatic mechanisms and is essential for drug target structures serving as
leads in *structure guided drug design*. In addition to providing accurate models of molecular structures, another advantage of crystallographic structure determination is that no limits of principle prevent the accurate description of very large molecular structures and molecular complexes, as evidenced in the nearly 2 MDa structure of the 50S ribosomal subunit containing 27 different proteins (~4000 residues) and the ribosomal 5S and 23S RNA, together comprising 2833 nucleotides. The even larger 66 MDa protein capsid of the bacteriophage head PRD1, is assembled from ~2000 subunits of 18 different proteins (illustrated in Figure 2-2). As the emphasis in modern biosciences shifts toward the understanding of entire biological systems, the capability to determine large, multi-unit complex structures, thereby providing a detailed picture of specific protein–protein interactions within their overall interaction networks, will become increasingly important.

**Crystallography provides the foundation of modern structural biology and structure-based drug discovery.** Exciting structures that increase our understanding of molecular form and function are determined at an ever-increasing pace (Figure 1-1). Whole virus particles, the F₁-ATPase, the potassium ion channel, fundamental structural information about cellular recognition by antigen presenting cells, immunity and T-cell signaling, or the functional insight provided by crystal structures of the ribosomal machinery are just a few examples that have tremendously extended our understanding of the structural basis of biomolecular function. At the same time, the quest for new therapeutic drugs has launched massive public and commercial efforts to determine the details of drug–target interactions, and to develop whole structure-guided drug discovery pipelines. The role and use of protein drug target crystallography is steadily increasing, and a growing number of therapeutic drugs, which are either the direct result of structure based discovery or where structure guided lead optimization has played at least a significant role, are reaching the market. Among those drugs are the well-publicized human immunodeficiency virus (HIV) protease inhibitors amprenavir (Agenerase®) and nelfinavir (Viracept®), which were developed with knowledge of the crystal structure of HIV protease; the influenza drug Tamiflu® (Oseltamivir); and the onco-therapeutic Gleevec® (Imatinib). Drug target structures are therefore frequently encountered as illustrative examples in this book (Sidebar 1-1).

**X-ray diffraction is a fundamental technique.** Analysis of the first X-ray diffraction patterns, obtained in 1912 by Walther Friedrich and Paul Knipping in Max von Laue’s laboratory (Sidebar 6-7) from crystals of simple compounds such as diamond, rock salt, or zinc sulfide, confirmed in a fundamental way the atomic constitution of matter and the interactions and bonding of the atoms. Given that the atom theory and particularly quantum mechanics were then in their infancy, these diffraction experiments provided the much needed crucial evidence and support for the development of the atomic theory of matter. Interestingly, shortly after X-rays were discovered by Wilhelm Conrad Röntgen in 1895 (Sidebar 6-1) and before the first diffraction patterns of simple compounds were recorded, it was already known that proteins could also form crystals. It would take more than three decades of technical developments until the first diffraction patterns of protein crystals could be recorded and another two decades until the successful determination of the first macromolecular structure, myoglobin, in 1957 (Sidebar 10-1). Many more relevant historic tidbits and anecdotes are provided in sidebars throughout this book.

**Macromolecular crystallography is no trivial pursuit.** Although we take modern biomolecular crystallography for granted as a mature tool, the method itself is less than 100 years old and only with the advance of computers has it become possible to tackle molecular structures the size of a protein. While today we are spoiled by recombinant DNA techniques, robotic automation, and easy access to powerful tunable synchrotron X-ray radiation sources, the first protein crystallographers were true pioneers and their efforts to obtain the first protein structure were no less than heroic. Not without reason, the field of crystallography and structural biology is rich with Nobel Laureates.
Molecular structure defines function

**Sidebar 1-1 List of selected macromolecular structure presented in illustrations.**

- Bovine pancreatic trypsin inhibitor, Figure 2-1
- Ribosome 50S subunit, Figure 2-1
- *Clostridium botulinum* neurotoxin (holotoxin), Figure 2-1
- PRD1 bacteriophage head, Figure 2-2
- Human apolipoprotein E4, Figure 2-11
- Concanavalin A, a plant lectin, Figure 2-15
- Ribonuclease A, Figure 2-16
- Ornithine decarboxylase complexed with DMFO, Figure 2-30
- Ribonuclease inhibitor, Figure 2-35
- Cyclooxygenase 2 (COX2), Figure 2-38
- Bacterial membrane porin, Figure 2-40, 3-44
- GABA<sub>2</sub>, pentameric neuroreceptor, Figure 2-41
- B-DNA decamer, Figure 2-44
- Zn-finger complexed with B-DNA, Figure 2-46
- p53 tumor suppressor complexed with DNA, Figure 2-47
- *Mycobacterium tuberculosis* InhA complexed with isoniazid, Figure 2-48
- Nucleosome core particle, Figure 2-49
- Ferritin, thumatin, Figure 3-16
- Trypsin complexed with benzamidine, Figure 3-41
- *Clostridium tetani* neurotoxin, ganglioside binding domain, Figure 3-42
- KcsA potassium channel, Figure 3-46, 3-47
- *Corynebacterium diphtheria*, DtxR repressor complexed with duplex DNA, Figure 3-48
- *Bacillus subtilis* organic hydperoxide-resistance protein OhrB, Figure 4-4
- Green fluorescence protein, engineered, Figure 4-14
- Apolipoprotein E3, 22 kDa LDL receptor binding domain, Se-Met mutant, Figure 4-15
- Mistic membrane insertion fusion protein, Figure 4-16
- OprP, a trimeric bacterial outer membrane protein, Figure 4-18
- *Bacillus anthracis* alanine racemase, Figure 4-19
- Quercetin 2,3-dioxygenase, glycosylated, Figure 4-20
- Hepatocytic growth factor complex, Figure 4-25
- Recombinant glucosylceramidase, defective in Gaucher’s disease, Figure 5-4 ff., 6-16
- *Mycobacterium tuberculosis* rhamnose epimerase RmlC, drug target, Figure 5-21, 11-6
- Chicken calmodulin, Figure 5-22, 11-7
- SARS coronavirus ancillary protein, Figure 5-33
- *Staphylococcus aureus* superantigen toxin-like fragment, Figure 5-34
- SH2 domain of the C-terminal human Src kinase (Csk), Figure 5-41
- European mistletoe viscotoxin A1, Figure 10-35
- Antibody F<sub>ab</sub> fragment against human LDL receptor, Figure 11-2
- GTP glycohydrolase I complex with its feedback regulatory protein GFRP, Figure 11-12
- *Rhodobacter spheroides* cytochrome c’ dimer, Figure 11-22
- *E. coli* medium chain length acyl-CoA thioesterase II (TesB), Figure 12-40
- Antibody F<sub>ab</sub> fragment against supersweetenter, Figure 13-13
- Cellular retinoic acid binding protein, CRABP, Figure 13-16
- *Streptomyces* D-alanyl carboxypeptidase-transpeptidase, Figure 13-14
- Human acetylcholinesterase, AChE, Figure 13-15
- *E. coli* small-conductance mechano-sensitve Channel, MscS, Figure 13-23
- Primary literature references are provided at their respective locations. Many more annotated and illustrated examples for important macromolecular molecular structures may be found in the “Molecule of the Month” column edited by David Goodsell on the Rutgers Protein Data Bank (RCSB-PDB) web site.

