GENES

BENJAMIN · LEWIN ·

III

GENES

THIRD EDITION

BENJAMIN LEWIN
Editor, *Cell*

JOHN WILEY & SONS

New York Chichester Brisbane Toronto Singapore

Illustrations by John Balbalis with the assistance of the Wiley Illustration Department

Library of Congress Cataloging-in-Publication Data:

Lewin, Benjamin.
Genes.

Includes bibliographies and index.
1. Genetics. I. Title.

QH430.L487 1987 575.1 86-18959
ISBN 0-471-83278-2

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CHAPTER 9

RNA POLYMERASE-PROMOTER INTERACTIONS CONTROL INITIATION

How did RNA come to be the intermediary between DNA and protein? Perhaps the first primitive cells made no distinction between types of nucleic acid, so that what passed for the genome was involved directly in both replication and translation. At some point, it may have become advantageous to separate translation from the genome, so that proteins were synthesized on messengers distinct from the genetic material itself.

It is impossible to say how this separation related in time to the development of the other ribonucleic acid components involved in translation, but it is striking that RNA is present in the ribosome as well as constituting the tRNA adaptor. (It would not be surprising if the role of RNA in the ribosome was more prominent in the past than is apparent today.) Perhaps RNA was the original nucleic acid, and its ubiquitous presence today is merely a recollection of its former activity.

All this speaks to the fact that RNA plays central roles in gene expression, not merely in constituting the messenger, but also in providing the means for its translation into protein. In a sense, these roles represent the various specific interests of an RNA conglomerate generally concerned with gene expression, but with several different functions.

The production of each type of RNA has a common origin: transcription of DNA. In the case of mRNA, the product is an intermediate whose function requires translation. In the case of tRNA and rRNA, the transcriptional product itself fulfills the final function.

To transcribe or not to transcribe: that is the question? Transcription is the principal stage at which gene expression is controlled. The first (and sometimes the only) step in control is the decision on whether or not to transcribe a gene. In considering the various stages of transcription, we should therefore keep in mind the opportunities they offer for regulating gene activity.

(Transcription is not the only means by which RNA can be synthesized. Viruses with RNA genomes specify enzymes able to synthesize RNA on a template itself consisting of RNA. Such reactions produce mRNAs coding for proteins needed in the infective cycle [RNA transcription] and provide genomic RNAs to perpetuate the infective cycle [RNA replication]. Yet a further reaction is possible with the retroviruses, in which viral RNA

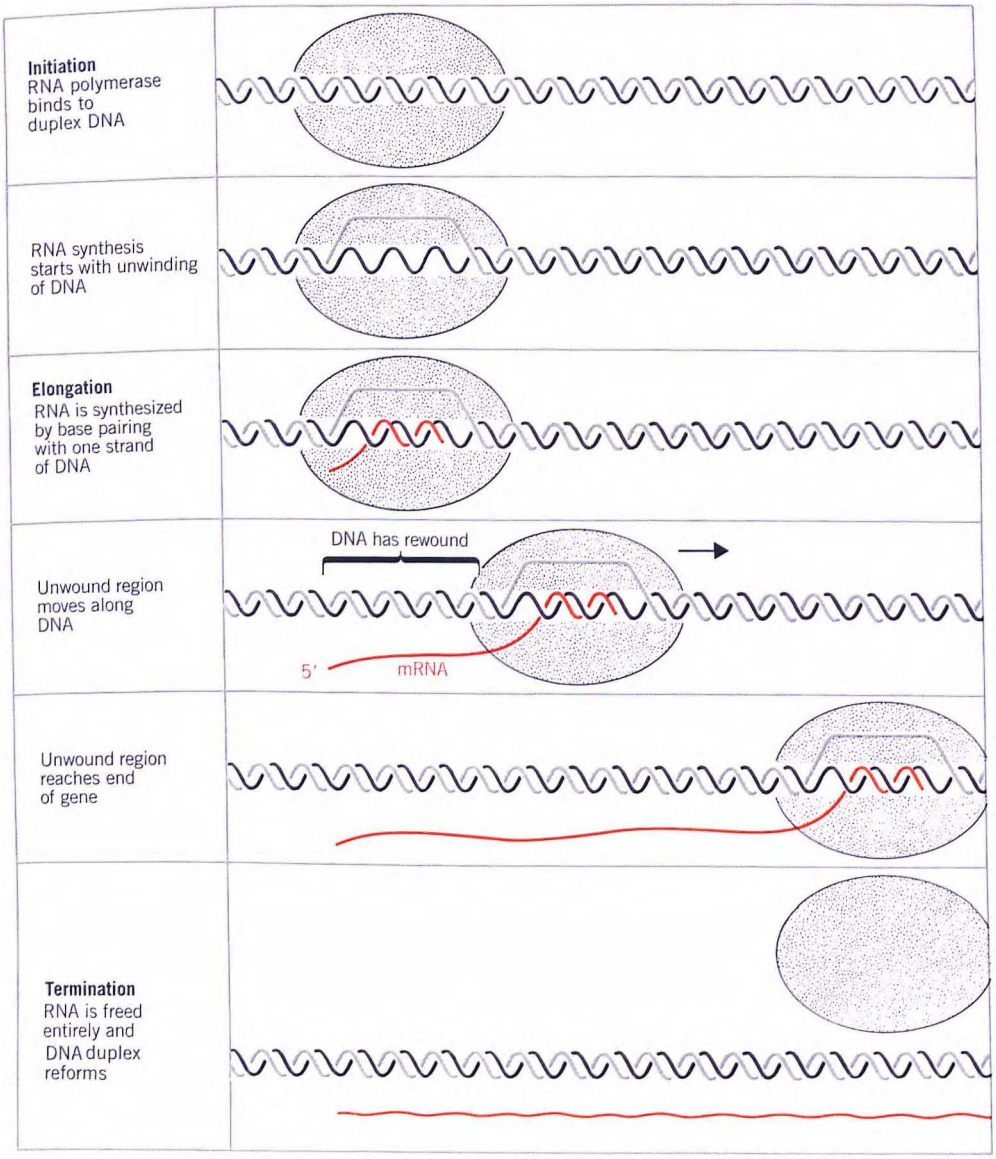


Figure 9.1
RNA is synthesized by base pairing with one strand of DNA in a region that is transiently unwound. As the region of unwinding moves, the DNA duplex reforms behind it, displacing the RNA in the form of a single polynucleotide chain.

serves as a template for reverse transcription to produce a DNA complement.)

TRANSCRIPTION IS CATALYZED BY RNA POLYMERASE

Transcription involves synthesis of an RNA chain representing one strand of a DNA duplex. By "representing" we mean that the RNA is identical in sequence with one strand of the DNA; it is complementary to the other strand, which provides the template for its synthesis.

Transcription takes place by the usual process of complementary base pairing, catalyzed by the enzyme **RNA polymerase.** The reaction can be divided into the three stages illustrated in **Figure 9.1.**

- **Initiation** begins with the binding of RNA polymerase to the double-stranded DNA. To make the tem-

plate strand available for base pairing with ribonucleotides, the strands of DNA must be separated. The unwinding is a local event that begins at the site bound by RNA polymerase. The initiation stage is completed when the first nucleotide is incorporated.

The entire sequence of DNA needed for these reactions is called the **promoter.** The site at which the first nucleotide is incorporated is called the **start-site** or **startpoint.**

- **Elongation** describes the phase during which nucleotides are covalently added to the 3′ end of the growing polynucleotide chain. Successive bases are added to the RNA chain, forming an RNA-DNA hybrid in the unwound region.

  To continue synthesis, the enzyme moves along the DNA, unwinding the double helix to expose a new segment of the template in single-stranded condition. As it moves, the RNA that was made previously is displaced from the DNA template strand, which pairs with its original partner to reform the double helix.

  Thus elongation involves the movement along DNA of a short segment that is transiently unwound, existing as a hybrid RNA-DNA duplex and a displaced single strand of DNA.

- **Termination** involves recognition of the point at which no further bases should be added to the chain. To terminate transcription, the formation of phosphodiester bonds must cease, and the transcription complex must come apart. When the last base is added to the RNA chain, the RNA-DNA hybrid is disrupted, the DNA reforms in duplex state, and the enzyme and RNA are both released from it. The sequence of DNA required for these reactions is called the **terminator.**

Originally defined simply by its ability to incorporate nucleotides into RNA under the direction of a DNA template, the enzyme RNA polymerase now is seen as part of a more complex apparatus involved in transcription. *The ability to catalyze RNA synthesis defines the minimum component that can be described as RNA polymerase.* It supervises the base pairing of the substrate ribonucleotides with DNA and catalyzes the formation of phosphodiester bonds between them.

But ancillary activities may be needed to initiate and to terminate the synthesis of RNA, when the enzyme must associate with, or dissociate from, a specific site on DNA. The analogy with the division of labors between the ribosome and the protein synthesis factors is obvious. Sometimes it is difficult to decide whether a particular protein that is involved in transcription at one of these stages should be considered as part of the "RNA polymerase" or as an ancillary factor.

All the components involved in elongation are necessary for initiation and termination. Genes may differ in their dependence on additional polypeptides at the initiation and termination stages. Some of these additional polypeptides may be needed at all genes, but others may be needed specifically for the initiation or termination of particular genes. An additional polypeptide needed to recognize all promoters (or terminators) is likely to be classified as part of the enzyme. A polypeptide needed only for the initiation (or termination) of particular genes is likely to be classified as an ancillary control factor.

With bacterial enzymes, it is possible to begin to define the roles of individual polypeptides in the stages of transcription. With eukaryotes, the enzymes are less well purified, and the actual enzymatic activities have yet to be completely resolved from the relatively crude preparations. Ironically enough, in eukaryotes we have begun to isolate ancillary factors needed to initiate or terminate particular genes, while the basic polymerase preparation itself remains rather poorly characterized.

BACTERIAL RNA POLYMERASE CONSISTS OF CORE ENZYME AND SIGMA FACTOR

A single type of RNA polymerase is responsible for all synthesis of mRNA, rRNA and tRNA in bacteria. The total number of RNA polymerase molecules present in an *E. coli* cell is ~7000. Many of them are actually engaged in transcription; probably between 2000 and 5000 enzymes are synthesizing RNA at any one time, the number depending on the growth conditions.

The best characterized RNA polymerase is that of *E. coli*, but its structure is similar in all other bacteria studied. The **complete enzyme** or **holoenzyme** has a molecular weight of ~480,000 daltons. Its subunit composition is summarized in **Table 9.1.**

Table 9.1
***E. coli* RNA polymerase has four types of subunit.**

Subunit	Gene	Number	Mass (daltons)	Location	Possible Functions
α	*rpoA*	2	40,000 each	core enzyme	promoter binding
β	*rpoB*	1	155,000	core enzyme	nucleotide binding
β'	*rpoC*	1	160,000	core enzyme	template binding
σ	*rpoD*	1	85,000	sigma factor	initiation

The α, β, and β' subunits have rather constant sizes in different bacterial species; the σ varies more widely, from 32,000 to 92,000. The enzyme has a rather elongated structure, with a maximum dimension of 15 nm (one turn of the DNA double helix is 3.4 nm).

The holoenzyme ($\alpha_2\beta\beta'\sigma$) can be separated into two components, the **core enzyme** ($\alpha_2\beta\beta'$) and the **sigma factor** (the σ polypeptide). The names reflect the fact that only the holoenzyme can initiate transcription; but then the sigma "factor" is released, leaving the core enzyme to undertake elongation. *Thus the core enzyme has the ability to synthesize RNA on a DNA template, but cannot initiate transcription at the proper sites.*

Core enzyme starts transcription at the separated DNA strands of an initiation complex. As the enzyme moves along the template extending the RNA chain, the region of local unwinding moves with it. The enzyme covers ~60 bp of DNA; the unwound segment comprises only a small part of this stretch, <17 bp according to the overall extent of unwinding.

As the DNA unwinds to free the template, each of its strands probably enters a separate site in the enzyme structure. As **Figure 9.2** indicates, the template strand will be free just ahead of the point at which the ribonucleotide is being added to the RNA chain, and it will exist as a DNA-RNA hybrid in the region where RNA has just been synthesized. The length of the hybrid region may be a little shorter than the stretch of unwound DNA. Probably the RNA-DNA hybrid is ~12 bp long.

As the enzyme leaves the area, the DNA duplex reforms, and the RNA is displaced as a free polynucleotide chain. About the last 50 ribonucleotides added to a growing chain are complexed with DNA and/or enzyme at any moment.

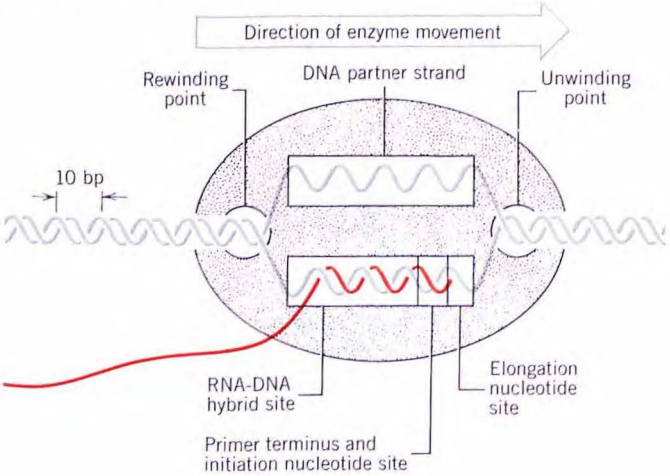
We still do not really understand the topology of unwinding and rewinding during transcription, but the ability of purified RNA polymerase to transcribe double-stranded DNA *in vitro* implies that the reaction depends on an intrinsic property of the enzyme. Unwinding and rewinding requires the strands of DNA to revolve about one another. One possibility is that the DNA revolves in the unwinding sense ahead of the enzyme, and revolves in the opposite sense behind it. This could require assistance *in vivo* from other enzymatic activities to adjust the topology of the DNA.

Figure 9.2
Bacterial RNA polymerase covers ~60 bp of DNA and has several active centers.

During elongation the reacting groups are held in two sites. The location occupied by the incoming nucleoside triphosphate is the **elongation nucleotide site.** The position of the last nucleotide added to the chain defines the **primer terminus site.**

When transcription is initiated, of course, there is no primer terminus, and the very first nucleotide (usually a purine) enters the **initiation nucleotide site,** which must largely overlap with the primer terminus site. The first nucleotide incorporated into the chain retains all its 5′ triphosphate residues.

The length of the region of unwinding is exaggerated (~4 fold) for the purposes of illustration.



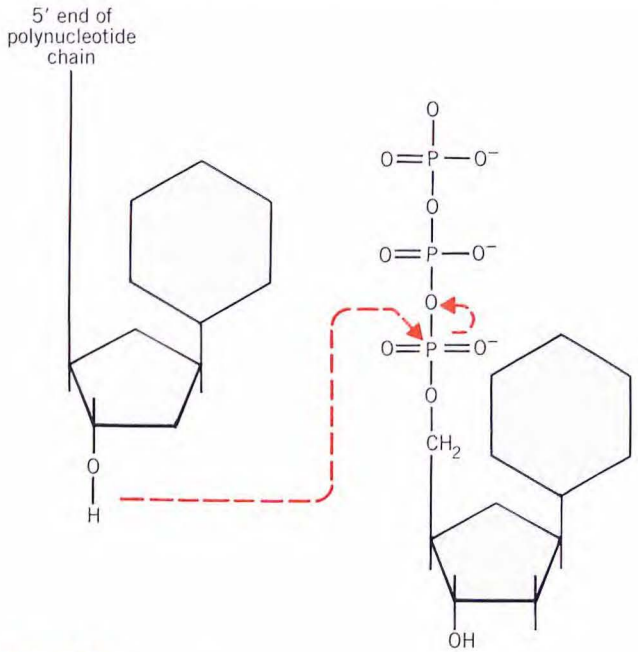


Figure 9.3
Phosphodiester bond formation involves a hydrophilic attack by the 3′—OH group of the last nucleotide of the chain on the 5′ triphosphate of the incoming nucleotide, with release of pyrophosphate.

All nucleic acids are synthesized from nucleoside 5′ triphosphate precursors. **Figure 9.3** shows the condensation reaction between the 5′ triphosphate group of the incoming nucleotide and the 3′—OH group of the last nucleotide to have been added to the chain. The incoming nucleotide loses its terminal two phosphate groups (γ and β); its α group is used in the phosphodiester bond linking it to the previous nucleotide.

The core enzyme must hold the two reacting groups in the proper apposition for phosphodiester bond formation; then, once they are covalently linked, it moves one base farther along the DNA template so that the reaction can be repeated. The reaction rate is fast, ~40 nucleotides/second at 37°C (see Chapter 8).

The acceptability of an incoming nucleotide is judged by its base pairing with the template strand of DNA, an action apparently supervised by the enzyme. Probably the site has a structure that allows phosphodiester bond formation to proceed only when the nucleotide is properly base paired with DNA. Presumably the nucleotide is expelled if its ability to base pair is deemed inadequate; then another can enter.

Our knowledge of the topology of the core enzyme is really very primitive, and the best we can do at present is to make a diagrammatic representation of the sites defined by the various enzymatic functions, as illustrated in Figure 9.2. None of these sites has yet been physically located on the polypeptide subunits. However, there is some general information about the roles of individual subunits.

Two types of antibiotic both act on the β subunit, as defined by the location of mutations conferring resistance (see Table 9.3). The **rifamycins** (of which rifampicin is the most used) prevent initiation, acting prior to formation of the first phosphodiester bond. **Streptolydigins** inhibit chain elongation. The β subunit is the target for both types of antibiotic; also it is labeled by certain affinity analogs of the nucleoside triphosphates. Together these results suggest that the β subunit may be involved in binding the nucleotide substrates.

