CHAPTER ONE

A General Introduction to 3-D Structures

PRIMER The three-dimensional structure of nucleic acids and proteins as it pertains to the mechanisms involved in gene regulation is the major focus of this book. Therefore, the reader will encounter many 3-D structures. The first chapter of the book presents the very basic ideas behind the three-dimensional aspects of biomolecules. The first part deals with the techniques used to determine 3-D structures. The presentation is virtually for the layperson. Then the basic structural elements found in proteins are examined. Having done this, we examine a particular 3-D structure (that includes both DNA and protein) presented with different modeling. This exercise will help you to become familiar with the different ways that scientists present their 3-D structures. We use different models because one aspect of structure and function can be better represented with one model, whereas another aspect is more suited to a different model.

This book deals with the three-dimensional aspects of gene regulation. The reader will encounter numerous three-dimensional structures, but this should not scare anybody away. Unfamiliar readers might think that interpreting these structures is difficult, but this is not true. All we need is a basic introduction into the three-dimensional aspects of proteins and nucleic acids and the way that it can be represented. The basic 3-D structure of a protein can be reduced to two elements: the alpha helix and the beta strand (and loops that connect them). The complicated 3-D structure of a protein is a combination of several of these elements. Also, depending on the presentation, the alpha helix or the beta strand might be shown with different styles. To get started, let us review the main elements of the 3-D structures, the different representation

styles, and the basic methods used in the determination of the three-dimensional structures.

Two methods are generally used to determine the 3-D structure of a biomolecule (nucleic acid or protein). One method is Nuclear Magnetic Resonance (NMR) spectroscopy, and the other is X-ray diffraction (or X-ray crystallography). NMR uses properties of the atomic nuclei to determine how closely they are positioned. The so-called nuclear Overhauser effect (NOE) is a nuclear relaxation effect. This intensity is a measure of the distance between two nuclei that are close together. The two nuclei might be far apart in the primary sequence, but they could be close in 3-D because the protein is folded. Gathering data from all atoms enables the researcher to create a 3-D model of the molecule under investigation. NOE(s) are detected by NOE spectroscopy (NOESY) NMR experiments. The intensity of NOESY determines the actual distance between two nuclei. A strong intensity indicates that the two nuclei are 3 Å apart, a medium intensity measures less than 4 Å, and weak intensity is less than 6 Å. Because structure determination by NMR is in solution, ends or loops of proteins, which are flexible, are sometimes not solved well. For this, more than 20 calculated structures should be received and superimposed. At this point, we should be able to see the regions that are not defined well. Finally, based on all calculated structures, an average structural model can be produced. These superimposed structures appear throughout the text. One limitation of this method is that it can resolve structures of small proteins (about 30 kDa). However, a few proteins of about 50 kDa have also been solved, and future developments might push these limits. Also, because NMR determines structures in solution, the protein should be stable in solution.

The other method, using X-ray diffraction, can be applied to large molecules or even complexes of them. When using X-ray diffraction, the protein must be crystallized. The crystal is then exposed to X-rays, and a picture is received on a film where the diffracted light from the crystal produces patterns, depending on the 3-D structure of the protein. For example, the celebrated 3-D structure of DNA, which is a periodic pattern, produced spots on the film, which were symmetrically arranged. This symmetry led Watson and Crick to deduce that the DNA must have two periodicities, one from base to base and the other every helical turn (nearly every 10 bases; see Chapter 2). Obviously, most complicated 3-D structures, such as the ones found in proteins, would produce a more elaborate pattern on the film, but algorithms and techniques have been developed to put these patterns into a 3-D structure. X-ray diffraction would provide very clear 3-D solutions and does not have the limitations with the flexible regions as in NMR. The only limitation is that not all proteins can be crystallized efficiently. When the same structure has been solved with both NMR and X-ray diffraction, the results usually match very well, indicating that both methods are quite reliable.