**Remote facility access.** New crystallographic techniques and methods, as well as easy access to remote synchrotron data collection, have made it possible for practically every structural biology laboratory—given proper training—to reliably pursue crystallographic structure determination projects. The apparent ease with which complex biomolecular structures can be obtained, with the help of remote synchrotron data collection from submitted crystals (referred to as “FedEx crystallography”) and advanced (but not foolproof) software tools, has occasionally given rise to the casual attitude that macromolecular crystallography is a simple biochemical technique on a par with running an SDS gel or recording an absorption spectrum. This notion is gravely mistaken (Sidebar 1-2), as the preview of the practical and fundamental challenges in the next section, as well as selected examples of cautionary tales from the literature dispersed in appropriate places throughout the text, will testify. The great technical advances
resulting from efforts of various Protein Structure Initiatives worldwide have even lead to the opinion that nowadays one may simply submit the amino acid sequence of the protein to one of the Protein Structure Centers, and in due course structure coordinates will be returned. While this approach may succeed for simple bacterial proteins, the complexity of eukaryotic or human proteins almost always requires significantly more effort. Some of the complications of eukaryotic proteins and membrane proteins and how to approach them are discussed in Chapter 4.

A major part of the process of any structure determination, and the key to success, is proper planning: a clear view is needed of the objectives of the structure determination, in particular what information at what level of detail will be necessary to answer your specific question or to test your specific biomedical hypothesis, and how to reach the specified goals efficiently. The following provides a brief introduction to the method of crystallographic structure determination, outlining key stages in a well-planned protein structure determination. The rewards of understanding the technical details of macromolecular crystallography become obvious when the process turns difficult and structure determination is less straightforward. Having command over this most powerful technique of structure determination will enhance your confidence and also allow you to take responsibility for the extraordinarily persuasive power conveyed by a crystallographic structure model.

**Box 1-1 Crystallography provides the foundation of modern structural biology.** The power of macromolecular crystallography results from the fact that highly accurate models of large molecular structures and molecular complexes can be determined, often at a near atomic level of detail. Crystallographic structure models have provided insight into molecular form and function, establishing the basis for structural biology and structure-guided drug discovery. Non-proprietary protein structure models are made available to the public by deposition in the Protein Data Bank, which holds more than 50 000 entries as of 2008.
1.2 Principle and challenges of crystallography: A preview

The basic single-crystal diffraction experiment is deceptively simple: A single crystal of the material of interest is placed into a finely focused X-ray beam, and the diffraction images are recorded. The electron density representing the atomic structure of the molecules in the crystal is reconstructed by Fourier methods from the diffraction data, and an atomic model of the structure is built into the electron density. Figure 1-2 shows the basic principle, while photographs of actual modern X-ray diffractometers can be found in Chapter 8. While this description appears quite straightforward, there are a number of practical and conceptual challenges that need to be addressed. Various nontrivial principles form the basis of the extraordinary power of crystallography, making it not just a most important tool of structural biology but an attractive field of study in its own right.

**Practical challenges.** The price to be paid for obtaining informative high resolution X-ray structures is that a well-diffracting protein crystal needs to be produced; the obvious practical issue is that without crystals there is no crystallography. Proteins, nucleic acids, or molecular complexes thereof—by their nature as large and flexible macromolecules—seldom self-assemble readily into the regular, periodically repeating arrangements typical for crystals. Indeed, growing well-diffracting protein crystals can prove quite challenging, we dedicate the whole of Chapter 3 to the principle and practice of growing protein crystals. Moreover, even obtaining the material to be crystallized in the first place is nontrivial; more often than not the protein needs to be modified or engineered so that it in fact can crystallize, as elaborated in Chapter 4. An additional complication during the data collection process (Chapter 8) originates from the fact that organic material, such as protein crystals, is highly sensitive to radiation damage by the intense ionizing radiation it is exposed to during the X-ray diffraction experiment. To prevent decay of the crystals during X-ray exposure, cooling of crystals to cryogenic temperatures (somewhat above the boiling point of liquid nitrogen of 77 K or \(-196^\circ C\)) is near universal practice. However, many crystals are difficult to flash-cool and it is hard to predict which cryoprotection conditions will work.

**Fundamental challenges.** The diffraction process is fundamentally different from microscopic imaging; crystallography is not an imaging technique. The crucial difference is that visible light scattered from objects can be focused through refractive lenses to create a magnified image of the object. This not the

---

**Figure 1-2 The principle of X-ray structure determination.** A crystal mounted on a goniostat with at least one rotatable axis is exposed to a finely collimated, intense X-ray beam in the 5–20 keV energy range (~2.3 to 0.6 Å wavelength). Individual diffraction images are recorded on an area detector during small rotation increments of the crystal and combined into a diffraction data set. However, the diffraction images are not direct images of the molecule. Diffraction images are transforms of the molecular shape into reciprocal space—equivalent to being transformed into “secret code.” This secret reciprocal space code must be deciphered for each structure by the crystallographer. The basic mathematical tool of back-transformation from reciprocal diffraction space into direct space is the Fourier transform (FT), which together with separately acquired phases for each diffraction spot allows synthesizing or reconstructing the electron density (blue grid) of the molecules self-assembled into the diffracting crystal. An atomic model of the structure, represented in the figure by a ribbon model, is then built into the 3-dimensional electron density. The absence of phases in the diffraction data is the origin of the phase problem in crystallography and a suitable phasing strategy needs to be developed for each structure determination.
Sidebar 1-2 So easy a monkey can do it? “A certain portion of the research community has tended to regard crystal structure analysis as entirely too easy and the value of crystallographic results has been challenged as being irrelevant beyond the solid state, despite countless examples of correlations between structural features observed in the solid state and the chemical, physical and biological properties of the same molecules in vitro and in vivo, as measured by a wide range of techniques. The rapidity with which new structures are determined and the rate at which crystallographic databases are expanding fuels the argument that crystallography is easy and crystallographic instrument manufacturers reinforce this opinion by claiming that with today’s instrument and software anyone can determine a crystal structure. Even Judith Howard, past president of the British Crystallographic Association, was quoted out of context as saying ‘instruments have been so advanced that you sometimes feel you can train a monkey to use it’. Unfortunately, a monkey so trained would know about as much about the underlying phenomena of diffraction, the proper use and analysis of diffraction data and the structural information it provides as many other current users who lack formal crystallographic training.”


