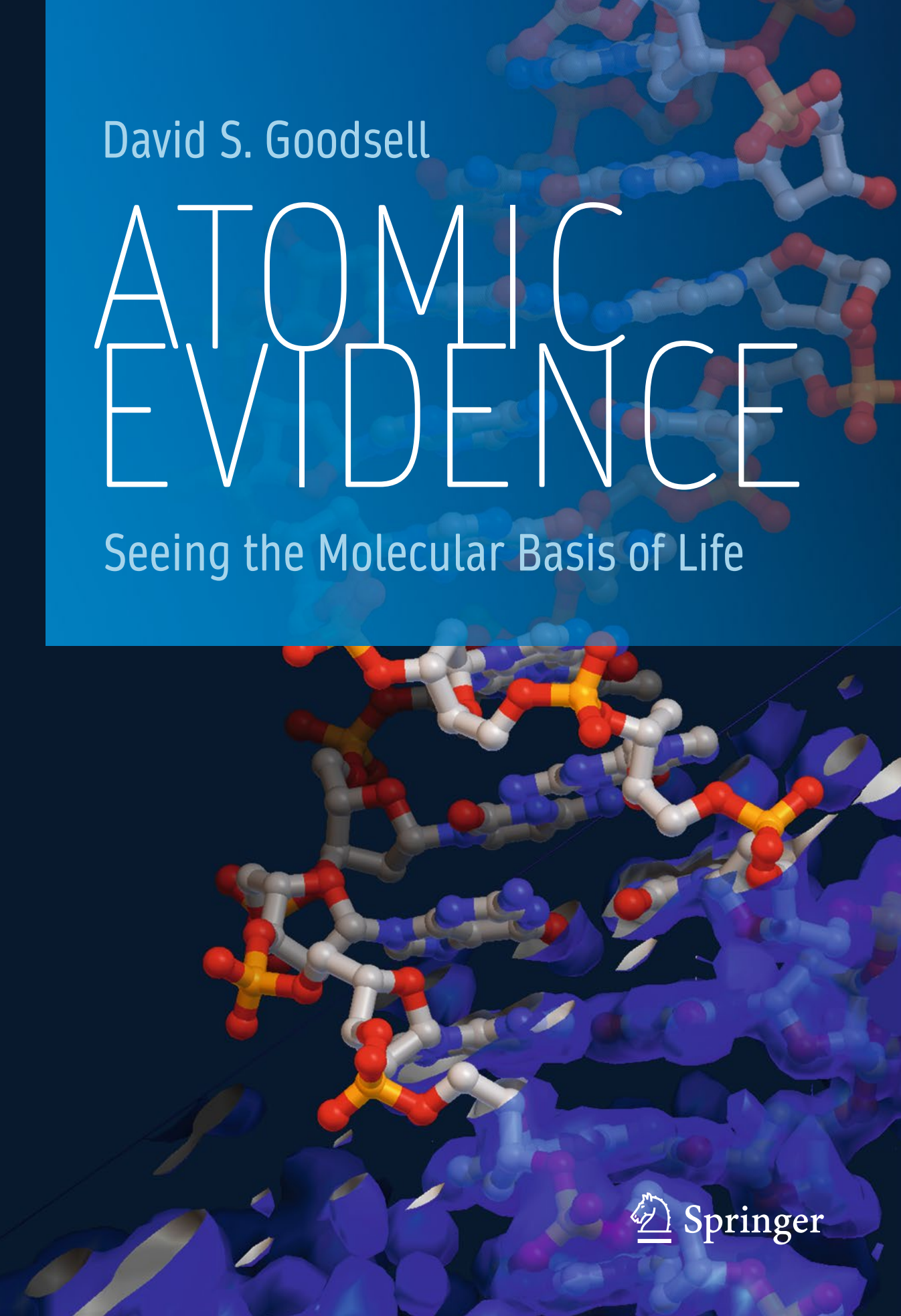
David S. Goodsell

ATOMIC EVIDENCE

Seeing the Molecular Basis of Life



Springer

Atomic Evidence

David S. Goodsell

Atomic Evidence

Seeing the Molecular Basis of Life

Springer

Copernicus Books is a brand of Springer

David S. Goodsell
The Scripps Research Institute
and RCSB Protein Data Bank
La Jolla, California
USA

ISBN 978-3-319-32508-8 ISBN 978-3-319-32510-1 (eBook)
DOI 10.1007/978-3-319-32510-1

Library of Congress Control Number: 2016943685

Printed on acid-free paper

This Copernicus imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

Why do I need a new flu shot each year? Should I be frightened by all the news about bacterial drug resistance? What about that new diet I just read about on the web? Biomolecular science is increasingly important in our everyday life, helping us answer questions like these, and giving us the knowledge to make critical choices about our diet, our health, and our wellness. How do fireflies light up? Why do plant and animal populations evolve over many generations? Biomolecular science also allows us to be curious, to look deeper into the natural world, and to be inspired by the complex inner workings of life.

