

MOLECULAR BIOLOGY OF THE CELL

SED

ALBERTS

JOHNSON

LEWIS

RAFF

ROBERTS

WALTER

F O U R T H E D I T I O N

MOLECULAR BIOLOGY OF
THE CELL
fourth edition

Bruce Alberts
Alexander Johnson
Julian Lewis
Martin Raff
Keith Roberts
Peter Walter

 **Garland Science**
Taylor & Francis Group

Garland

Vice President: Denise Schanck
Managing Editor: Sarah Gibbs
Senior Editorial Assistant: Kirsten Jenner
Managing Production Editor: Emma Hunt
Proofreader and Layout: Emma Hunt
Production Assistant: Angela Bennett
Text Editors: Marjorie Singer Anderson and Betsy Dileria
Copy Editor: Bruce Goatly
Word Processors: Fran Dependahl, Misty Landers and Carol Winter
Designer: Blink Studio, London
Illustrator: Nigel Orme
Indexer: Janine Ross and Sherry Granum
Manufacturing: Nigel Eyre and Marion Morrow

Cell Biology Interactive

Artistic and Scientific Direction: Peter Walter
Narrated by: Julie Theriot
Production, Design, and Development: Mike Morales

Bruce Alberts received his Ph.D. from Harvard University and is President of the National Academy of Sciences and Professor of Biochemistry and Biophysics at the University of California, San Francisco. **Alexander Johnson** received his Ph.D. from Harvard University and is a Professor of Microbiology and Immunology and Co-Director of the Biochemistry and Molecular Biology Program at the University of California, San Francisco.

Julian Lewis received his D.Phil. from the University of Oxford and is a Principal Scientist at the London Research Institute of Cancer Research UK. **Martin Raff** received his M.D. from McGill University and is at the Medical Research Council Laboratory for Molecular Cell Biology and Cell Biology Unit and in the Biology Department at University College London. **Keith Roberts** received his Ph.D. from the University of Cambridge and is Associate Research Director at the John Innes Centre, Norwich. **Peter Walter** received his Ph.D. from The Rockefeller University in New York and is Professor and Chairman of the Department of Biochemistry and Biophysics at the University of California, San Francisco, and an Investigator of the Howard Hughes Medical Institute.

© 2002 by Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter.

© 1983, 1989, 1994 by Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, and James D. Watson.

All rights reserved. No part of this book covered by the copyright hereon may be reproduced or used in any format in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or information storage and retrieval systems—without permission of the publisher.

Library of Congress Cataloging-in-Publication Data

Molecular biology of the cell / Bruce Alberts ... [et al.]. -- 4th ed.
p. cm
Includes bibliographical references and index.
ISBN 0-8153-3218-1 (hardbound) -- ISBN 0-8153-4072-9 (pbk.)
1. Cytology. 2. Molecular biology. I. Alberts, Bruce.
[DNLM: 1. Cells. 2. Molecular Biology.]
QH581.2 .M64 2002
571.6--dc21

2001054471 CIP

Published by Garland Science, a member of the Taylor & Francis Group,
29 West 35th Street, New York, NY 10001-2299

Printed in the United States of America

15 14 13 12 11 10 9 8 7 6 5 4 3 2

Front cover Human Genome: Reprinted by permission from *Nature*, International Human Genome Sequencing Consortium, 409:860–921, 2001 © Macmillan Magazines Ltd. Adapted from an image by Francis Collins, NHGRI; Jim Kent, UCSC; Ewan Birney, EBI; and Darryl Leja, NHGRI; showing a portion of Chromosome 1 from the initial sequencing of the human genome.

Chapter opener Portion of chromosome 2 from the genome of the fruit fly *Drosophila melanogaster*. (Reprinted with permission from M.D. Adams et al., *Science* 287:2185–2195, 2000. © AAAS.)

Back cover In 1967, the British artist Peter Blake created a design classic. Nearly 35 years later Nigel Orme (illustrator), Richard Denyer (photographer), and the authors have together produced an affectionate tribute to Mr Blake's image. With its gallery of icons and influences, its assembly created almost as much complexity, intrigue and mystery as the original. *Drosophila*, *Arabidopsis*, Dolly and the assembled company tempt you to dip inside where, as in the original, "a splendid time is guaranteed for all." (Gunter Blobel, courtesy of The Rockefeller University; Marie Curie, Keystone Press Agency Inc; Darwin bust, by permission of the President and Council of the Royal Society; Rosalind Franklin, courtesy of Cold Spring Harbor Laboratory Archives; Dorothy Hodgkin, © The Nobel Foundation, 1964; James Joyce, etching by Peter Blake; Robert Johnson, photo booth self-portrait early 1930s, © 1986 Delta Haze Corporation all rights reserved, used by permission; Albert L. Lehninger, (unidentified photographer) courtesy of The Alan Mason Chesney Medical Archives of The Johns Hopkins Medical Institutions; Linus Pauling, from Ava Helen and Linus Pauling Papers, Special Collections, Oregon State University; Nicholas Poussin, courtesy of ArtToday.com; Barbara McClintock, © David Micklos, 1983; Andrei Sakharov, courtesy of Elena Bonner; Frederick Sanger, © The Nobel Foundation, 1958.)

Preface

"There is a paradox in the growth of scientific knowledge. As information accumulates in ever more intimidating quantities, disconnected facts and impenetrable mysteries give way to rational explanations, and simplicity emerges from chaos."

Thus began the preface of our first edition, written 18 years ago. Much of what we wrote in that preface holds for the present edition too. Our goals have not changed: we want to make cell biology comprehensible. We aim, as before, to give readers a perspective both on what is known, and on what is unknown. We have written the book for a wide range of students, but it should also prove useful for scientists wishing to follow progress outside their own specialized fields. As in past editions, each chapter is a joint composition of multiple authors and has been reviewed by a number of experts. This helps to explain why this edition, like the first, has been "a long time in gestation—three times longer than an elephant, five times longer than a whale."

Although information in the life sciences is expanding rapidly, human brain capacity is not. This discrepancy creates an increasing challenge for textbook writers, teachers, and especially students. We have been forced to think harder than ever in deciding what facts and concepts are essential. We have had to summarize—omitting much detail that, in overabundance, can prevent understanding. But specific examples are still needed to bring the subject to life. Thus, the most difficult part of the revision has been deciding what to leave out, while the most rewarding has been the opportunity that new discoveries provide to strengthen the conceptual framework.

What is new in the 4th edition?

Genomics has given us a new perspective that has demanded a complete recasting and expansion of the material on molecular genetics (Chapter 1 and Chapters 4 through 8). These six chapters can now be used as a stand-alone text in molecular biology. As before, the book ends with a major section on Cells in Their Social Context, but we have added a new chapter on Pathogens, Infection, and Innate Immunity. This reflects the remarkable advances in the understanding of the cell biology of infection, as well as the renewed awareness that infectious disease remains one of the greatest unconquered dangers in our world.