Heparin is a polyanion that binds to the β′ subunit and inhibits transcription *in vitro*. Heparin competes with DNA for initially binding the polymerase. The β′ subunit is the most basic, which would fit with a role in template binding.

The α subunit has no known role. However, when phage T4 infects *E. coli*, the α subunit is modified by ADP-ribosylation of an arginine. The modification is associated with a reduced affinity for the promoters formerly recognized by the holoenzyme, so the α subunit might play a role in promoter recognition.

These assignments of individual functions are very primitive; probably each subunit contributes to the activity of the core enzyme as a whole, and we cannot compartmentalize its actions.

Why does bacterial RNA polymerase require a large, multimeric structure? The existence of much smaller RNA polymerases, comprising single polypeptide chains coded by certain phages, demonstrates that the apparatus required for RNA synthesis can be much smaller than that of the host enzyme.

These enzymes give some idea of the "minimum" apparatus necessary for transcription. They recognize a very few promoters on the phage DNA; and they have no ability to change the set of promoters to which they respond. Thus they are limited to the intrinsic abil-

ity to recognize a very few specific DNA binding sequences and to synthesize RNA. How complex are they?

The RNA polymerases coded by the related phages T3 and T7 are single polypeptide chains of ~11,000 daltons each. They synthesize RNA very rapidly (at rates of ~200 nucleotides/second at 37°C). The initiation reaction shows very little variation.

By contrast, the enzyme of the host bacterium can transcribe any one of many (>1000) transcription units. Some of these units are transcribed directly, with no further assistance. But many units can be transcribed only in the presence of further protein factors. Some of these factors are specific for a single transcription unit; others are involved in coordinating transcription from many units. Certain phages induce general changes in the affinity of host RNA polymerase, so that it stops recognizing host genes and instead initiates at phage promoters.

So the host enzyme requires the ability to interact with a variety of host and phage functions that modify its intrinsic transcriptional activities. The complexity of the enzyme may therefore at least in part reflect its need to interact with a multiplicity of other factors, rather than any demand inherent in its catalytic activity.

EUKARYOTIC RNA POLYMERASES CONSIST OF MANY SUBUNITS

The transcription apparatus of eukaryotic cells is more complex and less well defined than that of bacteria. There are three nuclear RNA polymerases, occupying different locations, each with a complex subunit structure. *Each enzyme is responsible for transcribing a different class of genes.* Their general properties are defined in **Table 9.2.**

The most prominent RNA-synthesizing activity is the enzyme RNA polymerase I, which resides in the nucleolus and is responsible for transcribing the genes coding for rRNA. It accounts for most cellular RNA synthesis.

The other major enzyme is RNA polymerase II, located in the nucleoplasm (the part of the nucleus excluding the nucleolus). It represents most of the rest of the cellular activity and is responsible for synthesizing heterogeneous nuclear RNA (hnRNA), the precursor for mRNA.

A minor enzyme activity is RNA polymerase III. This nucleoplasmic enzyme synthesizes tRNAs and many of the small nuclear RNAs.

Inhibitors of transcription have been useful in distinguishing between the enzymes. Different inhibitors act on prokaryotic and eukaryotic enzymes. The properties of some common inhibitors are summarized in **Table 9.3.**

A major distinction between the eukaryotic enzymes is drawn from their response to the bicyclic octapeptide **α-amanitin.** In cells from origins as divergent as animals, plants, and insects, the activity of RNA polymerase II is rapidly inhibited by low concentrations of α-amanitin. In cells from all origins, the RNA polymerase I enzyme is not inhibited. The response of RNA polymerase III to α-amanitin has not been so well conserved; in animal cells it is inhibited by high levels, but in yeast and insects it is not inhibited.

The crude enzyme activities all are large proteins, appearing as aggregates of 500,000 daltons or more. Their subunit compositions are complex. Each enzyme has two large subunits, generally one ~200,000 daltons and one ~140,000 daltons. There are <10 smaller subunits, ranging in size from 10,000 to 90,000 daltons. We do not know whether any of the subunits found in the different enzymes are the same.

Table 9.2
Eukaryotic nuclei have three RNA polymerases.

Enzyme	Location	Product	Relative Activity	α-Amanitin Sensitivity
RNA polymerase I	nucleolus	ribosomal RNA	50-70%	not sensitive
RNA polymerase II	nucleoplasm	hnRNA	20-40%	sensitive
RNA polymerase III	nucleoplasm	small RNA	~10%	species-specific

Table 9.3
Inhibitors of transcription act preferentially on particular enzymes.

Inhibitor	Target Enzyme	Inhibitory Action
Rifamycin	bacterial holoenzyme	binds to β to prevent initiation
Streptolydigin	bacterial core enzyme	binds to β to prevent elongation
Actinomycin D	eukaryotic pol I	binds to DNA & prevents elongation
α-Amanitin	eukaryotic pol II	binds to RNA polymerase II

Do the enzyme preparations represent the basic transcription apparatus, essentially similar in all cells and subject to regulation by further protein factors? Or do they include such factors as well as a basic catalytic apparatus?

Because it is not yet possible to reconstitute active RNA polymerase from the subunits of any of these enzymes, we have no evidence as to whether all of the protein subunits are integral parts of each enzyme. We do not know which subunits may represent catalytic activities and whether others may be involved in regulatory functions.

The route to investigating this question is to use isolated enzyme preparations to transcribe defined templates *in vitro*. Several heterologous reactions have been characterized, in which an RNA polymerase II preparation from one cell type and species is used to transcribe a gene that is active in a different cell type and species. The success of such experiments indicates that neither tissue- nor species-specific features are involved in promoter recognition *per se*.

Of course, this conclusion does not exclude the possibility that further protein factors or other sequences are involved in modulating the reaction (especially increasing its efficiency) in the natural situation. Factors have been found that are necessary to allow particular RNA polymerases to initiate transcription of a subset of their target genes. Sets of factors may be needed to transcribe particular groups of genes.

The overall complexity of the eukaryotic transcription apparatus is being defined only system by system; all we can say at present is that multimeric enzymes are needed for each of the three classes of RNA synthesis. Whether all the components of these enzymes are essential, and how many other proteins are needed, remains to be seen.

Because of these uncertainties, relatively crude, "dirty" systems may offer more chance of characterizing transcription *in vitro* than purified "clean" systems; too much purification may remove the very factors that we need to characterize!

The RNA polymerase activities of mitochondria and chloroplasts appear to be smaller and distinct from the nuclear enzymes. Of course, the organelle genomes are much smaller, the resident polymerase needs to transcribe only a few genes, and the control of transcription is likely to be very much simpler (if existing at all). So these enzymes may be analogous to the phage enzymes that have a single fixed purpose and do not need the ability to respond to a more complex environment.

BACTERIAL SIGMA FACTOR CONTROLS BINDING TO DNA

The function of the sigma factor is to ensure that bacterial RNA polymerase binds stably to DNA *only at promoters, not at other sites.*

The core enzyme itself has an affinity for DNA, in which electrostatic attraction between the basic protein and the acidic nucleic acid plays a major role. Probably this general ability to bind to any DNA, irrespective of its particular sequence, is a feature of all proteins that have specific binding sites on DNA (see Chapter 10).

Any sequence of DNA that is bound by RNA polymerase in this general binding reaction is described as a **loose binding site.** The enzyme·DNA complex is described as **closed,** because the DNA remains strictly in the double-stranded form. A closed complex is stable; the half-life for dissociation of the enzyme from DNA is ~60 minutes.

Sigma factor introduces a major change in the affinity of RNA polymerase for DNA. *The holoenzyme has a drastically reduced ability to recognize loose binding sites*—that is, to bind to any general sequence of DNA. The association constant for the reaction is reduced by a factor of ~10,000 and the half-life of the complex is <1 second. Thus sigma factor destabilizes the general binding ability very considerably.

But sigma factor also *confers the ability to recognize specific binding sites.* The holoenzyme binds to promoters very tightly, with an association constant increased from that of core enzyme by (on average) 1000 times and a half-life of several hours.

The association constant can be quoted only as an average, because there is (roughly) a hundredfold variation in the rate at which the holoenzyme binds to different promoter sequences; this is an important factor in determining the efficiency of a promoter in initiating transcription.

Recognition of promoters by holoenzyme passes through two stages, illustrated in **Figure 9.4.** The holoenzyme·promoter reaction starts in the same way as the loose binding reaction, by forming a closed complex. But then this complex is converted into an **open complex** by the "melting" of a short region of DNA within the sequence bound by the enzyme. The series of events leading to formation of an open complex is called **tight binding.**

Both the closed and open associations of RNA polymerase with DNA are described as **binary complexes.** The next step is to incorporate the first two nucleotides; then a phosphodiester bond forms between them. This creates a **ternary complex** of polymerase·DNA·nascent RNA. The ternary complex forms extremely rapidly when RNA polymerase finds a promoter, so the binary complex has an exceedingly transient existence.

Sigma factor is involved only in initiation. It is released from the core enzyme when RNA synthesis has been initiated.

The core enzyme in the ternary complex is very tightly bound to DNA. It is essentially "locked in" until elongation has been completed. When transcription terminates, the core enzyme is released from DNA as a free protein tetramer. It must then find another sigma factor in order to undertake a further cycle of transcription. The ratio of sigma factors to core enzymes is about one third.

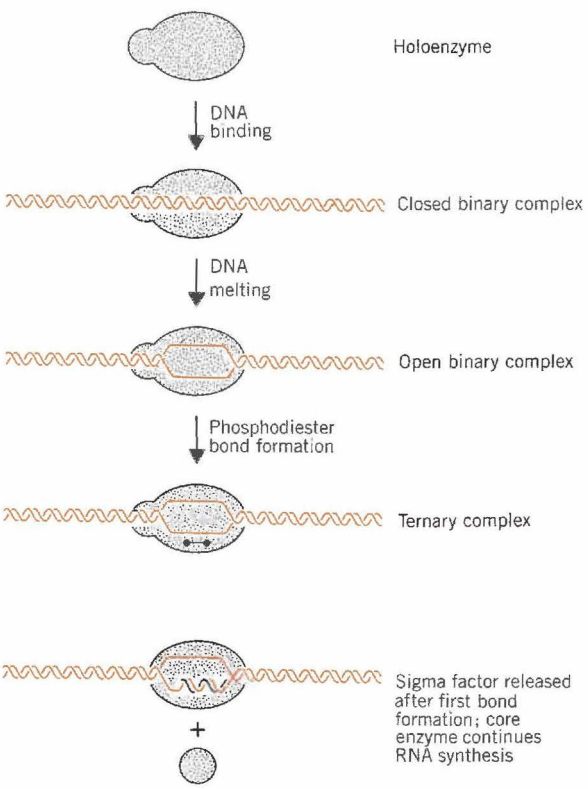
RNA polymerase may find promoters on DNA by the process of trial and error illustrated in **Figure 9.5.**

The excess core enzyme exists largely in the form of closed loose complexes, because the enzyme enters into them rapidly and leaves them slowly.

By contrast, the holoenzyme very rapidly associates with, and dissociates from, loose binding sites. So it is likely to continue to make and break a series of closed complexes in an agitated manner until (by chance) it encounters a promoter. Then its recognition of the specific sequence will allow tight binding to occur by formation of an open complex.

Three steps are needed for RNA polymerase to move from one binding site to another on DNA. It must dissociate from the first binding site, find the second site, and associate with it. Movement from one site to an-

Figure 9.4
Initiating transcription requires several steps, during which a closed binary complex is converted to an open form and then into a ternary complex.



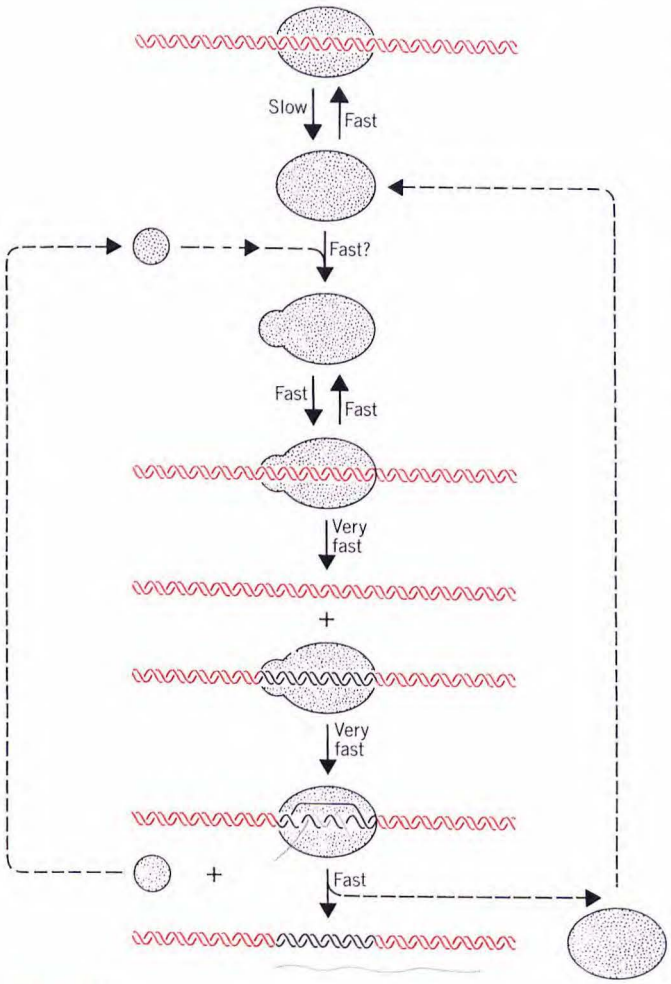


Figure 9.5

Sigma factor and core enzyme recycle at different points in transcription.

Sigma factor is released as soon as a ternary complex has formed at an initiation site; it becomes available for use by another core enzyme. The core enzyme is released at termination; it must either find a sigma and form a holoenzyme that can bind stably only at promoters or it must bind to loose sites on DNA.

other is limited by the speed of diffusion through the medium. The rate constant for binding promoters is very close to this limit, so close, in fact, that insufficient time is left for association and dissociation from loose binding sites during a random search cycle.

RNA polymerase may therefore use another means to seek its binding sites, possibly the direct displacement of one bound sequence by another. Instead of moving about by leaving one binding site and diffusing to another, the enzyme is likely to take hold of one sequence of DNA, exchange it very rapidly for another, and continue to exchange sequences in this promiscuous manner until a promoter is found. Then the enzyme forms a stable, open complex, after which initiation occurs. The search process becomes much faster because association and dissociation are virtually simultaneous, and time is not spent commuting between sites.

The existence of a cycle in which sigma factor and core enzyme come together only temporarily solves the dilemma of RNA polymerase in reconciling its needs for initiation with those for elongation. It is a dilemma because initiation requires tight binding *only* to particular sequences (promoters), while elongation requires close association with *all* sequences along which the enzyme must progress.

Core enzyme has a high intrinsic affinity for DNA, which is increased by the presence of nascent RNA. But its affinity for loose binding sites remains too high to allow the enzyme to find promoters efficiently; the associations and dissociations involved in the trial and error of finding a tight binding site could take many hours.

By reducing the stability of the loose complexes, sigma allows the process to occur much more rapidly; and by stabilizing the association at tight binding sites, the factor drives the reaction irreversibly into the formation of open complexes. To avoid becoming paralyzed by its specific affinity for the promoter, the enzyme releases sigma, and thus reverts to a general affinity for all DNA, irrespective of sequence, that suits it to continue transcription.

How does sigma change the enzyme so that promoters are specifically recognized? As an independent polypeptide, sigma does not seem to bind DNA, but when holoenzyme forms a tight binding complex, σ contacts the DNA in the region of the initial melting. The inability of free sigma factor to recognize promoter sequences may be important: if σ could freely bind to promoters, it might block holoenzyme from initiating transcription. We do not know what role the core subunits play in promoter recognition; sigma may change the conformation of core enzyme so that its ability to recognize DNA is altered, but the sequence specificity of this effect is not clear.

TRANSCRIPTION UNITS EXTEND FROM PROMOTERS TO TERMINATORS

Initiation of transcription is a critical point for controlling gene expression. Often the decision on whether or not to initiate at a particular promoter is the major step in determining whether a gene should be expressed. What controls the ability of RNA polymerase to initiate at a particular promoter? We can start by thinking about promoters in two general classes.

Some promoters can be recognized by RNA polymerase alone; in these cases, an accessible promoter will always be transcribed. Promoter availability may be determined by extraneous proteins, which may either act directly at the promoter to block access by RNA polymerase, or may function indirectly by controlling the structure of the genome in the region.

Other promoters are not by themselves adequate to support transcription; ancillary protein factors are needed for initiation to occur. The additional protein factors act by recognizing sequences of DNA that may be close to, or overlap with, the sequence bound by RNA polymerase itself.

As sequences of DNA whose function is to be *recognized by proteins,* a promoter and any adjacent control sites differ from other sequences whose role is exerted by being transcribed or translated. The information for promoter function is provided directly by the DNA sequence: its structure is the signal. By contrast, expressed regions gain their meaning only after the information is transferred into the form of some other nucleic acid or protein.

The key question in examining the interaction between an RNA polymerase and its promoter is how a protein can recognize a specific sequence of DNA. Does the enzyme have an active site that distinguishes the chemical structure of a particular sequence of bases in the DNA double helix? How specific are its requirements? Promoters vary in their affinities for RNA polymerase; this can be an important factor in controlling the frequency of initiation and thus the extent of gene expression.