In this book, I will take an evidence-based approach to current knowledge about the structure of biomolecules and their place in our lives, inviting us to explore how we know what we know and how current gaps in knowledge may influence our individual approach to the information. The book is separated into a series of short essays that present some of the foundational concepts of biomolecular science, with many examples of the molecules that perform the basic functions of life.

This book builds on my work with the RCSB Protein Data Bank, where I write a column each month that highlights atomic structures from the PDB archive. It has been a tremendous gift to have the opportunity to work on the Molecule of the Month, and I gratefully acknowledge Helen Berman, Stephen Burley, Christine Zardecki, and the entire RCSB team for their enthusiastic support over the past 15 years.

The molecular stories in this book are supported by a monumental body of work by scientists around the world. Throughout the book, I have included accession codes for structures at the PDB and EMDataBank. You can explore the structures directly at their websites (www.pdb.org and www.emdatabank.org). The database entries for each of these structures also include the primary journal publications that describe the detailed science supporting each structure.

David S. Goodsell
San Diego, CA, USA

Contents

1

The Protein Data Bank

D.S. Goodsell, *Atomic Evidence,* DOI 10.1007/978-3-319-32510-1_1

1

We're very lucky that today we can go to our computers and instantly start exploring a hundred thousand atomic structures of biomolecules. The structural biology community has spearheaded a comprehensive effort to make the results of biostructural research freely available to everyone. In 1971, a group of scientists at the Brookhaven National Laboratory started an archive of atomic structures, called the Protein Data Bank, as a way to make these structures available. The first archive contained the seven protein structures that were available at the time. Today the archive has grown to over a hundred thousand entries and is managed by centers around the world: RCSB and BMRB in the USA, PDBe in Europe, and PDBj in Japan. Together, they have created online interfaces to this massive archive, providing tools to deposit, curate, find, analyze, and visualize the structures.

This wasn't always the case, however. In the early days of structural science, many researchers chose to keep the primary results of their work, the atomic coordinates, secret. Instead of making these available, they published only pictures of their structures and descriptions of their own ideas about the structure and function. Arguably, this was justified: because these structures require so much effort to solve, these researchers wanted to have the freedom to analyze them completely themselves. Many researchers, however, felt that this policy went against the spirit of science, where results are made available and may be used by the entire community to build a more complete picture of our world. And perhaps more importantly, results need to be made available to allow other researchers to check their authenticity and reproduce any scientific insights gained from them.

For this reason, with the support of many researchers, Fred Richards drafted a letter in 1988 to the major government institutes funding science, requesting a policy that crystallographic data be made available, at least for all research supported by public funds. The effort was ultimately successful, and today, deposition of coordinates and data in a public database is typically a mandatory condition for funding of grants as well as for publication of results in many prominent journals.

The widespread availability of coordinates has transformed the study of molecular biology. Each structure is a window into a particular topic, allowing us to see the atomic details of biomolecular processes. But that is only the beginning. An entire field of structure-based drug design has been built upon these structures, allowing the discovery of new pharmaceuticals to fight everything from HIV to depression. Comparison of many different structures has led directly to new insights about the general principles for biomolecular structure and function and the evolution of these molecules, and these insights have blossomed into an entire field of protein design and biotechnology.

Today, we can download atomic structures for nearly any biological molecule we would be interested in exploring, from tiny hormones to huge viruses (Fig. 1.1). Most of the illustrations in this book are created directly from atomic coordinates from the PDB or, in some cases, from the experimental data supporting the atomic