As our book makes clear, the complete sequencing of the genomes of hundreds of organisms, from bacteria to humans, has revolutionized our understanding of living things and the relationships between them. At last we can see what is there: the set of genes and proteins is finite, and we can list them. But we also recognize that these components are combined for use in marvelously subtle and complex ways, even in the simplest of organisms. Therefore, the traditional explanatory cartoons that we show on nearly every page of the book generally represent only the primitive first step toward an explanation. These drawings cannot capture the enormous complexity of the networks of protein-protein interactions that are responsible for most intracellular processes, whose understanding will require new and more quantitative forms of analysis. Thus, we are no longer as confident as we were 18 years ago that simplicity will eventually

emerge from the complexity. The extreme sophistication of cellular mechanisms will challenge cell biologists throughout the new century, which is very good news for the many young scientists who will succeed us.

As never before, new imaging and computer technologies have changed the ways we can observe the inner workings of living cells. We have tried to capture some of the excitement of these observations in *Cell Biology Interactive*, a CD-ROM disk that is included with each book. It contains dozens of video clips, animations, molecular structures, and high-resolution micrographs, which complement the static material in individual book chapters. One cannot watch cells crawling, dividing, segregating their chromosomes, or rearranging their surface without feeling curious about the molecular mechanisms that underlie these processes. We hope that *Cell Biology Interactive* will stimulate this curiosity in students and thereby make the learning of cell biology easier and more rewarding. We also hope that instructors will use these visual resources in the classroom for the same purpose.

We are deeply indebted to the many people who have helped us with this revision. The experts who critically reviewed specific chapters are acknowledged separately on p. xxix. We are especially grateful to Julie Theriot, who largely wrote both Chapter 16 and Chapter 25; we have all benefited from her wisdom. We are also grateful to the other experts who made major contributions to individual chapters: Nancy Craig helped to revise part of Chapter 5; Maynard Olson drafted the section on genome evolution in Chapter 7; Peter Shaw helped to revise Chapter 9; David Morgan largely wrote Chapter 17; Lisa Satterwhite prepared initial drafts of Chapter 18; Robert Kypta largely revised Chapter 19; Cori Bargmann helped us to restructure Chapter 21; and Paul Edwards played a central part in the revision of Chapter 23. Karen Hopkin drafted large parts of Chapters 8 and 23. We are also very grateful to the many scientists who generously provided micrographs for the book and materials for the CD-ROM.

Finally we are indebted to the outstanding staff of Garland Publishing. Nigel Orme oversaw the production of the final artwork with remarkable skill and speed. Mike Morales put in endless hours tackling the diverse challenges of developing the CD-ROM. Emma Hunt skillfully and artfully set the entire text and figures into pages. Sarah Gibbs and Kirsten Jenner kept us organized. Adam Sendroff and Nasreen Arain connected us to our customers. Eleanor Lawrence prepared the Glossary, and Mary Purton collected many of the references. Through it all, Denise Schanck calmly directed the whole effort with great skill. Last, but not least, Tim Hunt and John Wilson once again created a masterful Problems Book that supplements the main text; they worked side-by-side with us and were a constant source of wise advice. Their clever problems illustrate how discoveries are made and provide a unique way of learning cell biology.

It goes without saying that our book could not have been written without the strong support of our families and the forbearance of our friends, colleagues, and students. For decades, they have had to put up with our absences at frequent and lengthy book meetings, where most of the writing was done. They have helped us in innumerable ways, and we are grateful to them all.

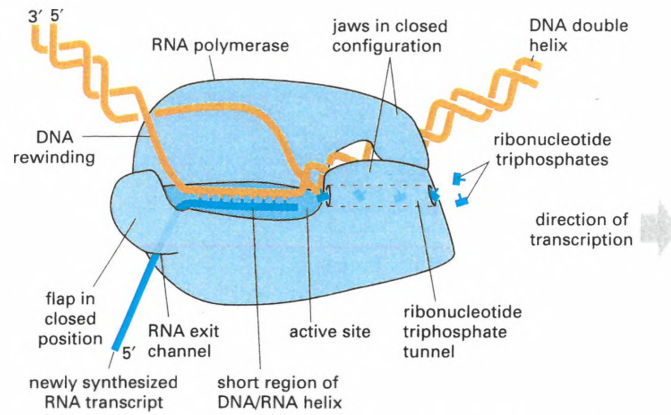


Figure 6-8 DNA is transcribed by the enzyme RNA polymerase. The RNA polymerase (pale blue) moves stepwise along the DNA, unwinding the DNA helix at its active site. As it progresses, the polymerase adds nucleotides (here, small "T" shapes) one by one to the RNA chain at the polymerization site using an exposed DNA strand as a template. The RNA transcript is thus a single-stranded complementary copy of one of the two DNA strands. The polymerase has a rudder (see Figure 6-11) that displaces the newly formed RNA, allowing the two strands of DNA behind the polymerase to rewind. A short region of DNA/RNA helix (approximately nine nucleotides in length) is therefore formed only transiently, and a "window" of DNA/RNA helix therefore moves along the DNA with the polymerase. The incoming nucleotides are in the form of ribonucleoside triphosphates (ATP, UTP, CTP, and GTP), and the energy stored in their phosphate-phosphate bonds provides the driving force for the polymerization reaction (see Figure 5-4). (Adapted from a figure kindly supplied by Robert Landick.)

The final product of a minority of genes, however, is the RNA itself. Careful analysis of the complete DNA sequence of the genome of the yeast *S. cerevisiae* has uncovered well over 750 genes (somewhat more than 10% of the total number of yeast genes) that produce RNA as their final product, although this number includes multiple copies of some highly repeated genes. These RNAs, like proteins, serve as enzymatic and structural components for a wide variety of processes in the cell. In Chapter 5 we encountered one of those RNAs, the template carried by the enzyme telomerase. Although not all of their functions are known, we see in this chapter that some *small nuclear RNA (snRNA)* molecules direct the splicing of pre-mRNA to form mRNA, that *ribosomal RNA (rRNA)* molecules form the core of ribosomes, and that *transfer RNA (tRNA)* molecules form the adaptors that select amino acids and hold them in place on a ribosome for incorporation into protein (Table 6-1).

Each transcribed segment of DNA is called a *transcription unit*. In eucaryotes, a transcription unit typically carries the information of just one gene, and therefore codes for either a single RNA molecule or a single protein (or group of related proteins if the initial RNA transcript is spliced in more than one way to produce different mRNAs). In bacteria, a set of adjacent genes is often transcribed as a unit; the resulting mRNA molecule therefore carries the information for several distinct proteins.

Overall, RNA makes up a few percent of a cell's dry weight. Most of the RNA in cells is rRNA; mRNA comprises only 3–5% of the total RNA in a typical mammalian cell. The mRNA population is made up of tens of thousands of different species, and there are on average only 10–15 molecules of each species of mRNA present in each cell.