Let us now familiarize ourselves with the basic structures in a protein. As noted earlier, the primary amino acid sequence can assume either a helical or a

beta strand conformation. Some amino acids are more likely than others to be in an alpha helix, and the same is true for amino acids found in beta strands. First, we will examine the basic structure of an amino acid and the peptide bond. All amino acids have a central carbon, C_α, to which a hydrogen atom, NH_2 (amino group), and COOH (carboxyl group) are attached. What discriminates the 20 different amino acids is the side chain, R, which is attached to the central carbon atom (Figure 1.1A). Amino acids are joined via the peptide bond to create polypeptides (Figure 1.1B). When amino acids are arranged in an alpha helix, there is hydrogen bonding between the C═O of a residue and

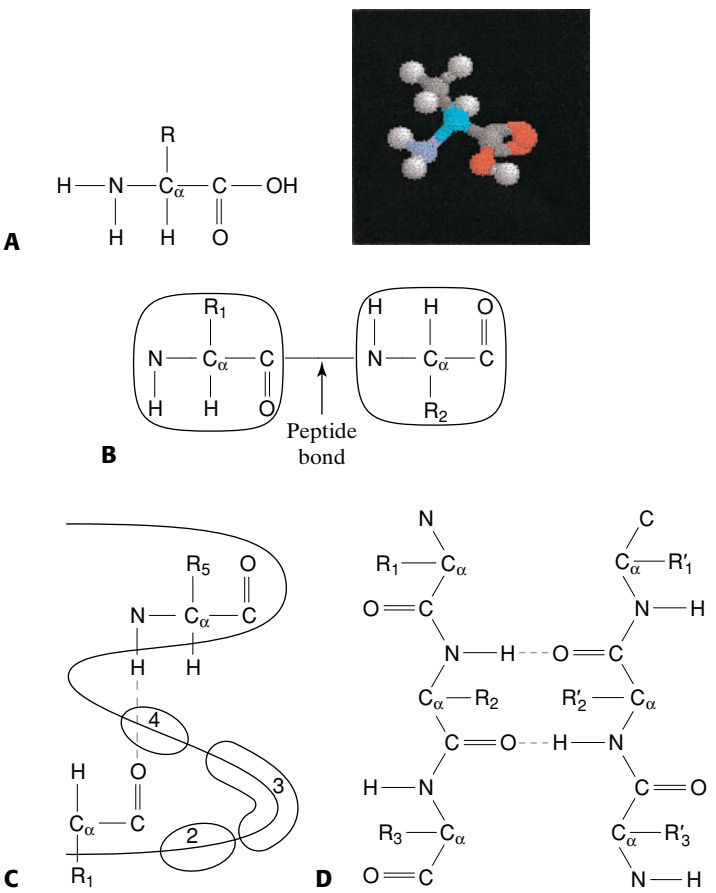

Figure 1.1. A: The basic chemical structure of an amino acid, indicating the standard H, NH_2, and COOH groups. R is the side chain that can vary in different amino acids. In the ball-and-stick representation, R is a CH_3 group and the amino acid is alanine. C is gray with the C_α cyan, N is blue, oxygen is red, and H is white. **B:** A dipeptide showing the creation of the peptide bond. **C:** An illustration of alpha helix. Note that residue 1 and residue 5 interact via hydrogen bonding (dashed line) using their C═O and NH groups, respectively. **D:** Two antiparallel beta strands creating a beta sheet via hydrogen bonding (dashed line) using an NH group from one strand and a C═O group from another. From F. R. Gorga, Protein Data Bank (PDB), Nucleic Acids Res. 28: 235–42.

the NH of another residue four positions away. In other words, there would be hydrogen bonds between residue 1 and 5, 2 and 6, and so on (Figure 1.1C). The alpha helix has 3.6 residues per turn, but variations exist with hydrogen bonds to residue $n + 5$ (pi helix) or $n + 3$ (3_{10} helix). Most of these helices are found at the end of alpha helices. The alpha helices are usually depicted as ribbons or cylinders in the 3-D structure of a protein.

The beta strand, and the resulting beta sheets from their interaction, is the second major element found in proteins. A beta strand contains 5 to 10 amino acids, which are in almost fully extended conformation. Interactions with adjacent beta strands can form a beta sheet. These interactions involve hydrogen bonding between the C═O of one strand and the NH group of another (Figure 1.1D). From such a configuration, we can see that the beta strands are pleated with C_α atoms successively above or below the plane of a sheet. The side chains follow this pattern as well. A beta sheet is called parallel when the strands run in the same direction or antiparallel when they do not. The example in Figure 1.1D is an antiparallel beta sheet. The beta strands are usually represented as arrows in the 3-D structure of a protein with the arrowhead pointing to the direction ($N \rightarrow C$).