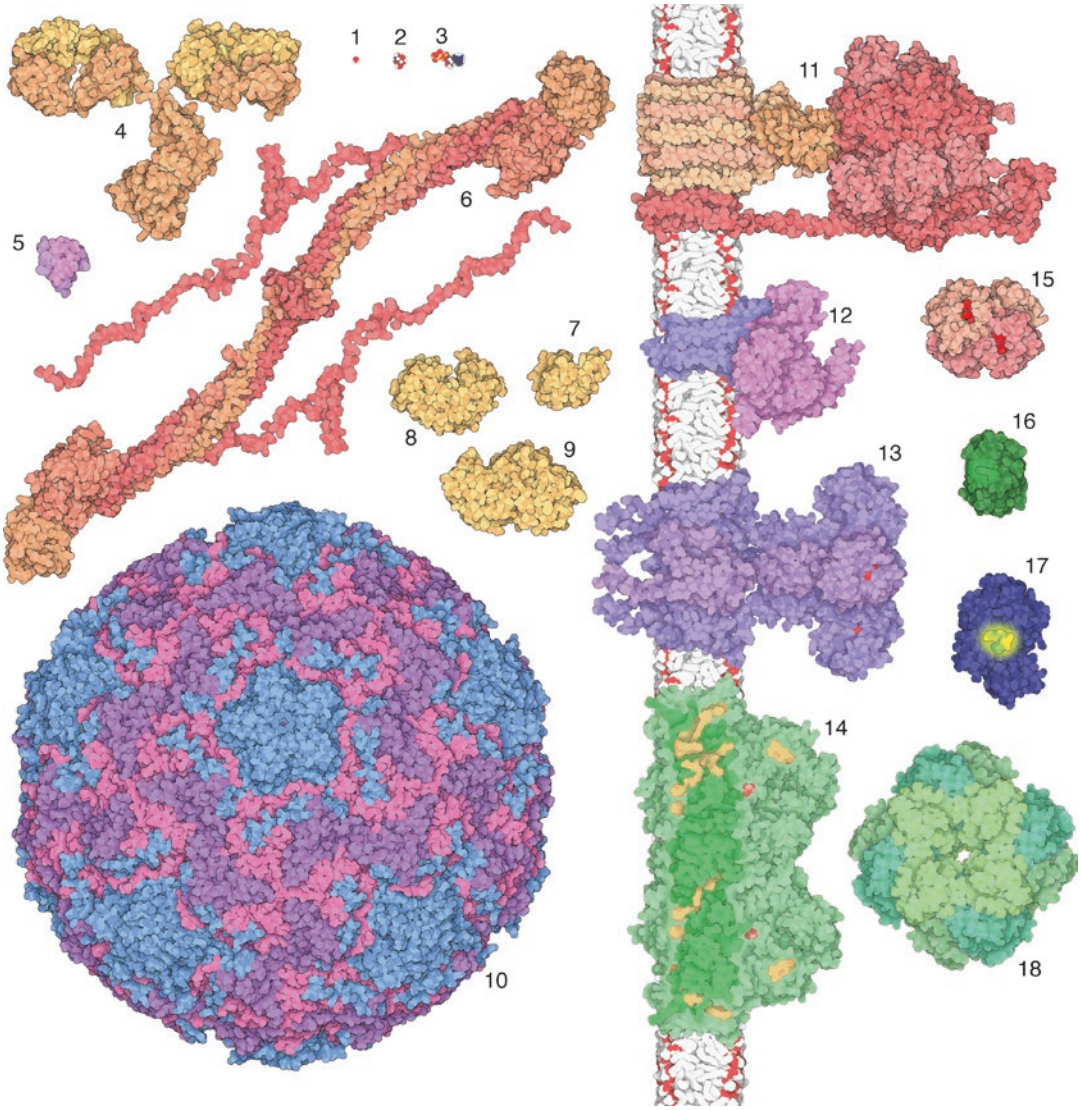


Fig. 1.1 Selected structures from the Protein Data Bank. The Protein Data Bank archives atomic structures of biomolecules such a proteins, DNA, and RNA. A few familiar examples are shown here. Three small molecules are shown for size comparison: (*1*) water, (*2*) glucose, and (*3*) ATP. Proteins in the blood: (*4*) antibody, (*5*) insulin, and (*6*) fibrinogen. Digestive enzymes: (*7*) lysozyme, (*8*) pepsin, and (*9*) amylase. A virus: (*10*) rhinovirus. Membrane-bound proteins: (*11*) ATP synthase, (*12*) adrenergic receptor and G-protein, (*13*) potassium ion channel, and (*14*) photosystem II. A few interesting proteins: (*15*) hemoglobin, (*16*) green fluorescent protein, (*17*) luciferase, and (*18*) ribulose-bisphosphate carboxylase oxygenase. Molecules involved in protein synthesis: (*19*) ribosome, (*20*) transfer RNA, (*21*) aminoacyl-tRNA synthetase, (*22*) protein chaperone GroEL/GroES, and (*23*) ubiquitin. A few enzymes: (*24*) catalase, (*25*) nitrogenase, and (*26*) leucine aminopeptidase. Proteins that bind to DNA: (*27*) repair protein DNA photolyase, (*28*) topoisomerase, (*29*) RNA polymerase, (*30*) lac repressor, (*31*) catabolite gene activator protein, and (*32*) transcription factor complex. Iron storage protein (*33*) ferritin and three enzymes involved in sugar metabolism: (*34*) hexokinase, (*35*) phosphofructokinase, and (*36*) pyruvate kinase (PDB entries 1igt, 2hiu, 1m1j, 2baf, 1lz1, 5pep, 1smd, 4rhv, 1e79, 1c17, 3sn6, 3lut, 1s5l, 4hhb, 1gfl, 2d1s, 1rcx, 1j5e, 1jj2, 4tna, 1ffy, 1aon, 1ubq, 1qqw, 1n2c, 1lap, 1tez, 1a36, 1tlf, 1efa, 1cgp, 1ais, 1hrs, 1dgk, 4pfk, 1a3w)