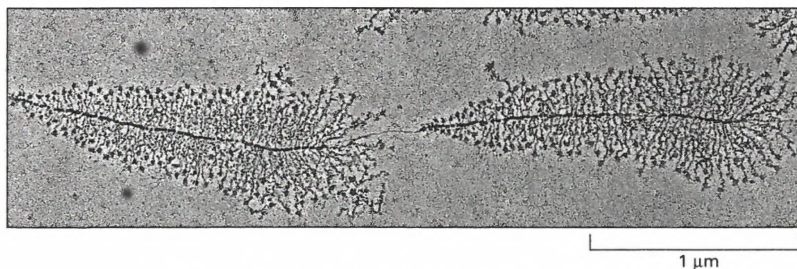


Figure 6-9 Transcription of two genes as observed under the electron microscope. The micrograph shows many molecules of RNA polymerase simultaneously transcribing each of two adjacent genes. Molecules of RNA polymerase are visible as a series of dots along the DNA with the newly synthesized transcripts (fine threads) attached to them. The RNA molecules (ribosomal RNAs) shown in this example are not translated into protein but are instead used directly as components of ribosomes, the machines on which translation takes place. The particles at the 5' end (the free end) of each rRNA transcript are believed to reflect the beginnings of ribosome assembly. From the lengths of the newly synthesized transcripts, it can be deduced that the RNA polymerase molecules are transcribing from left to right. (Courtesy of Ulrich Scheer.)

TABLE 6-1 Principal Types of RNAs Produced in Cells

TYPE OF RNA	FUNCTION
mRNAs	messenger RNAs, code for proteins
rRNAs	ribosomal RNAs, form the basic structure of the ribosome and catalyze protein synthesis
tRNAs	transfer RNAs, central to protein synthesis as adaptors between mRNA and amino acids
snRNAs	small nuclear RNAs, function in a variety of nuclear processes, including the splicing of pre-mRNA
snoRNAs	small nucleolar RNAs, used to process and chemically modify rRNAs
Other noncoding RNAs	function in diverse cellular processes, including telomere synthesis, X-chromosome inactivation, and the transport of proteins into the ER

Signals Encoded in DNA Tell RNA Polymerase Where to Start and Stop

To transcribe a gene accurately, RNA polymerase must recognize where on the genome to start and where to finish. The way in which RNA polymerases perform these tasks differs somewhat between bacteria and eucaryotes. Because the process in bacteria is simpler, we look there first.

The initiation of transcription is an especially important step in gene expression because it is the main point at which the cell regulates which proteins are to be produced and at what rate. Bacterial RNA polymerase is a multisubunit complex. A detachable subunit, called *sigma* (σ) *factor*, is largely responsible for its ability to read the signals in the DNA that tell it where to begin transcribing (Figure 6-10). RNA polymerase molecules adhere only weakly to the bacterial DNA when they collide with it, and a polymerase molecule typically slides rapidly along the long DNA molecule until it dissociates again. However, when the polymerase slides into a region on the DNA double helix called a **promoter**, a special sequence of nucleotides indicating the starting point for RNA synthesis, it binds tightly to it. The polymerase, using its σ factor, recognizes this DNA sequence by making specific contacts with the portions of the bases that are exposed on the outside of the helix (*Step 1* in Figure 6-10).

After the RNA polymerase binds tightly to the promoter DNA in this way, it opens up the double helix to expose a short stretch of nucleotides on each strand (*Step 2* in Figure 6-10). Unlike a DNA helicase reaction (see Figure 5-15), this limited opening of the helix does not require the energy of ATP hydrolysis. Instead, the polymerase and DNA both undergo reversible structural changes that result in a more energetically favorable state. With the DNA unwound, one of the two exposed DNA strands acts as a template for complementary base-pairing with incoming ribonucleotides (see Figure 6-7), two of which are joined together by the polymerase to begin an RNA chain. After the first ten or so nucleotides of RNA have been synthesized (a relatively inefficient process during which polymerase synthesizes and discards short nucleotide oligomers), the σ factor relaxes its tight hold on the polymerase and eventually dissociates from it. During this process, the polymerase undergoes additional structural changes that enable it to move forward rapidly, transcribing without the σ factor (*Step 4* in Figure 6-10). Chain elongation continues (at a speed of approximately 50 nucleotides/sec for bacterial RNA polymerases) until the enzyme encounters a second signal in the DNA, the **terminator** (described below), where the polymerase halts and releases both the DNA template and the newly made RNA chain (*Step 7* in Figure 6-10). After the polymerase has been released at a terminator, it reassociates with a free σ factor and searches for a new promoter, where it can begin the process of transcription again.

Several structural features of bacterial RNA polymerase make it particularly adept at performing the transcription cycle just described. Once the σ factor

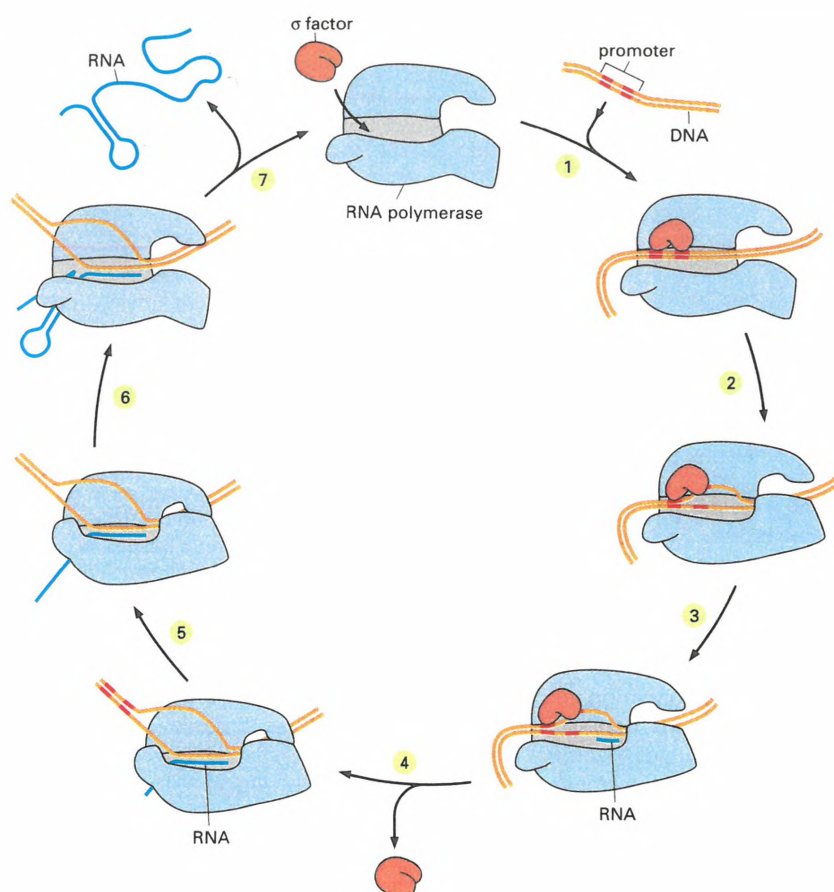


Figure 6-10 The transcription cycle of bacterial RNA polymerase. In step 1, the RNA polymerase holoenzyme (core polymerase plus σ factor) forms and then locates a promoter (see Figure 6-12). The polymerase unwinds the DNA at the position at which transcription is to begin (step 2) and begins transcribing (step 3). This initial RNA synthesis (sometimes called "abortive initiation") is relatively inefficient. However, once RNA polymerase has managed to synthesize about 10 nucleotides of RNA, σ relaxes its grip, and the polymerase undergoes a series of conformational changes (which probably includes a tightening of its jaws and the placement of RNA in the exit channel [see Figure 6-11]). The polymerase now shifts to the elongation mode of RNA synthesis (step 4), moving rightwards along the DNA in this diagram. During the elongation mode (step 5) transcription is highly processive, with the polymerase leaving the DNA template and releasing the newly transcribed RNA only when it encounters a termination signal (step 6). Termination signals are encoded in DNA and many function by forming an RNA structure that destabilizes the polymerase's hold on the RNA, as shown here. In bacteria, all RNA molecules are synthesized by a single type of RNA polymerase and the cycle depicted in the figure therefore applies to the production of mRNAs as well as structural and catalytic RNAs. (Adapted from a figure kindly supplied by Robert Landick.)

