Chapter 9. Assembly of the Transcription Initiation Complex

Learning outcomes
9.1. The Importance of DNA-binding Proteins
9.2. DNA-Protein Interactions During Transcription Initiation
9.3. Regulation of Transcription Initiation
Self study questions

Learning outcomes

When you have read Chapter 9, you should be able to:

1. Outline the various techniques that are used to locate the position at which a DNA-binding protein attaches to a DNA molecule
2. Explain how a DNA-binding protein is purified and how its structure can be determined
3. Describe the key structural motifs that enable proteins to make sequence-specific attachments to DNA molecules
4. Discuss the features of the double helix that are important in interactions between DNA and its binding proteins, and give details of the chemical events that underlie the interaction
5. Identify the key features of the various prokaryotic and eukaryotic RNA polymerases and describe the structures of the promoter sequences that they recognize
6. Give a detailed description of how the Escherichia coli transcription initiation complex is assembled, and discuss the various ways in which this process can be regulated
7. Give a detailed description of the assembly of the RNA polymerase II transcription initiation complex, and explain how assembly of this complex is influenced by proteins that activate or silence gene expression
8. Outline the processes of transcription initiation by eukaryotic RNA polymerases I and III

ONCE UPON A TIME the first stage in genome expression was described as 'transcription' or 'DNA makes RNA' (see Figure 3.2A), but we now realize that the process that leads from the genome to the transcriptome is much more complex than simply the synthesis of RNA. This part of genome expression is now divided into two key stages (Figure 9.1):

1. Initiation of transcription, which results in the complex of proteins, including the RNA polymerase enzyme and its various accessory proteins, that will subsequently copy the gene into an RNA transcript being assembled upstream of the gene. Inherent in this step are the events that determine whether or not the gene is actually expressed.
2. Synthesis and processing of RNA, which begins when the RNA polymerase leaves the initiation region and starts to make an RNA copy of the gene, and ends after completion of the processing and modification events that convert the initial transcript into a functional RNA.

This chapter deals with the initiation of transcription, and Chapter 10 covers RNA synthesis and processing. But before we move on to these topics we must do a little groundwork. The central players in many areas of molecular biology, including transcription, are DNA-binding proteins that attach to the genome in order to perform their biochemical functions. Histones are examples of DNA-binding proteins, and we will encounter many others later in this chapter when we look at assembly of the initiation complexes of prokaryotes and eukaryotes. There are also DNA-binding proteins that are involved in DNA replication, repair, and recombination, as well as a large group of related proteins that bind to RNA rather than DNA (Table 9.1). Many DNA-binding proteins recognize specific nucleotide sequences and bind predominantly to these target sites, whereas others bind non-specifically at various positions in the genome.
The mode of action of DNA-binding proteins is central to the initiation of transcription, and without a knowledge of how they function we can never hope to understand how the information in the genome is utilized. We will therefore spend some time examining what is known about DNA-binding proteins and how they interact with the genome.

![Figure 9.1. The two stages in the process that leads from genome to transcriptome](image)

**Table 9.1. Functions of DNA- and RNA-binding proteins**

<table>
<thead>
<tr>
<th>Function</th>
<th>Examples</th>
<th>Cross-reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA-binding proteins</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Genome expression</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Transcription initiation</td>
<td>Eukaryotic TATA-binding protein</td>
<td>Section 9.2.3</td>
</tr>
<tr>
<td>• RNA synthesis</td>
<td>RNA polymerases</td>
<td>Section 9.2.1</td>
</tr>
<tr>
<td>• Regulation of transcription</td>
<td>Eukaryotic activators and repressors</td>
<td>Section 9.3.2</td>
</tr>
<tr>
<td>DNA packaging</td>
<td>Eukaryotic histones</td>
<td>Section 2.2.1</td>
</tr>
<tr>
<td>DNA recombination</td>
<td>RecA</td>
<td>Section 14.3.1</td>
</tr>
<tr>
<td>DNA repair</td>
<td>DNA glycosylases, nucleases</td>
<td>Section 14.2.2</td>
</tr>
<tr>
<td>DNA replication</td>
<td>Origin recognition proteins</td>
<td>Section 13.2.1</td>
</tr>
<tr>
<td></td>
<td>DNA polymerases and ligases</td>
<td>Sections 4.1.1, 4.1.3 and 13.2.2</td>
</tr>
<tr>
<td></td>
<td>Single-strand binding proteins</td>
<td>Section 13.2.2</td>
</tr>
<tr>
<td></td>
<td>DNA topoisomerases</td>
<td>Section 13.1.2</td>
</tr>
<tr>
<td>Others</td>
<td>Prokaryotic restriction endonucleases</td>
<td>Section 4.1.2</td>
</tr>
<tr>
<td>RNA-binding proteins</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Genome expression</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intron splicing</td>
<td>snRNP proteins</td>
<td>Section 10.1.3</td>
</tr>
<tr>
<td>mRNA polyadenylation</td>
<td>CPSF, CstF</td>
<td>Section 10.1.2</td>
</tr>
<tr>
<td>mRNA editing</td>
<td>Adenosine deaminases</td>
<td>Section 10.2.2</td>
</tr>
<tr>
<td>RNA and RNA processing</td>
<td>Ribonucleases</td>
<td>Section 10.2.2</td>
</tr>
<tr>
<td>Translation</td>
<td>Aminocycl-RNA synthetases</td>
<td>Section 11.1.1</td>
</tr>
<tr>
<td></td>
<td>Translation factors</td>
<td>Sections 11.2.2 and 11.2.3</td>
</tr>
<tr>
<td>RNA degradation</td>
<td>Ribonucleases</td>
<td>Section 10.4</td>
</tr>
<tr>
<td>Ribosome structure</td>
<td>Ribosomal proteins</td>
<td>Section 11.2.1</td>
</tr>
</tbody>
</table>
9.1. The Importance of DNA-binding Proteins

As in all areas of molecular biology and genetics, the amount we know about a topic depends on the range and effectiveness of the methods available for its study. With regard to DNA-binding proteins we are fortunate in having a number of powerful techniques that can provide information on the interaction between a protein and the DNA sequence or sequences that it binds to. These techniques can be divided into three categories:

1. Methods for identifying the region(s) of a DNA molecule to which a protein binds;
2. Methods for purifying a DNA-binding protein;
3. Methods for studying the tertiary structure of a DNA-binding protein, including the complex formed when the protein is bound to DNA.

9.1.1. Locating the positions of DNA-binding sites in a genome

Often the first thing that is discovered about a DNA-binding protein is not the identity of the protein itself but the features of the DNA sequence that the protein recognizes. This is because genetic and molecular biology experiments, which we will deal with later in this chapter, have shown that many of the proteins that are involved in genome expression bind to short DNA sequences immediately upstream of the genes on which they act (Figure 9.2). This means that the sequence of a newly discovered gene, assuming that it includes both the coding DNA and the regions upstream of it, provides immediate access to the binding sites of at least some of the proteins responsible for expression of that gene. Because of this, a number of methods have been developed for locating protein binding sites within DNA fragments up to several kb in length, these methods working perfectly well even if the relevant DNA-binding proteins have not been identified.

Gel retardation identifies DNA fragments that bind to proteins

The first of these methods makes use of the substantial difference between the electrophoretic properties of a 'naked' DNA fragment and one that carries a bound protein. Recall that DNA fragments are separated by agarose gel electrophoresis because smaller fragments migrate through the pore-like structure of the gel more quickly than do larger fragments (see Technical Note 2.1). If a DNA fragment has a protein bound to it then its mobility through the gel will be impeded: the DNA-protein complex therefore forms a band at a position nearer to the starting point (Figure 9.3). This is called gel retardation (Garner and Revzin, 1981). In practice the technique is carried out with a collection of restriction fragments that span the region thought to contain a protein binding site. The digest is mixed with an extract of nuclear proteins (assuming that a eukaryote is being studied) and retarded fragments are identified by comparing the banding pattern obtained after electrophoresis with the pattern for restricted fragments that have not been mixed with proteins. A nuclear extract is used because at this stage of the project the DNA-binding protein has not usually been purified. If, however, the protein is available then the experiment can be carried out just as easily with the pure protein as with a mixed extract.
Gel retardation analysis. A nuclear extract has been mixed with a DNA restriction digest and a DNA-binding protein in the extract has attached to one of the restriction fragments. The DNA-protein complex has a larger molecular mass than the 'naked' DNA and so runs more slowly during gel electrophoresis. As a result, the band for this fragment is retarded and can be recognized by comparing with the banding pattern produced by restriction fragments that have not been mixed with the nuclear extract.

Protection assays pinpoint binding sites with greater accuracy

Gel retardation gives a general indication of the location of a protein binding site in a DNA sequence, but does not pinpoint the site with great accuracy. Often the retarded fragment is several hundred bp in length, compared with the expected length of the binding site of a few tens of bp at most, and there is no indication of where in the retarded fragment the binding site lies. Also, if the retarded fragment is long then it might contain separate binding sites for several proteins, or if it is quite small then there is the possibility that the binding site also includes nucleotides on adjacent fragments, ones that on their own do not form a stable complex with the protein and so do not lead to gel retardation. Retardation studies are therefore a starting point but other techniques are needed to provide more accurate information.

Modification protection assays can take over where gel retardation leaves off. The basis of these techniques is that if a DNA molecule carries a bound protein then part of its nucleotide sequence will be protected from modification. There are two ways of carrying out the modification:

- By treatment with a nuclease, which cleaves all phosphodiester bonds except those protected by the bound protein;
- By exposure to a methylating agent, such as dimethyl sulfate which adds methyl groups to G nucleotides. Any Gs protected by the bound protein will not be methylated.

The practical details of these two techniques are shown in Figures 9.4 and 9.5.

Both utilize an experimental approach called footprinting. In nuclease footprinting (Galas and Schmitz, 1978), the DNA fragment being examined is labeled at one end, complexed with binding protein (as a nuclear extract or as pure protein), and treated with deoxyribonuclease I (DNase I). Normally, DNase I cleaves every phosphodiester bond, leaving only the DNA segment protected by the binding protein. This is not very useful because it can be difficult to sequence such a small fragment. It is quicker to use the more subtle approach shown in Figure 9.4. The nuclease treatment is carried out under limiting conditions, such as a low temperature and/or very little enzyme, so that on average each copy of the DNA fragment suffers a single 'hit' - meaning that it is cleaved at just one position along its length. Although each fragment is cut just once, in the entire population of fragments all bonds are cleaved except those protected by the bound protein. The protein is now removed, the mixture electrophoresed, and the labeled fragments visualized. Each of these fragments has the label at one end and a cleavage site at the other. The result is a ladder of bands corresponding to fragments that differ in length.
by one nucleotide, the ladder broken by a blank area in which no labeled bands occur. This blank area, or ‘footprint’, corresponds to the positions of the protected phosphodiester bonds, and hence of the bound protein, in the starting DNA.

Figure 9.4. DNase I footprinting. The technique is described in the text. The restriction fragments used at the start of the procedure must be labeled at just one end. This is usually achieved by treating a set of longer restriction fragments with an enzyme that attaches labels at both ends, then cutting these labeled molecules with a second restriction enzyme, and purifying one of the sets of end fragments. The DNase I treatment is carried out in the presence of a manganese salt, which induces the enzyme to make random double-stranded cuts in the target molecules, leaving blunt-ended fragments.