Binding at the promoter is rapidly followed by initiation at the startpoint; then RNA polymerase continues along the template until it reaches a terminator sequence. This action defines a **transcription unit** that extends from the promoter to the terminator. The critical feature of the transcription unit, depicted in **Figure 9.6,** is that it constitutes a stretch of DNA *expressed via the production of a single RNA molecule.* A transcription unit may include only one or several genes.

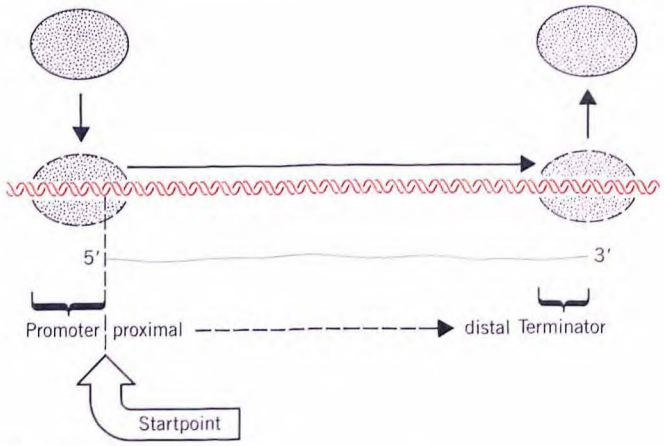


Figure 9.6
A transcription unit is a sequence of DNA transcribed into a single RNA, starting at the promoter and ending at the terminator.
Regions close to the promoter are described as **proximal,** while those toward the terminator are described as **distal.**

Sequences prior to the startpoint are described as **upstream** of it; those after the startpoint (within the transcribed sequence) are **downstream** of it. Sequences are conventionally written so that transcription proceeds from left (upstream) to right (downstream). This corresponds to writing the mRNA in the usual 5′ to 3′ direction.

Often the DNA sequence is written just to show the strand whose sequence is the same as the RNA. Base positions are numbered in both directions away from the startpoint, which is assigned the value +1; numbers are increased going downstream. The base before the startpoint is numbered −1, and the negative numbers increase going upstream.

The immediate product of transcription is called the **primary transcript.** It would consist of an RNA extending from the promoter to the terminator, possessing its original 5′ and 3′ ends. However, the primary transcript is almost always unstable and therefore very difficult to characterize *in vivo.* In prokaryotes, it is rapidly degraded (mRNA) or cleaved to give mature products (rRNA and tRNA). In eukaryotes, it is modified at

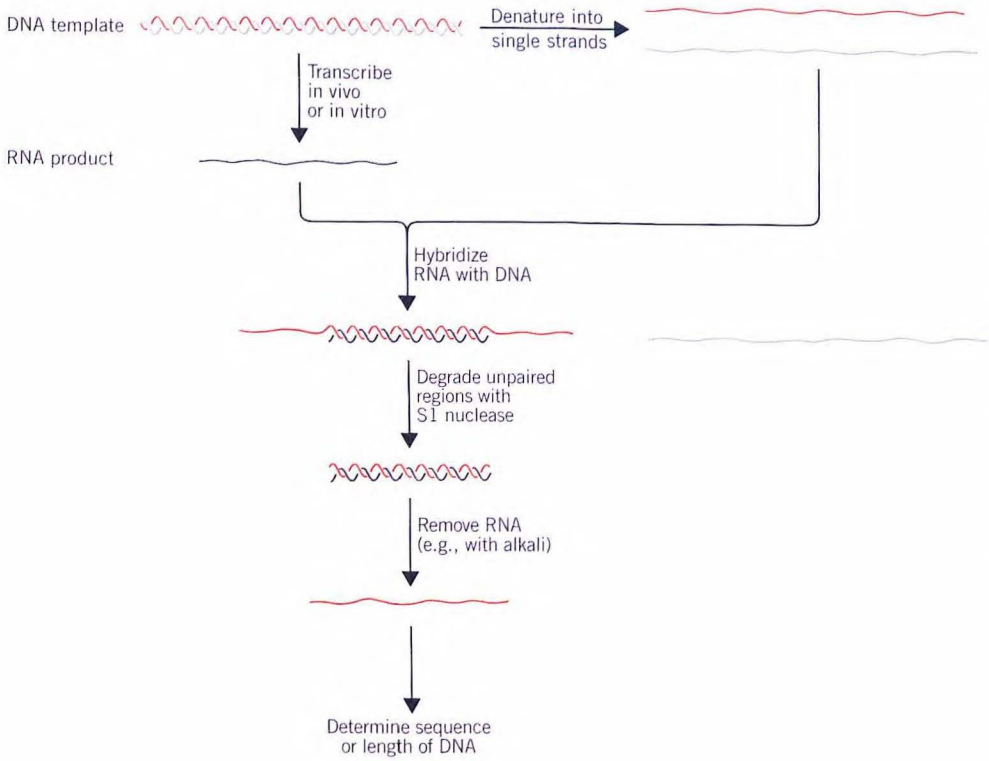


Figure 9.7
Startpoints can be determined by comparing the RNA product with the DNA template.
When the RNA is hybridized with the denatured strands of DNA, it forms an RNA-DNA hybrid with the strand that acted as its template; the other strand of DNA remains unpaired. Treatment with the enzyme S1 nuclease, which specifically degrades single-stranded DNA, destroys both the unpaired strand and regions of the template strand beyond the transcription unit. The RNA of the RNA-DNA hybrid can be removed. Then the sequence of the DNA can be determined or its length can be used to locate the position of the RNA on the template.

the ends (mRNA) and/or cleaved to give mature products (all RNA).

The startpoint is defined as the base pair in DNA corresponding to the first nucleotide incorporated into RNA. A common method for identifying the startpoint by hybridizing the transcript with its DNA template is illustrated in **Figure 9.7**. In principle, it consists of degrading all the DNA that cannot hybridize with the RNA, and then determining the sequence of the surviving DNA.

To define the startpoint, it is necessary to examine the 5′ end of the primary transcript. Because of the difficulty of isolating primary transcripts *in vivo*, most information about startpoints is provided by *in vitro* studies. However, in those cases in which the authentic 5′ end *has* been identified *in vivo* (by its possession of a triphosphate terminus), it is identical with the 5′ end generated by transcription *in vitro*. In particular, the capped terminus of eukaryotic mRNA appears to coincide with the startpoint.

Usually the startpoint is a unique base pair, sometimes it consists of either of two adjacent base pairs, and occasionally it may involve any one of several adjacent positions. It always represents a defined part of the promoter, usually within the sequence bound by RNA polymerase, in some cases at one extremity.

PROMOTERS INCLUDE CONSENSUS SEQUENCES

One way to design a promoter would be for an invariant sequence of DNA to be recognized by RNA polymerase. In the bacterial genome, the minimum length that could provide an adequate signal is 12 bp. (Any shorter sequence is likely to occur—just by chance—a sufficient number of additional times to provide false signals.) The 12 bp sequence need not be contiguous; and, in fact, if a specific number of base pairs separates two constant shorter sequences, their combined length could be less than 12 bp, since the *distance* of separation itself provides a part of the signal (even if the intermediate *sequence* is itself irrelevant).

Attempts to identify the features in DNA that are necessary for RNA polymerase binding started by comparing the sequences of different promoters. Any essential nucleotide sequence should be present in all the promoters. Such a sequence is said to be **conserved.** However, a conserved sequence need not necessarily be conserved at every single position; some variation may be permitted. How do we analyze a sequence of DNA to determine whether it is sufficiently conserved to constitute a recognizable signal?

Putative DNA recognition sites can be defined in terms of an idealized sequence that represents the base most often present at each position. A **consensus sequence** is defined by aligning all known examples so as to maximize their homology. For a sequence to be accepted as a consensus, each particular base must be reasonably predominant at its position, and most of the sequences must be related to the consensus by rather few substitutions, say, 1 or 2.

More than 100 promoters have been sequenced in *E. coli*, and a striking feature is the *lack of any extensive conservation of sequence* over the 60 bp associated with RNA polymerase. The sequence of much of the binding site may actually be irrelevant. But some short stretches within the promoter are conserved.

The startpoint is usually (>90% of the time) a purine. It is quite common for the startpoint to be the central base in the sequence CAT, but the conservation of this triplet is not great enough to regard it as an obligatory signal.

Just upstream of the startpoint, a 6 bp region is recognizable in almost all promoters. The consensus sequence is **TATAAT;** sometimes it is called the **Pribnow box**. The conservation of the base at each position of the Pribnow box varies from 45% to 100%. The consensus can be summarized in the form

$$\mathbf{T}_{80}\ \mathbf{A}_{95}\ \mathbf{t}_{45}\ \mathbf{A}_{60}\ \mathbf{a}_{50}\ \mathbf{T}_{96}$$

where the subscript denotes the percent occurrence of the most frequently found base.

(Capital letters are used to indicate bases conserved >54%; lower case letters are used to indicate bases not so well conserved, but nonetheless present more often than predicted from a random distribution. A position at which there is no discernible preference for any base would be indicated by N.)

If the frequency of occurrence indicates likely importance in binding RNA polymerase, we would expect the initial highly conserved TA and the final almost completely conserved T in the Pribnow box to be the most important bases.

The center of the Pribnow box generally is close to 10 bp upstream of the startpoint. Sometimes it is therefore called the **−10 sequence.** The center of the hexamer varies among promoters from position −18 to −10.

Similarities of sequence also occur at another location, centered ~35 bp upstream of the startpoint. This is called the **−35 sequence.** The consensus is **TTGACA;** in more detailed form, the conservation is

$$\mathbf{T}_{82}\ \mathbf{T}_{84}\ \mathbf{G}_{78}\ \mathbf{A}_{65}\ \mathbf{C}_{54}\ \mathbf{a}_{45}$$

The distance separating the −35 and −10 sites is between 16 and 18 bp in 90% of promoters; in the exceptions, it may be as little as 15 or as great as 20 bp. Although the actual sequence in the intervening region may be unimportant, the distance may be critical in holding the two sites at the appropriate separation for the geometry of RNA polymerase.

A major source of information about promoter function is provided by mutations. Mutations in promoters affect the level of expression of the gene(s) they control, without altering the gene products themselves. Most are identified in the form of bacterial mutants that have lost, or have very much reduced, transcription of the adjacent genes. They are known as **down mutations.** Less often, mutants are found in which there is increased transcription from the promoter. They are called **up mutations.**

Almost all of the point mutations that affect promoter function fall within the two consensus sequences. Occasionally a mutation is found just upstream of either consensus sequence (see Figure 9.11). The bases present at other positions in the vicinity are clearly much less important or even irrelevant in most promoters. In addition to these mutations, deletions or insertions between the consensus sequences may alter their separation.

It is important to remember that "up" and "down" mutations are defined relative to the *usual* efficiency with which a particular promoter functions. This varies widely. So a change that is recognized as a down mutation in one promoter might never have been isolated

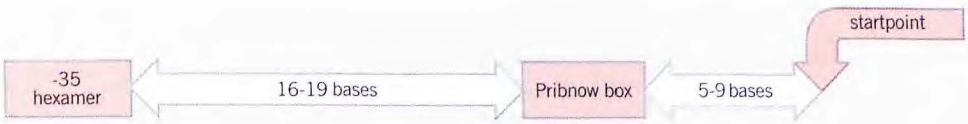


Figure 9.8
A typical promoter has three components, consisting of consensus sequences at −35 and −10, and the startpoint.

in another (which in its wild-type state could be even less efficient than the mutant form of the first promoter). Thus information gained from studies *in vivo* simply identifies the overall nature of the change caused by mutation.

A very few promoters lack a recognizable version of one of the consensus sequences. In at least some of these cases, the promoter cannot be recognized by RNA polymerase alone, for the reaction requires the intercession of ancillary proteins. Possibly their reaction with adjacent sequences overcomes the deficiency in the promoter.

We do not yet know all the details of the recognition reaction; all that can be said firmly now is that the "typical" promoter can use the −35 and −10 sequences to be recognized by RNA polymerase. From their absence from exceptional promoters, we realize that other means also can be used for recognition, but we do not know how many alternatives there are, nor exactly how they substitute for the absence of the consensus sequences.

Is the most effective promoter one that has the consensus sequences themselves? This expectation is borne out by the simple rule that up mutations usually increase homology with one of the consensus sequences or bring the distance between them closer to 17 bp. Down mutations usually decrease the resemblance of either site with the consensus or make the distance between them more distant from 17 bp. Down mutations tend to be concentrated in the most highly conserved positions, which confirms their particular importance as the main determinant of promoter efficiency.

Occasional exceptions to these rules demonstrate that promoter efficiency cannot be predicted entirely from homology with the consensus. Virtually all actual promoters vary from the consensus, so the neighbors of any particular base may differ from promoter to promoter, even if the base itself is conserved. We cannot predict the effects of context; it is possible that a nonconsensus base at some position can function effectively in one promoter but not in another. It is also possible that an important feature may be the *exclusion* of some base from a particular position, rather than the presence of one particular nucleotide.

With these caveats, however, *we can define the optimal promoter as a sequence consisting of the −35 hexamer, separated by 17 bp from the −10 hexamer, lying 7 bp upstream of the startpoint.* The structure of an optimal promoter is illustrated in **Figure 9.8.**

RNA POLYMERASE CAN BIND TO PROMOTERS *IN VITRO*

The tight binding sites where RNA polymerase forms stable initiation complexes lie within promoters. Their sequences can be recovered by the protocol summarized in **Figure 9.9,** in which RNA polymerase is bound *in vitro* to a DNA fragment containing a promoter. Then digestion with the enzyme DNAase is used to degrade all the regions of DNA that are not protected by the RNA polymerase.

The recovered fragments are ~44 bp long. The polymerase can initiate transcription if left attached to the protected fragment, to synthesize a short RNA of ~20 bases, terminated when the polymerase "runs off" the end of the fragment. This shows that the protected fragment corresponds to the sequence at which transcription is initiated, and locates the startpoint in the center of the binding site. The protected fragment extends from about −20 upstream to +20 downstream. However, RNA polymerase cannot *rebind* to the pro-

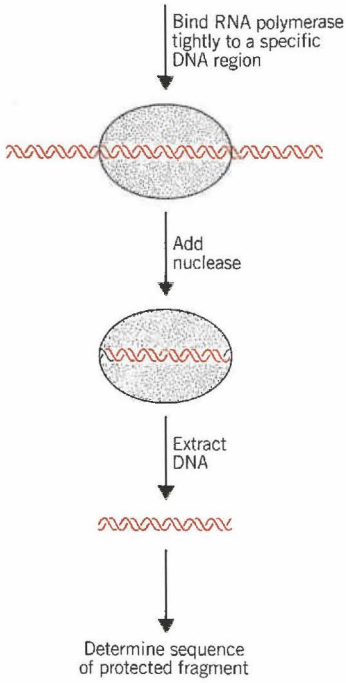


Figure 9.9

DNA binding sites for RNA polymerase can be recovered because they are protected by the enzyme against degradation by nucleases.

Figure 9.10

Footprinting identifies DNA binding sites for proteins by their protection against nicking.

The principle is the same as that involved in DNA sequencing; partial cleavage of an end-labeled molecule at a susceptible site creates a fragment of unique length. In a free DNA, *every* susceptible bond position is broken in one or another molecule. But when the DNA is complexed with a protein, the region covered by the DNA-binding protein is protected in every molecule. So two reactions are run in parallel: a control of pure DNA, and an experimental mixture containing the protein.

When the strands are separated and electrophoresed, a radioactive band is produced by each fragment that retains a labeled end. The position of the band corresponds to the number of bases in the fragment. The shortest fragments move the fastest, so distance from the labeled end is counted up from the bottom of the gel (see Figure 4.12). In the control, every bond is broken, generating a series of bands, one representing each base. In the figure, 31 bands can be counted. In the protected fragment, bonds cannot be broken in the region bound by the protein, so bands representing fragments of the corresponding sizes are not generated. The absence of bands 10–20 in the figure identifies a protein-binding site covering the region located 10–20 bases from the labeled end of the DNA.

Instead of using a single gel that analyzes DNA simply by length, both the control and experimental mixtures can be treated to generate four sequencing gels (see Figure 4.13). Comparison of the two gel sets allows the sequence to be "read off" directly, thus identifying the nucleotide sequence of the binding site.