positions the polymerase on the promoter and the template DNA has been unwound and pushed to the active site, a pair of moveable jaws is thought to clamp onto the DNA (Figure 6-11). When the first 10 nucleotides have been transcribed, the dissociation of σ allows a flap at the back of the polymerase to

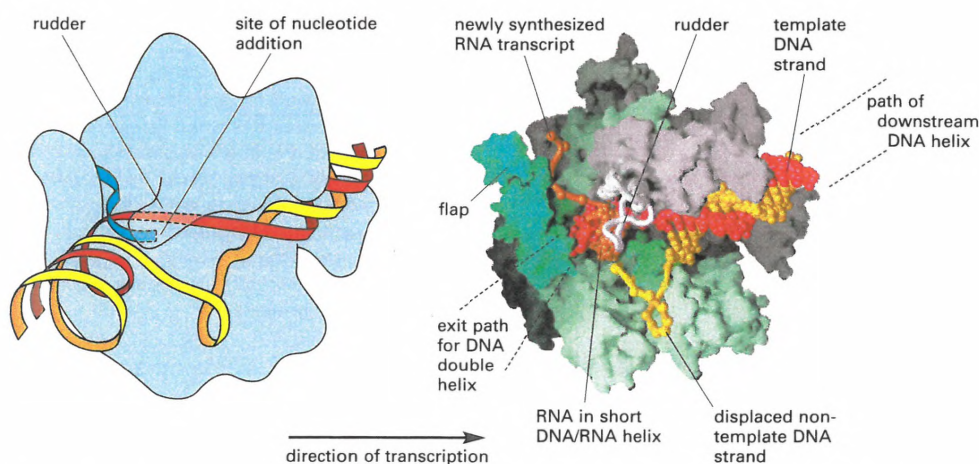


Figure 6-11 The structure of a bacterial RNA polymerase. Two depictions of the three-dimensional structure of a bacterial RNA polymerase, with the DNA and RNA modeled in. This RNA polymerase is formed from four different subunits, indicated by different colors (right). The DNA strand used as a template is red, and the non-template strand is yellow. The rudder wedges apart the DNA-RNA hybrid as the polymerase moves. For simplicity only the polypeptide backbone of the rudder is shown in the right-hand figure, and the DNA exiting from the polymerase has been omitted. Because the RNA polymerase is depicted in the elongation mode, the σ factor is absent. (Courtesy of Seth Darst.)

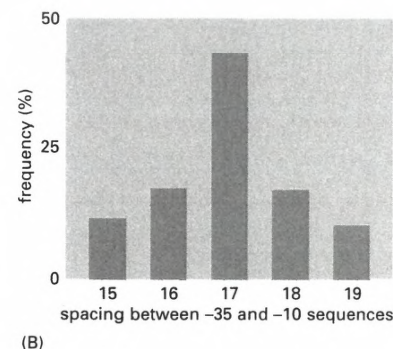
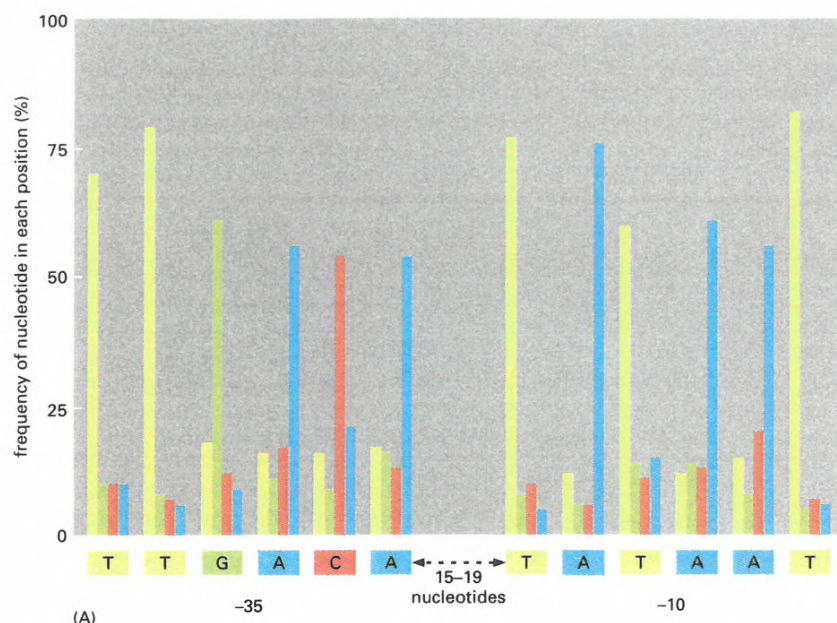


Figure 6-12 Consensus sequence for the major class of *E. coli* promoters.

(A) The promoters are characterized by two hexameric DNA sequences, the -35 sequence and the -10 sequence named for their approximate location relative to the start point of transcription (designated +1). For convenience, the nucleotide sequence of a single strand of DNA is shown; in reality the RNA polymerase recognizes the promoter as double-stranded DNA. On the basis of a comparison of 300 promoters, the frequencies of the four nucleotides at each position in the -35 and -10 hexamers are given. The consensus sequence, shown below the graph, reflects the most common nucleotide found at each position in the collection of promoters. The sequence of nucleotides between the -35 and -10 hexamers shows no significant similarities among promoters. (B) The distribution of spacing between the -35 and -10 hexamers found in *E. coli* promoters. The information displayed in these two graphs applies to *E. coli* promoters that are recognized by RNA polymerase and the major σ factor (designated σ^{70}). As we shall see in the next chapter, bacteria also contain minor σ factors, each of which recognizes a different promoter sequence. Some particularly strong promoters recognized by RNA polymerase and σ^{70} have an additional sequence, located upstream (to the left, in the figure) of the -35 hexamer, which is recognized by another subunit of RNA polymerase.

close to form an exit tunnel through which the newly made RNA leaves the enzyme. With the polymerase now functioning in its elongation mode, a rudder-like structure in the enzyme continuously pries apart the DNA-RNA hybrid formed. We can view the series of conformational changes that takes place during transcription initiation as a successive tightening of the enzyme around the DNA and RNA to ensure that it does not dissociate before it has finished transcribing a gene. If an RNA polymerase does dissociate prematurely, it cannot resume synthesis but must start over again at the promoter.