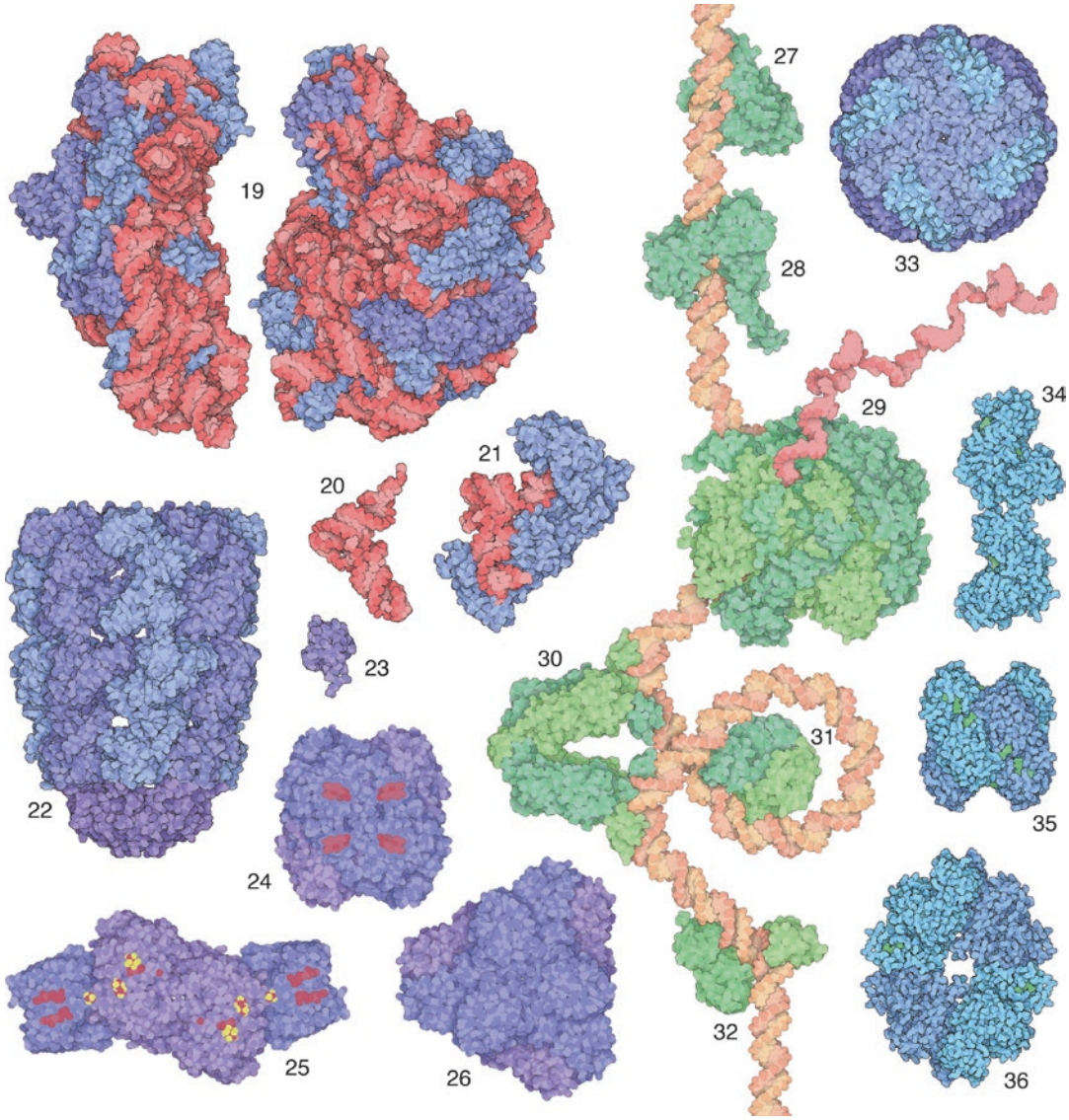


Fig. 1.1 (continued)

structure. For each illustration, I have included the accession code for the data at the PDB. With this information, you can easily explore the structure yourself using the tools at one of the PDB sites. The accession code also allows access to a variety of other information about the structure, for instance, the scientists who determined the structures, journal articles about the structure, and links to other databases related to the structure. So if a particular topic captures your interest, make a visit to the PDB to explore the molecules in more detail!