1.3 Protein molecules and the crystalline state

The native location of a protein molecule is not the ordered solid state but generally some fluid environment, often an aqueous solution. The intracellular cytosol and the extracellular environment are generally quite crowded. Cell membranes harboring receptors and channels are also fluid assemblies, with parts of the protein molecules extending into the cytosol and/or extracellular fluid. The question arises of how the fact that proteins have self-assembled into solid crystals affects the interpretation of the resulting crystal structures and their biological relevance with respect to their native solution state. As explained...
in great detail in Chapter 3, protein crystals are not rigid bricks. They actually are formed by a loose periodic network of weak, non-covalent interactions (Figure 3-5) and contain large solvent channels which allow relatively free diffusion of small molecules through the crystal. This solvent access maintains a solution environment (which is actually a requirement for protein crystals, once they dry out they almost invariably stop diffracting) that accommodates quite some conformational freedom for surface-exposed side chains or loops as well as some “breathing” motions of the structure core.

**Enzyme activity and the solid state.** Comparison of many nuclear magnetic resonance (NMR) solution structure ensembles with crystallographic structure models have shown that the core structure of protein molecules remains unchanged compared with the solution state during crystallization. In addition, enzymes packed in crystals even maintain biological activity. The maintained activity in crystals actually creates a challenge for the crystallographer, often necessitating the design of inactive enzyme substrate analogs or substitutes in order to dissect the molecular reaction mechanisms.

**Molecular flexibility.** The maintenance of the core structure and of enzymatic function shows that crystal structures are a very good approximation of the native protein solution structure. Nonetheless, highly flexible or mobile regions, frequently the amino- or carboxyl-termini of the protein chain or flexible loops connecting secondary structure elements, can be poorly defined or even absent in the electron density and thus can be modeled only with limited confidence. This situation is also reflected in the fact that in NMR structures, highly dynamic and flexible regions exhibit large deviations between the multiple models of an NMR ensemble. In either case, the absence of a well-ordered structure is a genuine reflection of the dynamic behavior of the protein molecules and not a weakness of either technique. NMR and X-ray crystallography in fact complement each other, and refinement of molecular models against data from both of the methods combined can ultimately give a more accurate structure than either method alone. The advantage of crystallography, however, is the more detailed information in well-defined parts of the molecule and in principle no size limitation, which restricts solution NMR to molecules below ~35 kDa.

**Crystal packing affects local regions.** In certain situations flexible and dynamic regions of a protein molecule can be rigidly fixed in a specific conformation as a result of crystal packing interactions. In most cases this represents just a snapshot of one possible conformation out of many and it must be understood that such a specific conformation may not locally represent the protein structure in solution. A simple safeguard against misinterpretation—which is usually assignment of certain biological relevance to regions where that is de facto is not warranted—is to display all neighboring, symmetry-related molecules in the crystal structure and examine if any intermolecular interactions are present that are a result of crystal packing. Such packing induced artifacts can also hamper for example drug discovery by altering or blocking binding sites and thus preventing an otherwise active substance from binding. Binding properties can be altered either directly through intermolecular contacts of key residues with neighboring protein molecules, or through normally flexible loops covering the binding site and thus preventing access; or indirectly by affecting allosteric binding sites of the protein. It is good practice to examine the possible presence of packing effects in a crystal structure before attempting any ligand docking or drug lead optimization.

**Large conformational changes destroy crystals.** The fact that protein molecules are periodically packed in a crystal lattice clearly places limitations on the direct observation of processes involving large conformational changes, which invariably destroy the delicate molecular packing arrangement of a protein crystal. Molecular transport processes or interactions involving extended conformational rearrangements therefore require multiple, stepwise “snapshot” structure determinations in order to dissect the details of such inherently dynamic processes. Enzymatic reactions involving limited, local structural changes in
Introduction: Preparing for your study

Dynamic behavior of molecules. Although no large-scale dynamic movements can take place in a crystal, there is in fact some information about the dynamic behavior of the molecules present in the crystal structure. High $B$-factors for an atom, a generic displacement measure indicating how well the model is defined in that region, are indicative of flexibility or disorder in that local region. Analysis of concerted movements of entire regions of the molecule by TLS parameterization during refinement can give indications for the propensity toward even substantial domain movements. The unqualified critique that protein structures are just a static snapshot is thus oversimplifying and often a manifestation of ignorance on the part of the model user. We will discuss the finer details of the implications and analysis of the dynamic behavior of proteins in Chapter 12.

Protein structure quality is locally defined. The most important fact to understand is that crystallographic protein structure models naturally have very clear and well-defined regions, while other parts reflect dynamic behavior, ranging from increased flexibility and displacement to complete disorder. Certain structures types, for example helix bundles or globular proteins, tend to form very stable and well-defined structures, while others, no matter how proficient the crystallographer, will always remain less well defined because of their inherent dynamic flexibility. Local analysis of protein crystal structures will be treated throughout the text and more deeply in Chapter 13.

1.4 Interlude: Molecular cryptanalysis

The computer-savvy student of today may find it entertaining to view crystallography as the cryptanalysis of reciprocal space—translating the secret molecular message carried by the diffraction data back into direct molecular space. A number of quite interesting parallels between crystallography and cryptanalysis exist, and in periods of despair it often helps to entertain the perspective of decoding a riddle or secret message. Technically speaking, each molecule is encoded by a single-use encryption key and each structure determination requires one to find its particular key. Sometimes finding the right key or methods may be straightforward, but quite often it will not be trivial at all.

The first step in code breaking just as in crystallography is to intercept the message. For us this means data collection—without diffraction data, there can be
Interlude: Molecular cryptanalysis

no structure determination. Let us assume we have intercepted the following cipher conveying a text message relevant to macromolecular crystallography:

19 17 17 19 14 21 17 16 18 24 16 19 15 18 24 08 22 03 12 18 06 03 04 06 22 12
18 14 07 12 19 08 18 24 12 19 14 18 06 18 26 17

The first requirement for cracking the code is that we understand its nature and its language. We readily recognize that the code elements are limited to numbers not exceeding 26, so it is reasonable to assume that we are dealing with some simple one-to-one encryption of an alphabetic message. Given that this text is in English, we furthermore assume that the message is also in English. The same necessity to understand the nature and language of the code exists in crystallography and Part II, *Fundamentals of Crystallography*, will establish the basics necessary for the understanding of crystal geometry, reciprocal space, scattering of X-rays by molecular crystals, and the analysis of the resulting X-ray diffraction data.