How do the signals in the DNA (termination signals) stop the elongating polymerase? For most bacterial genes a termination signal consists of a string of A-T nucleotide pairs preceded by a two-fold symmetric DNA sequence, which, when transcribed into RNA, folds into a "hairpin" structure through Watson-Crick base-pairing (see Figure 6-10). As the polymerase transcribes across a terminator, the hairpin may help to wedge open the movable flap on the RNA polymerase and release the RNA transcript from the exit tunnel. At the same time, the DNA-RNA hybrid in the active site, which is held together predominantly by U-A base pairs (which are less stable than G-C base pairs because they form two rather than three hydrogen bonds per base pair), is not sufficiently strong enough to hold the RNA in place, and it dissociates causing the release of the polymerase from the DNA, perhaps by forcing open its jaws. Thus, in some respects, transcription termination seems to involve a reversal of the structural transitions that happen during initiation. The process of termination also is an example of a common theme in this chapter: the ability of RNA to fold into specific structures figures prominently in many aspects of decoding the genome.

Transcription Start and Stop Signals Are Heterogeneous in Nucleotide Sequence

As we have just seen, the processes of transcription initiation and termination involve a complicated series of structural transitions in protein, DNA, and RNA molecules. It is perhaps not surprising that the signals encoded in DNA that specify these transitions are difficult for researchers to recognize. Indeed, a comparison of many different bacterial promoters reveals that they are heterogeneous in DNA sequence. Nevertheless, they all contain related sequences, reflecting in part aspects of the DNA that are recognized directly by the σ factor. These common features are often summarized in the form of a *consensus sequence* (Figure 6-12). In general, a consensus nucleotide sequence is derived

Figure 6-13 The importance of RNA polymerase orientation. The DNA strand serving as template must be traversed in a 3' to 5' direction, as illustrated in Figure 6-9. Thus, the direction of RNA polymerase movement determines which of the two DNA strands is to serve as a template for the synthesis of RNA, as shown in (A) and (B). Polymerase direction is, in turn, determined by the orientation of the promoter sequence, the site at which the RNA polymerase begins transcription.

by comparing many sequences with the same basic function and tallying up the most common nucleotide found at each position. It therefore serves as a summary or "average" of a large number of individual nucleotide sequences.

One reason that individual bacterial promoters differ in DNA sequence is that the precise sequence determines the strength (or number of initiation events per unit time) of the promoter. Evolutionary processes have thus finetuned each promoter to initiate as often as necessary and have created a wide spectrum of promoters. Promoters for genes that code for abundant proteins are much stronger than those associated with genes that encode rare proteins, and their nucleotide sequences are responsible for these differences.

Like bacterial promoters, transcription terminators also include a wide range of sequences, with the potential to form a simple RNA structure being the most important common feature. Since an almost unlimited number of nucleotide sequences have this potential, terminator sequences are much more heterogeneous than those of promoters.

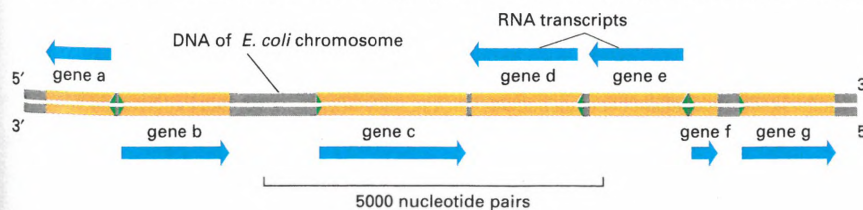
We have discussed bacterial promoters and terminators in some detail to illustrate an important point regarding the analysis of genome sequences. Although we know a great deal about bacterial promoters and terminators and can develop consensus sequences that summarize their most salient features, their variation in nucleotide sequence makes it difficult for researchers (even when aided by powerful computers) to definitively locate them simply by inspection of the nucleotide sequence of a genome. When we encounter analogous types of sequences in eucaryotes, the problem of locating them is even more difficult. Often, additional information, some of it from direct experimentation, is needed to accurately locate the short DNA signals contained in genomes.

Promoter sequences are asymmetric (see Figure 6-12), and this feature has important consequences for their arrangement in genomes. Since DNA is double-stranded, two different RNA molecules could in principle be transcribed from any gene, using each of the two DNA strands as a template. However a gene typically has only a single promoter, and because the nucleotide sequences of bacterial (as well as eucaryotic) promoters are asymmetric the polymerase can bind in only one orientation. The polymerase thus has no option but to transcribe the one DNA strand, since it can synthesize RNA only in the 5' to 3' direction (Figure 6-13). The choice of template strand for each gene is therefore determined by the location and orientation of the promoter. Genome sequences reveal that the DNA strand used as the template for RNA synthesis varies from gene to gene (Figure 6-14; see also Figure 1-31).

Having considered transcription in bacteria, we now turn to the situation in eucaryotes, where the synthesis of RNA molecules is a much more elaborate affair.

Transcription Initiation in Eucaryotes Requires Many Proteins

In contrast to bacteria, which contain a single type of RNA polymerase, eucaryotic nuclei have three, called *RNA polymerase I*, *RNA polymerase II*, and *RNA*



FROM DNA TO RNA

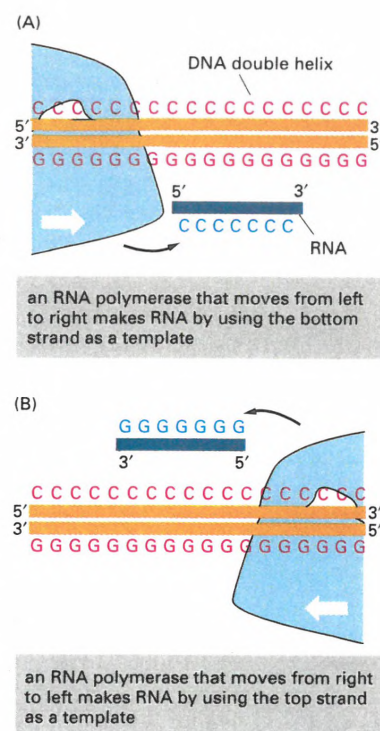


Figure 6-14 Directions of transcription along a short portion of a bacterial chromosome. Some genes are transcribed using one DNA strand as a template, while others are transcribed using the other DNA strand. The direction of transcription is determined by the promoter at the beginning of each gene (green arrowheads). Approximately 0.2% (9000 base pairs) of the *E. coli* chromosome is depicted here. The genes transcribed from left to right use the bottom DNA strand as the template; those transcribed from right to left use the top strand as the template.

TABLE 6-2 The Three RNA Polymerases in Eucaryotic Cells

TYPE OF POLYMERASE	GENES TRANSCRIBED
RNA polymerase I	5.8S, 18S, and 28S rRNA genes
RNA polymerase II	all protein-coding genes, plus snoRNA genes and some snRNA genes
RNA polymerase III	tRNA genes, 5S rRNA genes, some snRNA genes and genes for other small RNAs

polymerase III. The three polymerases are structurally similar to one another (and to the bacterial enzyme). They share some common subunits and many structural features, but they transcribe different types of genes (Table 6-2). RNA polymerases I and III transcribe the genes encoding transfer RNA, ribosomal RNA, and various small RNAs. RNA polymerase II transcribes the vast majority of genes, including all those that encode proteins, and our subsequent discussion therefore focuses on this enzyme.