This book contains numerous structures that are represented as different models. This approach is deliberate because some models can show a particular feature much better than others. To illustrate, Figure 1.2 presents a particular structure using four different models. The structure shows the interaction between the paired domain of the activator pax-6 with DNA. It is a good example

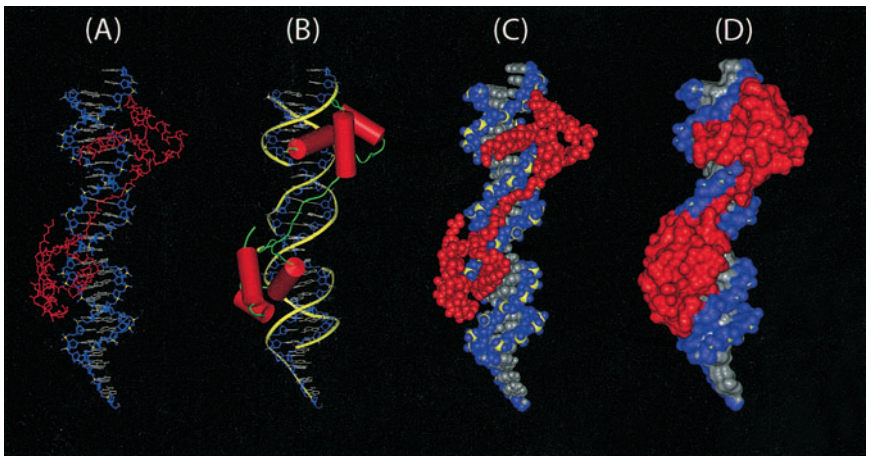

Figure 1.2. Different models of pax-6 bound to DNA. **A:** Ball-and-stick diagram. Phosphates are yellow, the sugar moiety is blue, nucleotide bases are gray, and pax-6 is red. **B:** Same as in A, but the helices of pax-6 are shown as red cylinders, and the connecting parts, as green strings. The DNA phosphates are yellow and have been traced to highlight the DNA. **C:** CPK (space-filling) model with same colors as in A. **D:** Surface representation with the same colors as in A. Images generated by E. Fuentes; Xu et al., Genes Develop. 13: 1263–75 (1999).

because we can examine both a nucleic acid and a protein. In Figure 1.2A, we can see the so-called ball-and-stick diagram, where atoms or groups are represented by balls (usually different colors; see also the model of alanine in Figure 1.1) and connected by sticks. In Figure 1.2B, we can see the same model as in Figure 1.2A, but the protein helices are represented as cylinders here. Helices can also be represented as ribbons. Both cylinder and ribbon models are used throughout this book. In Figure 1.2C, we can see the so-called CPK (Corey-Pauling-Kultun) model or spacefill (filling) model, which shows the surface of each atoms or group. Finally, in Figure 1.2D, we can see the solvent surface of the structure. This looks like the CPK model, but it represents the surface of the whole molecule instead of showing the surface of the atom or a group. This model is used mostly to represent the potential of a molecule, with red representing negative electrostatic potential and blue, positive. Variations of these models exist, but these models are the most common ones.

CHAPTER TWO

The Higher Organization of the Genome

PRIMER DNA is not free in the nucleus; it is bound by the proteins that package it. This fact is important because DNA must be accommodated in a small place. Also, as we will see in Chapter 6, the packaging provides a level of transcriptional control. In this chapter, the goal is to become familiar with the proteins that are involved in packaging and their effects on DNA. The first part of the chapter examines the 3-D structure of the nucleosome and its components, the histones. This structure is the highest order of DNA packaging. Other proteins (nonhistone), however, have the ability to bend DNA and, therefore, provide another degree of organization. Such architectural proteins are important because linear DNA is not very efficient when it comes to regulation. Regulatory elements are often separated by long distances, and they must come close for interaction. Also, bending DNA makes it more accessible for interactions with proteins. These nonhistone proteins possess different structural motifs that mediate interaction with DNA. Their structural characteristics and effects on DNA are presented. The importance of these architectural proteins is also stressed in later chapters.