We can take our first shot at deciphering the code by conducting a simple statistical analysis of the probability distribution of the code elements (Figure 1-3, A). We know that the most frequent letter in English text is the letter *e*, and we can assign in a first trial the letter *e* to the corresponding number 18, similar to a trial seeding position in crystallographic substructure solution methods. Notice how important it is to intercept the code correctly (i.e., to collect your data accurately); one simple transcription error swapping one 18 with 19 completely changes the trial assignment (Figure 1-3, B). Statistical methods play an important role in crystallographic analysis (Chapter 7) and phasing methods are particularly sensitive to poorly measured data—accurate data collection is important. Note, however, that a code number of say 58 instead of 18 would be easily recognizable as highly improbable value or error, equivalent to the concept of outlier detection commonly applied in crystallographic data processing.

A statistical technique we can apply now to improve our first trial “structure” is based on pattern recognition. Take for example the first four code numbers; they show a very peculiar pattern:

19 17 17 19 14 21 17 16 18 24 16 19 15 18 24 08 22 03 12 18 06 03 04 06 22 12
E 14 07 12 19 08 18 24 12 19 14 E 06 E 26 17

These symmetric code numbers very likely do not stand for all consonants (C) or all vowels (V) (you may struggle to construct a sensible combination) but rather a combination starting with CVVC or VCCV. We can search now through a dictionary and try all possible solutions—a multi-solution approach—and pick the one that makes the most sense (i.e., scores the highest). One of the first hits is ATTACK, which expands our code as follows:

ATTACK E 16 A 15 E 24 08 22 03 12 E 06 03 04 06 22 12
E 07 12 A 08 08 E 24 12 A E 06 E 26 T

A statistical technique we can apply now to improve our first trial “structure” is based on pattern recognition. Take for example the first four code numbers; they show a very peculiar pattern:

ATTACK E 16 A 15 E 24 08 22 03 12 E 06 03 04 06 22 12
E 07 12 A 08 08 E 24 12 A E 06 E 26 T

Figure 1-3 Statistical analysis of encoded data. The frequency of the code symbols is compared with the expectation values for letters. In panel (A), the most frequent code symbol 18 is correctly identified with the letter *e*. A simple transcription or measurement error (B) changing 18 to 19 completely changes the relative frequencies and serves as a reminder that diffraction data need to be accurately collected, which is particularly important for experimental phasing methods to succeed.
Pattern recognition and image seeking methods are found in various flavors in crystallography. The most common application is to **model building**, where, based on shape compatibility, the electron density is recognized as belonging to a certain secondary structure element such as an α-helix or β-sheet, or representing a **side chain** corresponding to a certain kind of **amino acid residue**. Automated methods, as well as the crystallographer during manual model building, rely on pattern recognition. While the individual often bests the computer in recognizing matching elements, the computer succeeds by brute force through trying many more possible solutions.

Applying pattern recognition techniques was a great step forward for our code cracking and now we apply prior knowledge about the context of the message to further analysis. Prior knowledge or the **prior probability** of a structure model plays an important role in crystallography because it helps us to judge how well our crystallographic model complies with the body of independently established scientific knowledge. This helps to make sure that despite all its new and exciting features, our structure model remains within physical reasonable bounds, complying with **stereochemical restraints** and other fundamental laws of nature.

Given that our encrypted message is of some importance for biomolecular crystallography, the sequence

\[
06\ 22\ 12\ \text{E}\ 07\ 12\ \text{A}\ 08
\]

attracts our attention—it probably means MOLECULAR. We further complete our clear text (the initial structure model) now and obtain the following message:

\[
\text{ATTACK}\ 16\ \text{E}\ 24\ 16\ \text{A}\ 15\ \text{E}\ 24\ \text{RO}\ 03\ \text{LEM}\ 03\ 04\ \text{MOLECULAR}\ 24\ \text{LACE}\ 26\ \text{M}
\]

What remains now is to complete, refine, and polish our model by filling the remaining gaps. It is left up to the reader to finish the cryptanalysis of this message referring to a common method of solving the crystallographic phase problem.

### 1.5 Planning and executing a protein structure determination

No more than two decades ago, the pursuit and successful determination of a single protein structure by **multiple isomorphous replacement** phasing could earn a Ph.D. and warrant a high impact publication. Today, the technique has matured to a point that in many cases structural biologists with little specific training in crystallography can quite successfully determine a protein structure. On the one hand, this has tremendously enriched our knowledge about fundamental biochemistry, reaction mechanisms, or the molecular basis for disease and therapeutic drug development. On the other hand, there have been cases where the structures have been rather poorly determined and incorrectly interpreted, albeit with great ambition. This is partly the result of poor understanding of the fundamentals and the capabilities of the crystallographic technique itself, and partly because of the resulting inability to analyze the structures properly in the context of the biological question that was to be addressed. As a result of the not inconsiderable effort involved in conducting a protein structure determination, the structure models are sometimes (ab)used to prove a hypothesis, instead of testing it. **Model bias**—the fact that with poor data and poor phases, the electron density tends to reflect the preconceptions of the model builder—combined with a desire to see one’s proposal established, can be a powerful and devastating combination, as discussed in Sidebar 13-1. Adequate planning of a structure determination in view of the ultimate objective of the study and awareness of potential pitfalls greatly increases the chance of success while minimizing the risk of subjectively over-interpreting weak evidence in the form of a biased or flawed structure model.
Project planning and alternatives

Any actual protein crystallography study starts with a considerable amount of information gathering and planning. Just as anywhere in research, “a year in the laboratory can save a day in the library” information gathering stage includes, besides technical searches and bioinformatics, the response to questions that need to be asked in the broader context of maximizing the return on the considerable investment of time and effort your structure determination project presents. Figure 1-4 illustrates the major stages of a protein structure determination project.