Modification interference identifies nucleotides central to protein binding
Modification protection should not be confused with modification interference, a different technique with greater sensitivity in the study of protein binding (Hendrickson and Schleif, 1985). Modification interference works on the basis that if a nucleotide critical for protein binding is altered, for example by addition of a methyl group, then binding may be prevented. One of this family of techniques is illustrated in Figure 9.6. The DNA fragment, labeled at one end, is treated with the modification reagent, in this case dimethyl sulfate, under limiting conditions so that just one guanine per fragment is methylated. Now the binding protein or nuclear extract is added, and the fragments electrophoresed. Two bands are seen, one corresponding to the DNA-protein complex and one containing DNA without bound protein. The latter contains molecules that have been prevented from attaching to the protein because the methylation treatment has modified one or more Gs that are crucial for the binding. To identify which Gs are modified, the fragment is purified from the gel and treated with piperidine, a compound that cleaves DNA at methylguanine nucleotides. The result of this treatment is that each fragment is cut into two segments, one of which carries the label. The length(s) of the labeled fragment(s), determined by a second round of electrophoresis, tells us which nucleotide(s) in the original fragment were methylated and hence identifies the position in the DNA sequence of Gs that participate in the binding reaction. Equivalent techniques can be used to identify the A, C and T nucleotides involved in binding.
Figure 9.5. The dimethyl sulfate (DMS) modification protection assay. The technique is similar to DNase I footprinting (see Figure 9.4). Instead of DNase I digestion, the fragments are treated with limited amounts of DMS so that a single guanine base is methylated in each fragment. Guanines that are protected by the bound protein cannot be modified. After removal of the protein the DNA is treated with piperidine, which cuts at the modified nucleotide positions. For simplicity, the diagram shows the double-stranded molecules being cleaved at this stage. In fact, they are only nicked, as piperidine only cuts the strand that is modified, rather than making a double-stranded cut across the entire molecule. The samples are therefore examined by 

Figure 9.5. The dimethyl sulfate (DMS) modification protection assay. The technique is similar to DNase I footprinting (see Figure 9.4). Instead of DNase I digestion, the fragments are treated with limited amounts of DMS so that a single guanine base is methylated in each fragment. Guanines that are protected by the bound protein cannot be modified. After removal of the protein the DNA is treated with piperidine, which cuts at the modified nucleotide positions. For simplicity, the diagram shows the double-stranded molecules being cleaved at this stage. In fact, they are only nicked, as piperidine only cuts the strand that is modified, rather than making a double-stranded cut across the entire molecule. The samples are therefore examined by denaturing gel electrophoresis so that the two strands are separated. The resulting autoradiograph shows the sizes of the strands that have one labeled end and one end created by piperidine nicking. The banding pattern for the control DNA strands - those not incubated with the nuclear extract - indicates the positions of Gs in the restriction fragment, and the footprint seen in the banding pattern for the test sample shows which Gs were protected.
Figure 9.6. Dimethyl sulfate (DMS) modification interference assay. The method is described in the text. See the legend to Figure 9.4 for a description of the procedure used to obtain DNA fragments labeled at just one end.

9.1.2. Purifying a DNA-binding protein

Once a binding site has been identified in a DNA molecule, this sequence can be used to purify the DNA-binding protein, as a prelude to more detailed structural studies. The purification techniques utilize the ability of the protein to bind to its target site. One possibility is to use a form of affinity chromatography (Figure 9.7A). A DNA fragment or synthetic oligonucleotide that contains a protein binding site is immobilized in a chromatography column, usually by attaching one end of the DNA to a silica particle (Kadonaga, 1991). The protein extract is then passed through the column in a low-salt buffer, which promotes binding of proteins to their target sites. The binding protein specific for the immobilized sequence is retained in the column while all other proteins pass through. Once these unwanted proteins have been completely
washed out, the column is eluted with a high-salt buffer, which destabilizes the DNA-protein complex. The pure binding protein can then be collected.

An alternative is to screen a cloning library (Singh et al., 1988). A library of cDNA clones, each synthesizing a different cloned protein from the organism being studied, is needed. These clones are blotted onto a nylon membrane in such a way that the protein content of each clone is retained (Figure 9.7B). The DNA fragment or oligonucleotide containing the protein binding site is labeled, and washed over the membrane. The DNA attaches to a blotted clone only if that clone has been synthesizing the appropriate DNA-binding protein. These clones are identified by detecting where the labeled DNA is located on the membrane. Samples of the clones can then be recovered from the master library and used to produce larger quantities of the binding protein.

Figure 9.7. Two ways of purifying a DNA-binding protein. (A) Affinity chromatography. DNA fragments or synthetic oligonucleotides containing the attachment site for the binding protein are attached to silica beads and these packed into a chromatography column. The binding protein is retained in the column when a nuclear extract is passed through, and can subsequently be recovered by eluting with a high-salt buffer. (B) Screening a clone library. cDNA clones are blotted onto a nylon filter and one that is expressing the binding protein identified through the ability of this colony to bind labeled DNA containing the protein attachment site (see Technical Note 4.3).
9.1.3. Studying the structures of proteins and DNA-protein complexes

The availability of a pure sample of a DNA-binding protein makes possible the analysis of its structure, in isolation or attached to its DNA-binding site. This provides the most detailed information on the DNA-protein interaction, enabling the precise structure of the DNA-binding part of the protein to be determined, and allowing the identity and nature of the contacts with the DNA helix to be elucidated. Two techniques - X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy - are central to this area of research.

X-ray crystallography has broad applications in structure determination

X-ray crystallography is a long-established technique whose pedigree stretches back to the late 19th century. Indeed, Nobel prizes were awarded as early as 1915 to William and Lawrence Bragg, father and son, for working out the basic methodology and using it to determine the crystal structures of salts such as sodium chloride and zinc sulfide. The technique is based on X-ray diffraction. X-rays have very short wavelengths - between 0.01 and 10 nm - which is 4000 times shorter than visible light and comparable with the spacings between atoms in chemical structures. When a beam of X-rays is directed onto a crystal, some of the X-rays pass straight through, but others are diffracted and emerge from the crystal at a different angle from which they entered (Figure 9.8A). If the crystal is comprised of many copies of the same molecule, all positioned in a regular array, then different X-rays are diffracted in similar ways, resulting in overlapping circles of diffracted waves which interfere with one another. An X-ray-sensitive photographic film or electronic detector placed across the beam reveals a series of spots (Figure 9.8B), an X-ray diffraction pattern, from which the structure of the molecule in the crystal can be deduced.

The challenge with X-ray crystallography lies with the complexity of the methodology used to deduce the structure of a molecule from its diffraction pattern. The basic principles are that the relative positioning of the spots indicates the arrangement of the molecules in the crystal, and their relative intensities provide information on the structure of the molecule. The problem is that the more complex the molecule, the greater the number of spots and the larger the number of comparisons that must be made between them. Even with computational help the analysis is difficult and time consuming. If successful, the result is an electron density map (Figure 9.8C and D) which, with a protein, provides a chart of the folded polypeptide from which the positioning of structural features such as α-helices and β-sheets can be determined. If sufficiently detailed, the R groups of the individual amino acids in the polypeptide can be identified and their orientations relative to one another established, allowing deductions to be made about the hydrogen bonding and other chemical interactions occurring within the protein structure. With luck, these deductions lead to a detailed three-dimensional model of the protein (Rhodes, 1999).

Permission to reproduce this figure in this web version of Genomes 2 is either pending or has not been granted.

Figure 9.8. X-ray crystallography. (A) An X-ray diffraction pattern is obtained by passing a beam of X-rays through a crystal of the molecule being studied. (B) The diffraction pattern obtained with crystals of ribonuclease. (C) Part of the electron-density map derived from this diffraction pattern. (D) If the electron-density map has sufficient resolution then it is possible to identify the R groups of individual amino acids, as shown here for tyrosine. (B) and (C) from Biology, 3rd edition, by Neil Campbell, copyright 1993 by Benjamin Cummings Publishing Company. Reprinted by permission. (D) reprinted with permission from Zubay G, Biochemistry, 4th edition. Copyright 1997 by The McGraw-Hill Companies.
The first protein structures to be determined by X-ray crystallography were for myoglobin and hemoglobin, resulting in further Nobel prizes, for Perutz and Kendrew in 1962. It still takes several months or longer to complete an X-ray crystallography analysis with a new protein, and there are many pitfalls that can prevent a successful conclusion being reached. In particular, it can often be difficult to obtain a suitable crystal of the protein. Despite these problems, the number of completed structures has gradually increased and now includes more than 50 DNA-binding proteins. An important innovation has been to crystallize DNA-binding proteins in the presence of their target sequences, the resulting protein-DNA structures revealing the precise positioning of the proteins relative to the double helix. It is from this type of information that most of our knowledge about the mode of action of DNA-binding proteins has been obtained.

**NMR gives detailed structural information for small proteins**

Like X-ray crystallography, NMR traces its origins to the early part of the 20th century, first being described in 1936 with the relevant Nobel prizes awarded in 1952. The principle of the technique is that rotation of a charged chemical nucleus generates a magnetic moment. When placed in an applied electromagnetic field, the spinning nucleus orientates in one of two ways, called $\alpha$ and $\beta$ (Figure 9.9), the $\alpha$-orientation (which is aligned with the magnetic field) having a slightly lower energy. In NMR spectroscopy the magnitude of this energy separation is determined by measuring the frequency of the electromagnetic radiation needed to induce the transition from $\alpha$ to $\beta$, the value being described as the resonance frequency of the nucleus being studied. The critical point is that although each type of nucleus (e.g. $^1$H, $^{13}$C, $^{15}$N) has its own specific resonance frequency, the measured frequency is often slightly different from the standard value (typically by less than 10 parts per million) because electrons in the vicinity of the rotating nucleus shield it to a certain extent from the applied magnetic field. This chemical shift (the difference between the observed resonance energy and the standard value for the nucleus being studied) enables the chemical environment of the nucleus to be inferred, and hence provides structural information. Particular types of analysis (called COSY and TOCSY) enable atoms linked by chemical bonds to the spinning nucleus to be identified; other analyses (e.g. NOESY) identify atoms that are close to the spinning nucleus in space but not directly connected to it.

![Figure 9.10. The helix-turn-helix motif](image)

The drawing shows the orientation of the helix-turn-helix motif (in blue) of the *Escherichia coli* bacteriophage 434 repressor in the major groove of the DNA double helix. ‘N’ and ‘C’ indicate the N- and C-termini of the motif, respectively. Reprinted from *DNA-Protein Interactions* by Andrew Travers, published by Chapman & Hall, 1993. Reprinted with kind permission of A. Travers.

Not all chemical nuclei are suitable for NMR. Most protein NMR projects are $^1$H studies, the aim being to identify the chemical environments and covalent linkages of every hydrogen atom, and from this information to infer the overall structure of the protein. These studies are frequently supplemented by analyses of substituted proteins in which at least some of the carbon and/or nitrogen atoms have been replaced with the rare isotopes $^{13}$C and $^{15}$N, these also giving good results with NMR.
When successful, NMR results in the same level of resolution as X-ray crystallography and so provides very detailed information on protein structure (Evans, 1995). The main advantage of NMR is that it works with molecules in solution and so avoids the problems that sometimes occur when attempting to obtain crystals of a protein for X-ray analysis. Solution studies also offer greater flexibility if the aim is to examine changes in protein structure, for example during protein folding or in response to addition of a substrate. The disadvantage of NMR is that it is only suitable for relatively small proteins. There are several reasons for this, one being the need to identify the resonance frequencies for each, or as many as possible, of the $^1$H or other nuclei being studied. This depends on the various nuclei having different chemical shifts so that their frequencies do not overlap. The larger the protein, the greater the number of nuclei and the greater the chances that frequencies overlap and structural information is lost. Although this limits the applicability of NMR, the technique is still very valuable. There are many interesting proteins that are small enough to be studied by NMR, and important information can also be obtained by structural analysis of peptides which, although not complete proteins, can act as models for aspects of protein activity such as nucleic acid binding.

9.1.4. The special features of DNA-binding proteins

Now that we have examined the methods used to study DNA-binding proteins, we can turn our attention to the proteins themselves. Our main interest lies with those proteins that are able to target a specific nucleotide sequence and hence bind to a limited number of positions on a DNA molecule, this being the type of interaction that is most important in expression of the genome. To bind in this specific fashion a protein must make contact with the double helix in such a way that the nucleotide sequence can be recognized, which generally requires that part of the protein penetrates into the major and/or minor grooves of the helix (see Figures 1.11A and 1.12) in order to achieve direct readout of the sequence (Section 9.1.5). This is usually accompanied by more general interactions with the surface of the molecule, which may simply stabilize the DNA-protein complex or which may be aimed at accessing indirect information on nucleotide sequence that is provided by the conformation of the helix.