PACKING DNA IN THE CHROMOSOMES

The genetic material, DNA, is packaged into chromosomes as chromatin, which is the DNA and chromosomal proteins. The way that this packaging is achieved is suitable for DNA accommodation in the nucleus. It also most likely plays a role in DNA regulation. The packing of DNA is characterized by coils, loops, and coils within the loops. Such an organization guarantees that the DNA is packed in an orderly fashion, making use of less space in the nucleus and, at the same time, avoiding extensive knotting, which could inhibit the

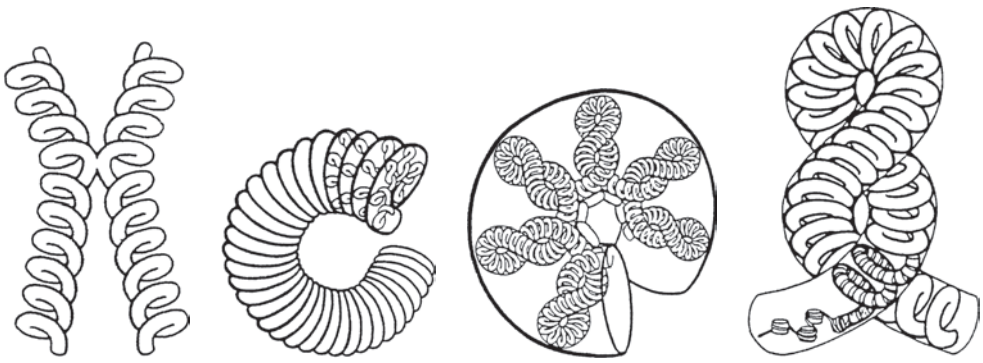

Figure 2.1. The different levels of chromatin folding into chromosomes. From P. R. Walker, EMBO J. 9: 1319–27 (1990).

necessary processes of replication and transcription. There are several levels of chromosomal folding. Starting from the higher level, we can observe that the chromosomes are coiled and that there appear to be nearly 10 coils per sister chromatid. Each coil is folded into rosettes, with 30 rosettes making up a coil. Each rosette is folded up into six loops with each loop containing up to 50 kbp of DNA. The loops are made up of 30-nm fiber, which is called solenoid. Each solenoid consists of 12 of the simplest units of packaging, the nucleosomes. At this level, we can observe that the DNA strands wrap around a proteinous core structure (Figure 2.1).

The proteins that make up a nucleosome are called histones, namely H2A, H2B, H3, and H4. The whole structure is an octamer of these histones and is constructed of two H2A-H2B and two H3-H4 dimers. The DNA (146 bp) wraps the histone complex in two turns (Figure 2.2).

The portion of the DNA that connects different nucleosomes depends on the species but is nearly 60 bp long and is bound by a different histone, H1. It seems that H1 binds DNA in such a fashion that a zigzag arrangement of the nucleosomes is possible. This arrangement is necessary for the nucleosomes to form the solenoids and has been observed in electron microscopy (Figure 2.3).

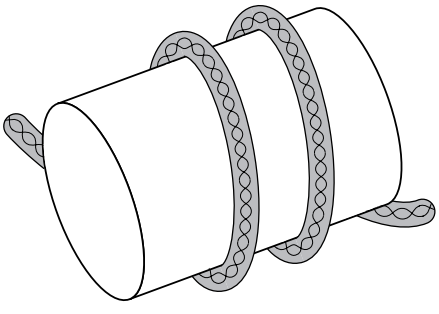

Figure 2.2. The nucleosome consisting of a histone complex (tube) and the wrapping DNA.