There is a priori no guarantee that a crystal structure study will succeed. It can fail at many stages and in addition to a sound scientific hypothesis to be proposed or tested by the structure model, any research plan should include alternate options (or projects) if the experiment cannot be completed as planned. While the up-front molecular biology and cloning have quite reasonable success rates, protein production and protein crystallization remain major challenges. In both cases, initial average success rates are around 30% and data from the NIH Protein Structure Initiative have shown that the chances of obtaining a crystal structure of an average globular bacterial protein in one straight path without modification to the initial strategy are only 10–20%. On the other hand, once a diffracting crystal is obtained, the chances are very good that structure determination will succeed given a proper phasing strategy. Ultimately, the analysis of the model will show whether the structure actually answers the initial question. Quite often, however, serendipity plays a significant role and many exciting breakthroughs (and Nature papers) in structural biology are the result of unexpected features discovered and supported by the analysis of a well-determined crystal structure model.

Certain categories of structure determination projects are inherently risky. The high impact of an integral membrane protein structure comes at a price and one must be prepared to spend perhaps years to determine just one single structure—if at all successful. It is advisable in such high risk/high impact situations to negotiate with the supervisor or laboratory head an alternate completion target for a thesis or assignment. On the other hand, it cannot be expected that a basic molecular replacement structure of a mutant of a well-known household

Figure 1-4 Overview of protein structure determination. The bar on the left side indicates which stages are performed using computers only (in silico, IS) and which depend largely on experimental work. The bar next to it indicates with shading the core crystallographic techniques covered in depth in this book. The results of the structure analysis frequently feed back into the design of a refined study, particularly in structure guided drug discovery. VLS: virtual ligand screening; SGD: structure guided drug discovery. Also consult Figure 1-8 for a more detailed diagram of key steps in structure determination and the corresponding Chapters in this book.
enzyme, of which there are already many structures deposited in the PDB, will yield a high impact publication or a Ph.D. thesis.

**Keeping track of the goal**

Even before technicalities of cloning, protein production, crystallization, and actual structure determination are detailed, the first step in planning is to assert what qualities the final structure model must have in order to address your underlying question. Not all crystallographic structure models are created equal: A drug discovery study will necessitate a much higher level of detail (see Figure 1-6) than the determination of the novel fold of a newly discovered protein or speculation about conformational rearrangements in a huge complex of a cellular secretion machinery. When the minutiae of a binding mechanism are to be determined, cofactors or substrate analogs generally must be present in a high quality structure. Comparison between the apo-structure and the complex with bound ligands may also be informative. Dynamic or flexible structures may need multiple structure determinations in different crystal forms to inform about the range of conformations that can be assumed, and large multi-domain structures may have to be expressed and crystallized as separate domains, while a hypothetical model of the holo-complex can be assembled in silico. Protein engineering (Chapter 4) is quite often a necessary prerequisite to facilitate crystallization. Some crystal forms may also exhibit packing contacts that can hinder the binding of ligands or make biological interpretation otherwise difficult.

**Automation and parallel approaches**

The major preparatory steps of a structural study, namely protein expression and protein crystallization, are experimentally driven and non-deterministic, in the sense that the outcome cannot be predicted from first principles. Multiple expression and crystallization trials under different conditions are inherently suitable for parallelization and automation. Robotic crystallization screening is already an established procedure in industry as well as in academic laboratories and parallel expression and purification strategies are being adopted in structural biology laboratories. The primary benefit of automated parallel approaches is the broad basis from which experiments are started, and a failure of an individual protein construct to express or crystallize is not fatal to the entire project. A multi-pronged, parallel approach also allows a quick selection of several promising leads, with higher probability that one of them may finally yield a structure. Early failures in parallel trials are comparatively cheap, as not many resources have yet been expended (Figure 1-5). In contrast, in serial approaches, a marginal lead such as a poorly expressing protein is sometimes pursued with great effort to the bitter end only to yield a weakly diffracting crystal that does not yield usable data. Broad screening and pursuit of multiple parallel constructs greatly increases the odds of success.
**Protein engineering**

Parallel approaches beginning with cloning and preparation of different protein constructs have additional advantages. Despite common perception, protein crystallization is not the biggest challenge in crystallography. The challenge begins much earlier, namely to produce a protein that actually can be crystallized. The crystallizability of a protein is to an overwhelming degree determined by the properties of the molecule itself. Intermolecular interactions between the molecules must be favorable toward self-assembly into a well-formed, periodic crystal lattice. If the protein either lacks suitably located surface residues, or has structural features such as large termini or flexibly tethered domains that prevent it inherently from crystallizing, then no amount of crystallization screening will ever yield usable crystals. Clearly, starting from a variety of protein constructs, usually with different affinity tags, possibly limited truncation, or stable sub-domains increases the chances of success. In a similar horizontal approach, orthologs from different species provide variants already engineered by nature as an alternate pathway. There is good reason to discuss protein production (Chapter 4) for the purpose of a crystallographic study after we have examined the principles of crystallization in Chapter 3. The requirements a protein must meet to be suitable for crystallization—particularly as regards conformational homogeneity—are generally much more stringent than they are for a functional or biochemical assay.

**Crystal growth**

Once a reasonably pure, soluble, and stable protein is obtained, crystallization screening begins. Despite the fundamental molecular complexity of the crystallization process outlined in Chapter 3, the actual setup of the common crystallization experiment is simple and easy to automate. Numerous crystallization techniques have been developed, each with different merits depending on the specific purpose (initial screening, optimization of growth, ease of harvesting, ease of automation) and type of research environment. Maximizing comprehensive sampling of crystallization space with the least amount of material plus ease of miniaturization favors a robotic setup of nano-drops and sampling statistics play a significant role in the design of efficient crystallization screening strategies.

Crystallization trials are observed under a microscope at regular time intervals and the experimenter, sometimes assisted by image recognition software, decides whether a potentially suitable crystal has been obtained. In many cases, the initial crystallization conditions need to be further refined or optimized before a well-diffracting crystal is obtained.

**Harvesting, cryocooling and mounting of crystals**

Another critical experimental step, and probably the least systematically explored, is harvesting and cryocooling of the crystals. Protein crystals that have grown in the crystallization trials must be harvested from their mother liquor and mounted on a diffractometer for data collection. There are numerous hazards associated with this procedure. Protein crystals are very small, ranging in size between few 100 μm to about 10 μm in size, and because of the high solvent content and weak, non-covalent intermolecular interactions they are very fragile. Like every biological material, they are susceptible to severe radiation damage in the intense X-ray beams needed for data collection. Preventing damage from exposure to the ionizing X-ray radiation is the foremost reason to cryocool the crystals to cryogenic temperatures, where damage from the ionizing radiation is at least greatly reduced. During the rapid flash-cooling (or quenching) to liquid nitrogen temperature, the formation of crystalline ice in the mother liquor surrounding the crystal must be avoided. Crystalline ice formation invariably destroys the delicate protein crystals.