Although eucaryotic RNA polymerase II has many structural similarities to bacterial RNA polymerase (Figure 6-15), there are several important differences in the way in which the bacterial and eucaryotic enzymes function, two of which concern us immediately.

1. While bacterial RNA polymerase (with σ factor as one of its subunits) is able to initiate transcription on a DNA template *in vitro* without the help of additional proteins, eucaryotic RNA polymerases cannot. They require the help of a large set of proteins called *general transcription factors*, which must assemble at the promoter with the polymerase before the polymerase can begin transcription.
2. Eucaryotic transcription initiation must deal with the packing of DNA into nucleosomes and higher order forms of chromatin structure, features absent from bacterial chromosomes.

RNA Polymerase II Requires General Transcription Factors

The discovery that, unlike bacterial RNA polymerase, purified eucaryotic RNA polymerase II could not initiate transcription *in vitro* led to the discovery and purification of the additional factors required for this process. These **general transcription factors** help to position the RNA polymerase correctly at the promoter, aid in pulling apart the two strands of DNA to allow transcription to begin, and release RNA polymerase from the promoter into the elongation mode once transcription has begun. The proteins are "general" because they assemble on all promoters used by RNA polymerase II; consisting of a set of interacting proteins, they are designated as *TFII* (for transcription factor for polymerase II),

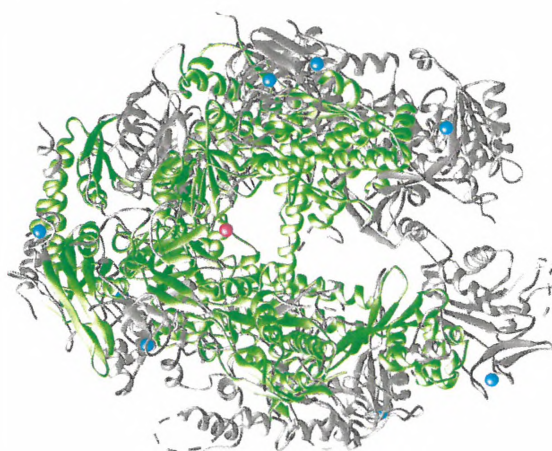
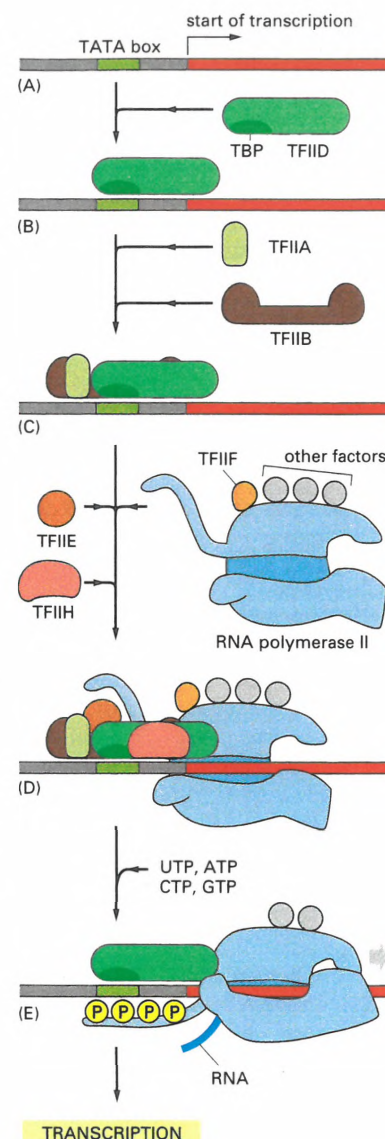


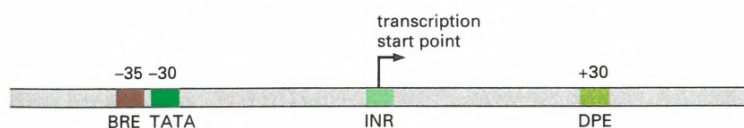
Figure 6-15 Structural similarity between a bacterial RNA polymerase and a eucaryotic RNA polymerase II. Regions of the two RNA polymerases that have similar structures are indicated in green. The eucaryotic polymerase is larger than the bacterial enzyme (12 subunits instead of 5), and some of the additional regions are shown in gray. The blue spheres represent Zn atoms that serve as structural components of the polymerases, and the red sphere represents the Mg atom present at the active site, where polymerization takes place. The RNA polymerases in all modern-day cells (bacteria, archaea, and eucaryotes) are closely related, indicating that the basic features of the enzyme were in place before the divergence of the three major branches of life. (Courtesy of P. Cramer and R. Kornberg.)

Figure 6-16 Initiation of transcription of a eucaryotic gene by RNA polymerase II. To begin transcription, RNA polymerase requires a number of general transcription factors (called TFIIA, TFIIB, and so on). (A) The promoter contains a DNA sequence called the TATA box, which is located 25 nucleotides away from the site at which transcription is initiated. (B) The TATA box is recognized and bound by transcription factor TFIID, which then enables the adjacent binding of TFIIB (C). For simplicity the DNA distortion produced by the binding of TFIID (see Figure 6-18) is not shown. (D) The rest of the general transcription factors, as well as the RNA polymerase itself, assemble at the promoter. (E) TFIIF then uses ATP to pry apart the DNA double helix at the transcription start point, allowing transcription to begin. TFIIF also phosphorylates RNA polymerase II, changing its conformation so that the polymerase is released from the general factors and can begin the elongation phase of transcription. As shown, the site of phosphorylation is a long C-terminal polypeptide tail that extends from the polymerase molecule. The assembly scheme shown in the figure was deduced from experiments performed *in vitro*, and the exact order in which the general transcription factors assemble on promoters in cells is not known with certainty. In some cases, the general factors are thought to first assemble with the polymerase, with the whole assembly subsequently binding to the DNA in a single step. The general transcription factors have been highly conserved in evolution; some of those from human cells can be replaced in biochemical experiments by the corresponding factors from simple yeasts.



and listed as TFIIA, TFIIB, and so on. In a broad sense, the eucaryotic general transcription factors carry out functions equivalent to those of the σ factor in bacteria.