2

Seeing Is Believing: Methods of Structure Solution

D.S. Goodsell, *Atomic Evidence,* DOI 10.1007/978-3-319-32510-1_2

2

Scientists are curious people. We're always asking questions and then trying to figure out ways to answer them. This is particularly tricky with molecular biology. There's no direct way to see individual molecules, at least in atomic detail, so we're forced to use a bunch of specialized methods that probe different aspects of the structure. Then, from this information, we can build up an understanding of the molecule and create images of the molecule that are consistent with the data. Take, for instance, the ribosome (Fig. 2.1). Researchers have been working for decades on this elusive subject, assembling information from many sources to build the detailed understanding we have today.

All of the methods currently used to determine the atomic structures of molecules rely on observing many copies of molecule. For this reason, the first step is to purify the molecule, separating it from its cellular context. This is a surprisingly big limitation with these studies, for several reasons. First, we can't really see how it is acting in the cell—we only observe how it behaves in an artificial, purified state. Second, a variety of noncellular conditions are often necessary to stabilize the molecule in its purified state. Fortunately, in the case of ribosomes, when they are purified and mixed with the proper partners, they happily go about their task of building proteins, so we have reasonable confidence that they act similarly when they are in their normal environment in the cell. Finally, once we have a purified (but still active) molecule, we can bring to bear the three major techniques for exploring biomolecular structure: electron microscopy, x-ray crystallography, and nuclear magnetic resonance (NMR) spectroscopy.

Much of the seminal work on ribosome structure was performed using electron microscopy. It is a satisfyingly visual method, capturing more or less directly an image of individual ribosomes. Early studies would spread ribosomes on a surface and stain them with heavy metals, gathering pictures of the outer shape of the molecules. Today, a field of molecules are frozen in a thin layer of ice, and an image is captured. Computer analyses of these many molecules, caught in different orientations, are combined and aligned to create a three-dimensional map of the molecule. As I write this book, the field of cryoelectron microscopy is undergoing a technical revolution, and for some large, well-behaved molecules, this process gives enough information to determine the location of each atom in the molecule.

Electron microscopy was used to discover all the basic features of ribosome structure and function: the shapes of the large and small subunits, the threading of messenger RNA between them and the location of the transfer RNA subsites, an exit tunnel out the back of the large subunit, association of ribosomes with protein transporters in the endoplasmic reticulum, and many other things. Today, researchers are using the detailed structures from cryo-EM to reveal piece by piece each step of protein synthesis and interactions with the many molecules that assist with the process.

X-ray crystallography is the least ambiguous, but perhaps the most artificial, method for atomic structure determination. A very pure solution of the molecule is coaxed to form crystals using a

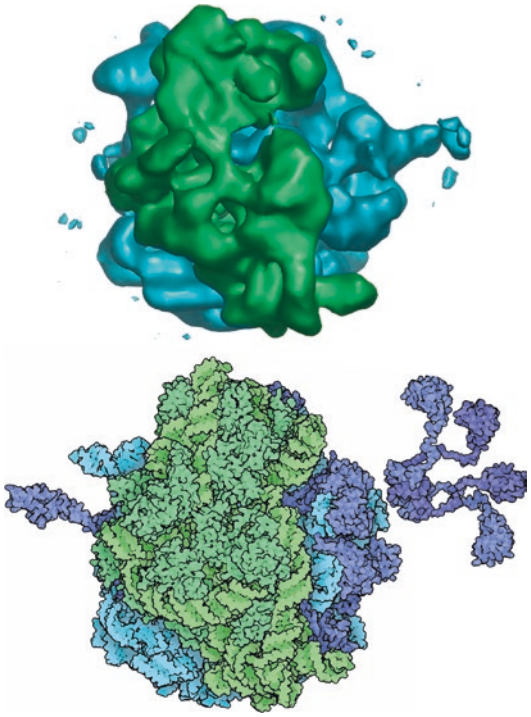


Fig. 2.1 Experimental views of a bacterial ribosome. The *upper image* shows a 3D reconstruction from electron microscopy, with the small subunit in *green* and the large subunit in *blue*. The *lower image* is an atomic structure from x-ray crystallography and an NMR structure of a flexible protein stalk that is not observed in the crystal structure (PDB entries 4v4q and 1rqv, EMDataBank entry EMD1110)

variety of unusual methods, such as concentrated solutions of salt or waxy polyethylene glycol. These crystals are then subjected to an intense beam of x-rays, which is diffracted into a characteristic pattern of spots by the many identically oriented copies of the molecule inside the crystal. Finally, these spots are analyzed to generate a three-dimensional map of the location of all of the electrons in the molecule. From this, the location of each atom is determined, provided that the crystal and diffraction are of high enough quality.

