Protein–nucleic acid interactions

Editorial overview

Jennifer A Doudna* and Timothy J Richmond†

Addresses
*Yale University, Department of Molecular Biophysics and Biochemistry and Howard Hughes Medical Institute, 260 Whitney Avenue, New Haven, CT 06520, USA; e-mail: doudna@csb.yale.edu
†ETH Zürich, Institut für Molekularbiologie und Biophysik, ETH-Hönggerberg, CH-8093 Zürich, Switzerland; e-mail: richmond@mol.biol.ethz.ch

Current Opinion in Structural Biology 2001, 11:11–13

0959-440X/01/$ – see front matter

Specific binding of proteins to a wide variety of nucleic acids underlies all aspects of gene expression, including genome replication, repair, transcription and RNA metabolism. During the past year, many exciting advances have increased our understanding of the structural and chemical basis for protein–nucleic acid interactions. Six reviews in this section of *Current Opinion in Structural Biology* focus on DNA–protein recognition involved in the bacterial restriction system, several aspects of transcriptional regulation and DNA mismatch repair. Two reviews discuss RNA–protein complexes involved in eukaryotic RNA metabolism and viral RNA packaging. A significant advance not included here is the recent determination of atomic resolution structures of both subunits of the bacterial ribosome, as well as medium-resolution structures of the intact ribosome, a major achievement that provides fertile ground for exploring both the principles of protein–RNA interactions and the evolutionary relationships between these complexes.

The bacterial restriction system relies on the ability of the restriction endonucleases to recognize and cleave short 4–8 base pair palindromic DNA sites, thus providing host cells with a means of destroying the DNA of invading pathogens. The exceptional specificity of these enzymes has led to their widespread use as tools of modern molecular biology, enabling the site-specific scission of DNA for cloning and analysis. Remarkably, a single base pair change within the recognition sequence results in more than a million-fold loss of activity. Several crystal structures solved in the early 1990s revealed that, despite little sequence homology, restriction enzymes share a similar α/β core motif and bind to their recognition sites as dimers. Nonetheless, it has proved difficult to alter existing enzymes to recognize even closely related sequences. Lukacs and Aggarwal (pp 14–18) discuss the basis for this immutability by comparing the co-crystal structures of *Bgl*II and *Mun*I, which recognize six-base-pair DNA sites that differ by only the outer base pairs from the recognition sites of two well-characterized enzymes, *Bam*HI and *Eco*RI, respectively. *Bgl*II and *Bam*HI recognize the central, common base pairs of their DNA recognition sites differently and in the *Bgl*II complex the target DNA is bent such that the outer base pairs do not superimpose on the positions of those in the *Bam*HI complex. *Mun*I and *Eco*RI recognize the common, inner base pairs of their recognition sequences using similar contacts in the DNA major groove, but contacts to the outer base pairs involve amino acids from different parts of each protein. The structures emphasize the importance of the entire protein in determining sequence specificity, explaining the resistance of these enzymes to specificity switches.

Structural studies of the bacterial Lac repressor protein continue to provide a wealth of information about both DNA–protein interactions and allosteric regulation of gene expression. The tetrameric repressor binds to three similar, but not identical, operator DNA sites using helix-turn-helix recognition motifs, down-regulating expression of *lac* mRNA in the absence of allolactose or similar ligands. Upon binding to activating ligands, the repressor changes conformation and releases the operator sites, enabling gene expression to occur. A recent crystal structure of a dimeric form of the *Escherichia coli* Lac repressor bound to its operator DNA site, reviewed by Bell and Lewis (pp 19–25), provides a detailed view of how the interactions at the monomer–monomer interface are altered in switching from the operator-bound to the inducer-bound conformation. Furthermore, an NMR structure reveals that the repressor's hinge helix, implicated in conformational switching, makes detailed interactions with the minor groove of the operator DNA. Biochemical experiments confirm previous suggestions that the repressor bends the wild-type operator similarly to the bending observed in complexes containing engineered symmetric operators, although this remains to be verified by a structure of the repressor bound to wild-type operator DNA.

Eukaryotic transcription involves DNA recognition by both general and sequence-specific transcription factors, and recent progress in understanding these interactions is reviewed by Müller (pp 26–32) and by Rastinejad (pp 33–38). As for restriction enzymes, specific sequence recognition is achieved by proteins with similar folds that use surprisingly different modes of binding. The helix-turn-helix (HTH) motif common to many transcription factors can be altered to recognize different target sequences through changes in sidechain conformations, water structure or interdomain linkers. Transcription factors belonging to the immunoglobulin-fold superfamily use variable loops connecting strands of the central β-barrel core to contact different DNA sequences. For nuclear hormone receptors, binding of the hormone enables recognition of a six-base-pair DNA sequence upstream of

hormone-responsive genes. In this case, differential spacing of tandem target sites creates response element diversity by producing binding sites for a range of hormone receptor homodimers and heterodimers. Two recent crystal structures of the retinoid X receptor (RXR) bound to DNA reveal that the protein–protein dimer contacts form exclusively at the minor groove of the spacer between the recognition elements. The length of the spacer, typically between one and five base pairs, governs the protein–protein contacts that can form. Consequently, protein dimerization occurs only in the presence of the appropriate DNA response element. Furthermore, the pattern of site selectivity based on target spacing implies that the signaling pathways of different hormones could be interchanged simply by adding or removing single base pairs from the spacing of the target site repeats.