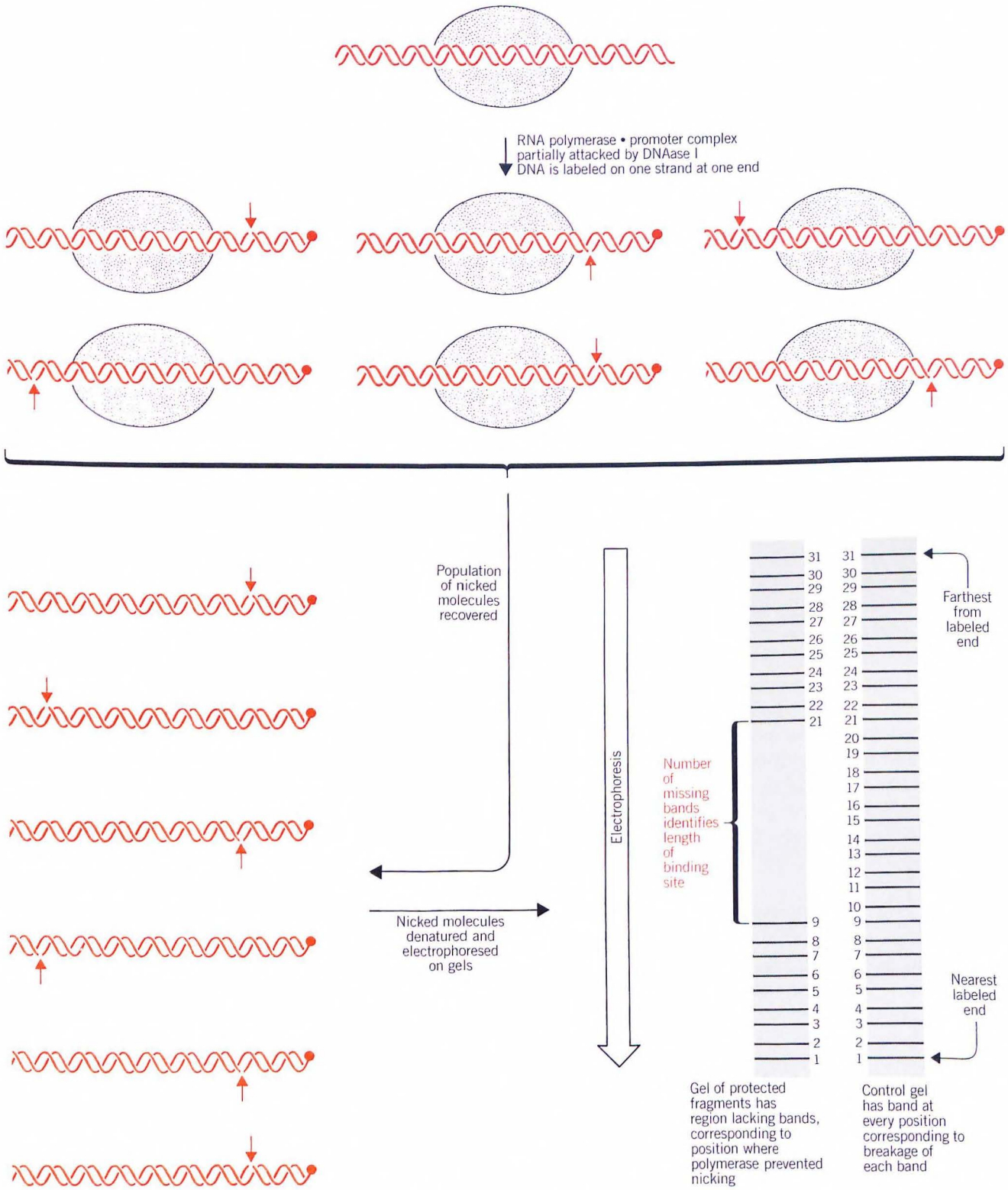
tected (−20 to +20) fragments, so additional sequences must be needed for the initial binding.

The ability of RNA polymerase to recognize DNA can be characterized in more detail by **footprinting.** A sequence of DNA bound to RNA polymerase (or any other protein) is *partially* digested with an endonuclease, an enzyme that attacks individual phosphodiester bonds *within* a nucleic acid. Under appropriate conditions, every phosphodiester bond that is in principle accessible to the nuclease is broken in some, but not in all, molecules of DNA. Only if RNA polymerase blocks access of the nuclease to DNA will a particular bond fail to be broken at all.

The positions that are cleaved are recognized by using DNA labeled on one strand at one end only. As **Figure 9.10** shows, following the nuclease treatment, the broken DNA fragments are recovered and electrophoresed on a gel that separates them according to length. For every susceptible bond position, a band is found on the gel, corresponding to the distance from the site of breakage to the labeled end. For every position protected against cleavage by RNA polymerase, a band is missing. Each of the two strands of DNA can be analyzed separately by using it as the labeled strand. By combining footprinting with DNA sequencing, the nucleotide sequence of the binding site can be determined as well as its position.

Analyzed by this technique, the RNA polymerase binding site turns out to be more extensive than detected by the recovery of protected fragments. It extends for roughly an extra 30 bp upstream, that is, from about −50 to +20. This stretch includes all the sequences needed for binding and initiation. The −10 consensus sequence lies in the center of the tightly protected region; the −35 consensus sequence lies toward one end of the binding site, but is not tightly protected.

The two strands of DNA are not equally well protected in all regions of the promoter, especially at the ends of the binding site. This implies that RNA poly-


RNA polymerase • promoter complex
partially attacked by DNAase I
DNA is labeled on one strand at one end
Population of nicked molecules recovered
Nicked molecules denatured and electrophoresed on gels
Electrophoresis
Number of missing bands identifies length of binding site
31
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
Farthest from labeled end
Nearest labeled end
Gel of protected fragments has region lacking bands, corresponding to position where polymerase prevented nicking
Control gel has band at every position corresponding to breakage of each band

merase has an asymmetric conformation when bound to DNA, which accords with the need to transcribe only one strand of DNA.

Somewhat similar results can be obtained with digestion using an **exonuclease**, an enzyme that attaches to an end of DNA and degrades it continuously. If a protein is bound to the DNA, digestion stops when the nuclease encounters the protein. The limits of the promoter can be defined in terms of the point on each side at which the exonuclease is blocked from proceeding, about −44 upstream and +20 downstream.

So RNA polymerase binds asymmetrically to a region of DNA stretching from <50 bp upstream to a point ~20 bp downstream. The terminal 25–30 bp of the binding site upstream must be more loosely associated with the enzyme. This region is bound sufficiently well to be protected against gentle endonucleolytic or exonucleolytic cleavage (in footprinting), but cannot withstand the stronger conditions used to retrieve intact fragments. These data argue in favor of a model in which RNA polymerase binds to a single, continuous sequence of DNA where some points of contact are more important than others.

To determine the absolute effects of promoter mutations, we must measure the affinity of RNA polymerase for wild-type and mutant promoters *in vitro*. There is ~100 fold variation in the rate at which RNA polymerase binds to different promoters *in vitro,* which correlates well with the frequency of transcription when their genes are expressed *in vivo*. Taking this analysis further, we can investigate the stage at which a mutation influences the capacity of the promoter. Does it change the affinity of the promoter for binding RNA polymerase? Does it leave the enzyme able to bind but unable to initiate? Is the influence of an ancillary factor altered?

By measuring the rate constants for formation of a closed complex and its conversion to an open complex, we can dissect the two stages of the initiation reaction. Down mutations in the −35 sequence reduce the rate of closed complex formation, but do not inhibit the conversion to an open complex. On the other hand, down mutations in the −10 sequence do not slow the initial formation of a closed complex, but they slow its conversion to the open form. These results suggest that *the function of the −35 sequence is to provide the signal for recognition by RNA polymerase, while the −10 sequence allows the complex to convert from closed to open form.*

The consensus sequence of the −10 site consists exclusively of A-T base pairs, which may assist the initial melting of DNA into single strands. The lower energy needed to disrupt A-T pairs compared with G-C pairs means that a stretch of A-T pairs demands the minimum amount of energy for strand separation.

The points at which RNA polymerase contacts the promoter can be identified by treating RNA polymerase·promoter complexes with reagents that modify particular bases. The presence of the enzyme may either increase or decrease the availability of a particular base (relative to a control consisting of the DNA by itself). These changes in sensitivity reveal the geometry of the complex, as summarized in **Figure 9.11.**

The common feature of all the types of modification is that *they allow a breakage to be made at the corresponding bond in the polynucleotide chain.* The site of breakage can be identified by the same approach used in footprinting with endonucleases (see Figure 9.10). By labeling DNA at one end of one strand, each breakage generates an electrophoretic band of corresponding length. When the susceptibility of an RNA polymerase·DNA complex is compared with free DNA, some bands disappear, identifying sites at which the enzyme has protected the promoter against modification. Other bands may increase in intensity, identifying sites at which the DNA must be held in a conformation in which it is more exposed.

The reverse experiment can be performed by modifying the DNA *first;* then it is bound to RNA polymerase. Those DNA molecules that cannot bind RNA polymerase are recovered and treated in the usual way to generate strand breakages whose positions can be identified. This locates points at which prior modification *prevents* RNA polymerase from binding to DNA.

Such experiments show that the regions at −35 and −10 contain most of the contact points for the enzyme. Within these regions, the same sets of positions tend both to prevent binding if previously modified, and to show increased or decreased susceptibility to modification after binding. Figure 9.11 compares the points of contact with sites of mutation; although they do not coincide completely, they occur in the same limited

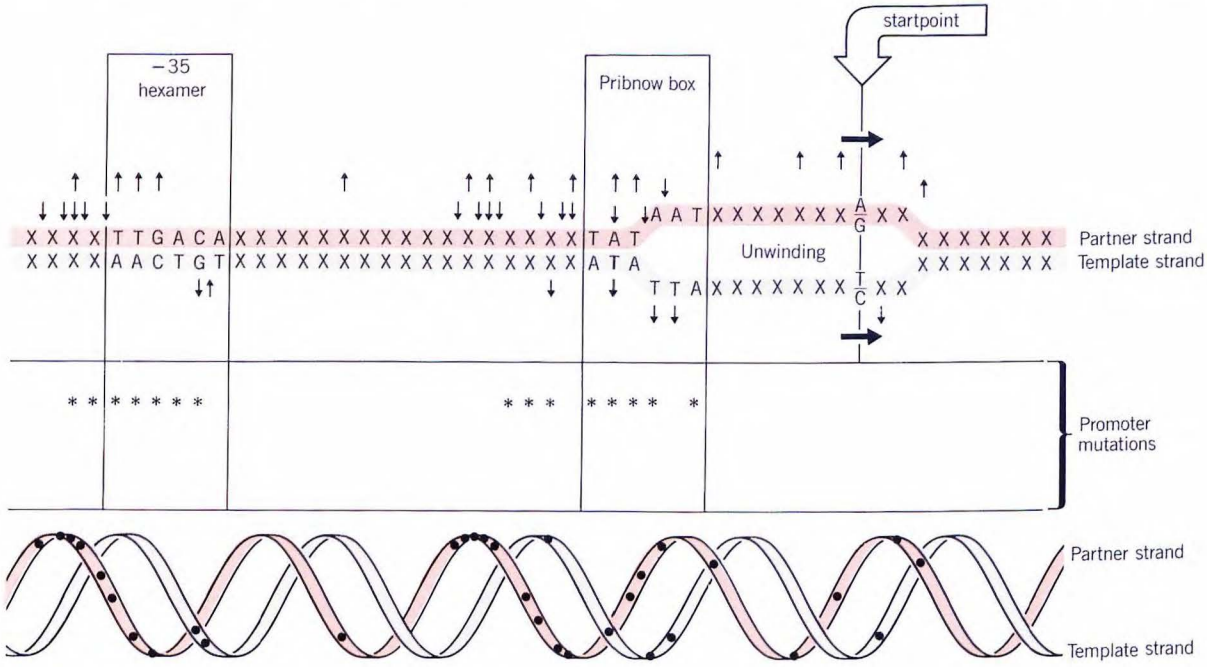


Figure 9.11

One face of the promoter contains the contact points for RNA.

The DNA sequence shows a typical promoter, with consensus sequences at −35 and −10, and a region for initial unwinding extending from within the Pribnow box to just past the startpoint.

Upper. Sites at which modification prevents RNA polymerase binding are shown by arrows pointing toward the double helix. Sites at which the DNA is protected by RNA polymerase against modification are shown by the arrows pointing away from the double helix. Arrows pointing at bases indicate modification of the base itself; arrows pointing between bases indicate modification of the connecting phosphodiester bond.

Center. Sites at which mutations affect promoter function are indicated by asterisks.

Lower. When the DNA is drawn as a double helix viewed from one side, as indicated diagrammatically, all the contact points lie on one face. Most lie on the partner strand (that is, not on the template strand).

region. At both consensus sites, the region of contact extends for 12–15 bp, somewhat longer than the conserved region.

It is noteworthy that the same *positions* in different promoters may provide the contact points, even though a different base is present. This indicates that there may be a common mechanism for RNA polymerase binding, although the reaction does not depend on the presence of particular bases at some of the points of contact. This model may explain why some of the points of contact are not sites of mutation. Also, not every mutation lies in a point of contact; could some influence the neighborhood without actually being touched by the enzyme?

It is especially significant that the experiments with prior modification identify *only* sites in the same region that is protected by the enzyme against subsequent modification. These two experiments measure different things. The first identifies all those sites that the enzyme must recognize in order to bind to DNA. The second recognizes all those sites that actually make contact in the binary complex. The protected sites include all the recognition sites and also some additional positions, which suggests that the enzyme first rec-

ognizes a set of sites necessary for it to "touch down," and then extends its points of contact to further sites.

A modification experiment allows the region of DNA that is unwound in the binary complex to be identified directly. When the strands of DNA are separated, the unpaired bases may become susceptible to reagents that cannot reach them in the double helix. The susceptibility of sites in the RNA polymerase·DNA binary complex therefore indicates that they lie in an unpaired region. Experiments using methylation of adenine or cytosine have implicated positions between −9 and +3 in the initial melting reaction. The region unwound during initiation therefore includes the right end of the −10 sequence and extends just past the startpoint. (This measure for the extent of strand separation is less than the 17 bp estimated by the overall degree of unwinding.)

Viewed in three dimensions, the points of contact upstream of the −10 sequence all lie on one face of DNA, as illustrated in Figure 9.11. These bases could be recognized in the initial formation of a closed binary complex. This would make it possible for RNA polymerase to approach DNA from one side and recognize that face of the DNA. As DNA unwinding commences, further sites that originally lay on the other face of DNA might be recognized and bound.

The importance of strand separation in initiating transcription is emphasized by the effects of supercoiling. Both prokaryotic and eukaryotic RNA polymerases can initiate transcription more efficiently *in vitro* when the template is supercoiled, presumably because the supercoiled structure requires less free energy for the initial melting of DNA in the initiation complex.

The involvement of this effect in controlling promoter activity in bacteria is shown by the effects of interfering with enzymes that influence the degree of supercoiling. Among the relevant enzymes are DNA gyrase, which *introduces* negative supercoils, and topoisomerase I, which *relaxes* (removes) negative supercoils (see Chapter 28). Inhibitors of DNA gyrase reduce transcription; mutations in topoisomerase I may increase transcription. Both effects are seen only at some promoters.

The nature of these effects is not entirely clear; bacteria evidently endeavor to set the degree of supercoiling between certain limits, because mutations in one enzyme that alter the level may be compensated by mutations in another to restore the balance. However, it does seem that the efficiency of some promoters may be dependent on a certain degree of supercoiling.

Why should some promoters be influenced by the extent of supercoiling while others are not? One possibility is that every promoter has a characteristic dependence on supercoiling, determined by its sequence. This would predict that some promoters have sequences that are easier to melt (and are therefore less dependent on supercoiling), while others have more difficult sequences (and have a greater need to be supercoiled). An alternative is that the location of the promoter might be important, if different regions of the bacterial chromosome have different degrees of supercoiling.

SUBSTITUTION OF SIGMA FACTORS MAY CONTROL INITIATION

E. coli RNA polymerase transcribes virtually all the bacterial genes, and must therefore recognize a wide spectrum of promoters, whose associated genes are transcribed at different levels and on different occasions. The ability to initiate at particular promoters is controlled by a seemingly endless series of ancillary factors, assisting or interfering with the enzyme.

In almost none of these instances is any change made in the subunits of the enzyme itself. Presumably such a mechanism is not favored, because a change that allowed the RNA polymerase specifically to recognize one type of promoter might prevent it from finding another. This would reduce the flexibility of the enzyme.

Yet the division of labors between a core enzyme that undertakes chain elongation and a sigma factor involved in site selection immediately raises the question of whether there could be more than one type of sigma, each specific for a different class of promoters.

Changes in sigma factors appear to occur usually only when there is a wholesale reorganization of transcription. For a long time, *E. coli* was thought to have only a single sigma factor. Now others have been discovered. One is utilized under conditions of "heat shock," when the bacteria change their transcription pattern as the result of an increase in temperature.

A common type of response to heat shock occurs in many organisms, prokaryotic and eukaryotic. Upon a shift up in temperature, synthesis of the proteins currently being made is turned off or down, and a new set of proteins is synthesized. The new proteins are the products of the **heat shock genes.** Little is known about their functions, but presumably they play some role in protecting the cell against environmental stress. They may be synthesized in response to other conditions as well as heat shock.

In *E. coli,* the expression of 17 heat shock proteins is triggered by changes at transcription. The gene *htpR* is a regulator needed to switch on the heat shock response. Its product is a 32,000 dalton protein that functions as an alternate sigma factor.

Multiple sigma factors are denoted σ^{00}, where "00" indicates the molecular weight of the factor. Thus the heat shock sigma is called σ^{32}; the sigma factor that functions under normal conditions is called σ^{70}.

σ^{32} directs core enzyme to initiate at the promoters of heat shock genes. Do these promoters have some special sequence that identifies them to holoenzyme containing α^{32}? They have an extended and slightly different -35 sequence compared with the usual consensus, but have an entirely different consensus at -10. **Table 9.4** compares the consensus sequences of promoters recognized via σ^{70} with the heat shock consensus recognized by σ^{32}.

Heat shock promoters contain a -35 consensus sequence that shares a tetramer with the -35 sequence of general promoters. The difference in promoter recognition between RNA polymerase directed by σ^{70} and heat shock RNA polymerase containing σ^{32} may therefore lie in recognition of the -10 consensus sequence by the sigma factor.

Selection of the genes transcribed during heat shock must in some way depend on a balance between the general sigma factor (called α^{70}) and the heat shock σ^{32}. We do not yet know how σ^{32} is activated and what controls its association with core enzyme relative to the interaction with σ^{70}. However, σ^{32} is unstable, which should allow its amount to be increased or decreased rapidly by regulating its expression.