Figure 9.11. The homeodomain motif. The first three helices of a typical homeodomain are shown with helix 3 orientated in the major groove and helix 1 making contacts in the minor groove. Helices 1–3 run in the N→C direction along the motif. Reprinted from DNA-Protein Interactions by Andrew Travers, published by Chapman & Hall, 1993. Reprinted with kind permission of A. Travers.

When the structures of sequence-specific DNA-binding proteins are compared, it is immediately evident that the family as a whole can be divided into a limited number of different groups on the basis of the structure of the segment of the protein that interacts with the DNA molecule (Table 9.2; Luisi, 1995). Each of these DNA-binding motifs is present in a range of proteins, often from very different organisms, and at least some of them probably evolved more than once. We will look at two in detail - the helix-turn-helix (HTH) motif and the zinc finger - and then briefly survey the others.
Figure 9.12. The Cys$_2$His$_2$ zinc finger. This particular zinc finger is from the yeast SWI5 protein. The zinc atom is held between two cysteines within the β-sheet of the motif and two histidines in the α-helix. The solid green lines indicate the R groups of these amino acids. ‘N’ and ‘C’ indicate the N- and C-termini of the motif, respectively. Reprinted from DNA-Protein Interactions by Andrew Travers, published by Chapman & Hall, 1993. Reprinted with kind permission of A. Travers.

<table>
<thead>
<tr>
<th>Motif</th>
<th>Examples of proteins with this motif</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sequence-specific DNA-binding motifs</strong></td>
<td></td>
</tr>
<tr>
<td>Helix-turn-helix family</td>
<td></td>
</tr>
<tr>
<td>Standard helix-turn-helix</td>
<td>Escherichia coli lactose repressor, tryptophan repressor</td>
</tr>
<tr>
<td>Homeodomain</td>
<td>Drosophila Antennapedia protein</td>
</tr>
<tr>
<td>Paired homeodomain</td>
<td>Vertebrate Pax transcription factors</td>
</tr>
<tr>
<td>POU domain</td>
<td>Vertebrate regulatory proteins PIT-1, OCT-1 and OCT-2</td>
</tr>
<tr>
<td>Winged helix-turn-helix</td>
<td>GABP regulatory protein of higher eukaryotes</td>
</tr>
<tr>
<td>High mobility group (HMG) domain</td>
<td>Mammalian sex determination protein SRY</td>
</tr>
<tr>
<td><strong>Zinc-finger family</strong></td>
<td></td>
</tr>
<tr>
<td>Cys$_2$His$_2$ finger</td>
<td>Transcription factor TFIIIA of eukaryotes</td>
</tr>
<tr>
<td>Multi-cysteine zinc finger</td>
<td>Steroid receptor family of higher eukaryotes</td>
</tr>
<tr>
<td>Zinc binuclear cluster</td>
<td>Yeast GAL4 transcription factor</td>
</tr>
<tr>
<td>Basic domain</td>
<td>Yeast GCN4 transcription factor</td>
</tr>
<tr>
<td>Ribon helix-helix</td>
<td>Bacterial MetU, Arc and Mnt repressors</td>
</tr>
<tr>
<td>TBP domain</td>
<td>Eukaryotic TATA-binding protein</td>
</tr>
<tr>
<td>β-Barrel dimer</td>
<td>Papilomavirus E2 protein</td>
</tr>
<tr>
<td>Rel homology domain (RHB)</td>
<td>Mammalian transcription factor NF-κB</td>
</tr>
<tr>
<td><strong>Non-specific DNA-binding motifs</strong></td>
<td></td>
</tr>
<tr>
<td>Histone fold</td>
<td>Eukaryotic histones</td>
</tr>
<tr>
<td>HU/IHF motif*</td>
<td>Bacterial HU and IHF proteins</td>
</tr>
<tr>
<td>Polymerase cleft</td>
<td>DNA and RNA polymerases</td>
</tr>
</tbody>
</table>

*The HU/IHF motif is a non-specific DNA-binding motif in bacterial HU proteins (the nucleoid packaging proteins, Section 2.3.1) but directs sequence-specific binding of the IHF (integration host factor) protein.

**The helix-turn-helix motif is present in prokaryotic and eukaryotic proteins**

The HTH motif was the first DNA-binding structure to be identified (Harrison and Aggarwal, 1990). As the name suggests, the motif is made up of two α-helices separated by a turn (Figure 9.10). The latter is not a random conformation but a specific structure, referred to as a β-turn,
Many prokaryotic and eukaryotic DNA-binding proteins utilize an HTH motif. In bacteria, HTH motifs are present in some of the best studied regulatory proteins, which switch on and off the expression of individual genes. An example is the lactose repressor, which regulates expression of the lactose operon (Sections 2.3.2 and 9.3.1). The various eukaryotic HTH proteins include many whose DNA-binding properties are important in the developmental regulation of genome expression, such as the homeodomain proteins, whose roles we will examine in Section 12.3.3. The homeodomain is an extended HTH motif possessed by each of these proteins. It is made up of 60 amino acids which form four α-helices, numbers 2 and 3 separated by a β-turn, with number 3 acting as the recognition helix and number 1 making contacts within the minor groove (Figure 9.11). Other versions of the HTH motif found in eukaryotes include:

- The POU domain, which is usually found in proteins that also have a homeodomain, the two motifs probably working together by binding different regions of a double helix. The name 'POU' comes from the initial letters of the names of the first proteins found to contain this motif (Herr et al., 1988).
- The winged helix-turn-helix motif, which is another extended version of the basic HTH structure, this one with a third α-helix on one side of the HTH motif and a β-sheet on the other side.

Many proteins, prokaryotic and eukaryotic, possess an HTH motif, but the details of the interaction of the recognition helix with the major groove are not exactly the same in all cases. The length of the recognition helix varies, generally being longer in eukaryotic proteins, the orientation of the helix in the major groove is not always the same, and the position within the recognition helix of those amino acids that make contacts with nucleotides is different.

**Zinc fingers are common in eukaryotic proteins**

The second type of DNA-binding motif that we will look at in detail is the zinc finger, which is rare in prokaryotic proteins but very common in eukaryotes (Mackay and Crossley, 1998). There appear to be more than 500 different zinc-finger proteins in the worm Caenorhabditis elegans, out of a total 19 000 proteins (Clarke and Berg, 1998), and it is estimated that 1% of all mammalian genes code for zinc-finger proteins.

There are at least six different versions of the zinc finger. The first to be studied in detail was the Cys2His2 finger, which comprises a series of 12 or so amino acids, including two cysteines and two histidines, which form a segment of β-sheet followed by an α-helix. These two structures, which form the ‘finger’ projecting from the surface of the protein, hold between them a bound zinc atom, coordinated to the two cysteines and two histidines (Figure 9.12). The α-helix is the part of the motif that makes the critical contacts within the major groove, its positioning within the groove being determined by the β-sheet, which interacts with the sugar-phosphate backbone of the DNA, and the zinc atom, which holds the sheet and helix in the appropriate positions relative to one another. Other versions of the zinc finger differ in the structure of the finger, some lacking the sheet component and consisting simply of one or more α-helices, and the precise way in which the zinc atom is held in place also varies. For example, the multicysteine zinc fingers lack histidines, the zinc atom being coordinated between four cysteines.
An interesting feature of the zinc finger is that multiple copies of the finger are sometimes found on a single protein. Several have two, three or four fingers, but there are examples with many more than this - 37 for one toad protein. In most cases, the individual zinc fingers are thought to make independent contacts with the DNA molecule, but in some cases the relationship between different fingers is more complex. In one particular group of proteins - the nuclear or steroid receptor family - two α-helices containing six cysteines combine to coordinate two zinc atoms in a single DNA-binding domain, larger than a standard zinc finger (Figure 9.13). Within this motif it appears that one of the α-helices enters the major groove whereas the second makes contacts with other proteins.

Figure 9.13. The steroid receptor zinc finger. The R groups of the amino acids involved in the interactions with the zinc atoms are shown as solid green lines. 'N' and 'C' indicate the N- and C-termini of the motif, respectively. Reprinted from DNA-Protein Interactions by Andrew Travers, published by Chapman & Hall, 1993. Reprinted with kind permission of A. Travers.

Other DNA-binding motifs
The various other DNA-binding motifs that have been discovered in different proteins include: The basic domain, in which the DNA recognition structure is an α-helix that contains a high number of basic amino acids (e.g. arginine, serine and threonine). A peculiarity of this motif is that the α-helix only forms when the protein interacts with DNA: in the unbound state the helix has a disorganized structure. Basic domains are found in a number of eukaryotic proteins involved in transcription of DNA into RNA.

The ribbon-helix-helix motif, which is one of the few motifs that achieves sequence-specific DNA binding without making use of an α-helix as the recognition structure. Instead, the ribbon (i.e. two strands of a β-sheet) makes contact with the major groove (Figure 9.14). Ribbon-helix-helix motifs are found in some gene-regulatory proteins in bacteria.

The TBP domain has so far only been discovered in the TATA-binding protein (Section 9.2.3), after which it is named (Kim et al., 1993). As with the ribbon-helix-helix motif, the recognition structure is a β-sheet, but in this case the main contacts are with the minor, not major, groove of the DNA molecule.

9.1.5. The interaction between DNA and its binding proteins
In recent years our understanding of the part played by the DNA molecule in the interaction with a binding protein has begun to change. It has always been accepted that proteins that recognize a specific sequence as their binding site can locate this site by forming contacts with chemical groups attached to the nitrogenous bases that are exposed within the major and minor grooves that spiral around the double helix (see Figure 1.11A). It is now recognized that the nucleotide sequence also influences the precise conformation of each region of the helix, and that these conformational features represent a second, less direct way in which the DNA sequence can influence protein binding.
Direct readout of the nucleotide sequence

It was clear from the double helix structure described by Watson and Crick (Section 1.1.3) that although the nucleotide bases are on the inside of the DNA molecule, they are not entirely buried, and some of the chemical groups attached to the purine and pyrimidine bases are accessible from outside the helix. Direct readout of the nucleotide sequence should therefore be possible without breaking the base pairs and opening up the molecule.

In order to form chemical bonds with groups attached to the nucleotide bases, a binding protein must make contacts within one or both of the grooves on the surface of the helix. With the B-form of DNA, the identity and orientation of the exposed parts of the bases within the major groove is such that most sequences can be read unambiguously, whereas within the minor groove it is possible to identify if each base pair is A-T or G-C but difficult to know which nucleotide of the pair is in which strand of the helix. Direct readout of the B-form therefore predominantly involves contacts in the major groove. With other DNA types there is much less information on the contacts formed with binding proteins, but the picture is likely to be quite different. In the A-form, for example, the major groove is deep and narrow and less easily penetrated by any part of a protein molecule (see Table 1.1). The shallower minor groove is therefore likely to play the main part in direct readout. With Z-DNA, the major groove is virtually non-existent and direct readout is possible to a certain extent without moving beyond the surface of the helix.

The nucleotide sequence has a number of indirect effects on helix structure

The recent change in our view of DNA structure concerns the influence of the nucleotide sequence on the conformation of the helix at different positions along its length. Originally it was thought that cellular DNA molecules have fairly uniform structures, made up mainly of the B-form of the double helix. Some short segments might be in the A-form, and there might be some Z-DNA tracts, especially near the ends of a molecule, but the majority of the length of a double helix would be unvarying B-DNA. We now recognize that DNA is much more polymorphic, and that it is possible for the A-, B- and Z-DNA configurations, and intermediates between them, to coexist within a single DNA molecule, different parts of the molecule having different structures. These conformational variations are sequence dependent, being largely
the result of the base-stacking interactions that occur between adjacent base pairs. As well as being responsible, along with base-pairing, for the stability of helix, the base-stacking also influences the amount of rotation that occurs around the covalent bonds within individual nucleotides and hence determines the conformation of the helix at a particular position. The rotational possibilities in one base pair are influenced, via the base-stacking interactions, by the identities of the neighboring base pairs. This means that the nucleotide sequence indirectly affects the overall conformation of the helix, possibly providing structural information that a binding protein can use to help it locate its appropriate attachment site on a DNA molecule. At present this is just a theoretical possibility as no protein that specifically recognizes a non-B form of the helix has been identified, but many researchers believe that helix conformation is likely to play some role in the interaction between DNA and protein.