Crystallography has revealed the inner secrets of the ribosome in glorious detail. For many years, researchers studied the individual proteins by crystallography, slowly building up a picture of the whole molecule. Then, in 2000, three labs presented atomic structures of the intact ribosomal subunits. One major insight from these structures was the discovery that the ribosome is a ribozyme, with one particular nucleotide in the RNA catalyzing the protein-building reaction. The structures also revealed how the small subunit positions the messenger RNA, the details of the tunnel where the newly synthesized protein exits from the construction site, and a host of other interesting details.

NMR spectroscopy captures biological molecules in a more cell-like environment. A solution of the purified molecule is subjected to a radio field, and a series of characteristic resonances are

measured. By tailoring the types of fields, information is obtained on the local conformation of the molecular chain, and atoms that are close to one another may be identified. This information is then used to create an atomic model of the molecule that is consistent with all the observations. The complexity of NMR spectra typically limits the method to smallish proteins and nucleic acids, at least if entire atomic structures are going to be determined, but NMR excels at study of flexible molecules, which typically thwart structure determination by microscopy or crystallography. For instance, a recent structure of the L7/L12 stalk of the ribosome was solved by NMR methods, revealing how it changes conformations to organize the interaction of the ribosome with the many protein factors that guide each step of protein synthesis.

The structural biology community is currently very excited about the concept of "integrative" structural biology. The idea is to approach large and difficult problems by throwing everything we have at it. This approach is opening many doors that were previously closed for study, particularly for large and flexible assemblies. For instance, the integrative approach has been essential for all aspects of the study of the ribosome. Electron microscopy was used for years (and still is) to define the overall shape and evolution of ribosomes and to discover all of the basic mechanisms of protein synthesis. The recent atomic structures have revealed the details of ribosomes and many aspects of the peptide-forming reaction and interaction with drugs. But the integration of EM and crystallography is still essential for defining how the many protein helpers guide each of the steps and some of the more mobile aspects of the structure.

The underlying foundation of the scientific method tells us to question everything, and when we use the results of science, we always need to be critical. Do the experimental data support the structures or are we building them based on our biases or imagination? Are our discoveries about the function of the molecules based on what we have observed or on our preconceived notions? When we go to the PDB looking for a structure, we have to watch out for a few potential pitfalls.

Fortunately, the overall validity of structures in the PDB is not typically at question. Scientists are highly critical people, and there are usually at least two or three different groups competing with one another on a particular topic. We continually question our own work and that of our competitors, making sure that the results are supported by evidence. The PDB site, as well, contains a variety of methods for validating structures and assessing the quality of the underlying data. For instance, the quality of crystallographic data is often measured by the resolution of the electron density maps, which determine how much detail can be seen. Structures in the PDB range from structures where every atom may be clearly seen to elusive structures where only the general shape is observed (Fig. 2.2).

Each of these experimental methods has distinct advantages but also characteristic weaknesses. For instance, x-ray crystallography is typically able to determine very exact positions of heavy atoms