The reason why cryocooling procedures are rarely systematically investigated, is easy to understand: Once a cryocondition is found that yields a diffracting crystal and good data are collected, there is little incentive to mount the remaining crystals or to grow or flash-cool more of them for systematic studies. In
addition, the delicate micro-manipulations vital to crystal harvesting are hard to automate. Remember that crystals are expensive; losing them amounts to a correspondingly costly late-stage failure.

**Diffraction data collection**

Once a diffracting crystal has been harvested and mounted in a cryo-loop, the actual process of the structure determination begins. In contrast to expression and crystallization screening of biological material, which more or less depend on trial and error experimentation (regardless of the clever screening strategies deployed), the path from data to structure can be understood from physical first principles. Once the data are collected, all subsequent steps—data processing, phasing calculations, electron density reconstruction, model building, structure refinement, validation and analysis—are conducted *in silico* with computer programs.

**Information content of data.** The diffraction pattern images recorded on area detectors are indexed, integrated and scaled, and unit cell and space group are determined as discussed in Chapter 8. A reduced and hopefully complete data set, essentially representing a periodically sampled reciprocal space transform of the molecules in the crystal, is obtained. The extent to which the crystal diffracts (well-ordered crystals diffract to a higher resolution than less poorly ordered ones) directly determines how detailed the final reconstruction of the electron density will be, and hence, ultimately and without mercy, how detailed the resulting model can be. While the exact correlation between diffraction limit and sampling density will be derived in Chapter 9, it is easy to understand the obvious benefits of high resolution. Figure 1-6 emphasizes the qualitative connection between the extent of diffraction and amount of detail discernible in the electron density reconstruction (discussed in detail in Chapter 9). Note again that crystallography does not image atoms or molecules, it reveals the...
Planning and executing a protein structure determination

mathematically reconstructed electron density of the molecules that form the protein crystal.

**Phase determination**

The measured intensities of the diffraction spots (or reflections) that form the data set provide only the magnitude or **amplitude** of the diffracted X-rays, but the phase relations between the reflections are absent (Figure 1-7). The phases for each reflection, however, are necessary to reconstruct the electron density. Hence, because of the lack of phase information in the diffraction patterns, direct reconstruction of the electron density of the molecules via Fourier transforms from the intensity or amplitude data alone is generally not possible (known as the **phase problem** in crystallography). Obtaining the missing phases is the conceptually most challenging part in the *in silico* segment of a structure determination project. In protein crystallography, there are two major avenues to obtain the missing phases:

- **Molecular replacement.** If a previously determined, structurally similar model is available, it can be used to calculate initial phases, which are then used in the initial construction of the electron density. The task involves finding the correct position of the search molecule in the crystal and thus the method is named molecular replacement, in the sense of repositioning (not substituting) the molecule. Although the method allows a quick determination of initial phases, it can introduce severe model **phase bias:** Because phases dominate the reconstruction of the electron density, the initial structure will largely reflect the features of the search model and not the true structure, and phase bias removal methods are extensively used to build the correct model structure. Overall, about three-quarters of all structures deposited in the protein data bank are solved by molecular replacement techniques exploiting a structurally similar structure model as a source of initial phases. Chapter 11 discusses molecular replacement in depth.
Experimental phasing. In the absence of a suitable known structure model, phases must be determined de novo by a separate phasing experiment, hence the name experimental phasing. Experimental phasing methods are generally applicable and depend on the determination of a marker atom substructure by exploiting—often very weak—intensity differences between isomorphous data sets (maintaining the same crystal structure). Isomorphous differences are found between native and derivative crystals that contain a heavy atom soaked into the crystal, forming the basis for traditional isomorphous replacement phasing methods. Anomalous differences in data sets from crystals containing anomalous scatterers—generally atoms heavier than H, C, N, O—are most frequently used for experimental phase determination. Even the minuscule anomalous signal from natively present sulfur can be used in certain cases and, when measured carefully, to determine the sulfur atom substructure. The heavy atom substructure then provides the initial phases required to reconstruct the electron density. One popular phasing technique exploits the fact that the amino acid methionine can be replaced with seleno-methionine by overexpressing the protein in a suitable expression system (Chapter 4). The Se atoms then provide a site-specific source of anomalous phasing signal.

Experimental phasing accounts for about a quarter of all structures deposited in the PDB. A variety of experimental phasing methods are available (Chapter 10), and together with density modification techniques even a single anomalous data set may suffice to solve a protein structure. Anomalous phasing methods dominate, and even traditional isomorphous replacement methods are routinely supplemented by orthogonal phase information from anomalous diffraction data. Table 10-1 contains an overview of the most common phasing techniques (including their bewildering and sometimes MAD abbreviations).

In almost all cases, the initial protein structure phases obtained during the substructure phasing stage are further improved by various density modification techniques providing substantially improved electron density maps into which the initial protein structure model is built.

Electron density interpretation and model building

Once an interpretable electron density map (a 3-dimensional contour grid of the electron density) is obtained from improved experimental phases, a model of the protein structure must be built into the electron density. The model building is carried out using programs that graphically display the electron density and allow placement and manipulation of protein backbone markers and residues. Various electron density fitting and geometry refinement tools as well as automated model building programs greatly accelerate the process. Automated model building programs often provide a quite reasonable starting model, which can then be completed and polished by hand as discussed in Chapter 12.

**Figure 1-7 The crystallographic phase problem.** In order to reconstruct the electron density of the molecule, two quantities need to be provided for each reflection (data point): the structure factor amplitude, $F_{hk}$, which is directly obtained through the experiment and is proportional to the square root of the measured intensity of the diffraction spot or reflection; and the phase angle of each reflection, $\phi_{hk}$, which is not directly observable and must be supplied by additional phasing experiments. The methods and mathematics of electron density reconstruction by Fourier methods are extensively treated in Chapter 9.
Model building is often perceived as the most dreaded part of the structure determination. Model building certainly constitutes the most intense involvement in the protein structure and with poor electron density a large degree of chemical intuition and experience are required. It is here where excellent quality data, good phases, and high resolution deliver substantial payoff: In addition to revealing much higher detail and allowing greater confidence when building the model, the tedium of model building is also greatly reduced.

**Restrained reciprocal space refinement**

Despite most careful model building in real space, the initial model will, in addition to missing certain parts, have many small errors such as incorrect bond lengths and angles, poor backbone geometry, or improbable torsion angles. These small errors are corrected during the course of reciprocal space refinement. The refinement program adjusts the atomic position coordinates and B-factors of the model so that the differences between observed and calculated diffraction data are minimized. The diffraction data are a reciprocal space representation of the molecule, hence the name reciprocal space refinement. Various other corrections such as anisotropic scaling and bulk solvent corrections are also applied during the refinement. The global measure of the agreement between calculated and observed structure factor amplitudes is a linear residual, the crystallographic $R$-value and its cross-validation equivalent, $R$(free). (discussed in detail in Chapters 7 and 12).