Figure 9.15. Recognition of an A-T base pair in the B-form double helix. An A-T base pair is shown in outline (see Figure 1.11B), with arrows indicating the chemical features that can be recognized by accessing the base pair via the major groove (above) and minor groove (below). In the major groove the chemical features are asymmetric and the orientation of the A-T pair can be identified by a binding protein. For some time it was believed that this is not possible in the minor groove, because only the two features shown in black were thought to be present, and these are symmetric. Using these two features, the binding protein could recognize the A-T base pair but not know which nucleotide is in which strand of the helix. The asymmetric features shown in green have recently been discovered, suggesting that the orientation of the pair might in fact be discernible via the minor groove. Abbreviations: a, hydrogen bond acceptor; d, hydrogen bond donor; vDW, van der Waals interaction (see Box 3.3). Reprinted with permission from Kielkopf CL, et al., Science, 282, 111–115. Copyright 1998 American Association for the Advancement of Science.

Table 1.1. Features of different conformations of the DNA double helix

<table>
<thead>
<tr>
<th>Feature</th>
<th>B-DNA</th>
<th>Conformation A-DNA</th>
<th>Z-DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of helix</td>
<td>Right-handed</td>
<td>Right-handed</td>
<td>Left-handed</td>
</tr>
<tr>
<td>Helical diameter (nm)</td>
<td>2.37</td>
<td>2.55</td>
<td>1.84</td>
</tr>
<tr>
<td>Rise per base pair (nm)</td>
<td>0.34</td>
<td>0.29</td>
<td>0.37</td>
</tr>
<tr>
<td>Distance per complete turn (pitch) (nm)</td>
<td>3.4</td>
<td>3.2</td>
<td>4.5</td>
</tr>
<tr>
<td>Number of base pairs per complete turn</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>Topology of major groove</td>
<td>Wide, deep</td>
<td>Narrow, deep</td>
<td>Flat</td>
</tr>
<tr>
<td>Topology of minor groove</td>
<td>Narrow, shallow</td>
<td>Broad, shallow</td>
<td>Narrow, deep</td>
</tr>
</tbody>
</table>
A second type of conformational change is DNA bending (Travers, 1995). This does not refer to the natural flexibility of DNA which enables it to form circles and supercoils, but instead to localized positions where the nucleotide sequence causes the DNA to bend. Like other conformational variations, DNA bending is sequence dependent. In particular, a DNA molecule in which one polynucleotide contains two or more groups of repeated adenines, each group comprising 3–5 As, with individual groups separated by 10 or 11 nucleotides, will bend at the 3’ end of the adenine-rich region (Young and Beveridge, 1998). As with helix conformation, it is not yet known to what extent DNA bending influences protein binding, although protein-induced bending at flexible sites has a clearly demonstrated function in the regulation of some genes (e.g. Falvo et al., 1995; Section 9.3.2).

Contacts between DNA and proteins
The contacts formed between DNA and its binding proteins are non-covalent. Within the major groove, hydrogen bonds form between the nucleotide bases and the R groups of amino acids in the recognition structure of the protein, whereas in the minor groove hydrophobic interactions are more important. On the surface of the helix, the major interactions are electrostatic, between the negative charges on the phosphate component of each nucleotide and the positive charges on the R groups of amino acids such as lysine and arginine, although some hydrogen bonding also occurs. In some cases, hydrogen bonding on the surface of the helix or in the major groove is direct between DNA and protein; in others it is mediated by water molecules. Few generalizations can be made: at this level of DNA-protein interaction each example has its own unique features and the details of the binding have to be worked out by structural studies rather than by comparisons with other proteins.

Most proteins that recognize specific sequences are also able to bind non-specifically to other parts of a DNA molecule. In fact it has been suggested that the amount of DNA in a cell is so large, and the numbers of each binding protein so small, that the proteins spend most, if not all, of their time attached non-specifically to DNA (Stormo and Fields, 1998). The distinction between the non-specific and specific forms of binding is that the latter is more favorable in thermodynamic terms. As a result, a protein is able to bind to its specific site even though there are literally millions of other sites to which it could attach non-specifically. To achieve this thermodynamic favorability, the specific binding process must involve the greatest possible number of DNA-protein contacts, which explains in part why the recognition structures of many DNA-binding motifs have evolved to fit snugly into the major groove of the helix, where the opportunity for DNA-protein contacts is greatest. It also explains why some DNA-protein interactions result in conformational changes to one or other partner, increasing still further the complementarity of the interacting surfaces, and allowing additional bonding to occur.

The need to maximize contacts in order to ensure specificity is also one of the reasons why many DNA-binding proteins are dimers, consisting of two proteins attached to one another. This is the case for most HTH proteins and many of the zinc-finger type. Dimerization occurs in such a way that the DNA-binding motifs of the two proteins are both able to access the helix, possibly with some degree of cooperativity between them, so that the resulting number of contacts is greater than twice the number achievable by a monomer. As well as their DNA-binding motifs, many proteins contain additional characteristic domains that participate in the protein-protein contacts that result in dimer formation. One of these is the leucine zipper, which is an α-helix that coils more tightly than normal and presents a series of leucines on one of its faces. These can form contacts with the leucines of the zipper on a second protein, forming the dimer (Figure 9.16). A second dimerization domain is, rather unfortunately, called the helix-loop-helix motif, which is distinct from, and should not be confused with, the helix-turn-helix DNA-binding motif.
9.2. DNA-Protein Interactions During Transcription Initiation

Now that we have established that DNA-protein interactions are the key to understanding the initiation of transcription, we can move on to begin our examination of the events involved in the assembly of the initiation complex. We will do this in two stages. First, we will study the DNA-protein interactions that are involved in transcription initiation. Then, in Section 9.3, we will investigate how assembly of the initiation complex, and its ability to initiate transcription, can be controlled by various additional proteins that respond to stimuli from inside or outside the cell and ensure that the correct genes are transcribed at the appropriate times.

9.2.1. RNA polymerases

In Section 3.2.2 we learnt that the enzymes responsible for transcription of DNA into RNA are called DNA-dependent RNA polymerases. Transcription of eukaryotic nuclear genes requires three different RNA polymerases: RNA polymerase I, RNA polymerase II and RNA polymerase III. Each is a multi-subunit protein (8–12 subunits) with a molecular mass in excess of 500 kDa. Structurally, these polymerases are quite similar to one another, the three largest subunits being closely related and some of the smaller ones being shared by more than one enzyme; functionally, however, they are quite distinct. Each works on a different set of genes, with no interchangeability (Table 9.3). Most research attention has been directed at RNA polymerase II, as this is the one that transcribes genes that code for proteins. It also works on a set of genes specifying the small nuclear RNAs that are involved in RNA processing. RNA polymerase III transcribes other genes for small RNAs, including those for transfer RNAs (tRNAs). RNA polymerase I transcribes the multicycopy repeat units containing the 28S, 5.8S and 18S rRNA genes. The functions of all these RNAs were summarized in Section 3.2.1 and are described in detail in Chapters 10 and 11.

Archaea possess a single RNA polymerase that is very similar to the eukaryotic enzymes (Bult et al., 1996). But this is not typical of the prokaryotes in general because the bacterial RNA polymerase is very different, consisting of just five subunits, described as $\alpha_2\beta\beta'$ (two $\alpha$...
subunits, one each of $\beta$ and the related $\beta'$, and one of $\sigma$). The $\alpha$, $\beta$, and $\beta'$ subunits are equivalent to the three largest subunits of the eukaryotic RNA polymerases, but the $\sigma$ subunit has its own special properties, both in terms of its structure and, as we will see in the next section, its function.

Table 9.3. Functions of the three eukaryotic nuclear RNA polymerases

<table>
<thead>
<tr>
<th>Polymerase</th>
<th>Genes transcribed*</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA polymerase I</td>
<td>28S, 5.8S and 18S ribosomal RNA (rRNA) genes</td>
</tr>
<tr>
<td>RNA polymerase II</td>
<td>Protein-coding genes; most small nuclear RNA (snRNA) genes</td>
</tr>
<tr>
<td>RNA polymerase III</td>
<td>Genes for transfer RNAs (tRNA), 5S rRNA, U6-snRNA, small nucleolar (sno) RNAs, small cytoplasmic (sc) RNA</td>
</tr>
</tbody>
</table>

*See Section 3.2.1 for descriptions of these RNA types.

9.2.2. Recognition sequences for transcription initiation

It is essential that transcription initiation complexes are constructed at the correct positions on DNA molecules. These positions are marked by target sequences that are recognized either by the RNA polymerase itself or by a DNA-binding protein which, once attached to the DNA, forms a platform to which the RNA polymerase binds (see Figure 3.6).

Figure 3.6. Two ways in which RNA polymerases bind to their promoters. (A) shows the direct recognition of the promoter by the RNA polymerase, as occurs in bacteria. (B) shows recognition of the promoter by a DNA-binding protein which forms a platform onto which the RNA polymerase binds. This indirect mechanism occurs with eukaryotic and archaeal RNA polymerases.

Bacterial RNA polymerases bind to promoter sequences

In bacteria, the target sequence for RNA polymerase attachment is called the promoter. This term was first used by geneticists in 1964 to describe the function of a locus immediately upstream of the three genes in the lactose operon (Figure 9.17). When this locus was inactivated by mutation, the genes in the operon were not expressed; the locus therefore appeared to promote expression of the genes. We now know that this is because the locus is the binding site for the RNA polymerase that transcribes the operon.

The sequences that make up the *Escherichia coli* promoter were first identified by comparing the regions upstream of over 100 genes. It was assumed that promoter sequences would be very similar for all genes and so should be recognizable when the upstream regions were compared.
Figure 9.17. The promoter for the lactose operon of *Escherichia coli*. The promoter is located immediately upstream of *lacZ*, the first gene in the operon. The DNA sequence shows the positions of the -35 and -10 boxes, the two distinct sequence components of the promoter. Compare these sequences with the consensus sequences described in the text. See also Table 9.4. For more information on the lactose operon see Figure 2.20A.

These analyses showed that the *E. coli* promoter consists of two segments, both of six nucleotides, described as follows (see Figure 9.17):

\[
\begin{align*}
-35 \text{ box} & \quad 5'-\text{TTGACA}-3' \\
-10 \text{ box} & \quad 5'-\text{TATAAT}-3'
\end{align*}
\]

These are consensus sequences and so describe the ‘average’ of all promoter sequences in *E. coli*; the actual sequences upstream of any particular gene might be slightly different (Table 9.4). The names of the boxes indicate their positions relative to the point at which transcription begins. The nucleotide at this point is labeled ‘+1’ and is anything between 20 and 600 nucleotides upstream of the start of the coding region of the gene. The spacing between the two boxes is important because it places the two motifs on the same face of the double helix, facilitating their interaction with the DNA-binding component of the RNA polymerase (Section 9.2.3).

Table 9.4. Sequences of *Escherichia coli* promoters

<table>
<thead>
<tr>
<th>Promoter</th>
<th>-35 box</th>
<th>-10 box</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consensus</td>
<td>5’-TTGACA-3’</td>
<td>5’-TATAAT-3’</td>
</tr>
<tr>
<td>Lactose operon</td>
<td>5’-TTTACA-3’</td>
<td>5’-TATGTT-3’</td>
</tr>
<tr>
<td>Tryptophan operon</td>
<td>5’-TTGACA-3’</td>
<td>5’-TTAAT-3’</td>
</tr>
</tbody>
</table>

**Eukaryotic promoters are more complex**

In eukaryotes, the term ‘promoter’ is used to describe all the sequences that are important in initiation of transcription of a gene. For some genes these sequences can be numerous and diverse in their functions, including not only the core promoter, sometimes called the basal promoter, which is the site at which the initiation complex is assembled, but also one or more upstream promoter elements which, as their name implies, lie upstream of the core promoter. Assembly of the initiation complex on the core promoter can usually occur in the absence of the upstream elements, but only in an inefficient way. This indicates that the proteins that bind to the upstream elements include at least some that are activators of transcription, and which therefore ‘promote’ gene expression. Inclusion of these sequences in the ‘promoter’ is therefore justified.
Each of the three types of eukaryotic RNA polymerase recognizes a different type of promoter sequence; indeed, it is the difference between the promoters that defines which genes are transcribed by which polymerases. The details for vertebrates are as follows (Figure 9.18): RNA polymerase I promoters consist of a core promoter spanning the transcription start point, between nucleotides -45 and +20, and an upstream control element about 100 bp further upstream.