General and sequence-specific transcriptional regulation involves large multicomponent complexes that interact with RNA polymerase II. Together with recently determined crystal structures of RNA polymerase, low-resolution three-dimensional reconstructions of the multisubunit transcription factor TFIID provide a basis for understanding the dynamic arrangement of factors during transcription initiation. The horseshoe-shaped structure of TFIID contains a deep groove large enough to accommodate a DNA duplex and antibody tagging has revealed the positions of specific factors within the complex. The X-ray structure of the double bromodomain from the $TAF_{II}250$ component of TFIID calls attention to the possible involvement of TFIID in chromatin modification. It is clear from these studies that a combination of biophysical techniques, coupled with biochemical analysis, will be essential to unravel the complexities of transcriptional regulation *in vivo*.

Evolution has produced a variety of structural motifs incorporating the specific chelation of a divalent zinc ion by combining four variably spaced cysteine and histidine amino acids. The first of these motifs relevant to DNA binding was discovered in transcription factor IIIA (TFIIIA), was named the zinc finger and comprises a $\beta\beta\alpha$ module containing about 30 amino acids. It appeared nine times in TFIIIA and the occurrence of multiple copies of zinc fingers has become the rule, rather than the exception, for DNA-binding domains containing them. Although DNA recognition by zinc fingers has been well studied, the effect of the linker polypeptide connecting modules is only recently becoming understood. The evidence for a dynamically flexible linker becoming ordered on DNA binding and thereby making an important contribution to specificity is reviewed by Laity, Lee and Wright (pp 39–46) in this section. They then delve into structural variations on the zinc finger theme, both natural and engineered, and show how variant $\beta\beta\alpha$ modules, as well as other secondary structure motifs, maintain zinc binding based on the cysteine and histidine sidechains. They illustrate how this diversity in zinc module structure extends to function as well. Although clear functional examples exist for DNA binding by transcription factors and for RNA interaction within the ribosome, it is also likely that zinc modules mediate protein–protein interactions and, possibly, protein–lipid interactions.

Faithful replication of the genome requires detection and repair of DNA mismatches during DNA synthesis. At the center of the repair pathway is MutS, a dimeric protein highly conserved from bacteria to humans that is responsible for recognition of mismatches. In humans, mutations in MutS homolog (MSH) proteins give rise to a common form of hereditary colon cancer, underscoring the importance of the repair pathway in normal DNA biosynthesis. Two recent crystal structures of bacterial MutS proteins, reviewed by Sixma (pp 47–52), suggest that recognition of a mismatched base results from specific hydrogen bonding to the mismatch, together with shape recognition of bent DNA, which depends upon the greater flexibility of mismatched DNA. In each MutS monomer, two domains are involved in DNA binding and induce a 60° bend in the DNA. The insertion of a functionally critical phenylalanine residue adjacent to the mismatched base implies a mechanism in which the MutS dimer scans substrate DNA for mismatches by localizing to flexible (i.e. mismatched) helical regions that are susceptible to bending and phenylalanine intercalation.

Specific recognition of RNA sequences and structures is central to every aspect of RNA metabolism, and Pérez-Cañadillas and Varani (pp 53–58) review recent advances in understanding how this is achieved. RNA-binding proteins typically have a modular structure and contain RNA-binding domains of 70–150 amino acids in size. Three major classes of eukaryotic RNA-binding proteins have been identified: the RNA-recognition motif (RRM), the double stranded RNA binding domain and the K-homology domain. Structures of representative members of each class share a similar $\alpha\beta$ fold, but recognize RNA in completely different ways. Many proteins contain multiple RNA-binding domains connected by short flexible linkers and these tandem domains enable greater specificity and affinity for target RNAs. Nearly all of the characterized RNA–protein complexes involve induced fit of the interacting partners and Pérez-Cañadillas and Varani discuss the biological implications of this observation. As for many other biological processes, future efforts will focus on understanding how these isolated domains function as part of multicomponent assemblies.

In RNA viruses, protein–RNA interactions direct the assembly of the viral capsid, resulting in highly ordered RNA structures and disposition within virion particles. Larson and McPherson (pp 59–65) review recent crystallographic studies of satellite tobacco mosaic virus in which close to 80% of the single-stranded RNA genome can be localized in electron density maps. Nonidentical stem-loop elements occur at every edge of the icosahedral virion, a

compact fold well suited for efficient RNA packaging, but incompatible with RNA substructures required during the viral life cycle. Hence, the structure implies that significant rearrangements of the RNA occur during viral infection, perhaps requiring multiple cycles of RNA folding and unfolding. Furthermore, the structure shows that variable lengths of stems, loops and connecting strands can be accommodated within the capsid, possibly allowing for evolution of the viral genome.

Diversity of detail seems to be the general message from the reviews presented here. It is exciting to see brand new structures, such as the DNA mismatch repair protein MutS for the first time, particularly when they are seen in complex with interacting partners that are requisite for function. They stimulate the imagination of possible mechanisms of biological action, as well as evolutionary relationships among protein families. But although imagination is a guide, the devil is in the detail. For that reason, key results from long-term studies designed to elucidate mechanism are intellectually satisfying. The classic structure presented here is, of course, the Lac repressor, for which the atomic mechanism of allosteric induction of DNA release is finally coming to light merely three decades after the protein was first available for study. No less important are the comparative studies, such as for the *Bgl*II and *Mun*I endonucleases, which demonstrate that the balance of life hangs on very subtle differences, with only a few atomic contacts or hydrogen bonds having important functional consequences. Although today there is much emphasis on filling databases, at least as significant are the efforts to reveal the complexity of interacting partners in macromolecular assemblies and the pain taken to ferret out the details of mechanisms one atom at a time until they are completely understood