Figure 6-16 shows how the general transcription factors assemble *in vitro* at promoters used by RNA polymerase II. The assembly process starts with the binding of the general transcription factor TFIID to a short double-helical DNA sequence primarily composed of T and A nucleotides. For this reason, this sequence is known as the TATA sequence, or **TATA box**, and the subunit of TFIID that recognizes it is called TBP (for TATA-binding protein). The TATA box is typically located 25 nucleotides upstream from the transcription start site. It is not the only DNA sequence that signals the start of transcription (Figure 6-17), but



element	consensus sequence	general transcription factor
BRE	G/C G/C G/A C G C C	TFIIB
TATA	T A T A A/T A A/T	TBP
INR	C/T C/T A N T/A C/T C/T	TFIID
DPE	A/G G A/T C G T G	TFIID

Figure 6-17 Consensus sequences found in the vicinity of eucaryotic RNA polymerase II start points. The name given to each consensus sequence (first column) and the general transcription factor that recognizes it (last column) are indicated. N indicates any nucleotide, and two nucleotides separated by a slash indicate an equal probability of either nucleotide at the indicated position. In reality, each consensus sequence is a shorthand representation of a histogram similar to that of Figure 6-12. For most RNA polymerase II transcription start points, only two or three of the four sequences are present. For example, most polymerase II promoters have a TATA box sequence, and those that do not typically have a "strong" INR sequence. Although most of the DNA sequences that influence transcription initiation are located "upstream" of the transcription start point, a few, such as the DPE shown in the figure, are located in the transcribed region.

for most polymerase II promoters, it is the most important. The binding of TFIID causes a large distortion in the DNA of the TATA box (Figure 6-18). This distortion is thought to serve as a physical landmark for the location of an active promoter in the midst of a very large genome, and it brings DNA sequences on both sides of the distortion together to allow for subsequent protein assembly steps. Other factors are then assembled, along with RNA polymerase II, to form a complete *transcription initiation complex* (see Figure 6-16).

After RNA polymerase II has been guided onto the promoter DNA to form a transcription initiation complex, it must gain access to the template strand at the transcription start point. This step is aided by one of the general transcription factors, TFIIF, which contains a DNA helicase. Next, like the bacterial polymerase, polymerase II remains at the promoter, synthesizing short lengths of RNA until it undergoes a conformational change and is released to begin transcribing a gene. A key step in this release is the addition of phosphate groups to the "tail" of the RNA polymerase (known as the CTD or C-terminal domain). This phosphorylation is also catalyzed by TFIIF, which, in addition to a helicase, contains a protein kinase as one of its subunits (see Figure 6-16, D and E). The polymerase can then disengage from the cluster of general transcription factors, undergoing a series of conformational changes that tighten its interaction with DNA and acquiring new proteins that allow it to transcribe for long distances without dissociating.

Once the polymerase II has begun elongating the RNA transcript, most of the general transcription factors are released from the DNA so that they are available to initiate another round of transcription with a new RNA polymerase molecule. As we see shortly, the phosphorylation of the tail of RNA polymerase II also causes components of the RNA processing machinery to load onto the polymerase and thus be in position to modify the newly transcribed RNA as it emerges from the polymerase.

Polymerase II Also Requires Activator, Mediator, and Chromatin-modifying Proteins

The model for transcription initiation just described was established by studying the action of RNA polymerase II and its general transcription factors on purified DNA templates *in vitro*. However, as discussed in Chapter 4, DNA in eucaryotic cells is packaged into nucleosomes, which are further arranged in higher-order chromatin structures. As a result, transcription initiation in a eucaryotic cell is more complex and requires more proteins than it does on purified DNA. First, gene regulatory proteins known as *transcriptional activators*

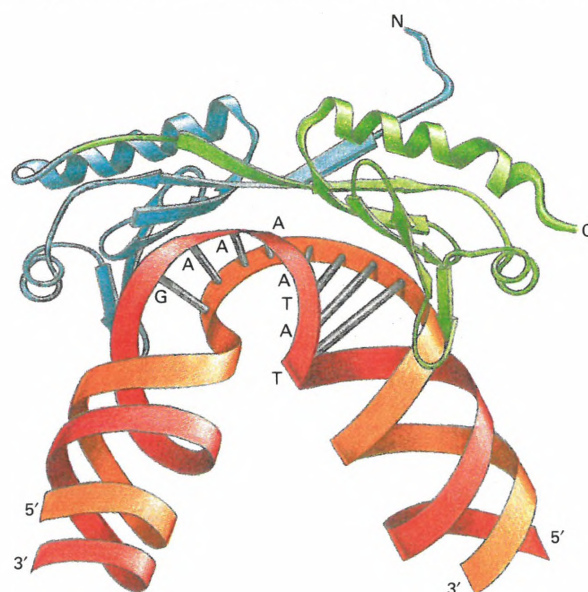


Figure 6-18 Three-dimensional structure of TBP (TATA-binding protein) bound to DNA. The TBP is the subunit of the general transcription factor TFIID that is responsible for recognizing and binding to the TATA box sequence in the DNA (red). The unique DNA bending caused by TBP—two kinks in the double helix separated by partly unwound DNA—may serve as a landmark that helps to attract the other general transcription factors. TBP is a single polypeptide chain that is folded into two very similar domains (blue and green). (Adapted from J.L. Kim et al., *Nature* 365:520–527, 1993.)

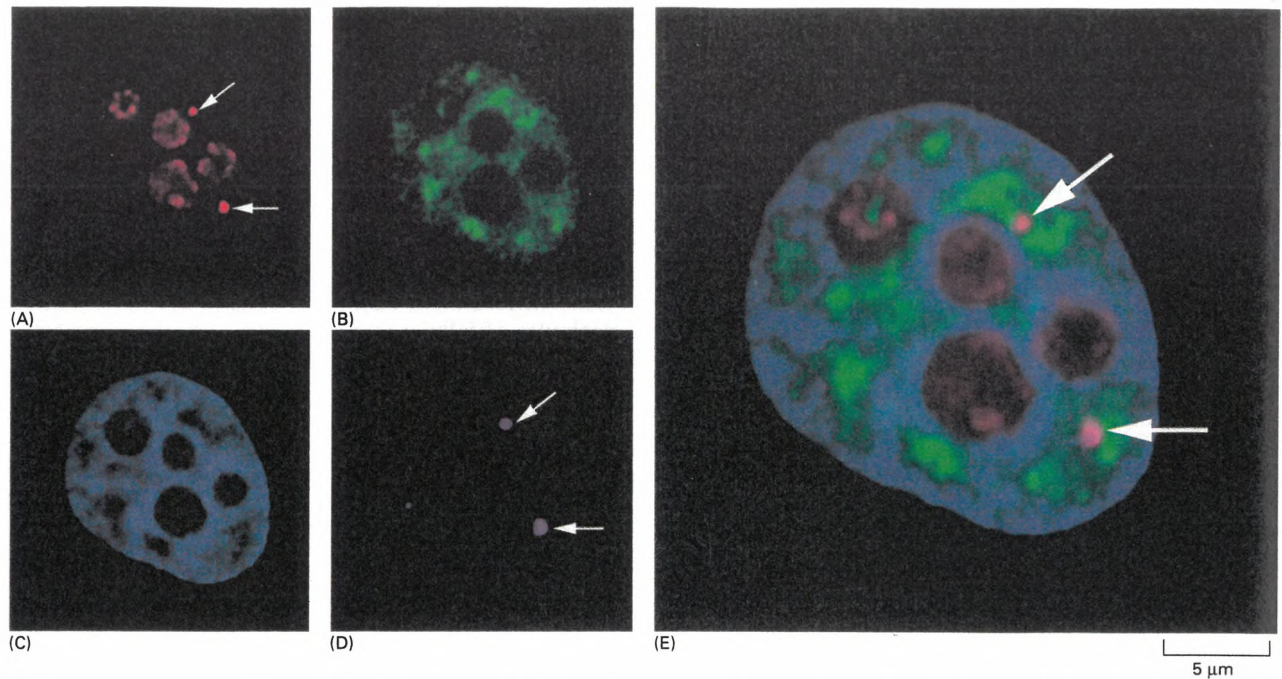


Figure 6-48 Visualization of chromatin and nuclear bodies. (A)–(D) show micrographs of the same human cell nucleus, each processed differently to show a particular set of nuclear structures. (E) shows an enlarged superposition of all four individual images. (A) shows the location of the protein fibrillarin (a component of several snoRNPs), which is present at both nucleoli and Cajal bodies, the latter indicated by arrows. (B) shows interchromatin granule clusters or “speckles” detected by using antibodies against a protein involved in pre-mRNA splicing. (C) is stained to show bulk chromatin. (D) shows the location of the protein coilin, which is present at Cajal bodies (indicated by arrows). (From J.R. Swedlow and A.I. Lamond, *Gen. Biol.* 2:1–7, 2001; micrographs courtesy of Judith Sleeman.)