![Figure 9.18. Structures of eukaryotic promoters.](image)

RNA polymerase II promoters are variable and can stretch for several kilobases upstream of the transcription start site. The core promoter consists of two segments: the -25 or TATA box (consensus 5'-TATAWAW-3', where W is A or T) and the initiator (Inr) sequence (consensus 5'-YYCARR-3', where Y is C or T, and R is A or G) located around nucleotide +1. Some genes transcribed by RNA polymerase II have only one of these two components of the core promoter, and some, surprisingly, have neither. The latter are called 'null' genes. They are still transcribed, possibly through interactions between the RNA polymerase and a sequence called MED-1 which lies within the gene (Novina and Roy, 1996), although the start position for transcription is more variable than for a gene with a TATA and/or Inr sequence. As well as the core promoter, genes recognized by RNA polymerase II have various upstream promoter elements, the functions of which are described in Section 9.3.2.

RNA polymerase III promoters are unusual in that they are located within the genes whose transcription they promote. These promoters are variable, falling into at least three categories. Usually the core promoter spans 50–100 bp and comprises two sequence boxes. One category of RNA polymerase III promoter is similar to those for RNA polymerase II, having a TATA box and a range of upstream promoter elements. Interestingly, this arrangement is seen with the U6 gene, which is one of a family of genes for small nuclear RNAs, all the other members of which are transcribed by RNA polymerase II.

9.2.3. Assembly of the transcription initiation complex

In a general sense, initiation of transcription operates along the same lines with each of the four types of RNA polymerase that we have been considering (Figure 9.19). The bacterial polymerase and the three eukaryotic enzymes all begin by attaching, directly or via accessory proteins, to their promoter or core promoter sequences. Next this closed promoter complex is converted into an open promoter complex by breakage of a limited number of base pairs around the transcription initiation site. Finally, the RNA polymerase moves away from the promoter. This last step is more complicated than it might appear because some attempts by the polymerase to achieve promoter clearance are unsuccessful and lead to truncated transcripts that are degraded soon after they are synthesized. The true completion of the
initiation stage of transcription is therefore the establishment of a stable transcription complex that is actively transcribing the gene to which it is attached.

Although the scheme shown in Figure 9.19 is correct in outline for all four polymerases, the details are different for each one. We will begin with the more straightforward events occurring in E. coli and other bacteria, and then move on to the ramifications of initiation in eukaryotes.

Figure 9.19. Generalized scheme for the events occurring during initiation of transcription. The core promoter is shown in blue and the transcription initiation site is indicated by a green dot. After RNA polymerase attachment, the closed complex is converted into the open complex by breakage of base pairs within a short region of the DNA double helix. RNA synthesis begins but successful initiation is not achieved until the polymerase moves away from the promoter region.

**Transcription initiation in E. coli**

In E. coli, a direct contact is formed between the promoter and RNA polymerase. The sequence specificity of the polymerase resides in its σ subunit: the 'core enzyme', which lacks this component, can only make loose and non-specific attachments to DNA.

Mutational studies of E. coli promoters have shown that changes to the sequence of the -35 box affect the ability of RNA polymerase to bind, whereas changes to the -10 box affect the conversion of the closed promoter complex into the open form. These results led to the model for E. coli initiation shown in Figure 9.20, where recognition of the promoter occurs by an interaction between the σ subunit and the -35 box, forming a closed promoter complex in which the RNA polymerase spans some 60 bp from upstream of the -35 box to downstream of the -10 box. This is followed by breaking of the base pairs within the -10 box to produce the
open complex. The model is consistent with the fact that the -10 boxes of different promoters are comprised mainly or entirely of A-T base pairs, which are weaker than G-C pairs, being linked by just two hydrogen bonds as opposed to three (see Figure 1.11B).

Opening up of the helix involves contacts between the polymerase and the non-template strand (i.e. the one that is not copied into RNA), again with the σ subunit playing a central role (Marr and Roberts, 1997). However, the σ subunit is not all-important because it dissociates soon after initiation is complete, converting the holoenzyme to the core enzyme which carries out the elongation phase of transcription (Section 10.1.1).

![Figure 9.20. Initiation of transcription in *Escherichia coli*. The E. coli RNA polymerase recognizes the -35 box as its binding sequence. After attachment to the DNA, the transition from closed to open complex is initiated by breakage of base pairs in the AT-rich -10 box. Note that although the polymerase is shown as a single structure it is the σ subunit that possesses sequence-specific DNA binding activity and which therefore recognizes the -35 sequence. For subsequent events leading to promoter clearance, refer to Figure 9.19.](image)

**Transcription initiation with RNA polymerase II**

How does the easily understandable series of events occurring in *E. coli* compare with the equivalent processes in eukaryotes? RNA polymerase II will show us that eukaryotic initiation involves more proteins and has added complexities.

The first difference between initiation of transcription in *E. coli* and eukaryotes is that eukaryotic polymerases do not directly recognize their core promoter sequences. For genes transcribed by RNA polymerase II, the initial contact is made by the general transcription factor (GTF) TFIID, which is a complex made up of the TATA-binding protein (TBP) and at least 12 TBP-associated factors or TAFs. TBP is a sequence-specific protein that binds to DNA via its unusual TBP domain (Section 9.1.4) which makes contact with the minor groove in the region of the TATA box. X-ray crystallography studies of TBP show that it has a saddle-like shape that wraps partially around the double helix (Chasman et al., 1993), forming a platform onto which the remainder of the initiation complex can be assembled. The TAFs are intriguing proteins that appear to play a variety of roles during initiation of transcription and also during other events that involve assembly of multiprotein complexes onto the genome. Five of the yeast TAFs are also present in SAGA, one of the histone acetyltransferase complexes that we met in Section...
8.2.1 (Grant et al., 1998), and TAFs have also been implicated in control of the cell cycle in various eukaryotes (Green, 2000) and in regulation of the developmental changes that result in formation of gametes in animals (Verrizer, 2001). During transcription, TAFs assist in attachment of TFIIID to the TATA box and, in conjunction with other proteins called TAF- and initiator-dependent cofactors (TICs), possibly also participate in recognition of the Inr sequence, especially at those promoters that lack a TATA box. A clue as to how TAFs carry out their multifarious roles has been provided by structural studies which have shown that at least three of them contain a histone fold - a DNA-binding domain which, as the name suggests, is also present in histone proteins (Table 9.2; Research Briefing 9.1). It has been proposed that these TAFs might be able to form a DNA-binding structure resembling a nucleosome (Burley and Roeder, 1996), but this idea may not be entirely correct because the TAFs lack certain amino acids that are looked on as essential for stabilizing the contacts between real nucleosomes and DNA (Gangloff et al., 2001).

After TFIIID has attached to the core promoter, the pre-initiation complex (PIC) is formed by attachment of the remaining GTFs (Table 9.5). Test-tube experiments suggest that these GTFs bind to the complex in the order TFIIA, TFIIB, TFIIIF/RNA polymerase II, TFIIE and TFIIH (Figure 9.21) but it is now thought that in vivo assembly involves a more complex set of interactions than indicated by this step-by-step sequence (Lee and Young, 2000).

Figure 9.21. Assembly of the RNA polymerase II pre-initiation complex. The first step in assembly of the pre-initiation complex is recognition of the TATA box and possibly the Inr sequence by the TATA-binding protein (TBP), in conjunction with the TBP-associated factors (TAFs). Test-tube experiments suggest that the other components of the pre-initiation complex assemble in the order TFIIA, TFIIB, TFIIIF/RNA polymerase II, TFIIE and TFIIH. The drawing is purely schematic with regard to the sizes and shapes of the components and their relative positioning within the pre-initiation complex, although account is taken of the protein-protein contacts that are known to occur. The C-terminal domain (CTD) of the largest subunit of RNA polymerase II, which must be phosphorylated before the polymerase can leave the promoter, is depicted as a spike projecting from the bottom of the complex.
Within the overall process, three events are particularly important:

1. Attachment of TBP induces formation of a bend in the DNA in the region of the TATA box.
2. The bend provides a recognition structure for TFIIB, which ensures correct positioning of RNA polymerase II relative to the transcription start site.
3. The disruption to the base pairing needed to form the open promoter complex is brought about by TFIIH (Kim et al., 2000).

The final step in assembly of the initiation complex is the addition of phosphate groups to the C-terminal domain (CTD) of the largest subunit of RNA polymerase II. In mammals, this domain consists of 52 repeats of the seven-amino-acid sequence Tyr-Ser-Pro-Thr-Ser-Pro-Ser. Two of the three serines in each repeat unit can be modified by addition of a phosphate group, causing a substantial change in the ionic properties of the polymerase. Once phosphorylated, the polymerase is able to leave the pre-initiation complex and begin synthesizing RNA. Phosphorylation might be carried out by TFIIH, which has the appropriate protein kinase capability, or it might be the function of the mediator (Section 9.3.2), which transduces signals from activator proteins that regulate expression of individual genes (Lee and Young, 2000). After departure of the polymerase, at least some of the GTFs detach from the core promoter, but TFIID, TFIIA and TFIIH remain, enabling re-initiation to occur without the need to rebuild the entire assembly from the beginning (Yudkovsky et al., 2000). Re-initiation is therefore a more rapid process than primary initiation, which means that once a gene is switched on, transcripts can be initiated from its promoter with relative ease until such a time as a new set of signals switches the gene off.

**Transcription initiation with RNA polymerases I and III**

Initiation of transcription at RNA polymerase I and III promoters involves similar events to those seen with RNA polymerase II, but the details are different. One of the most striking similarities is that TBP, first identified as the key sequence-specific DNA-binding component of the RNA polymerase II pre-initiation complex, is also involved in initiation of transcription by the two other eukaryotic RNA polymerases.

The RNA polymerase I initiation complex involves four protein complexes in addition to the polymerase itself. One of these, UBF, is a dimer of identical proteins that interacts with both the core promoter and the upstream control element (see Figure 9.18). UBF is another protein which, like some of the RNA polymerase II TAFs, resembles a histone and may form a nucleosome-like structure in the promoter region (Wolffe, 1994). A second protein complex, called SL1 in humans and TIF-I in mice, contains TBP and, together with UBF, directs RNA polymerase I and the last two complexes, TIF-I and TIF-IC, to the promoter. Originally it was thought that the initiation complex was built up in a stepwise fashion, but recent results suggest that RNA polymerase I binds the four protein complexes before promoter recognition, the entire assembly attaching to the DNA in a single step (Seither et al., 1998).

RNA polymerase III promoters are variable in structure (see Figure 9.18) and this is reflected by a non-uniformity of the processes through which they are recognized. Initiation at the different categories of RNA polymerase III promoter requires different sets of GTFs, but each type of initiation process involves TFIIB, one of whose subunits is TBP. With promoters of the type seen with the U6 gene, which contain a TATA sequence, TBP probably binds directly to the DNA. At other RNA polymerase III promoters, which have no TATA sequence, binding is probably via a second protein, the latter making the direct DNA contact.

**9.3. Regulation of Transcription Initiation**

As we progress through the next few chapters we will encounter a number of strategies that organisms use to regulate expression of individual genes. We will discover that virtually every step in the pathway from genome to proteome is subject to some degree of control. Of all
these regulatory systems, it appears that transcription initiation is the stage at which the critical controls over the expression of individual genes (i.e. those controls that have greatest impact on the biochemical properties of the cell) are exerted. This is perfectly understandable.