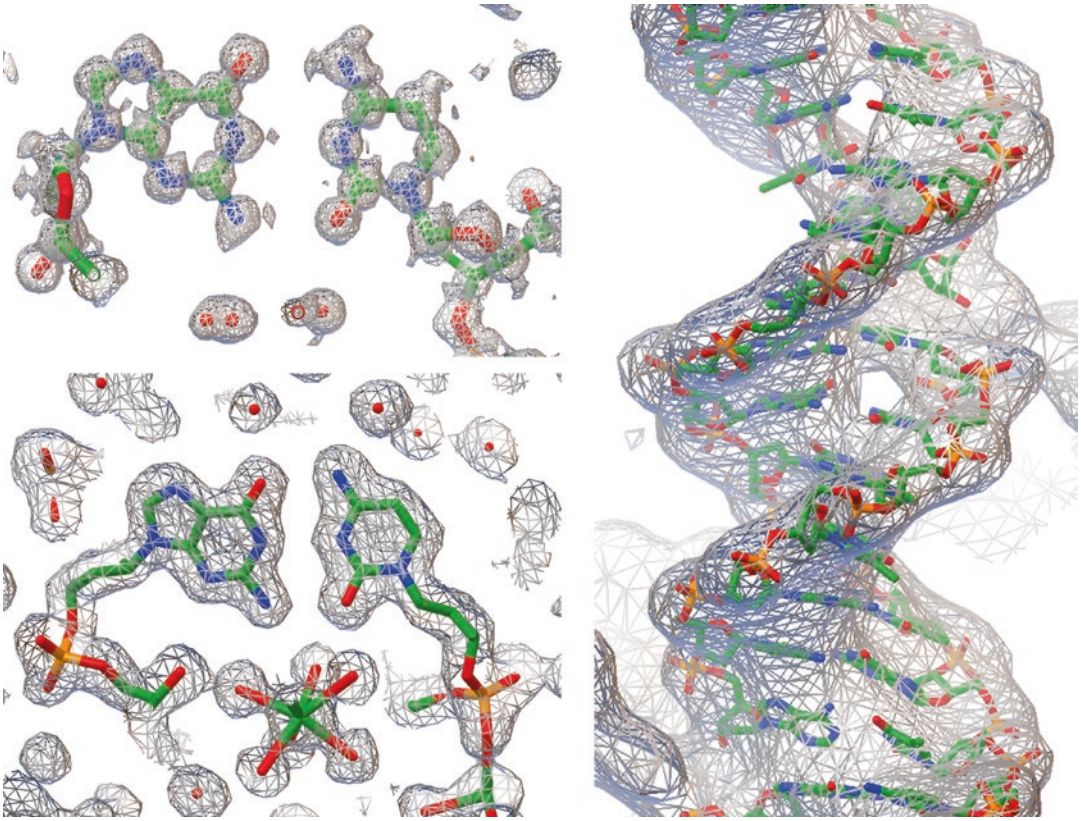


Fig. 2.2 Resolution of crystallographic electron density maps. Three electron density maps of DNA are shown here. At the *upper left* is a very high-resolution structure, where every atom is resolved, and we can even see hints of hydrogen positions. At the *lower left* is a more typical map, similar in resolution to most of the structures in the PDB. The overall shape of the bases and backbone, as well as a beautiful hydrated magnesium ion, is easily discernable, but individual atoms are not resolved. At the right is a low-resolution structure, which is sufficient to place the overall shape of the double helix, but not resolve the individual nucleotides (PDB entries 4hig, 196d, 3gbi, maps taken from the Uppsala EDS server)

(carbon, nitrogen, oxygen, etc.) in a protein molecule but rarely resolves the many tiny hydrogen atoms. For this reason, most of the structures in the PDB are missing their hydrogens, and if they are important for the study of the molecule, they need to be modeled based on the known geometry. NMR spectroscopy, on the other hand, observes the relative location of hydrogen atoms in a structure and infers much of the remaining structure based on the known chemical structure of the molecule.

Atomic structures are difficult to determine, and researchers often have to do drastic things to the molecules they study (Fig. 2.3). For instance, flexible molecules are often cut into smaller, more rigid pieces, and each piece is studied separately. To understand the function of the whole protein, we then need to reassemble the pieces in the computer to model the entire assembly. Proteins are often engineered to make them easier to study, with strings of histidine amino acids that are easy to purify or selenium atoms that have a distinctive signal in crystallographic experiments. In most cases,

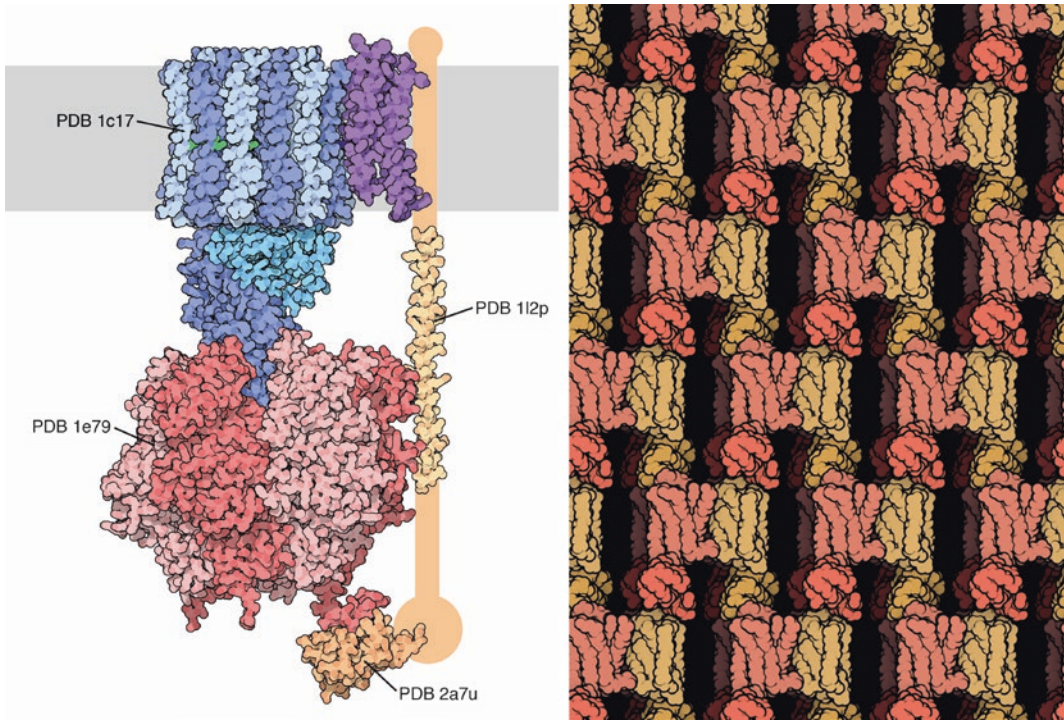


Fig. 2.3 Pitfalls of the PDB. ATP synthase (*left*) is a rotary motor with several moving parts. The whole assembly has not been crystallized yet, but structures have been obtained by cutting it into several more or less rigid pieces. G-protein-coupled receptors were an elusive target for many years, until researchers engineered a version with an entire lysozyme protein grafted into one loop. The lysozyme assists in the formation of crystals (*right*) (PDB entries 1c17, 1e79, 1l2p, 2a7u, 2rh1)