Another sigma factor may be used under conditions of nitrogen starvation. *E. coli* cells contain a small amount of the protein now known as σ^{60}, which is activated when ammonia is absent from the medium. In these conditions, genes are turned on to allow utilization of alternative nitrogen sources. The σ^{60} factor may allow the core enzyme to recognize promoters that have a distinct consensus sequence, with a conserved element at -10 and another close by at -20 (given in the "-35" column of Table 9.4).

A more extensive example of switches in sigma factors is provided by *Bacillus subtilis,* which contains multiple sigma factors with different specificities. And in certain circumstances, when a drastic change occurs in the life style of the cell, there is a massive switch in gene expression. Then there is evidently less impediment to introducing permanent changes in the RNA polymerase itself, and *B. subtilis* substitutes its original sigma factor by others with different promoter specificities.

The major RNA polymerase found in *B. subtilis* cells engaged in normal vegetative growth has the same structure as that of *E. coli*, $\alpha_2\beta\beta'\sigma$. The sigma factor has a mass of 43,000 daltons and is described as σ^{43}. It recognizes promoters with the same consensus sequences used by the *E. coli* enzyme under direction from σ^{70}. Variants of the enzyme that contain other

Table 9.4
***E. coli* sigma factors recognize promoters with different consensus sequences.**

Factor	Gene	Use	−35 Sequence	Separation	−10 Sequence
σ^{70}	*rpoD*	general	TTGACA	16-18 bp	TATAAT
σ^{32}	*rpoH*	heat shock	CNCTTGAA	13-15 bp	CCCCATNT
σ^{60}	*rpoN*	nitrogen	CTGGNA	6 bp	TTGCA

Genes coding for sigma factors have the designation *rpo* for RNA polymerase subunit, followed by a letter to identify the individual gene (D for the original sigma factor, H for heat shock, N for nitrogen starvation.)

sigma factors are found in much smaller amounts. The variant enzymes recognize different promoters.

Substitutions of sigma factors that cause a transition from expression of one set of genes to expression of another set occur during bacteriophage infection. **Lytic infection** is the process by which a phage (bacterial virus) takes over a bacterium and destroys it in the process of reproducing more phage particles. In all but the very simplest cases, the development of the phage involves shifts in the pattern of transcription. In *E. coli*, these shifts are accomplished by the synthesis of a phage-coded RNA polymerase or by the efforts of phage-coded ancillary factors that control the bacterial RNA polymerase. During infection of *B. subtilis* by phage SPO1, however, two new sigma factors are elaborated.

The infective cycle of SPO1 passes through three stages of gene expression. Immediately on infection, the **early** genes of the phage are transcribed. After 4–5 minutes, the early genes cease transcription and the **middle** genes are transcribed. Then at 8–12 minutes, middle gene transcription is replaced by transcription of **late** genes.

The early genes are transcribed by the holoenzyme of the host bacterium. They are essentially indistinguishable from host genes whose promoters have the intrinsic ability to be recognized by the RNA polymerase $\alpha_2\beta\beta'\sigma^{43}$.

Expression of phage genes is required for the transitions to middle and late gene transcription. Three regulatory genes, named *28*, *33*, and *34*, control the course of transcription. Their functions are summarized in **Figure 9.12.** The pattern of regulation creates a **cascade,** in which the host enzyme transcribes an early gene whose product is needed to transcribe the middle genes; and then two of the middle genes code for products that are needed to transcribe the late genes.

Mutants in the early gene *28* cannot transcribe the middle genes. The product of gene *28* (called gp28) is a protein of 26,000 daltons that replaces the sigma factor on the core enzyme. *This substitution is the sole event required to make the transition from early to middle gene expression.* It creates a complete enzyme that can no longer transcribe the host genes, but instead specifically transcribes the middle genes. We do not know how gp28 displaces σ^{43}, or what happens to the host sigma polypeptide. Probably gp28 has a greater affinity for the core enzyme.

Two of the middle genes are involved in the next transition. Mutations in either gene *33* or *34* prevent transcription of the late genes. The products of these genes are proteins of 13,000 and 24,000 daltons, respectively, that replace gp28 on the core polymerase. Again, we do not know how gp33 and gp34 exclude gp28 (or any residual host σ^{43}), *but once they have bound to the core enzyme, it is able to initiate transcription only at the promoters for late genes.*

The successive replacements of sigma factor have dual consequences. Each time the subunit is changed, the RNA polymerase becomes able to recognize a new class of genes, *and* no longer recognizes the previous class. These switches therefore constitute global changes in the activity of RNA polymerase. Probably all or virtually all of the core enzyme becomes associated with the sigma factor of the moment; and the change is irreversible.

New sigma factors are utilized also during **sporulation**, an alternative life style available to some bacteria. At the end of the **vegetative phase**, logarithmic growth ceases because nutrients in the medium become depleted. This triggers sporulation; DNA is replicated, a genome is segregated at one end of the cell, and eventually it is surrounded by the tough spore coat. The process takes ~8 hours.

Sporulation involves a drastic change in the biosynthetic activities of the bacterium, and many genes are involved. The basic level of control lies at transcription. Some of the genes that functioned in the vegetative phase are turned off during sporulation, but most continue to be expressed. In addition, the genes specific for sporulation are expressed only during this period. At the end of sporulation, ~40% of the bacterial mRNA is sporulation-specific.

New forms of the RNA polymerase become active in sporulating cells; they contain the same core enzyme as vegetative cells, but have different proteins in place of the vegetative σ^{43}. These changes in transcriptional specificity are summarized in **Figure 9.13.**

At the start of sporulation, σ^{43} is replaced by σ^{37}, a smaller protein that actually is present in vegetative cells, although then it is not associated with the core enzyme. Under the direction of σ^{37}, RNA polymerase transcribes the first set of sporulation genes instead of the vegetative genes it was previously transcribing.

What controls the timing of the replacement? Mu-

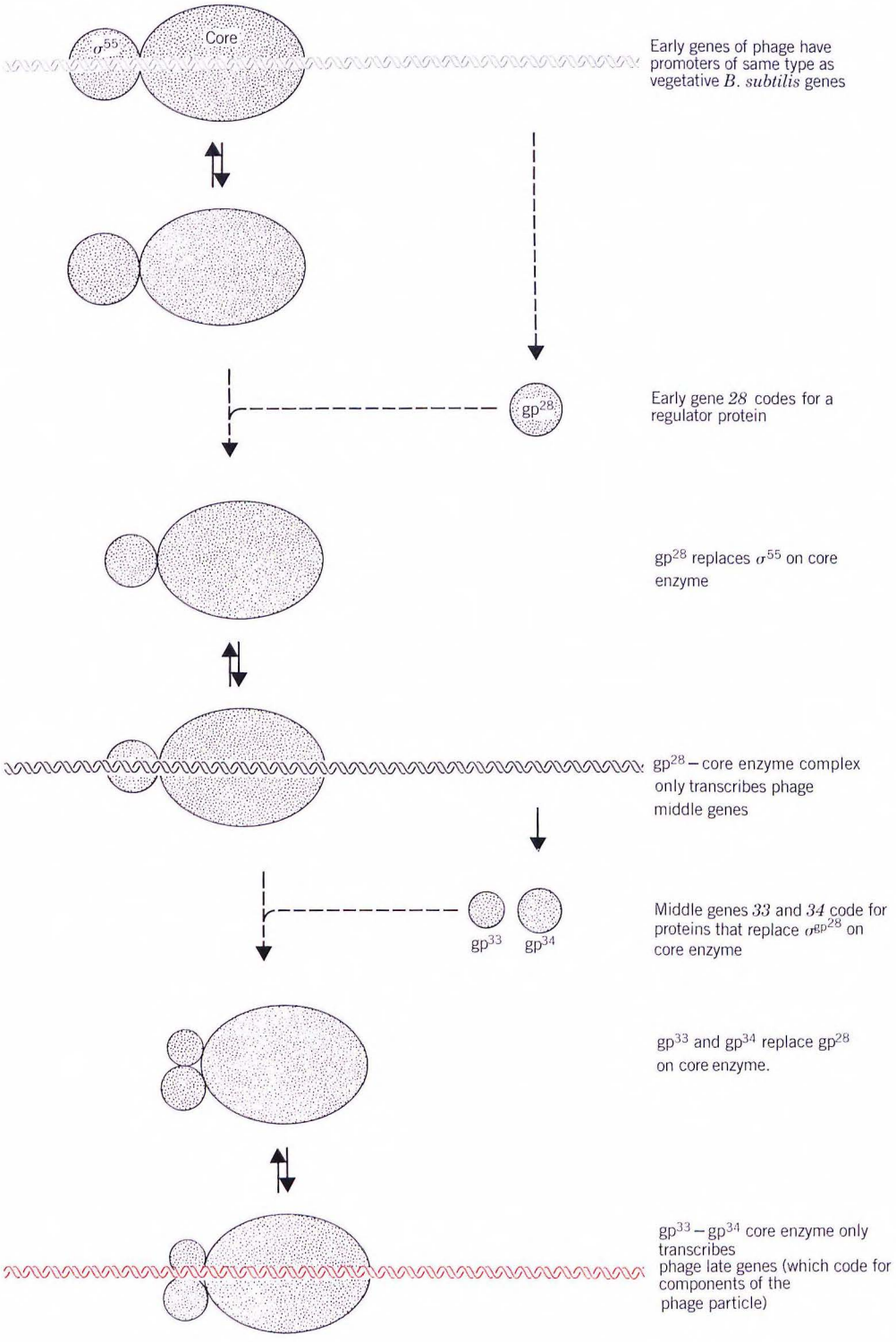


Figure 9.12
Transcription of phage SPO1 genes is controlled by two successive substitutions of the sigma factor that change the initiation specificity.

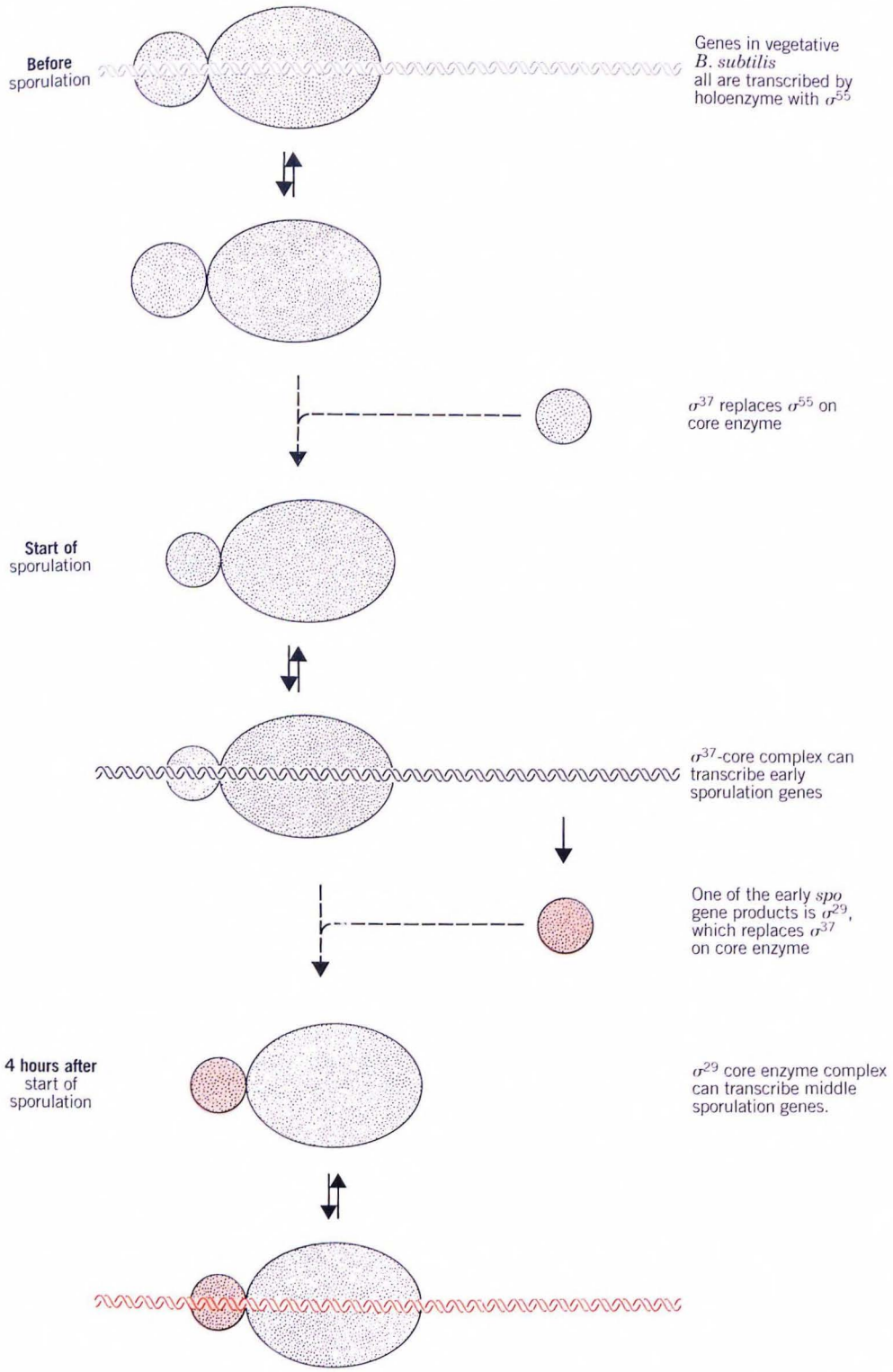


Figure 9.13
Sporulation involves successive changes in the sigma factor that control the initiation specificity of RNA polymerase.

tations in any one of eight genes can block transcription of the early sporulation genes that are expressed via the $\alpha_2\beta\beta'\sigma^{43}$ enzyme, so the process of substituting σ^{37} in place of σ^{43} may be quite complex, involving additional proteins. The replacement reaction affects only part of the RNA polymerase population, since there is enough σ^{37} to bind only ~10% of the core enzyme. Probably some vegetative enzyme remains present during sporulation. The displaced σ^{43} is not destroyed, but can be recovered from extracts of sporulating cells.

At least one other sigma factor, σ^{32}, is present in vegetative cells and becomes active early in sporulation. It is rather rare, being found in amounts less than 1% of the core enzyme population. Again it directs transcription from distinct promoters.

Another form of RNA polymerase appears in cells ~4 hours after the start of sporulation. It contains σ^{29}, a new sigma factor that allows the enzyme to transcribe yet another set of genes. The σ^{29} factor is not present in vegetative cells and is probably the product of one of the sporulation genes transcribed under the direction of σ^{37}. Again, it is a mystery how σ^{29} is able to replace σ^{37} or σ^{43} on the core enzyme.

Factor σ^{28} is associated with core enzyme in vegetative cells, where this form of RNA polymerase represents a very small proportion of the total enzyme activity. It *ceases* to be active when sporulation begins. Its transcripts are absent from vegetative cells of certain mutants that cannot sporulate. It is possible that σ^{28} is part of some signalling system, responsible for expressing genes whose products detect nutritional deprivation and initiate the sporulative response.

Sporulation may therefore be controlled by a pattern in which successive sigma factors are activated, each directing the synthesis of a particular set of genes. As new sigma factors become active, old sigma factors may be displaced, so that transitions in sigma factors may turn genes off as well as on. The incorporation of each factor into RNA polymerase dictates when its set of target genes is expressed; and the amount of factor available may have something to do with the level of gene expression.

What distinguishes the different classes of promoters recognized by the various sigma factors? The host enzyme, containing σ^{43}, recognizes promoters with the same −35 and −10 sequences described in *E. coli*. However, these consensus sequences are not found in the promoters recognized by any of the other sigma factors. Perhaps each set of promoters has its own characteristic consensus. We have not characterized enough target promoters for each sigma factor to draw up all the consensus sequences, but some possible conserved sequences are summarized in **Table 9.5.**

A significant feature of the promoters for each enzyme is that *they have the same size and location relative to the startpoint, and they show conserved sequences only around the usual centers of −35 and −10.* The consensus at −10 is usually A-T-rich; the consensus at −35 also shows a tendency in this di-

Table 9.5
Each *B. subtilis* sigma factor uses a set of promoters with a characteristic consensus.

Sigma Factor	Source & Use	−35 Region	−10 Region
σ^{43}	vegetative	TTGACA	TATAAT
σ^{37}	used in sporulation	AGGNTTT	GGNATTGNT
σ^{32}	used in sporulation	AAATC	TANTGTTNTA
σ^{29}	synthesized in sporulation	TTNAAA	CATATT
σ^{28}	not used in sporulation	CTNAAA	CCGATAT
gp^{28}	SPO1 middle expression	AGGAGA	TTTNTTT
$gp^{33\text{-}34}$	SPO1 late expression	CGTTAGA	GATATT

Most of these consensus sequences are rather speculative; only the sequences for the host enzyme (σ^{43}) and the two phage enzymes rest on a sufficient number of target promoters for the consensus to be reliable.

rection. We do not know how widespread is the ability to use different sigma factors to direct recognition of different promoters; it is rare in *E. coli,* but characteristic of *B. subtilis*, although both host holoenzymes recognize promoters by the same criteria.

Because the holoenzyme can display (at least) seven different specificities in *B. subtilis*, it seems unlikely that each sigma factor could induce a particular conformation in the core enzyme that causes it to recognize a distinct set of promoters. It is easier to think in terms of sigma factors that themselves directly recognize features of the promoters, although this poses the problem of how a small polypeptide can contact sites spanning more than 20 bp of DNA.