**Restraints and cross-validation.** While matching observed and calculated intensities by shifting atomic parameters, the refinement program also restrains the model to conform to certain stereochemical expectation values. Some geometry values, particularly covalent bond lengths and bond angles, are well known a priori, and there is no reason to assume that they would be any different in each protein structure. These stereochemical restraints also address a general problem in protein structure refinement, which is the low data to parameter ratio. In this situation, refinement is generally not stable against experimental data alone. Restraint terms creating penalties for deviations from known geometry target values serve as additional observations, and safeguards against overfitting are necessary. Overfitting is the introduction and variation of additional model parameters, which artificially improves the fit between observed and calculated data, but actually does not improve the structure model. Common sources of overfitting are the introduction of too many water molecules or placing various kinds of solvent molecules (or ligands) into spurious density. Cross-validation against a small subset of unused (free) data, monitored by the $R$(free)-value for the free data set is standard practice.

**Structure validation**

Even after careful model correction and refinement, certain errors persist or are not reported by the refinement program’s diagnostics, or they are just overlooked. A protein structure can be huge and several thousand residues must be built and refined to obtain the final structure model. It is quite possible (but not good practice) that some less interesting parts of the structure have been neglected. During refinement and rebuilding the model is thus constantly subjected to an array of validation techniques, ranging from basic geometry checks to detailed chemical and folding plausibility checks based on prior knowledge (Chapter 13).

We have already mentioned that the quality of a protein structure is a local property and global agreement indicators such as the $R$-values cannot be specific as far as the local quality of a structure model is concerned. Local geometry validation programs evaluate the local geometry on a per-residue basis and flag outliers. Extended stretches of consistently high deviations in either case are indicative of serious “problem zones” within the model. A very powerful method of assessing the local quality of a protein structure is the real space correlation coefficient of the model against a bias minimized electron density map, which will be discussed in detail together with additional validation tools Chapter 13.
To fix model errors, the electron density around the questionable residues is inspected, necessary corrections are made, and the structure is again refined. This process is repeated until all significant errors are corrected and the difference electron density shows no more offending high-level features. It is important that errors are corrected as much as possible (time and commitment permitting) because fixing many small errors can result in a significant overall improvement in structure quality. Once reasonably polished the structure is ready for detailed analysis and the structure factor amplitudes and the model coordinates are deposited with the Protein Data Bank and released to the public.

Analysis and description of the structure
One of the most exciting parts of any crystallographic study is the first analysis of the new structure. One is already quite familiar with the molecule from the preceding model building and refinement steps, and often the story unfolds during model building. The detailed analysis then depends to a large degree on the purpose of the structure study and on which hypothesis the structure model was intended to test. Always remember that the structure model itself is a hypothesis; it must withstand scrutiny against both the experimental evidence in the form of electron density and the entire pool of established prior knowledge and all laws of nature. Unconventional and unusual features always require strong and convincing support through experimental evidence, including all the additional supporting biochemical evidence that may exist in addition to the crystallographic structure model.

Figure 1-8 provides a review of the key steps in a protein structure determination, including references to the corresponding book chapters.
1.6 Crystallographic models and coordinate files

We have thus far discussed the basic steps to obtain a crystallographic protein structure model, but not how to represent the model. Protein structure models at minimum must contain the atomic coordinates of each atom contained in the asymmetric unit cell, generally provided in Cartesian world coordinates in dimensions of Å. In addition, the basic crystallographic information in the form of cell parameters (unit cell dimensions $a$, $b$, $c$, and angles $\alpha$, $\beta$, $\gamma$ and space group (Chapter 5) allows a graphic display program to generate all symmetry related molecules to assemble the packing environment of the molecules in the crystal structure. The coordinates and additional information are deposited in the Protein Data Bank (PDB).

The Protein Data Bank

Models of protein structures created by experimental X-ray crystallography or NMR are saved in the form of coordinate data files, headed by additional method-specific information. The authoritative public repository and archive for these experimentally derived data files is the Protein Data Bank, PDB.\(^{32}\) The PDB archives contain atomic coordinates, bibliographic citations, primary and secondary structure information, as well as crystallographic structure factors and NMR experimental data. Purely computational models are no longer accepted by the PDB. In addition to collecting, annotating,\(^{33}\) and curating data files, the PDB provides deposition and validation services,\(^{34}\) cross-links to other databases, and also provides a variety of analysis tools and structure viewers. The PDB, however, is not responsible for the correctness of deposited structures. It is the ultimate responsibility of a proficient crystallographer to deposit a model as free of errors as reasonably achievable (Chapters 12 and 13 will expand in great detail on the practical aspects of what “reasonably achievable” means in the nontrivial context of macromolecular refinement).

Format of protein structure coordinate files

A PDB file is identified by a four character PDB identification code. The first letter is a number, which was originally intended to indicate revisions to the file by incrementing the numbers, followed by three alphanumeric characters. A PDB file has two main sections: a header and the actual atom coordinate records. The two most common file formats are: a fixed-length, 80-character per line, key-worded record format descending from the FORTRAN computer programming language standards; and a more modern, variable length record format called the macromolecular Crystallographic Information Format, or mmCIF.

Sidebar 1-4 A brief history of the Protein Data Bank and the Molecular Structure Database. The Protein Data Bank (PDB) was formally created 1971 and initially maintained at the Brookhaven National Laboratory. In 1999 it moved to the Research Collaboratory for Structural Biology (RCSB) at Rutgers, NY, under direction of Helen Berman, John Westbrook, Phil Bourne at UC San Diego, and others, where it has grown into a formidable collection of data and tools.\(^{17}\) Since 2003 three sites harbor the world wide PDB (wwPDB). In addition to the RCSB, the wwPDB is also mirrored in Japan at the Institute for Protein Research in Osaka and at the European Bioinformatics Institute, EBI, in Hinxton (close to Cambridge), UK. The EBI also hosts the Macromolecular Structure Database (MSD) site, which includes an additional array of structural bioinformatics analysis tools and provides deeper annotation and cross-referencing of the PDB files which goes beyond the standard PDB services. Particularly useful in the EBI/MSD site is the PDBsum database\(^{35}\) providing an at-a-glance overview of deposited PDB structures including topology drawings, motif analysis, sequence alignments, binding pocket analysis, interface analysis, and more. The MSD also harbors a mirror of the Electron Density Server (EDS), a powerful validation tool that provides real space correlation plots of structures on a per-residue basis. Such plots, which we will discuss in Chapter 13 in great detail, compare the model with the actual electron density and provide quite accurate information about the local quality of the structure. However, the EDS plots can only be created if the crystallographer has also deposited the corresponding data file in the form of the structure factor amplitudes against which the model was refined. It may come as a surprise that crystallographers were not required until early 2008 to deposit their primary data in the form of measured structure factor amplitudes together with the model coordinates.
Translators between the formats and to extendable mark-up language (XML) format are available from the PDB sites. Despite its shortcomings, the fixed record PDB file format is persistent, largely because it is quite easy to read. The file format and the dictionary items are published on the PDB web site, and a brief introduction, with particular emphasis on the fixed-format CRYS'T and ATOM records and their peculiarities, is provided in Appendix A.1.