these modifications don't seriously perturb the function of the protein, but this needs to be validated through experiment to make sure we're getting a biologically relevant view.

Given the evidence-based approach of this book, I will show only the portions of the molecules that have been observed in experiment and use a schematic approach to show the portions that are inferred. Fortunately, science is a forever-growing field, and scientists continue to shed light into these currently shadowy areas.

3

Visualizing the Invisible World of Molecules

D.S. Goodsell, *Atomic Evidence*, DOI 10.1007/978-3-319-32510-1_3

3

In my career, I have had the great pleasure to be able to combine two of my interests: science and art. I started my studies at a serendipitous time, when the field of molecular visualization was just getting off the ground. When I started my studies as a graduate student, computer graphics was brand new, and those of us who knew how to use the hardware, and how to write the software to make it work, had a monopoly on the new technology. Scientists routinely came to us to create figures for papers, or movies for talks, or just to sit and explore their molecules. It was a wonderfully exciting time—we were making things up as we went, developing new methods for viewing molecules and trying to make them practical enough that we could use them in research (Fig. 3.1).

I'm happy to say that this has all changed now. Sophisticated computer graphics hardware is available on everybody's desktop, and even on our phones, and we have dozens of user-friendly molecular graphics programs to help us look at our molecules. Today, researchers produce most of their images themselves, without needing me to act as middleman between them and their molecules (Fig. 3.2).

Computer graphics images are our primary way of exploring and understanding the structure of biological molecules, and the pictures we create are the evidence that we use to document our discoveries. So, it is critically important that we use visual methods that are accurate and capture relevant aspects of the molecule's structure and function. Over the years, researchers have developed a number of useful ways to create images of molecules based on the experimental atomic structures. Initially, these images were created by clever scientists, often with the help of an artist. Today, nearly all molecular images are created with computer graphics. This has the

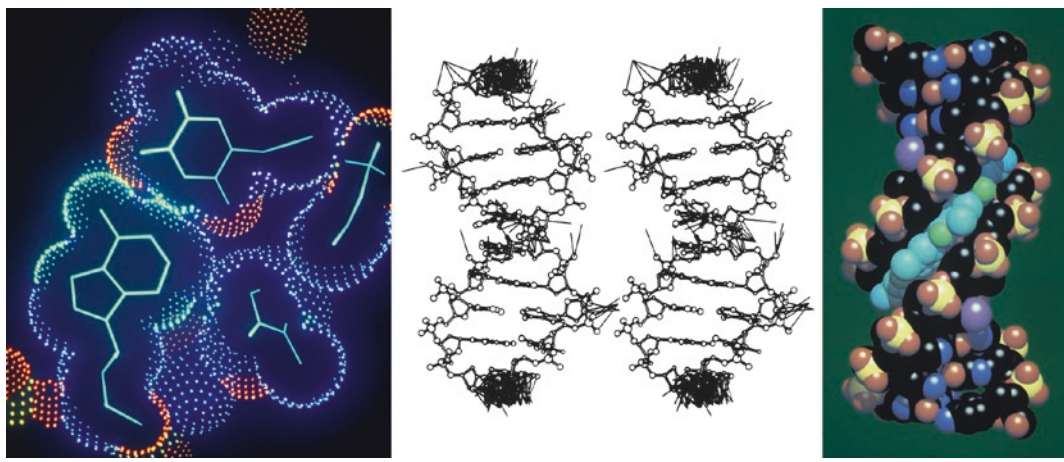


Fig. 3.1 Some experiments in molecular visualization. *Left*: the Evans and Sutherland Multipicture System allowed interactive display of *dots* and *lines* and was widely used by crystallographers to interpret their experimental electron density maps. This image shows a cross section through DNA molecule, with *lines* to show the bonds between atoms and *dots* to show the surface of the molecule. *Center*: pen plotters were used to create illustrations for journal publications, where most figures were printed in black and white. This illustration shows all of the sites of interaction between this DNA molecule and its neighbors in the crystal lattice. We often printed stereopairs like this to provide (with a little practice) a three-dimensional view. *Right*: raster images, which are used for almost everything today, were quite slow when they were first developed. This illustration of DNA took almost an hour to calculate