It makes sense that transcription initiation, being the first step in genome expression, should be the stage at which ‘primary’ regulation occurs, this being the level of regulation that determines which genes are expressed. Later steps in the pathway might be expected to respond to ‘secondary’ regulation, the function of which is not to switch genes on or off but to modulate expression by making small changes to the rate at which the protein product is synthesized, or possibly by changing the nature of the product in some way (Figure 9.22).

![Figure 9.22. Primary and secondary levels of gene regulation. According to this scheme, ‘primary’ regulation of genome expression occurs at the level of transcription initiation, this step determining which genes are expressed in a particular cell at a particular time and setting the relative rates of expression of those genes that are switched on. ‘Secondary’ regulation involves all steps in the gene expression pathway after transcription initiation, and serves to modulate the amount of protein that is synthesized or to change the nature of the protein in some way, for example by chemical modification.

In Chapter 8 we looked at how chromatin structure can influence gene expression by controlling the accessibility of promoter sequences to RNA polymerase and its associated proteins. This is just one way in which initiation of transcription can be regulated. To obtain a broader picture we will establish some general principles with bacteria, and then examine the events in eukaryotes.

9.3.1. Strategies for controlling transcription initiation in bacteria

In bacteria such as *E. coli*, we recognize two distinct ways in which transcription initiation is controlled:

- **Constitutive control**, which depends on the structure of the promoter;
- **Regulatory control**, which depends on the influence of regulatory proteins.

**Promoter structure determines the basal level of transcription initiation**

The consensus sequence for the *E. coli* promoter (Section 9.2.2) is quite variable, with a range of different motifs being permissible at both the -35 and -10 boxes (see Table 9.4). These variations, together with less well-defined sequence features around the transcription start site and in the first 50 or so nucleotides of the transcription unit, affect the efficiency of the promoter. Efficiency is defined as the number of productive initiations that are promoted per second, a productive initiation being one that results in the RNA polymerase clearing the promoter and beginning synthesis of a full-length transcript. The exact way in which the sequence of the promoter affects initiation is not known, but from our discussion of the events involved in transcription initiation (Section 9.2.3) we might, intuitively, expect that the precise sequence of the -35 box would influence recognition by the σ subunit and hence the rate of
attachment of RNA polymerase, that the transition from the closed to open promoter complex might be dependent on the sequence of the -10 box, and that the frequency of abortive initiations (ones that terminate before they progress very far into the transcription unit) might be influenced by the sequence at, and immediately downstream of, nucleotide +1. All this is speculation but it is a sound ‘working hypothesis’. What is clear is that different promoters vary 1000 fold in their efficiencies, the most efficient promoters (called strong promoters) directing 1000 times as many productive initiations as the weakest promoters. We refer to these as differences in the basal rate of transcription initiation.

![Figure 9.23. Recognition of an Escherichia coli heat shock gene by the σ^{32} subunit](image)

(A) An E. coli heat shock gene

Heat shock promoter

Heat shock gene

-44
-36
-10

...CTGCCACCCTC...CATNT...

(B) Recognition by the σ^{32} subunit

σ^{32} polymerase cannot bind

σ^{32} polymerase binds to the heat shock promoter

Figure 9.23. Recognition of an Escherichia coli heat shock gene by the σ^{32} subunit. (A) The sequence of the heat-shock promoter is different from that of the normal E. coli promoter (compare with Table 9.4 ). (B) The heat-shock promoter is not recognized by the normal E. coli RNA polymerase containing the σ^{70} subunit, but is recognized by the σ^{32} RNA polymerase that is active during heat shock. Abbreviation: N, any nucleotide. For more details of the use of novel σ factors by bacteria see Section 12.3.1.

Note that the basal rate of transcription initiation for a gene is preprogrammed by the sequence of its promoter and so, under normal circumstances, cannot be changed. It could be changed by a mutation that alters a critical nucleotide in the promoter, and undoubtedly this happens from time to time, but it is not something that the bacterium has control over. The bacterium can, however, determine which promoter sequences are favored by changing the σ subunit of its RNA polymerase. The σ subunit is the part of the polymerase that has the sequence-specific DNA-binding capability (Section 9.2.3), so replacing one version of this subunit with a different version with a slightly different DNA-binding motif, and hence an altered sequence specificity, would result in a different set of promoters being recognized. In E. coli, the standard σ subunit, which recognizes the consensus promoter sequence shown on page 255 and hence directs transcription of most genes, is called σ^{70} (its molecular mass is approximately 70 kDa). E. coli also has a second σ subunit, σ^{32}, which is made when the bacterium is exposed to a heat shock. During a heat shock, E. coli, in common with other organisms, switches on a set of genes coding for special proteins that help the bacterium withstand the stress (Figure 9.23). These genes have special promoter sequences, ones specifically recognized by the σ^{32} subunit. The bacterium is therefore able to switch on a whole range of different genes by making one simple alteration to the structure of its RNA polymerase. This system is common in bacteria; for example, Klebsiella pneumoniae uses it to
control expression of genes involved in nitrogen fixation, this time with the σ54 subunit, and *Bacillus* species use a whole range of different σ subunits to switch on and off groups of genes during the changeover from normal growth to formation of spores (Section 12.3.1).

**Regulatory control over bacterial transcription initiation**

Promoter structure determines the basal level of transcription initiation for a bacterial gene but, with the exception of alternative σ subunits, does not provide any general means by which the expression of the gene can respond to changes in the environment or to the biochemical requirements of the cell. Other types of regulatory control are needed.

The foundation of our understanding of regulatory control over transcription initiation in bacteria was laid in the early 1960s by François Jacob, Jacques Monod, and other geneticists who studied the lactose operon and other model systems (Biran and Gayon, 1999). We have already seen how this work led to discovery of the promoter for the lactose operon (Section 9.2.2). It also resulted in identification of the operator, a region adjacent to the promoter and which regulates initiation of transcription of the operon (Figure 9.24A). The original model envisaged that a DNA-binding protein - the lactose repressor - attached to the operator and prevented the RNA polymerase from binding to the promoter, simply by denying it access to the relevant segment of DNA (Figure 9.24B). Whether the repressor binds depends on the presence in the cell of allolactose, an isomer of lactose, the latter being the substrate for the biochemical pathway carried out by the enzymes coded by the three genes in the operon. Allolactose is an inducer of the lactose operon. When allolactose is present it binds to the lactose repressor, causing a slight structural change which prevents the HTH motifs of the repressor from recognizing the operator as a DNA-binding site. The allolactose-repressor complex therefore cannot bind to the operator, enabling the RNA polymerase to gain access to the promoter. When the supply of lactose is used up and there is no allolactose left to bind to the repressor, the repressor re-attaches to the operator and prevents transcription. The operon is therefore expressed only when the enzymes coded by the operon are needed.

Some repressors respond not to an inducer but to a co-repressor. An example is provided by the tryptophan operon of *E. coli*, which codes for a set of genes involved in synthesis of tryptophan (see Figure 2.20B). In contrast to the lactose operon, the regulatory molecule for the tryptophan operon is not a substrate for the relevant biochemical pathway, but the product, tryptophan itself (Figure 9.25). Only when tryptophan is attached to the tryptophan repressor can the latter bind to the operator. The tryptophan operon is therefore switched off in the presence of tryptophan, and switched on when tryptophan is needed.

Some DNA-binding proteins are activators rather than repressors of transcription initiation. The best studied example in *E. coli* is the catabolite activator protein, which binds to sites upstream of several operons, including the lactose operon, and increases the efficiency of transcript initiation, probably by forming a direct contact with the RNA polymerase. The biological role of the catabolite activator protein is described in Section 12.1.1.

Other DNA-binding proteins work singly or together to increase or repress transcription of genes to which they are not closely linked. These enhancers and silencers are not common in bacteria but a few examples are known, including an enhancer that acts on the *E. coli* heat-shock genes whose promoters are recognized by the σ32 version of the RNA polymerase. Because they are so far from the genes that they control, they can only form a contact with the RNA polymerase if the DNA forms a loop. A characteristic feature is that a single enhancer or silencer can control expression of more than one gene.
Figure 9.24. Regulation of the lactose operon of *Escherichia coli*. (A) The operator sequence lies immediately downstream of the promoter for the lactose operon. (B) In the original model for lactose regulation, the lactose repressor is looked on as a simple blocking device that binds to the operator and prevents the RNA polymerase gaining access to the promoter. The three genes in the operon are therefore switched off. This is the situation in the absence of lactose, although transcription is not completely blocked because the repressor occasionally detaches, allowing a few transcripts to be made. Because of this basal level of transcription, the bacterium always possesses a few copies of each of the three enzymes coded by the operon (see Figure 2.20A), probably less than five of each. This means that when the bacterium encounters a source of lactose it is able to transport a few molecules into the cell and split these into glucose and galactose. An intermediate in this reaction is allolactose, an isomer of lactose, which induces expression of the lactose operon by binding to the repressor, causing a change in the conformation of the latter so it is no longer able to attach to the operator. This allows the RNA polymerase to bind to the promoter and transcribe the three genes. When fully induced, approximately 5000 copies of each protein product are present in the cell. When the lactose supply is used up and allolactose is no longer present, the repressor re-attaches to the operator and the operon is switched off. Note that the shapes of the repressor and polymerase structures shown here are purely schematic.

Most of the original scheme for regulation of the lactose operon has been confirmed by DNA sequencing of the control region and by structural studies of the repressor bound to its operator. The one complication has been the discovery that the repressor has three potential binding sites at nucleotide positions -82, +11 and +412, and attachment at only one of these, +11, would be expected to prevent access of the polymerase to the promoter. The repressor is a tetramer of four identical proteins which work in pairs to attach to a single operator, so it is possible that the repressor has the capacity to bind to two of the three operator sites at once. It is also possible that the repressor can bind to an operator sequence in such a way that it does not block attachment of the polymerase to the promoter, but does prevent a later step in initiation, such as formation of the open promoter complex.
Figure 9.25. Regulation of the tryptophan operon of *Escherichia coli*. Regulation occurs via a repressor-operator system in a similar way to that described for the lactose operon (Figure 9.24) but with the difference that the operon is repressed by the regulatory molecule, tryptophan, which is the product of the biochemical pathway specified by the genes in the operon (see Figure 2.20B). When tryptophan is present, and so does not need to be synthesized, the operon is switched off because the repressor-tryptophan complex binds to the operator. In the absence of tryptophan the repressor cannot attach to the operator, and the operon is expressed.

9.3.2. Control of transcription initiation in eukaryotes

With bacteria, it is possible to make a clear distinction between constitutive and regulatory forms of control over transcription initiation. The former depends on promoter structure and determines the basal rate of transcription initiation; the latter depends on the activity of regulatory proteins and changes the rate of transcription initiation if the basal rate is inappropriate for the prevailing conditions. With eukaryotes, categorization of different types of control system is less easy. This is because of a fundamental difference in transcription initiation between bacteria and eukaryotes. In bacteria, the RNA polymerase has a strong affinity for its promoter and the basal rate of transcription initiation is relatively high for all but the weakest promoters. With most eukaryotic genes, the reverse is true. The RNA polymerase II and III pre-initiation complexes do not assemble efficiently and the basal rate of transcription initiation is therefore very low, regardless of how ‘strong’ the promoter is. In order to achieve effective initiation, formation of the complex must be activated by additional proteins. Some of these could be defined as ‘constitutive’ activators, in that they work on many different genes and seem not to respond to any external signals; others could be termed ‘regulatory’ activators because they target a limited number of genes and do respond to external signals. But there are gray areas between the two types and it is unwise to use this categorization as anything other than a guide to the types of event that occur.