PROMOTERS FOR RNA POLYMERASE II ARE UPSTREAM OF THE STARTPOINT

We can apply several criteria in identifying the sequence components of a promoter (or any other site in DNA):

- A consensus sequence may be present in many examples of the site.
- Mutations in the site may prevent function *in vitro* or *in vivo.* (Many techniques now exist for introducing point mutations at particular base pairs, and in principle every site in a consensus sequence or any other region can be mutated, and the mutant sequence tested *in vitro* or *in vivo.*)
- Proteins involved in use of the site (such as RNA polymerase or accessory factors for promoters) may be footprinted on it.

Attempts to define the promoters for eukaryotic RNA polymerases have taken advantage of the precedents established with bacterial RNA polymerase. However, eukaryotic systems suffer from two particular limitations:

- Virtually no promoter mutations have been identified *in vivo*, so we start without any prior information on the location of the promoter.
- We have not yet been able to retrieve the DNA sequences bound by any of the RNA polymerases. Recovery of protected fragments is rendered difficult by the complexity of the enzyme preparations and the lack of information on exactly what constitutes the active structure of the enzyme. Of course, it is only a matter of time until a successful system is developed that will allow us to recover the DNA binding site from an initiation complex.

A less direct assay, but one that actually is more informative in some ways, is to define the promoter in terms of its ability to initiate transcription in a suitable test system. Three types of system have been used:

- The ***in vitro* system** takes the classic approach of purifying all the components and manipulating conditions until faithful initiation is seen. "Faithful" initiation is defined as production of an RNA starting at the site corresponding to the 5′ end of mRNA (or rRNA or tRNA precursors). Systems for each of the three RNA polymerases are now in various stages of purification; ultimately each will consist of a preparation in which all of the components have been defined. Then we shall be able to compare the activities *in vitro* of RNA polymerases from different tissues or species.
- The **oocyte system** follows the principles established for translation, and relies on injection of a suitable DNA template, this time into the nucleus of the *X. laevis* oocyte. The RNA transcript can be recovered and analyzed. The main limitation of this system is that it is restricted to the conditions that prevail in the oocyte.
- ***In vivo* systems** do not (as the name might imply) necessarily involve characterizing the ability of a cell to transcribe one of its usual products. What is done here is to follow the ability of a cultured cell to transcribe a template introduced by transfection (a procedure that allows an exogenous DNA to enter a cell and be expressed; the events involved are discussed in Chapter 31). The system is genuinely *in vivo* in the sense that transcription is accomplished by the same apparatus responsible for expressing the cell's own genome; but it differs from the natural situation because the template may consist of a gene that would not usually be transcribed in the host cell.

The approach to characterizing the promoter is the same in all three systems. We seek to manipulate the template *in vitro* before it is submitted to the system for transcription. Sometimes this is called "surrogate genetics."

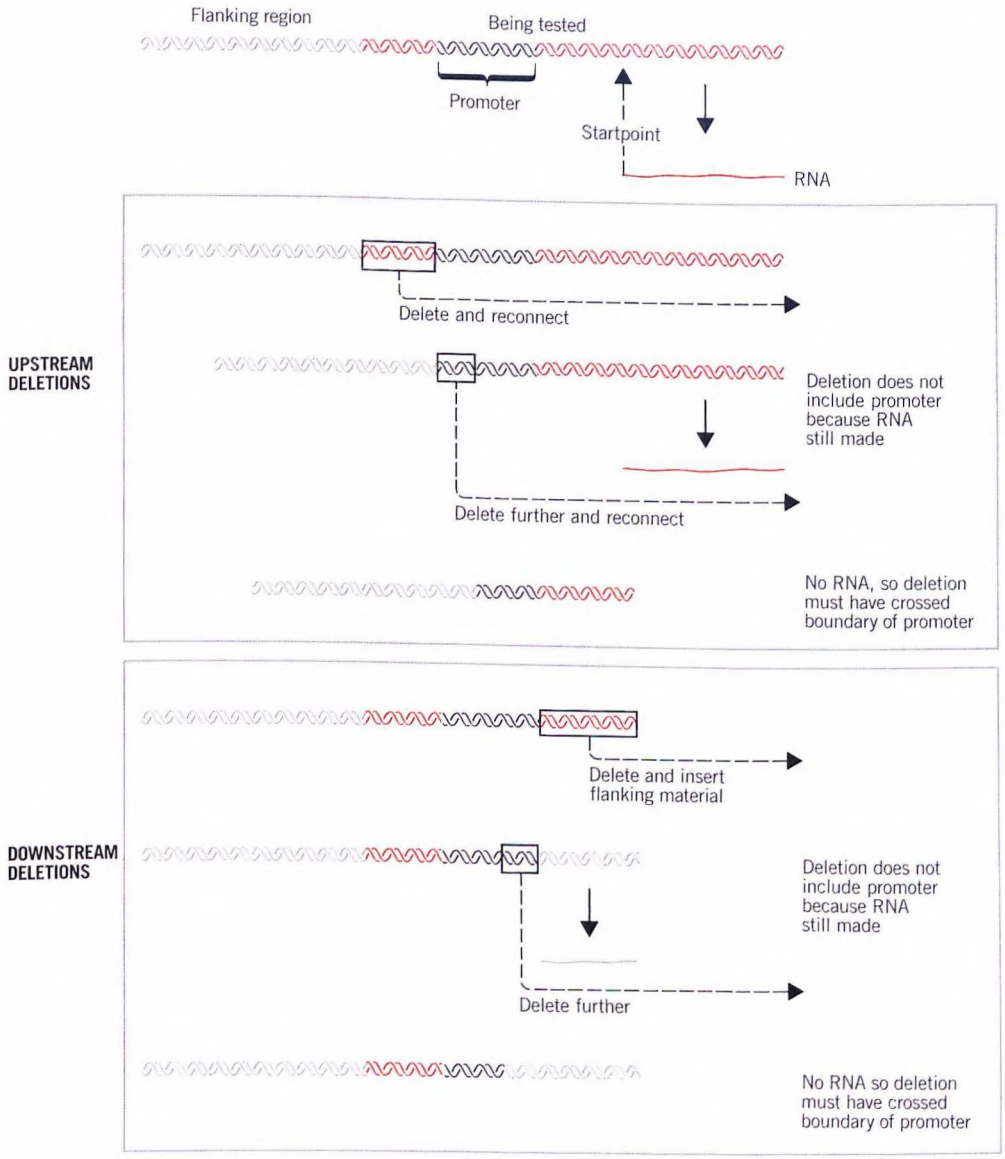


Figure 9.14
Promoter boundaries can be determined by deletions that approach from either side.
Each deletion removes the material on one side of the unit being tested and extends farther into the unit than the last. When one deletion fails to prevent RNA synthesis but the next stops transcription, the boundary of the promoter must lie between them.

When a particular fragment of DNA can be used to initiate transcription, it is taken to include a functional promoter. Then the boundaries of the sequence constituting this promoter can be determined by reducing the length of the fragment from either end, until at some point it ceases to be active. The type of protocol is illustrated in **Figure 9.14.** The boundary upstream can be identified by progressively removing material from this end until promoter function is lost. For the boundary downstream, it is necessary to reconnect the shortened promoter to the sequence to be transcribed (since otherwise there is no product to assay).

We have to take several precautions to avoid extraneous effects. To ensure that the promoter is always in the same context, the same long upstream sequence is always placed next to it. Because termination may not occur properly in the *in vitro* systems, the template may be cut at some distance from the pro-

moter (usually ~500 bp downstream), to ensure that all polymerases "run off" at the same point, generating an identifiable transcript.

Once the boundaries of the promoter have been defined, the importance of particular bases within it can be determined by introducing point mutations or other rearrangements in the sequence. As with bacterial RNA polymerase, these can be characterized as *up* or *down* mutations. Some of these rearrangements may affect only the *rate* of initiation; others may influence the *site* at which initiation occurs, as seen in a change of the startpoint. To be sure that we are dealing with comparable products, in each case it is necessary to characterize the 5′ end of the RNA, as described previously in Figure 9.7.

One useful technique for analyzing sequences needed for promoter function is provided by the technique of **linker scanning,** which allows clusters of mutations to be introduced at particular sites. **Figure 9.15** illustrates the protocol. Deletion mutants are made from both sides, removing the regions either 5′ or 3′ to the relevant site. A "linker sequence" is added to the end of each deletion; the linker consists of a short synthetic oligonucleotide that includes the sequence recognized by some restriction enzyme. Both fragments are cleaved with the enzyme, and then they are joined crosswise (a reaction that is characteristic of ends cleaved by some restriction enzymes; see Chapter 15).

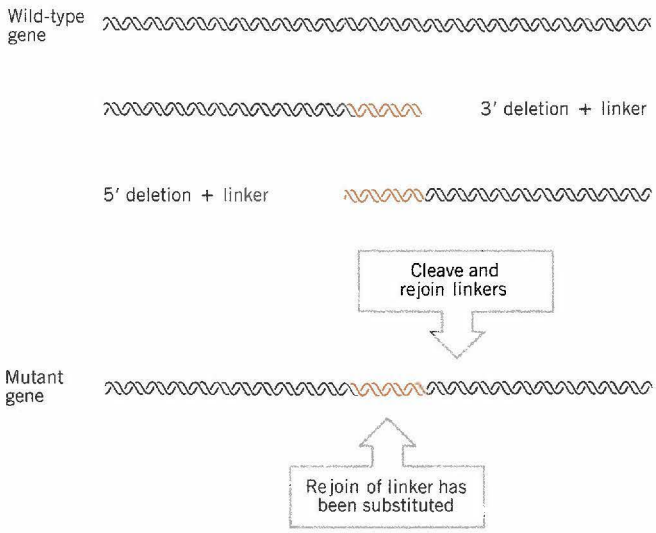


Figure 9.15
The linker-scanning technique allows a short sequence of wild-type gene to be replaced by the linker sequence at the point corresponding to the ends of the matching deletions.

This reaction inserts the sequence of the linker in place of the sequence originally located at the point where the deletions meet. By using matching 5′ and 3′ deletions that end at a series of sites along the wild-type sequence, the entire sequence can be "scanned" for its sensitivity to mutation.

In the *in vitro* systems, boundaries for the promoter can be defined in the vicinity of the startpoint. A relatively short sequence is needed to initiate transcription.

Proceeding from the upstream direction, sequences can be progressively removed without any effect, until reaching a point located somewhere between −45 and −30. The limit defines the left boundary of the promoter. Once the limit is transgressed, transcription is reduced twentyfold or more.

Proceeding from the downstream direction, the right boundary of the promoter is usually close to the startpoint. It may as far beyond it as +6 or as far before it as −10 or −12.

When the startpoint is deleted, initiation occurs at a point in the template that is the *same distance* from the promoter as the original startpoint, except that it may be adjusted by a base or two in order to find a purine (usually A) with which to initiate. The apparent failure always to include the startpoint in the promoter could mean that the geometry of the complex may determine where initiation occurs, presumably because the enzyme stretches a certain distance downstream from its binding site. However, the efficiency of initiation may be somewhat reduced by the absence of the usual startpoint.

As with bacterial promoters, homologies in the regions near the startpoint are restricted to rather short sequences. Homologies of potential significance have been noted in three regions: the startpoint, a sequence centered at about −25, and a region farther upstream, centered at about −75.

At the startpoint, there is no extensive homology of sequence, but there is a tendency for the first base of mRNA to be A, flanked on either side by pyrimidines. (This description is also valid for the CAT start sequence of bacterial promoters.)

Most promoters have a sequence called the **TATA** or **Hogness box.** between 19 and 27 bp upstream of

the startpoint. It has been found in mammals, birds, amphibians, and insects. From all known cases (irrespective of species), the consensus is

$$T_{82}\ A_{97}\ T_{93}\ A_{85}\ \begin{matrix} A_{63} \\ T_{37} \end{matrix}\ A_{83}\ \begin{matrix} A_{50} \\ T_{33} \end{matrix}$$

The consensus sequence consists entirely of A-T base pairs (at two positions the orientation is variable), and in only a minority of actual cases is a G-C pair present. The TATA box tends to be surrounded by G-C rich sequences, which could be a factor in its function. It is almost identical with the Pribnow box found in bacterial promoters; in fact, it could pass for one except for the difference in its location at −25 instead of −10.

The TATA box is contained within the promoter defined *in vitro*. Its importance is confirmed by the fact that single base substitutions in it act as strong down mutations. One such mutation reversed the orientation of an A-T pair, so the base composition alone is not sufficient to fulfill the function of the sequence.

Thus the TATA box comprises an element whose behavior is analogous to our concept of the bacterial promoter: a short, well-defined sequence just upstream of the startpoint that is necessary for transcription.

RNA POLYMERASE II PROMOTERS ARE MULTIPARTITE

The TATA box may be adequate *in vitro*, but with *in vivo* or *in oocyte* systems, we see in addition a strong dependence on sequences farther upstream. Some short consensus sequences have been identified, and accessory factors found that bind to them.

Their exact location varies, but they lie more than 40 bp upstream of the TATA box. When this region is deleted, initiation at the usual startpoint is reduced to ~2% of its previous level. By contrast, if the TATA box is removed, initiation continues to occur, but the startpoint varies from its usual precise location.

By using the linker scanning technique, three distinct sequence elements have been identified in the thymidine kinase (TK) gene of herpes virus and in the β-globin gene. The region around the TATA box is needed for accurate initiation. Two other separate regions, the *middle region* located between −50 and −70, and the *distal region* located between −80 to −110, are needed for efficient initiation. The introduction of double mutations shows that the middle and distal regions are concerned with the same function, for mutations in either one alone have as much effect as mutations in both together.

This multipartite structure means that we cannot investigate the internal structure of the promoter simply by making deletions within it; deletions could abolish function by changing critical distances between components, even though the deleted sequences are themselves irrelevant. So we need to make small deletions that are replaced by other sequences of the same length.

The regions between the three sequence elements are not involved in promoter function. It turns out that the distance between the three sites is moderately flexible; the separation between the two most upstream elements can be increased by >15 bp before they become unable to function; and their distance from the TATA box can be increased by >30 bp before the promoter becomes affected.

The corresponding elements in the two promoters appear to play equivalent roles, and can be exchanged. For example, a functional promoter can be constructed by joining the distal TK component to the middle and TATA β-globin components. This suggests that the promoter may be "modular," in the sense that its individual components can be provided by alternate sequence elements.

In vivo the promoter seems to have two types of function:

- Frequency of initiation is strongly influenced by the *upstream function* provided by the distal and middle elements. They probably have a major effect on the binding of RNA polymerase. However, the residual transcription that occurs in their absence does initiate at the proper startpoint.
- Choice of startpoint depends on the *near function*, conveyed by an element close to the startpoint that surrounds the highly conserved TATA box. Its deletion causes the site of initiation to become erratic, although any overall reduction in transcription is relatively small. The role of this sequence could be to align the RNA polymerase so that it initiates at the proper site.

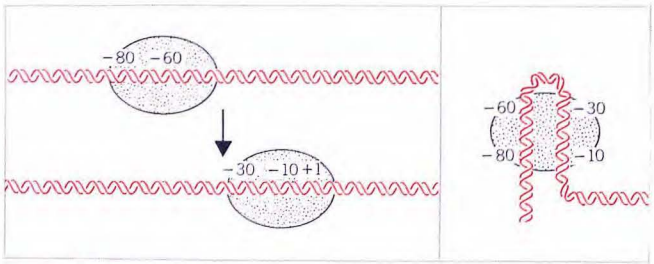


Figure 9.16
One model to reconcile the size of the promoter with the size of RNA polymerase supposes that the enzyme moves. Another supposes that the DNA is compactly organized.

How can a promoter consist of separated elements that stretch over a distance of DNA greater than RNA polymerase could contact? Two models are illustrated in **Figure 9.16.**

One possibility is that RNA polymerase initially contacts the site farther upstream and then moves to the site nearer the startpoint. Another view requires us to remember that *in vivo* DNA is not stretched out in linear fashion; its compact organization could bring into juxtaposition sites that are separated on the duplex molecule (see Chapter 26). The binding site for RNA polymerase could consist of DNA sequences that are not contiguous, but are held together by proteins that bind DNA. This model implies further that the *spacing* (rather than the exact sequence) between the promoter components be important.

In either type of model, we can see that the region farther upstream could be the most important for initially binding RNA polymerase, but the region nearer the startpoint could be required to hold the enzyme in a configuration that allows it to recognize the exact startpoint. Thus when the sequence farther upstream is absent, the ability of RNA polymerase to bind to the DNA is much reduced; but those enzyme molecules that do bind can initiate accurately because the TATA box is present. On the other hand, when the TATA box is deleted, RNA polymerase remains able to bind efficiently to the upstream sequence, but its contacts in the region around the startpoint are less precise, allowing initiation to occur at more than one point.

How are we to explain the discrepancy between the need *in vivo* for the sequences upstream of −50 and their apparent lack of importance *in vitro*? A major responsibility for this effect may lie with the differing efficiencies of transcription in the two circumstances. The *in vitro* system is relatively inefficient; less than 1% of the templates actually are transcribed. We do not know just what proportion of the templates are utilized *in vivo* or *in oocyte*, but it could very well be much higher. The *in vitro* level could even correspond to the residual expression that we see *in vivo* in the absence of sequences upstream of −60!

At all events, the simpler structure of the template *in vitro*, as a DNA molecule not organized in the usual proteinaceous structure, may mean that its recognition by RNA polymerase is different, certainly less efficient, perhaps because it uses only some features of the promoter.