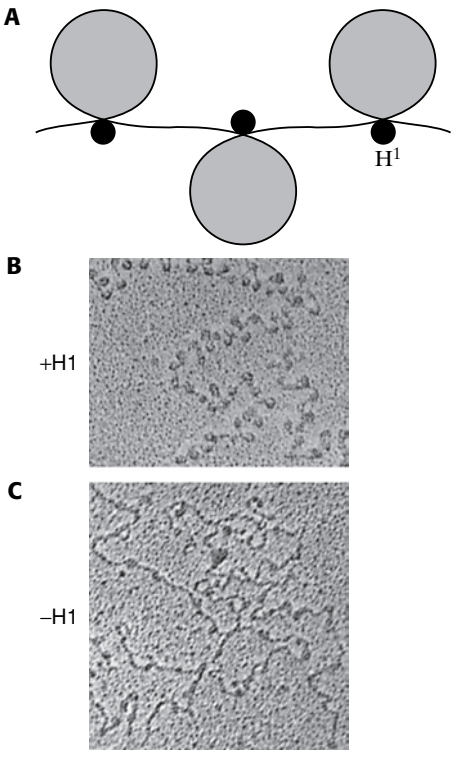

Figure 2.3. A: H1 binds free DNA. **B** and **C:** Binding of chromatin as observed by electron microscopy with (**B**) or without (**C**) H1. Note that the zigzag arrangement disappears in C. B and C reproduced from A. Klug, J. Cell Biol. 83: 410 (Nov. 1979), with copyright permission of the Rockefeller University Press.

THE THREE-DIMENSIONAL STRUCTURE OF THE NUCLEOSOME

The higher order of folding has not been observed through 3-D images, and speculation of the packaging order has been based on electron microscopy and biochemical studies. Therefore, our view of the higher order might change in the future when more accurate 3-D images will be available. However, the 3-D structure of the nucleosome has been solved at a level where accurate atomic interactions can be observed. These images have provided unique insights on the organization and structure of nucleosomes. As mentioned, the nucleosome is composed of a histone octamer wrapped by 146 bp of DNA. The specific arrangement of histones is characterized by the two H2A-H2B and two H3-H4 dimers. Before we examine the structure of the nucleosome, however, let us become familiar with the basic 3-D structure of histones. Each histone is composed of three helices that form the characteristic histone fold (Figure 2.4), and two tails with no particular structure.

Interactions between DNA and proteins can be seen in Figure 2.5, which is a cross section of a nucleosome (one H2A-H2B and one H3-H4 dimer with one turn of DNA). Several important features can be observed in Figure 2.5. H3 specifically binds DNA at the entrance and exit site. The wrapping DNA is contacted every 10 bases, at intervals at minor grooves that face the proteins. The phosphate backbone of the DNA is contacted by main-chain atoms of

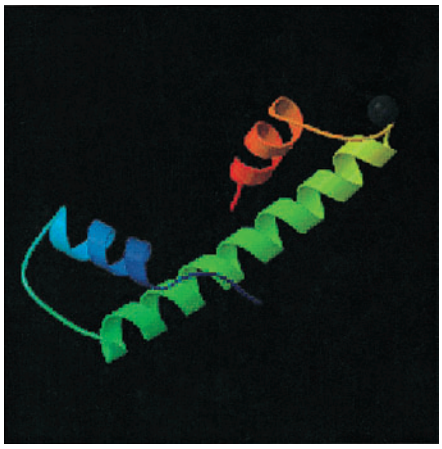

Figure 2.4. The characteristic structure of histones indicating the histone fold. The histone fold is characterized by one long alpha helix flanked on each side by a loop and a shorter alpha helix. This fold is for *Methanothermus ferridus* H3. From U. Heinemann, PDB file 1 HTA, K. Decanniere et al., J. Mol. Biol. 303: 35–47 (2000).

the proteins. This interaction is not sequence specific (because all different sequences of DNA wrap the nucleosomes) and is necessary to keep DNA tightly bound. Arginine residues penetrate all 14 minor grooves (7 in the cross section) and face the core of the protein arrangement.