Graphical display of structure models

The first low-resolution protein structure models were made from clay displaying only secondary structure elements, later followed by complicated atomic brass ball models on steel rod assemblies (Sidebar 12-1). Given the size and complexity of protein molecules, all serious analysis is now carried out with the assistance of graphical display programs.

All display programs have in common that they can read at least the atom coordinate records and useful ones can also interpret the crystallographic information and generate symmetry related molecules and display crystal packing. Note that displaying only the asymmetric unit as contained in the coordinate file does not suffice for full analysis of crystal contacts; you need to display all neighboring molecules including translationally related symmetry copies (detailed in Chapter 5). As there is no specific connectivity record in the PDB files for standard residues, the display programs generally generate the bond information based on atom type and distance to neighboring atoms. The secondary structure is in most cases computationally assigned by DSSP or a similar program. This means that occasionally a display program cannot interpret and properly draw ligands or non-standard residues, and poor or unrefined models tend to have missing bonds and only partial secondary structure assignment. Note that the deposited model does not always reveal the biologically relevant molecular assembly, as explained in Section 2.9.

Ball and stick-models, space filling models, and secondary structure cartoons or ribbons are practically always available in molecular display programs for informative rendering, and depending on the purpose, various other properties can be mapped onto certain mesh objects or onto surface representations. Charge distribution, hydrophobicity, and binding pocket surfaces are common properties to display. Most of the crystallographic web browser plug-ins or Java-based display applets are quite useful for a first overview, but despite some of the applets allowing quite reasonable display options, fully detailed analysis usually requires specific software. We will encounter figures prepared with several programs throughout the text (Sidebar 13-7 describes how the figures were generated) and a summary of popular molecular graphic programs and their availability is given in the online supplement.

Sidebar 1-5 Not all models are created equal. While graphic display programs can generate beautiful pictures of protein structures, these images of great persuasive power lack crucial quality information. The structure representations—commonly ribbon diagrams or ball and stick models—look equally convincing and convey a deceptive sense of precision regardless of whether the model coordinates originate from a poor homology model, a low resolution 3.5 Å X-ray structure, or from a high quality 1.2 Å atomic resolution structure. The true proof is in the image of the electron density and how well it is matched by the model. A real space correlation plot provides a rapid overview of the fit between model and electron density for the entire structure, and it is good practice to show the model together with clear and properly contoured electron density whenever discussing an important feature of the structure. The majority of figures of structure details in this book will thus show the electron density together with parts of the structure model.


1.7 Crystallographic computer programs

Protein crystallography depends heavily on computational methods. Crystallographic computing has made substantial progress, largely as a result of abundant and cheap high performance computing. It is now possible to determine and analyze complex crystal structures entirely on inexpensive laptop or desktop computers with a few GB of memory. Automation and user interfaces have reached a high level of sophistication (although compatibility and integration issues remain). As a result, the actual process of structure solution, although the theoretically most sophisticated part in a structure determination, is commonly not considered a bottleneck in routine structure determination projects. Given reliable data of decent resolution (~2.5 Å or better) and no overly large or complex molecules, many structures can in fact be solved de novo and refined (although probably not completely polished) within several hours. Automated model building programs—many of them available as web services—have removed much of the tedium of initial model building. Several of the crystallographic programs and program suites will be used or introduced in the corresponding chapters, and a current compilation can be found in the
online supplement. The supplemental web material also contains links to the program web sites, which often contain useful tutorial materials and to web sites providing sample data and additional exercises.

1.8 Key concepts

- The power of macromolecular crystallography lies in the fact that highly accurate models of large molecular structures and molecular complexes can be determined at often near atomic level of detail.
- Crystallographic structure models have provided insight into molecular form and function, and provide the basis for structural biology and structure guided drug discovery.
- Non-proprietary protein structure models are made available to the public by deposition in the Protein Data Bank, which holds more than 57 000 entries as of May 2009.
- Proteins are generally difficult to crystallize; without crystals there is no crystallography.
- Preparing the material and modifying the protein by protein engineering so that it can actually crystallize is nontrivial.
- Radiation damage by ionizing X-ray radiation requires cryocooling of crystals, and many crystals are difficult to flash-cool.
- The X-ray diffraction patterns are not a direct image of the molecular structure.
- The electron density of the scattering molecular structure must be reconstructed by Fourier transform techniques.
- Both structure factor amplitude and relative phase angle of each reflection are required for the Fourier reconstruction.
- While the structure factor amplitudes are readily accessible, being proportional to the square root of the measured reflection intensities, the relative phase angles must be supplied by additional phasing experiments.
- The absence of directly accessible phases constitutes the phase problem in crystallography.
- The nonlinear refinement of the structure model is nontrivial and prior stereochemical knowledge must generally be incorporated into the restrained refinement.
- Protein crystals are formed by a loose periodic network of weak, non-covalent interactions and contain large solvent channels.
- The solvent channels allow relatively free diffusion of small molecules through the crystal and also provide conformational freedom for surface-exposed side chains or loops.
- The core structure of protein molecules in solution as determined by NMR is identical to the crystal structure.
- Even enzymes generally maintain activity in protein crystals.
- Crystal packing can affect local regions of the structure where surface-exposed side chains or flexible surface loops form intermolecular crystal contacts.
- Large conformational movements destroy crystals and cannot be directly observed though a single crystal structure.
- Limited information about the dynamic behavior of molecules can be obtained from analysis of the $B$-factors as a measure of local displacement or by analysis of correlated displacement by TLS (Translation-Libration-Screw) analysis.
- The quality of a protein structure is a local property. Surface exposed residues or mobile loops may not be traceable in electron density, no matter how well defined the rest of the structure is.

1.10 Additional reading