What are the active sequence components of the promoter modules? The function of each module is determined by a rather short consensus sequence, although the sequence surrounding this "core" may in-

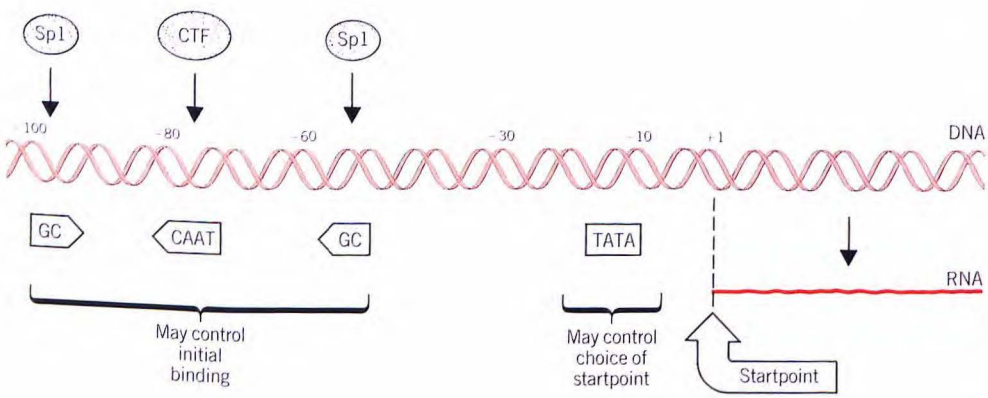


Figure 9.17
A promoter for RNA polymerase II contains separate sequence components with different roles. The sequences between the components are not important. The orientation of the conserved sequences is indicated by the direction of the arrows. Transcription factors bind to the conserved sequence motifs.

fluence its effectiveness. The consensus sequences all are described in terms of "boxes," such as the TATA box.

The **CAAT box** is found in both the TK and β-globin promoters and is conserved in several (but not all) known promoters. It has the consensus

$$GG\begin{matrix}C\\T\end{matrix}CAATCT$$

It is present as written in the β-globin middle region, and is present in reverse orientation in the TK distal region.

Another conserved element is the **GC box,** which contains the sequence GGGCGG. Often multiple copies are present in the promoter, and they occur in either orientation.

The organization of these elements in the thymidine kinase promoter is summarized in **Figure 9.17.** Proceeding upstream from the startpoint, the near region consists of the TATA box, the middle region contains a GC box, and the far region contains a CAAT box and a second GC box. The two GC boxes are in opposite orientation, and the CAAT box is in the reverse orientation from usual. One of the puzzles of promoter organization is that the promoter conveys directional information (transcription proceeds only in the downstream direction), but the GC and CAAT boxes seem to be able to function in either orientation.

None of these elements is found in every promoter. Some promoters lack a TATA box, and in these initiation does not occur at a unique startpoint, but occurs at any of a cluster of startpoints. Others lack a CAAT box and/or have no GC boxes. Examples of the components of some promoters are summarized in **Figure 9.18.**

Promoters for RNA polymerase II in yeast have a TATA box like the higher eukaryotic promoters we have just described, but they differ in allowing its location to vary widely, from ~40–90 bp upstream of the startpoint. The TATA box is essential for transcription, but sequences at or near the startpoint are needed to fix the precise site of initiation.

TRANSCRIPTION FACTORS RECOGNIZE PARTICULAR CONSENSUS SEQUENCES

Our criterion for saying that a protein is part of the transcription apparatus is that it is needed for transcription to initiate in an *in vitro* system. The proteins of the transcription apparatus can be divided into three groups:

- Subunits of RNA polymerase are needed for some or all of the stages of transcription, but they are not specific for individual promoters.
- Transcription factors may bind RNA polymerase when or after it forms an initiation complex, although they are not part of the free enzyme. These factors are likely to be needed for transcription to initiate at all promoters or (for example) to terminate. Immediately

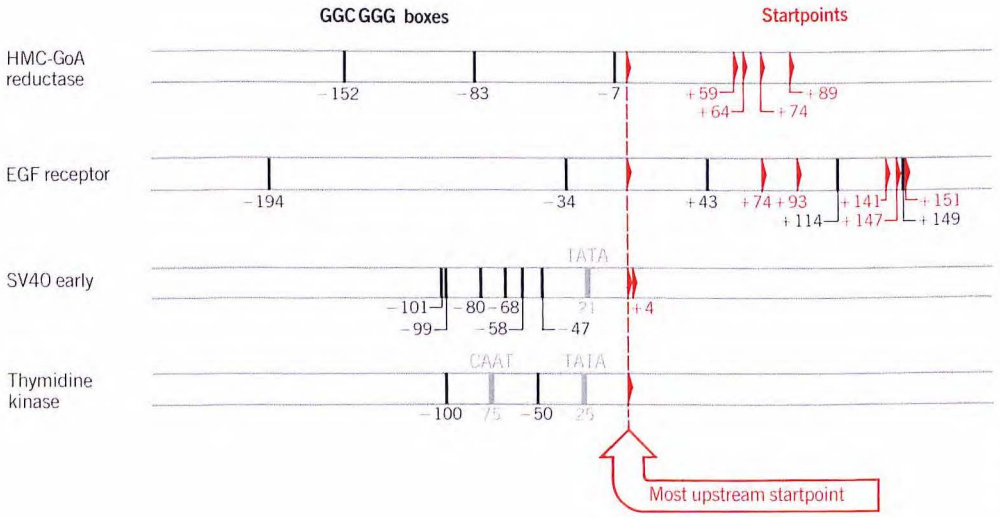


Figure 9.18
Promoters contain different combinations of TATA boxes, CAAT boxes, and GC boxes.

this brings us back to the question of which polypeptides are subunits of RNA polymerase and which are accessory factors.

- Transcription factors may bind specific sequences in the target promoters. If the sequences were present in all promoters, the factors would be part of the general transcription apparatus. If some sequences are present only in certain classes of promoters, factors that recognize them could be needed specifically to initiate at those promoters.

Most efforts to characterize transcription factors have focused on the last class, and we now know of some factors that are needed for transcription specifically of promoters that contain particular consensus sequences. They are summarized in **Table 9.6.**

Characterizing the binding sites for these factors provides us with another criterion for identifying promoter components. We expect a factor to bind to a DNA region that includes a sequence identified as essential for transcription *in vitro* by mutational analysis.

At least two factors are needed for *D. melanogaster* RNA polymerase to initiate transcription *in vitro.* One is the B factor, which binds to a region containing the TATA box. The human factor TFIID binds to a similar region. *Some of the characteristic features of promoters therefore may interact directly with factors (as distinguished chromatographically) rather than with RNA polymerase II itself.*

Similarly the CTF (CAT binding factor) binds to the region of the major late adenovirus promoter homologous to the CAAT box; it may therefore be needed for transcription of promoters with this consensus sequence.

Another factor binding in the upstream region is USF (also known as MLTF), which recognizes a sequence located around −55 in the late adenovirus promoter.

The presence of the GC box is associated with the ability to bind the sp1 transcription factor. In the SV40 promoter, the multiple boxes between −70 and −110 all are bound, so that the whole region is protected by sp1. Probably an individual sp1 binding unit contacts one strand of DNA over a ~20 bp binding site that includes at least one GC box. We don't yet know the active form of the protein, which has a monomeric subunit of 100,000 daltons.

Although the GC box is essential for binding, different sequences containing GC boxes are bound with different affinities, so the flanking sequence may influence recognition by sp1. We do not know what function sp1·GC box binding plays in transcription, or how and why GC boxes can function at varying distances from the startpoint.

The sp1 factor could be involved in allowing RNA polymerase II to recognize a certain class of promoters. There has been some speculation that the GC boxes may identify genes that are expressed constitutively.

The results obtained with these factors suggest that promoters may fall into general groups identified by the presence of particular consensus sequences. Factors that recognize these sequences would in effect be responsible for coordinately recognizing classes of promoters.

Recognition by multiple factors may be needed to initiate transcription. In the thymidine kinase promoter, sp1 binds to the GC boxes and CTF binds to the CAT

Table 9.6
Some transcription factors bind to sequence elements in RNA polymerase II promoters.

Factor	Species	Promoter	Site Recognized
B	*Drosophila*	TATA promoters	TATA box/startpoint
TFIID	man	adeno	TATA box
CTF	man	adeno, globin, TK	CAAT box
USF/MLTF	man	adeno late	GGCCACGTGACC
sp1	man	see Figure 9.18	GC box
HSTF	*Drosophila*	heat shock	heat shock consensus

box to cover almost all of the middle and distal regions, as indicated in Figure 9.17. In the late adenovirus promoter, TFIID binds to the TATA box and USF binds to a region nearby.

When multiple factors bind, their interactions with the promoter may be cooperative, that is, the binding of one factor may assist another factor to bind. We may think of their interactions at the promoter as building a structure that, together with RNA polymerase, constitutes an initiation complex.

These factors are needed for general transcription, that is to say, we do not implicate them in a regulatory capacity. Other factors may be used to regulate the transcription of particular groups of genes. A good example is provided by the heat shock genes.

The heat shock response is common to a wide range of prokaryotes and eukaryotes and may involve multiple controls of gene expression: an increase in temperature turns off transcription of some genes, turns on transcription of the **heat shock genes,** and may also cause changes in the translation of mRNAs. Eukaryotic heat shock genes possess a consensus sequence of ~15 bp upstream of the startpoint. The HSTF factor is active only in heat-shocked cells; it binds to a site including the heat shock consensus sequence. The activation of this factor therefore provides a means to initiate transcription at the specific group of ~20 genes that contains the appropriate target sequence at its promoter.

We do not understand the mechanism of initiation in enough detail to know what role the factors play, but in terms of the overall effect, we may think of the division between RNA polymerase II and the factors as analogous to the division between bacterial core enzyme and sigma factor. A difference is that initiation at any given eukaryotic promoter may require several factors, whereas bacterial sigma factor is a single polypeptide. And the use of different factors to support transcription of different classes of promoters seems to be more widespread in the eukaryotic nucleus. We see the parallel most clearly in the case of the heat shock genes, where a specific transcription factor is needed in *Drosophila* cells to recognize the heat shock consensus, while a new sigma factor is needed in bacteria to recognize an alternate −10 sequence.

Eukaryotic transcription factors may have a wide range of specificities. The most common are probably those needed to recognize all promoters containing TATA boxes. The least common may be factors employed to turn on transcription of small groups of genes responding to some common signal. Thus the factors may range from relatively common components of the transcription apparatus to relatively rare regulatory proteins.

This idea may be extended to a general principle. *A gene may be regulated by a sequence at the promoter that is recognized by a specific protein. The protein functions as a transcription factor needed for RNA polymerase to initiate. It is available only under conditions when the gene is to be expressed; its absence ensures that the promoter cannot be used.* Why is the promoter dependent on recognition of this sequence? Perhaps it lacks a full set of the modules needed for recognition by RNA polymerase II, and factor binding at the regulatory sequence compensates for this deficiency.

We can take this principle further. *When a promoter is regulated in more than one way, each regulatory event may depend on binding of its own protein to a particular sequence.* For example, one of the *Drosophila* heat shock genes is expressed in two circumstances: in all tissues at elevated temperature; and in the ovaries at normal temperature. Deletion analysis shows that sequences from −49 to −85 are needed for initiation during heat shock, while sequences upstream of −341 are needed for ovarian expression.

These regulatory elements are usually rather short. For example, a ~12 bp sequence allows the metallothionein promoter to be activated in the presence of metal ions. We may think of promoters for RNA polymerase II, therefore, as containing both basic modules (such as the TATA box) that are involved in a general interaction with RNA polymerase, and a series of special sequences each of which allows RNA polymerase to bind and initiate transcription under particular conditions.

ENHANCERS ARE BIDIRECTIONAL ELEMENTS THAT ASSIST INITIATION

We have considered the promoter so far essentially as an isolated region responsible for binding RNA polymerase. But eukaryotic promoters do not necessarily

function alone. In at least some cases, the activity of a promoter is enormously increased by the presence of another sequence, known as an **enhancer.**

An enhancer is distinguished from the promoter itself by two characteristics: its position relative to the promoter need not be fixed, but can vary substantially; and it can function in either orientation. An enhancer is not restricted to assisting a particular promoter, but can stimulate any promoter placed in its vicinity.

The SV40 enhancer is located in a region of the genome that contains two identical sequences of 72 bp each, repeated in tandem ~200 bp upstream of the startpoint of a transcription unit. These **72 bp repeats** lie in a region with an unusual nucleoprotein structure, apparently one that is more exposed than usual (see Chapter 27). Deletion mapping shows that either one of these repeats is adequate to support normal transcription; but removal of both repeats greatly reduces transcription *in vivo*.

By this type of criterion, we might argue that the repeated region constitutes an upstream component of the promoter. But reconstruction experiments in which the 72 bp sequence is removed from the DNA and then is inserted elsewhere show that normal transcription can be sustained so long as it is present *anywhere* on the DNA molecule. In fact, if a β-globin gene is placed on a DNA molecule that contains a 72 bp repeat, its transcription is increased *in vivo* more than 200-fold, even when the 72 bp sequence is as much as 1400 bp upstream or 3300 bp downstream of the startpoint. And these are simply the limits that have been tested so far; we have yet to discover at what distance the 72 bp sequence fails to work. The sequence can be inverted and replaced; and it still works.

The moral of these results is clear. We need to be careful in defining the components of the promoter. It is not enough just to show that deletion of a particular sequence reduces transcription; we must also investigate the importance of the *location* of the deleted sequence.

So what is a promoter? *If we use a working definition that it constitutes a sequence or sequences of DNA that must be in a (relatively) fixed location relative to the startpoint, the TATA box and other upstream elements are included, but the enhancer is excluded.*

Several viral genomes include enhancers. One of particular interest is carried by retroviruses, viruses whose insertion into a host genome may activate genes in the vicinity, possibly via the presence of an enhancer. Some viral enhancers show specificity for cell type in their function. The ability of the enhancer to function in a particular cell may be partly responsible for determining the host range of the virus. In the case of polyoma, a mutation in the enhancer extends the range of cells that the virus can infect.

Cellular enhancers have been discovered in the form of elements in the genome that stimulate the use of a nearby promoter in a specific tissue. Such enhancers may provide part of the regulatory network by which gene expression is controlled. One case is represented by immunoglobulin genes, which carry enhancers *within* the transcription unit. Thus the enhancer is downstream of the promoter that it stimulates. The immunoglobulin enhancers appear to be active only in the B lymphocytes in which the immunoglobulin genes are expressed (see Chapter 32).

We have yet to delineate the boundaries of the enhancers. Base substitutions create down mutations over a distance of ~100 bp spanning an SV40 72 bp repeat. Short consensus sequences can be deduced from the sequences of known enhancers, but evidence for their function is rarely clear.

The best evidence for the involvement of particular sequences is provided by experiments to footprint enhancers *in vivo.* When the methylating agent dimethyl sulfate (DMS) is applied *in vivo,* it reacts with DNA in the same way as *in vitro.* Used *in vivo,* it identifies a protein-bound sequence of DNA by virtue of the protection or enhancement of particular G residues relative to the reaction with free DNA *in vitro.* Of course, this does not tell us *which protein* is bound to the DNA *in vivo.*

An immunoglobulin enhancer shows a cluster of protections and enhancements at variants of an octameric consensus sequence, CAGGTGGC. Four octamers are present in the enhancer region. The reaction occurs only in B lymphocytes, which suggests that the enhancer may be activated by a protein specifically made in these cells. A sequence related to this consensus is found in several enhancers.

Enhancers may have modular structures. The SV40 enhancer can be separated into two halves, neither of which functions by itself, but which are active even when separated by some distance. Mutational analysis

identifies three domains: mutation of any one domain inactivates the enhancer, but can be reverted by duplicating either of the two remaining domains. Although the domains are different in sequence, they appear to play similar roles, since an active enhancer can be created by the combination of a sufficient *number* of wild-type domains, irrespective of their types.

How can an enhancer stimulate initiation at a promoter that can be located on either side of it at apparently any distance? Some possibilities are that it might be concerned with structure, location, or enzyme binding:

- An enhancer could change the overall structure of the template—for example, by influencing the DNA-protein organization of chromatin, or by changing the density of supercoiling. Against this notion is the result that enhancers continue to function under conditions when supercoiling cannot be propagated along the double helix.

  All the enhancers so far characterized include a stretch of alternating pyrimidine-purine residues, just the sort of sequence likely to form Z-DNA (see Chapter 3). One possibility is that enhancers function by forming a short length of Z-DNA, although we remain mystified as to what effect the Z-DNA has on local transcription. However, a structural effect of this nature would explain why the enhancer can affect a promoter on either side.
- If an enhancer provides an attachment site, it could be responsible for locating the template at a particular place within the cell—for example, attaching it to the nuclear matrix.
- An enhancer could provide a bidirectional "entry site," a point at which RNA polymerase (or some other essential protein) associates with chromatin. The polymerase must then move to the promoter.

Elements somewhat similar to enhancers, called upstream activator sequences (UAS), are found in yeast. They can function in either orientation, at variable distances upstream, of the promoter, but cannot function when located downstream. They have a regulatory role: in the *gal* system, the UAS is bound by the regulatory protein that activates the genes.

If a site bound by another protein, or a sequence that causes transcription to terminate, is introduced between the UAS and the promoter of the gene, the activation effect is abolished. This would be consistent with a "tracking" model in which some protein, perhaps RNA polymerase, physically moves to the promoter in a manner sensitive to physical blocks (by other proteins) or regulatory signals (to terminate movement).