The lactose operon illustrates the basic principle of regulatory control of transcription initiation: attachment of a DNA-binding protein to its specific recognition site can influence the events involved in assembly of the transcription initiation complex and/or initiation of productive RNA synthesis by an RNA polymerase. Several variations on this theme are seen with other bacterial genes:

**Activators of eukaryotic transcription initiation**

Any protein that stimulates transcription initiation is called an activator. Initially it was imagined that all activators were sequence-specific DNA-binding proteins, some recognizing upstream promoter elements and influencing transcription initiation only at the promoter to which these elements are attached, and others targeting sites within enhancers and influencing transcription of several genes at once (Figure 9.26). As with bacteria, eukaryotic enhancers can be some distance from their genes; their target specificity is ensured by the presence of insulators at either side of each functional domain, preventing the enhancers within that domain from influencing gene expression in adjacent domains (Section 8.1.2). Whether bound to an upstream promoter element or to a more distant enhancer, the activator, according to the traditional view, stabilizes the pre-initiation complex by making contact with it.
Figure 9.26. Activators of eukaryotic transcription initiation. The blue activator is attached to a regulatory module upstream of a gene, and influences transcription initiation only at that single gene. The green activator is attached to a site within an enhancer and is influencing transcription of all three genes.

This traditional view still holds for the majority of activators that have been identified but cannot be looked upon as all-encompassing. We have already seen that some proteins that were initially identified as activators are now recognized as components of chromatin modification complexes such as SAGA and Swi/Snf (Section 8.2.1). Other proteins classed as activators influence gene expression by introducing bends and other distortions into DNA (Thomas and Travers, 2001), possibly as a prelude to chromatin modification, or possibly to bring together proteins attached to non-adjacent sites, enabling the bound factors to work together in a structure that has been called an enhanceosome. An example of an activator that works in this way is SRY, which is the primary protein responsible for determining sex in mammals (Wolffe, 1995). Still other activators have no DNA-binding properties and they stimulate transcription simply by forming protein-protein contacts with the pre-initiation complex. As more and more activators are discovered, our appreciation of their diversity will undoubtedly grow (Lee and Young, 2000).

Activators have been looked upon as important in initiation by RNA polymerases II and III, but their role at RNA polymerase I promoters has been less well defined. RNA polymerase I is unusual in that it transcribes just a single set of genes: the multiple copies of the transcription unit containing the 28S, 5.8S and 18S rRNA sequences (Section 3.2.1). These genes are expressed continuously in most cells, but the rate of transcription varies during the cell cycle and is subject to a certain amount of tissue-specific regulation. The regulatory mechanism has not been described in detail but recent research has suggested a role for the RNA polymerase I termination factor. This factor, called TTF-1 in mice and Reb1p in Saccharomyces cerevisiae, was first identified as an activator of RNA polymerase II transcription. It appears that the termination factor may also activate RNA polymerase I transcription, a binding site for it having been located immediately upstream of the promoter for the rRNA transcription unit (Reeder and Leng, 1997).

Contacts between activators and the pre-initiation complex

A critical feature of the ‘traditional’ type of activator - those that bind to upstream promoter elements or to enhancers - is the contact that is formed with the pre-initiation complex. The part of the activator that makes this contact is called the activation domain. Structural studies have shown that although activation domains are variable, most of them fall into one of three categories:

- **Acidic domains** are ones that are relatively rich in acidic amino acids (aspartic acid and glutamic acid). This is the commonest category of activation domain.
- **Glutamine-rich domains** are often found in activators whose DNA-binding motifs are of the homeodomain or POU type (Section 9.1.4).
- **Proline-rich domains** are less common.

Details of the interaction between activators and the pre-initiation complex were obscure for several years, with apparently conflicting evidence coming from work with different organisms. A number of protein-protein interaction studies had suggested that direct contacts could be made between different activators and various parts of the complex, with TBP, various TAFs, TFIIB, TFIH and RNA polymerase II all implicated as partners in different interactions. An alternative possibility was raised when a large protein complex called the mediator was identified in yeast. The mediator forms a physical contact between activators and the C-
terminal domain of RNA polymerase II (Figure 9.27; Thompson et al., 1993; Kim et al., 1994), suggesting that rather than direct interaction between an activator and the pre-initiation complex, the signal is transduced by the mediator. This hypothesis was strengthened when it was shown that the mediator possesses a protein kinase activity that enables it to phosphorylate the CTD of RNA polymerase II, stimulating promoter clearance (Section 9.2.3). The importance of the mediator in yeast transcription initiation was further underlined by the discovery that several of its components were previously looked upon as coactivators, proteins that are needed for full activation of the pre-initiation complex but which do not themselves respond directly to any of the external signals that modulate genome expression (as described in the last section in this chapter).

Figure 9.27. The role of the mediator

For a few years it appeared that the mediator might not be a common feature of eukaryotic pre-initiation complexes in general, but eventually an equivalent structure was identified in mammalian cells (Kingston, 1999; Malik and Roeder, 2000). Subsequent work has shown that there are several different versions of the mammalian mediator, each one responding to a different, although possibly overlapping, set of activators. Current opinion tends to the view that a mediator is an obligatory component of the RNA polymerase II pre-initiation complex, and that the stimulatory effects of all activators pass through the mediator. The possibility that some activators bypass the mediator and have a direct effect on one or other part of the pre-initiation complex cannot, however, be discounted.

Repressors of eukaryotic transcription initiation

Most of the research on regulation of transcription initiation in eukaryotes has concentrated on activation, partly because the low level of basal initiation occurring at RNA polymerase II and III promoters suggests that the repression of initiation, which is so important in bacteria (Section 9.3.1), is unlikely to play a major part in control of eukaryotic transcription. This view is probably incorrect because a growing number of DNA-binding proteins that repress transcription initiation are being discovered, these proteins binding to upstream promoter elements or to more distant sites in silencers. Some influence genome expression in a general way through histone deacetylation or DNA methylation (Section 8.2.2), but others have more specific effects at individual promoters. The yeast repressors called Mot1 and NC2, for example, inhibit assembly of the pre-initiation complex by binding directly to TBP and disrupting its activity. Mot1 causes TBP to dissociate from the DNA, and NC2 prevents further assembly of the complex on the bound TBP (Lee and Young, 2000). Both of these repressors have a broad spectrum of activity, inactivating a large set of genes, as does the Ssn6-Tup1 repressor, which is one of the main gene silencers in the yeast Schizosaccharomyces pombe, and which has homologs in many other eukaryotes (Smith and Johnson, 2000).

Another indication of the importance of repression in eukaryotic transcription comes from the demonstration that some proteins can exert both activating and repressing effects, depending on the circumstances. NC2, for example, represses initiation of transcription from promoters with a TATA box but has an activating effect on promoters that lack the TATA sequence (Willy et al., 2000). Pit-1, which is the first of the three proteins after which the POU domain is named (Section 9.1.4), activates some genes and represses others, depending on the sequence of its DNA-binding site (Scully et al., 2000). The presence in this site of two additional nucleotides induces a change in the conformation of Pit-1, enabling it to interact with a second protein called N-CoR and repress transcription of the target gene (Figures 9.28 and 9.29).
Figure 9.28. Conformation of the POU domains of the Pit-1 activator bound to its target sites upstream of the prolactin and growth-hormone genes. Pit-1 is a dimer, and each monomer has two POU domains (Section 9.1.4). The two domains of one monomer are shown in red and the two domains of the other monomer are shown in blue. The barrels are α-helices, with α3 being the recognition helix of each domain. Note the difference between the conformations of the domains when bound to the two binding sites. The more open structure adopted at the growth hormone site enables the Pit-1 dimer to interact with N-CoR and other proteins to repress transcription of the growth hormone gene. Reprinted with permission from Scully et al., Science, 290, 1127–1131. Copyright 2000 American Association for the Advancement of Science.

Figure 9.29. Pit-1 can activate or repress transcription initiation depending on the sequence of its DNA-binding site. Pit-1 activates transcription of the prolactin gene but represses transcription of the growth hormone gene. The drawing shows the contacts made between the amino acids of the POU domain (blue boxes; the numbers indicate the positions of the amino acids in the protein) and the nucleotides at the DNA-binding sites. The lines indicate potential hydrogen bonds and van der Waals contacts (see Box 3.3). The growth hormone site contains two extra base pairs (shown as open boxes) in the middle of the contact region. The presence of these two base pairs leads to the change in the conformation of the Pit-1 dimer, allowing the interaction with N-CoR. From Scully et al. 2000.

Relatively little is known about the precise interactions occurring between repressors and the pre-initiation complex. A variety of inhibition domains (the converse of an activation domain) have been identified in eukaryotic repressors, several of which are rich in prolines, but no general patterns have emerged (Hanna-Rose and Hansen, 1996). The direct interactions with TBP displayed by Mot1 and NC2 argue against the involvement in repression of a complex equivalent to the mediator that is required for gene activation.
Controlling the activities of activators and repressors

The operation of individual activators and repressors must be controlled in order to ensure that the appropriate set of genes is expressed by a cell. We will return to this topic in Chapter 12, when it will form the central theme of our study of the ways in which genome activity is regulated in response to extracellular signals and during differentiation and development.

There are several ways in which an activator or repressor could be regulated. One possibility is to control its synthesis, but this does not permit rapid changes in genome expression because it takes time to accumulate an activator or repressor in the cell, or to destroy it when it is not needed. This type of control is therefore associated with activators and repressors responsible for maintaining stable patterns of genome expression, for example those underlying cellular differentiation and some aspects of development. An alternative way of controlling an activator or repressor is by chemical modification, for example by phosphorylation, or by inducing a change in its conformation. These changes are much more rapid than de novo synthesis, and enable the cell to respond to extracellular signaling compounds that induce transient changes in genome expression. We will examine the details of these various regulatory mechanisms in Chapter 12.

BOX

9.1. RNA-binding motifs

RNA-binding proteins also have specific motifs that form the attachment with the RNA molecule. The most important of these are as follows:

- The ribonucleoprotein (RNP) domain comprises four β-strands and two α-helices in the order β-α-β-α-β-β-α. The two central β-strands make the critical attachments with the RNA molecule. The RNP domain is the commonest RNA-binding motif and has been found in more than 250 proteins.
- The double-stranded RNA binding domain (dsRBD) is similar to the RNP domain but with the structure α-β-β-β-α-β-α. The RNA-binding function lies between the β and α at the end of the structure. As the name implies, the motif is found in proteins that bind double-stranded RNA (Fierro-Monti and Mathews, 2000).
- The κ-homology domain has the structure β-α-α-β-β-β-α-β-α-α. With the binding function between the pair of α-helices. It is relatively uncommon but present in at least one nuclear RNA-binding protein.
- The DNA-binding homeodomain may also have RNA-binding activity in some proteins. One ribosomal protein uses a structure similar to a homeodomain to attach to rRNA, and some homeodomain proteins such as Bicoid of Drosophila melanogaster (Section 12.3.3) can bind both DNA and RNA.

9.2. Can sequence specificity be predicted from the structure of a recognition helix?

An intriguing question is whether the specificity of DNA binding can be understood in sufficient detail for the sequence of a protein's target site to be predicted from examination of the structure of the recognition helix of a DNA-binding motif. To date, this objective has largely eluded us, but it has been possible to deduce some rules for the interaction involving certain types of zinc finger (Choo and Klug, 1997). In these proteins, four amino acids, three in the recognition helix and one immediately adjacent to it, form critical attachments with the nucleotide bases of the target site. Some of these attachments involve a single amino acid with a single base; others involve two amino acids with one base. By comparing the sequences of amino acids in the recognition helices of different zinc fingers with the sequences of nucleotides at the binding sites, it has been possible to identify a set of rules governing the interaction. These enable the nucleotide sequence specificity of a new zinc finger protein to be predicted, admittedly with the possibility of some ambiguity, once the amino acid composition of its recognition helix is known.
9.3. Mitochondrial and chloroplast RNA polymerases

The RNA polymerases that transcribe organelle genes are unlike their counterparts in the nucleus, reflecting the bacterial origins of mitochondria and chloroplasts (Section 2.2.2).