The helical periodicity (average number of bases per helical turn of DNA) is 10.6. However, around the nucleosome, the periodicity is 10.2. In addition to

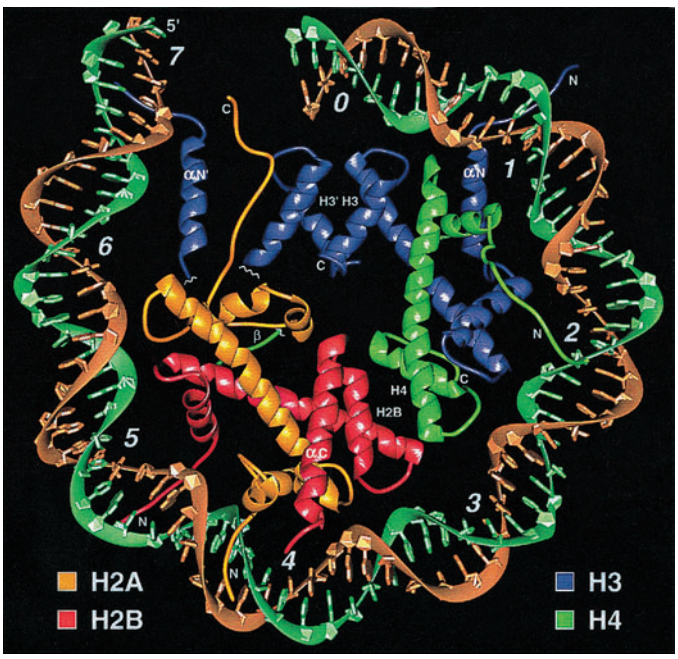

Figure 2.5. 3-D structure of the nucleosome and interactions of histones and DNA. From T. J. Richmond, Nature 389: 251–60 (1977). Reprinted by permission from Nature, Macmillan Magazines Ltd.

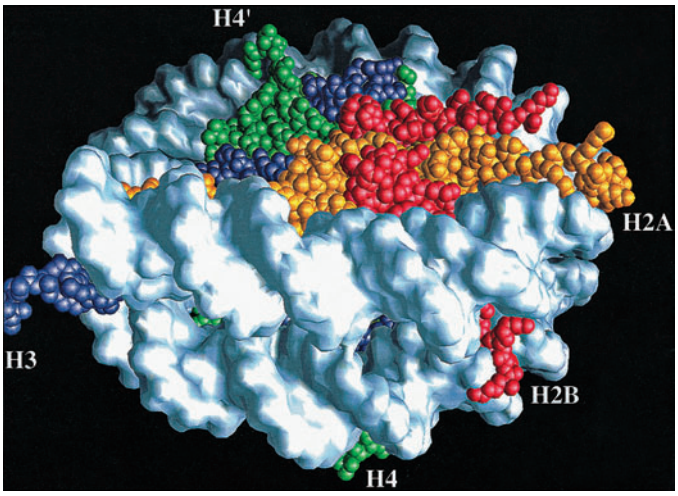

Figure 2.6. Histone tails between DNA turns in the nucleosome. Note that the tails of H3 and H2B pass the channels in the DNA superhelix (white, surface representation). Courtesy of Dr. T. J. Richmond. T. J. Richmond, Science 277: 1763 (1997).

the fact that the DNA superhelix core wrapping is not uniform but distorted by bends, this periodicity positions the minor and major grooves from neighboring turns in such a way as to line up and form channels through which histone tails can pass (Figure 2.6). This remarkable arrangement can actually enable different nucleosomes to interact and pack tightly together.

OTHER PROTEINS

With the exception of this kind of packaging in the nucleosomes, other processes of DNA require higher order nucleoprotein complexes. These are mainly replication, transcription, recombination, and transposition. For such processes to take place, DNA distortions, such as bending wrapping and looping, are necessary. Such distortions are facilitated by other sequence-nonspecific DNA-binding proteins. Some of the most important such proteins include the prokaryotic HU protein and the eukaryotic HMG (High Mobility Group). These proteins have been implicated in replication and transcription, and they are also called DNA flexers. The 3-D structure of the *Escherichia coli* HU protein and its interactions have been solved and provide a view of how these proteins bend DNA. HU is a heterodimer of two highly conserved and related subunits, creating a base consisting of two alpha helices from which antiparallel beta sheet arms are extended (Figure 2.7A). The arms contain highly charged amino acids (as in histones). HU proteins bind DNA in the minor groove in a sequence-nonspecific manner. Flexing of DNA can be achieved by interactions of the arms of the HU heterodimer as shown in Figure 2.7B.