The generality of enhancement is not yet clear. We do not know what proportion of cellular promoters usually rely on an enhancer to achieve their customary level of expression. Nor do we know how often an enhancer provides a target for regulation. Some enhancers are activated only in the tissues in which their genes function, but others could be active in all cells.

In one particularly striking case, an enhancer is responsible for a hormonal response. Transcription of the mouse mammary tumor virus DNA is stimulated by steroid hormones. The element responsible for the hormone response is located ~100 bp upstream of the startpoint, binds a complex consisting of the hormone and its protein receptor, and can stimulate the function of other genes when placed in either orientation at variable distances from their promoters. In short, it behaves as an enhancer associated with a hormone-binding sequence.

Enhancer activation may provide the general mechanism by which steroids regulate a set of target genes. **Figure 9.19** illustrates the processes involved for glucocorticoids, about which we have the most informa-

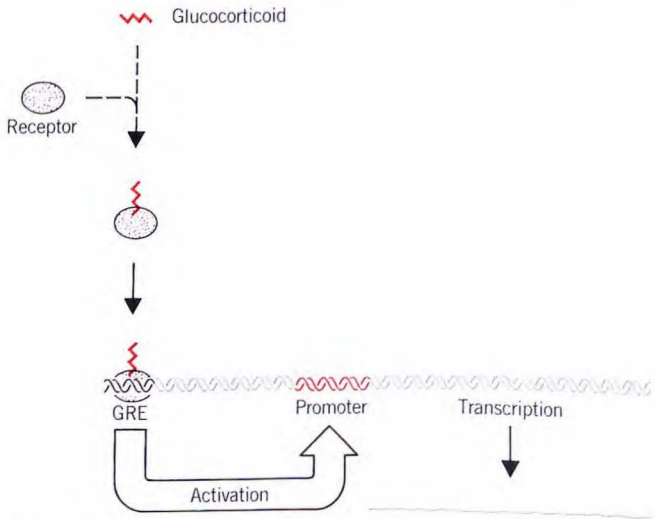


Figure 9.19
Glucocorticoids regulate gene transcription by causing their receptor to bind to an enhancer whose action is needed for promoter function.

tion. The glucocorticoid enters the cell and binds the glucocorticoid receptor. Binding activates the receptor, which then recognizes a consensus sequence present in enhancers located near genes that respond to glucocorticoids. When the glucocorticoid-receptor complex binds to the enhancer, the promoter nearby is activated, and transcription initiates there.

Although a consensus sequence is common to glucocorticoid-regulated enhancers, the length of the sequence bound by the receptor and its location relative to the promoter are variable. The consensus sequence is essential for binding, but surrounding sequences also are necessary. The glucocorticoid response element (GRE) may be several kilobases upstream or downstream of the promoter.

In one case, the activity of an enhancer is specifically inhibited, so it possible that gene activity may sometimes be controlled by preventing an enhancer from functioning.

RNA POLYMERASE III HAS A DOWNSTREAM PROMOTER

Before the promoter of the genes coding for 5S RNA in *X. laevis* was identified, all attempts to identify promoter sequences assumed that they would lie upstream of the startpoint. But in the 5S RNA genes, transcribed by RNA polymerase III, the promoter lies well *within* the transcription unit, more than 50 bases downstream of the startpoint.

A 5S RNA gene can be transcribed when a plasmid carrying it is used as template for a nuclear extract obtained from *X. laevis* oocytes, the tissue in which the gene usually is expressed. The promoter was located by using plasmids in which deletions extended into the gene from either direction. The 5S RNA product continues to be synthesized when the entire sequence upstream of the gene is removed.

When the deletions continue into the gene, a product very similar in size to the usual 5S RNA continues to be synthesized so long as the deletion ends before about base +55. The first part of the RNA product represents plasmid DNA; the second part represents whatever segment remains of the usual 5S RNA sequence. But when the deletion extends past +55, transcription does not occur. Thus the promoter lies *downstream of position +55*, but causes RNA polymerase III to initiate transcription a more or less fixed distance away. The wild-type startpoint is unique; in deletions that lack it, transcription initiates at the purine base nearest to the position 55 bp upstream of the promoter.

When deletions extend into the gene from its distal end, transcription is unaffected so long as the first 80 bp remain intact. Once the deletion cuts into this region, transcription ceases. This places the downstream boundary position of the promoter at about position +80.

So the promoter for 5S RNA transcription lies between positions +55 and +80 within the gene. A fragment containing this region can sponsor initiation of any DNA in which it is placed, at a position ~55 bp upstream. How does RNA polymerase initiate transcription upstream of its promoter? The most likely explanation is that the enzyme binds to the promoter, but is large enough to contact regions 55 bp away.

As with RNA polymerase II, the geometry of binding to the promoter must dictate the position of the startpoint, subject to the reservation that pyrimidines cannot be used for initiation. The difference between the enzymes is that RNA polymerase II reaches forward to the startpoint from its promoter, whereas RNA polymerase III reaches backward.

The boundaries of some RNA polymerase III promoters are summarized in **Figure 9.20.** They always lie within the transcription unit. In tRNA genes, the promoter lies in two separate parts, both within the gene. Deletion mapping shows that the sequences of both the *A block,* lying between +8 and +30, and the *B block,* lying between +51 and +72, must be present. Changes in the sequence between the blocks have no effect. Any deletion that reduces their separation prevents initiation; the separation can be increased by <30 bp without effect, but longer insertions may inhibit initiation.

A tRNA promoter therefore consists of two separate regions of ~20 bp each, which must be separated by >20 bp and may not be located too far apart. The sequences of these regions are highly conserved in eukaryotic tRNAs, a fact that had been interpreted solely in terms of tRNA function, but that now also may be attributed to the needs of the promoter. (The tRNA promoter is related to the 5S RNA promoter as seen by homology between the A block and the 5′ end of the 5S RNA promoter.)

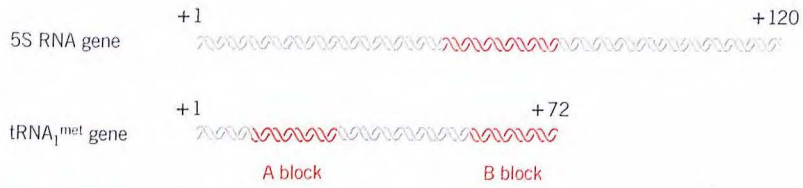


Figure 9.20
Genes for RNA polymerase III all contain internal promoters (regions in red). The internal control region may be a single block (as in 5S genes) or two separate blocks (as in tRNA genes). The numbers indicate the first and last bases of the gene.

The internal location of the promoter poses an important question. When a promoter lies outside the transcription unit itself, presumably it can evolve freely just to meet the needs of the enzyme. But the sequences needed for initiation of the RNA polymerase III transcription units lie within rather different types of genes, and are therefore constrained to meet the needs of products as diverse as 5S RNA, VA RNA, tRNA, and small nuclear RNAs. How then are they able also to provide whatever features are needed for recognition by RNA polymerase III?

An alternative to imposing a requirement that the enzyme can recognize a wide variety of promoter sequences is to suppose that it acts via an ancillary factor. A different factor could be responsible for allowing the enzyme to bind to each type of promoter. Factors specific for particular genes have been found in several systems.

RNA polymerase III can transcribe the *Xenopus* 5S RNA genes only in the presence of an added factor, a 37,000 dalton protein (TFIIIA) that binds to the region from +45 to +96. (The protein serves a dual purpose; it also binds the 5S product in the oocyte.) Two other factors are also needed, but they are necessary for the transcription of *all* class III genes, whereas the 37,000 dalton protein is specific for the 5S genes.

Thus initiation may depend on three classes of protein: RNA polymerase III itself, factor(s) involved generally in initiation, and factor(s) specific for an individual gene or genes.

The factors form a **preinitiation complex** at the promoter. At the 5S promoter in *Xenopus,* a single copy of the 37,000 dalton protein is bound to the control region. The complex also includes the other two factors.

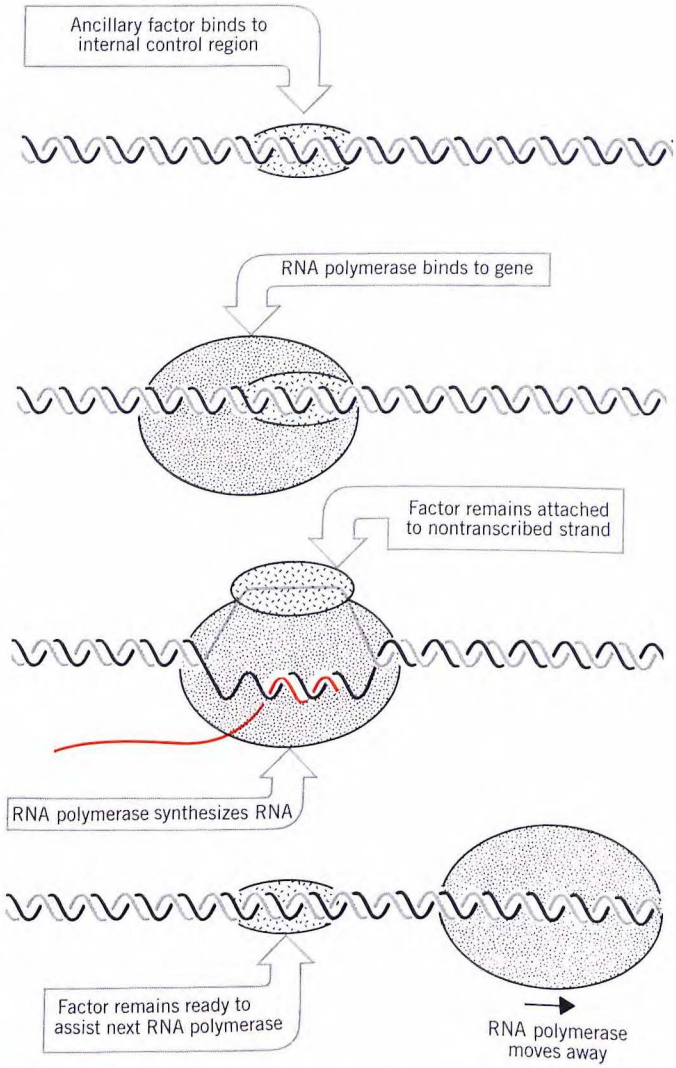


Figure 9.21
The ancillary control factor for the 5S gene binds to the noncoding strand, so it can remain attached to the DNA through many cycles of transcription.

Three factors, A, B, and C, are also needed to initiate 5S RNA transcription in human cells. But with tRNA genes, two of the factors, B and C, form the preinitiation complex; and with the VA1 gene, factor C alone forms the complex (although factor B remains necessary for initiation). In yeast, two factors, B and C, are needed to transcribe tRNA genes; but individual genes vary in their requirements for each factor. Both the A and B sequence blocks are needed for factor binding, and their separation also is important.

The existence of a preinitiation complex signals that the gene is in an "active" state, ready to be transcribed. RNA polymerase can bind to the gene only when the factor(s) have previously bound. The complex is stable, and may remain in existence through many cycles of replication. The ability to form a preinitiation complex could be a general regulatory mechanism. By binding to a promoter to make it possible for RNA polymerase in turn to bind, the factor in effect switches the gene on.

Figure 9.21 presents a model to explain how one molecule of factor could stimulate many successive rounds of transcription from the promoter. By remaining bound to the noncoding strand through the initiation process, the factor is not displaced. We may surmise that other factors of this sort remain to be characterized and will prove to be important in promoter selection.

Although the ability to transcribe these genes is conferred by the internal promoter, the startpoint does have some influence. Changes in the region immediately upstream of the startpoint can alter the efficiency of transcription. Thus the primary responsibility for recognition lies with the internal promoter; but some responsibility for establishing the frequency of initiation lies with the region at the startpoint.

FURTHER READING

A source for many reviews and original research articles is the volume edited by **Losick & Chamberlin,** *RNA Polymerase* (Cold Spring Harbor Laboratory, New York, 1976). Two chapters that cover the general matters dealt with here were written by **Chamberlin**, giving first a general overview (pp. 17–68) and then an account of the interactions of the bacterial enzyme with its template (pp. 159–192). The molecular interaction of bacterial RNA polymerase with its promoters has been well described by **Siebenlist, Simpson & Gilbert** (*Cell* **20,** 269–281, 1980). The stages of the interaction have been reviewed by **Von Hippel et al.** (*Ann. Rev. Biochem.* **53,** 389–449, 1984) and **McClure** (*Ann. Rev. Biochem.* **54,** 171–204, 1985). Initiation and its control has been reviewed by **Reznikoff et al.** (*Ann. Rev. Genet.* **19,** 355–387, 1985. Changes induced by sporulation were reviewed by **Losick & Pero** (*Cell* **25,** 582–584, 1981).

Techniques in characterizing eukaryotic RNA polymerase II promoters were discussed by **Corden et al.** (*Science* **209,** 1406–1414, 1981). The discrepancy between promoter function *in vivo* and *in vitro* was brought to light by **McKnight et al.** (*Cell* **25,** 385–398, 1981), who also dissected promoter components (*Science* **217,** 316–324, 1982; *Cell* **31,** 355–365, 1982), as did **Dierks et al.** (*Cell* **32,** 695–706, 1983). Promoter components and factor binding have been reviewed by **McKnight & Tjian** (*Cell* **46,** 795–805, 1986). A thoughtful analysis of the SV40 enhancer has been provided by **Banerji, Rusconi & Schaffner** (*Cell* **27,** 299–308, 1981); regulation of enhancers by steroid receptors has been reviewed by **Yamamoto** (*Ann. Rev. Genet.* **19,** 209–252, 1985). Transcription of 5S genes has been reviewed by **Korn** (*Nature* **295,** 101–105, 1982). The nature of tRNA split promoters has been reviewed by **Hall et al.** (*Cell* **29,** 3–5, 1982).

Primer is a short sequence (often of RNA) that is paired with one strand of DNA and provides a free 3′—OH end at which a DNA polymerase starts synthesis of a deoxyribonucleotide chain.

Primosome describes the complex of proteins involved in the priming action that initiates synthesis of each Okazaki fragment during discontinuous DNA replication; the primosome may move along DNA to engage in successive priming events.

Prokaryotic organisms (bacteria) lack nuclei.

Processed pseudogene is an inactive gene copy that lacks introns, contrasted with the interrupted structure of the active gene. Such genes presumably originate by reverse transcription of mRNA and insertion of a duplex copy into the genome.

Processive enzymes continue to act on a particular substrate, that is, do not dissociate between repetitions of the catalytic event.

Promoter is a region of DNA involved in binding of RNA polymerase to initiate transcription.

Proofreading refers to any mechanism for correcting errors in protein or nucleic acid synthesis that involves scrutiny of individual units *after* they have been added to the chain.

Prophage is a phage genome covalently integrated as a linear part of the bacterial chromosome.

Proto-oncogenes are the normal counterparts in the eukaryotic genome to the oncogenes carried by some retroviruses. They are given names of the form *c-onc*.

Provirus is a duplex DNA sequence in the eukaryotic chromosome corresponding to the genome of an RNA retrovirus.

Pseudogenes are inactive but stable components of the genome derived by mutation of an ancestral active gene.

Puff is an expansion of a band of a polytene chromosome associated with the synthesis of RNA at some locus in the band.

Pulse-chase experiments are performed by incubating cells very briefly with a radioactively labeled precursor (of some pathway or macromolecule); then the fate of the label is followed during a subsequent incubation with a nonlabeled precursor.

Quaternary structure of a protein refers to its multimeric constitution.

Quick-stop *dna* mutants of *E. coli* cease replication immediately when the temperature is increased to 42°C.

R loop is the structure formed when an RNA strand hybridizes with its complementary strand in a DNA duplex, thereby displacing the original strand of DNA in the form of a loop extending over the region of hybridization.

Rapid lysis (*r*) mutants display a change in the pattern of lysis of *E. coli* at the end of an infection by a T-even phage.

Reading frame is one of three possible ways of reading a nucleotide sequence as a series of triplets.

Reassociation of DNA describes the pairing of complementary single strands to form a double helix.

RecA is the product of the *recA* locus of *E. coli;* a protein with dual activities, acting as a protease and also able to exchange single strands of DNA molecules. The protease activity controls the SOS response; the nucleic acid handling facility is involved in recombination-repair pathways.

Recessive allele is obscured in the phenotype of a heterozygote by the dominant allele, often due to inactivity or absence of the product of the recessive allele.

Recessive lethal is an allele that is lethal when the cell is homozygous for it.

Reciprocal recombination is the production of new genotypes with the reverse arrangements of alleles according to maternal and paternal origin.

Reciprocal translocation exchanges part of one chromosome with part of another chromosome.

Recombinant progeny have a different genotype from that of either parent.

Recombinant joint is the point at which two recombining molecules of duplex DNA are connected (the edge of the heteroduplex region).

Recombination nodules (nodes) are dense objects present on the synaptonemal complex; could be involved in crossing-over.

Recombination-repair is a mode of filling a gap in one strand of duplex DNA by retrieving a homologous single strand from another duplex.

Regulatory gene codes for an RNA or protein product whose function is to control the expression of other genes.

Relaxed mutants of *E. coli* do not display the stringent response to starvation for amino acids (or other nutritional deprivation).

Relaxed replication control refers to the ability of some plasmids to continue replicating after bacteria cease dividing.

Release (termination) factors respond to nonsense codons to cause release of the completed polypeptide chain and the ribosome from mRNA.

Renaturation is the reassociation of denatured complementary single strands of a DNA double helix. Also used to describe recovery of structure by denatured protein.

Repeating unit in a tandem cluster is the length of the sequence that is repeated; appears circular on a restriction map.