- The mitochondrial RNA polymerase consists of a single subunit with a molecular mass of 140 kDa. Interestingly, this enzyme is more closely related to the RNA polymerases of certain bacteriophages than it is to the standard bacterial version. The polymerase works in conjunction with at least two additional proteins, one of which is a transcription factor - called mtTF1 in mammals and ABF2 in *Saccharomyces cerevisiae* - which binds to mitochondrial promoters and enhances initiation of transcription. These promoters are variable in sequence but most contain an AT-rich region.
- Chloroplast RNA polymerases appear to be much more similar to the bacterial enzymes and their promoters follow the pattern described for *Escherichia coli*, consisting of two boxes at the -35 and -10 positions relative to the start point for transcription.

9.4. Initiation of transcription in the archaea

Initiation of transcription is one of the key areas in which the archaea differ from the bacteria, emphasizing the major distinction between these two types of prokaryote. The archaeal RNA polymerase is made of 11 subunits and closely resembles the RNA polymerases of eukaryotes. Promoters are recognized by a TATA-binding protein, and RNA polymerase attachment is mediated by a transcription factor similar to the eukaryotic TFIIB. Archaeal homologs of the other eukaryotic general transcription factors have not been identified, and searches of the completed archaeal genome sequences suggest that they may not be present.

9.5. Cis and trans

'Cis' and 'trans' are two important terms relevant to the genetic study of gene regulation in bacteria and other organisms.

A locus is *cis*-acting on a second locus if it must be on the same DNA molecule in order to have an effect. The operator is a *cis*-acting element because it works only when physically attached to the gene whose expression it regulates.

A locus is *trans*-acting if it can affect a second locus even when on a different DNA molecule. The gene for the lactose repressor (lacI) is *trans*-acting because it can regulate expression of the lactose operon even when removed from the *Escherichia coli* chromosome and placed on a plasmid.

To a molecular biologist, a *cis*-acting regulatory element is usually a target site for a DNA-binding protein, upstream of the gene whose expression is being regulated. A *trans*-acting element is the regulatory protein itself, which can diffuse through the cell from its site of synthesis to its DNA-binding site.

9.6. The modular structures of RNA polymerase II promoters

The promoter for a gene transcribed by RNA polymerase II can be looked on as a series of modules, each comprising a short sequence of nucleotides and each acting as the binding site for a protein that influences assembly of the transcription initiation complex. Many different genes are transcribed by RNA polymerase II (approximately 35 000 in humans) but there are only a limited number of promoter modules. The expression pattern for a gene is therefore determined not by an individual module but by the combination of modules within its promoter, and possibly by their relative positions. The amount of transcription initiation that occurs is dependent on which modules are occupied by their binding proteins at a particular time.
The modules can be categorized in various ways. One scheme is as follows:

- **The core promoter modules** (Section 9.2.2). The most important of these are the TATA box (consensus 5'-TATAAW-3', where W is A or T) and the Inr sequence (consensus 5'-YYCARR-3', where Y is C or T, and R is A or G).

- **Basal promoter elements** are modules that are present in many promoters and set the basal level of transcription initiation, without responding to any tissue-specific or developmental signals. These include: the CAAT box (consensus 5'-GGCCAATCT-3'), recognized by the activators NF-1 and NF-Y; the GC box (consensus 5'-GGCCGG-3') recognized by the Sp1 activator; and the octamer module (consensus 5'-ATGCAAAT-3'), recognized by Oct-1.

- **Response modules** are found upstream of various genes and enable transcription initiation to respond to general signals from outside of the cell. Examples are: the cyclic AMP response module CRE (consensus 5'-WCGTCA-3'), recognized by the CREB activator; the heat-shock module (consensus 5'-CTNGAATNTTCTAGA-3'), recognized by HSP70 and other activators; and the serum response module (consensus 5'-CCWWWWWWGG-3'), recognized by the serum response factor.

- **Cell-specific modules** are located in the promoters of genes that are expressed in just one type of tissue. Examples include: the erythroid module (consensus 5'-WGATAR-3') which is the binding site for the GATA-1 activator; the pituitary cell module (consensus 5'-ATATTCAT-3'), recognized by Pit-1; the myoblast module (consensus 5'-CAACTGAC-3'), recognized by MyoD; and the lymphoid cell module or 8 site (consensus 5'-GGGACTTTCC-3'), recognized by NF-8. Note that in lymphoid cells the octamer module is recognized by the tissue-specific Oct-2 activator.

- **Modules for developmental regulators** mediate expression of genes that are active at specific developmental stages. Two examples in *Drosophila* are (Section 12.3.3): the Bicoid module (consensus 5'-TCTTAATCC-3') and the Antennapedia module (consensus 5'-TAATAATAATATAGGG-3').

The human insulin gene (shown below) illustrates the modular structure of a promoter for RNA polymerase II.

As well as the modules in the region immediately upstream of the gene, the same and other modules can also be contained within enhancers, which are 200–300 bp in length and can be located some distance upstream or downstream of their target gene. A single enhancer can influence transcription initiation of a variety of genes located within a single functional domain (Section 8.1.2). Silencers are similar to enhancers but, as their name suggests, their modules have a negative rather than enhancing influence on transcription initiation.

The modular interpretation of promoter structure is a well-established and useful concept, but its limitations must always be kept in mind. It emphasizes the role of sequence-specific DNA-binding proteins in genome expression and downplays the influence of proteins that either do not bind specifically to the genome, or whose binding sites lie distant from the genes whose expression they influence. The module concept therefore draws attention away from histone-modifying enzymes, nucleosome remodeling complexes, and other proteins that play a critical role in genome expression.
Research Briefing

9.1. Similarities between TFIID and the histone core octamer

An intriguing insight into the interaction between TFIID and the RNA polymerase II promoter was provided by the discovery that several TAFs have structural similarities with histones.

A key step in initiation of transcription by RNA polymerase II is attachment of the general transcription factor TFIID to the core promoter. The resulting structure forms the platform onto which the pre-initiation complex is constructed. Intuitively, one anticipates that this platform must be a stable structure in order for the bulky pre-initiation complex to be positioned correctly so that productive initiation of transcription can occur. To understand the nature of the attachment, attention has been focused on the TBP-associated factors (TAFs) that, together with the TATA-binding protein, make up the TFIID protein complex.

**Homology searching reveals similarities between TAFs and histones**

The first link between TAFs and histones was established by homology searching (Section 7.2.1). In 1993-94, a number of TAF genes were cloned, enabling nucleotide sequences to be obtained. The nucleotide sequences were translated into the amino acid sequences of the TAF proteins, and the amino acid sequences used to search the databases for interesting homologies with other proteins. Weak similarities were discovered with the histones of the core octamer of the nucleosome (Section 2.2.1). Part of one of the sequence alignments, between the central regions of TAF62 and histone H4, is shown below. In this central region of 55 amino acids there are 15 positions at which the same amino acid is present in both sequences, and a further 15 positions where the amino acids are ‘similar’, these being ones which, although different, have equivalent chemical properties. The same kind of pattern is seen when TAF42 is compared with histone H3. The degree of similarity between these two sets of proteins is not so great that a structural and/or functional relationship can immediately be inferred, but it is sufficiently striking to suggest that more detailed comparisons between TAFs and histones should be made.

These more detailed comparisons were carried out by X-ray crystallographic analyses (Section 9.1.3) of TAF42 and TAF62. The results were remarkable, both TAFs being shown to possess a typical histone fold (Table 9.2), a non-sequence-specific DNA-binding motif, comprising a long α-helix flanked on either side by two shorter α-helices, that is a characteristic feature of histones. Even more striking was the discovery that in the complex formed between TAF42 and TAF62, the two histone folds are orientated in almost exactly the same way as in the histone H3/H4 dimer found in the central tetramer of the nucleosome core particle:

![Histone H4 and TAF42 alignment](image)

Can these results be taken as evidence that TFIID forms a nucleosome-like structure at the core promoter, this pseudonucleosome acting as the platform for assembly of the pre-initiation complex? The conclusion is attractive but premature. More research is needed to determine if the similarities between TAFs and histones extend to the details of the attachment between each pre-initiation complex and its DNA target. An alternative is that the similarities merely reflect an equivalence in the protein-protein interactions within the two complexes, and are not directly relevant to the DNA attachment.

Permission to reproduce this figure in this web version of Genomes is either pending or has not been granted.
Self study questions

1. Explain why DNA-binding proteins are central to genome expression.
2. Describe how gel retardation can be used to study DNA-protein interactions. What are the limitations of this technique?
3. Draw diagrams to illustrate the modification protection and modification interference techniques. Indicate the key differences and describe how these differences underlie the specific applications of the two techniques.
4. Explain how affinity chromatography is used to purify a DNA-binding protein.
5. Write short essays on (a) X-ray crystallography, and (b) nuclear magnetic resonance spectroscopy, emphasizing the use of these techniques in the study of DNA-binding proteins.
6. Describe, with examples, how proteins that contain the helix-turn-helix motif bind to DNA. List, again with examples, the modified versions of the helix-turn-helix motif that are found in eukaryotic proteins.
7. Using examples, distinguish between two or more types of zinc finger.
8. Compare and contrast the structures used by proteins to bind to DNA and/or RNA molecules.
9. What features of the double helix are important in determining the nature of the interaction between DNA and a binding protein?
10. Describe the types of contact made between DNA and a binding protein. Why are many DNA-binding proteins dimeric?
11. Distinguish between the three nuclear RNA polymerases of eukaryotes. How is the Escherichia coli RNA polymerase similar to or different from the eukaryotes enzymes?
12. Define the term ‘promoter’. Draw annotated diagrams to illustrate the structures of the promoters for the three eukaryotic RNA polymerases and for the Escherichia coli enzyme.
13. Explain the roles of the two components of the Escherichia coli promoter during initiation of transcription. Be sure to make clear the difference between the closed and open versions of the promoter-RNA polymerase complex.
14. Write an essay on ‘Assembly of the RNA polymerase II initiation complex’. As part of your essay, compile a table giving the names of the main proteins or groups of proteins involved in assembly of this complex, along with a summary of the role of each one.
15. What is the importance of the C-terminal domain of the largest subunit of RNA polymerase II?
16. How does the TATA-binding protein provide a link between the initiation processes of all three eukaryotic RNA polymerases?
17. Describe how promoter structure influences gene expression in Escherichia coli.
18. Using examples, outline how the use of alternative σ subunits enables a bacterium to alter its pattern of genome expression.
19. Draw a series of diagrams to show how initiation of transcription is regulated at the lactose and tryptophan operons of Escherichia coli. Indicate the key differences between these two control mechanisms.
20. What is an activator? How do activators influence assembly of the RNA polymerase II initiation complex?
21. Explain why the discovery of a mammalian mediator was looked upon as a critical breakthrough in understanding the control of transcription initiation.
22. Describe our current knowledge of proteins that repress eukaryotic transcription initiation.
Problem-based learning

1. The methods for locating the positions of protein binding sites described in Section 9.1.1 assume that these sites are located in the region upstream of a gene. Is this assumption justified?

2. Use your knowledge of DNA chip and microarray technologies (Technical Note 5.1) to devise a method for identifying the attachment sites for a DNA-binding protein across the entire genome, as opposed to just within the region upstream of a single gene.

3. Write a report that elaborates on, and extends, the discussion presented in Box 9.2 (page 252), concerning the possibility that the amino acid sequence of a recognition helix can be used to deduce the nucleotide sequence of the DNA-binding site for a protein that contains that helix.

4. Construct a hypothesis to explain why eukaryotes have three RNA polymerases. Can your hypothesis be tested?

5. A model for control of transcription of the lactose operon in Escherichia coli was first proposed by François Jacob and Jacques Monod in 1961 (Jacob F and Monod J [1961] Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol., 3, 318–356). Explain the extent to which their work, which was based almost entirely on genetic analysis, provided an accurate description of the molecular events that are now known to occur.

6. To what extent is E. coli a good model for the regulation of transcription initiation in eukaryotes? Justify your opinion by providing specific examples of how extrapolations from E. coli have been helpful and/or unhelpful in the development of our understanding of equivalent events in eukaryotes.

7. Assess the accuracy and usefulness of the module concept for the structure of an RNA polymerase II promoter.