to be caused by a subtle defect in snRNP assembly and subsequent pre-mRNA splicing. More severe defects would be expected to be lethal.

Given the importance of nuclear subdomains in RNA processing, it might have been expected that pre-mRNA splicing would occur in a particular location in the nucleus, as it requires numerous RNA and protein components. However,

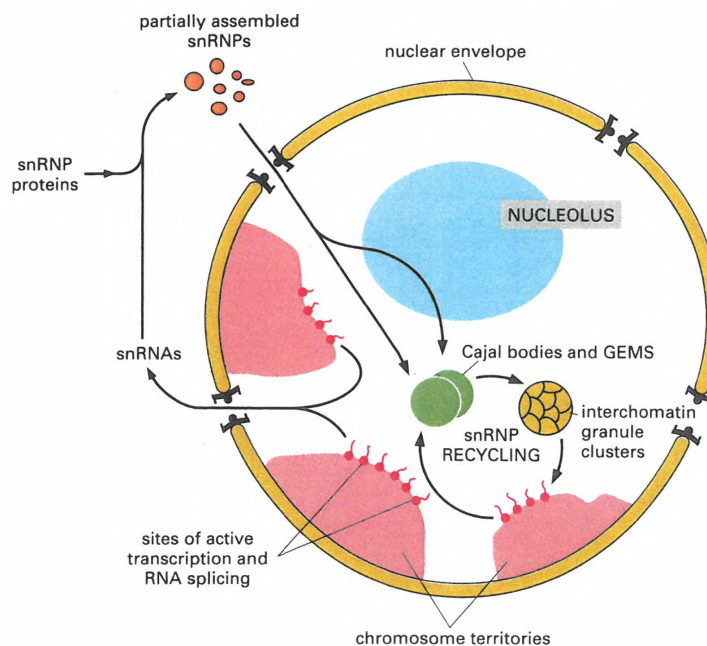


Figure 6-49 Schematic view of subnuclear structures. A typical vertebrate nucleus has several Cajal bodies, which are proposed to be the sites where snRNPs and snoRNPs undergo their final modifications. Interchromatin granule clusters are proposed to be storage sites for fully mature snRNPs. A typical vertebrate nucleus has 20–50 interchromatin granule clusters.

After their initial synthesis, snRNAs are exported from the nucleus, after which they undergo 5' and 3' end-processing and assemble with the seven common snRNP proteins (called Sm proteins). These complexes are reimported into the nucleus and the snRNPs undergo their final modification in Cajal bodies. In addition, the U6 snRNP requires chemical modification by snoRNAs in the nucleolus. The sites of active transcription and splicing (approximately 2000–3000 sites per vertebrate nucleus) correspond to the “perichromatin fibers” seen under the electron microscope. (Adapted from J.D. Lewis and D. Tollervey, *Science* 288:1385–1389, 2000.)

we have seen that the assembly of splicing components on pre-mRNA is co-transcriptional; thus splicing must occur at many locations along chromosomes. We saw in Chapter 4 that interphase chromosomes occupy discrete territories in the nucleus, and transcription and pre-mRNA splicing must take place within these territories. However, interphase chromosomes are themselves dynamic and their exact positioning in the nucleus correlates with gene expression. For example, transcriptionally silent regions of interphase chromosomes are often associated with the nuclear envelope where the concentration of heterochromatin components is believed to be especially high. When these same regions become transcriptionally active, they relocate towards the interior of the nucleus, which is richer in the components required for mRNA synthesis. It has been proposed that, although a typical mammalian cell may be expressing on the order of 15,000 genes, transcription and RNA splicing may be localized to only several thousand sites in the nucleus. These sites themselves are highly dynamic and probably result from the association of transcription and splicing components to create small "assembly lines" where the local concentration of these components is very high. As a result, the nucleus seems to be highly organized into subdomains, with snRNPs, snoRNPs, and other nuclear components moving between them in an orderly fashion according to the needs of the cell (Figure 6-49).

Summary

Before the synthesis of a particular protein can begin, the corresponding mRNA molecule must be produced by transcription. Bacteria contain a single type of RNA polymerase (the enzyme that carries out the transcription of DNA into RNA). An mRNA molecule is produced when this enzyme initiates transcription at a promoter, synthesizes the RNA by chain elongation, stops transcription at a terminator, and releases both the DNA template and the completed mRNA molecule. In eucaryotic cells, the process of transcription is much more complex, and there are three RNA polymerases—designated polymerase I, II, and III—that are related evolutionarily to one another and to the bacterial polymerase.

Eucaryotic mRNA is synthesized by RNA polymerase II. This enzyme requires a series of additional proteins, termed the general transcription factors, to initiate transcription on a purified DNA template and still more proteins (including chromatin-remodeling complexes and histone acetyltransferases) to initiate transcription on its chromatin template inside the cell. During the elongation phase of transcription, the nascent RNA undergoes three types of processing events: a special nucleotide is added to its 5' end (capping), intron sequences are removed from the middle of the RNA molecule (splicing), and the 3' end of the RNA is generated (cleavage and polyadenylation). Some of these RNA processing events that modify the initial RNA transcript (for example, those involved in RNA splicing) are carried out primarily by special small RNA molecules.

For some genes, RNA is the final product. In eucaryotes, these genes are usually transcribed by either RNA polymerase I or RNA polymerase III. RNA polymerase I makes the ribosomal RNAs. After their synthesis as a large precursor, the rRNAs are chemically modified, cleaved, and assembled into ribosomes in the nucleolus—a distinct subnuclear structure that also helps to process some smaller RNA-protein complexes in the cell. Additional subnuclear structures (including Cajal bodies and interchromatin granule clusters) are sites where components involved in RNA processing are assembled, stored, and recycled.

FROM RNA TO PROTEIN

In the preceding section we have seen that the final product of some genes is an RNA molecule itself, such as those present in the snRNPs and in ribosomes. However, most genes in a cell produce mRNA molecules that serve as intermediaries on the pathway to proteins. In this section we examine how the cell converts the information carried in an mRNA molecule into a protein molecule. This feat of translation first attracted the attention of biologists in the late 1950s,

Motif Exhibit 1042. Page 16 of 20

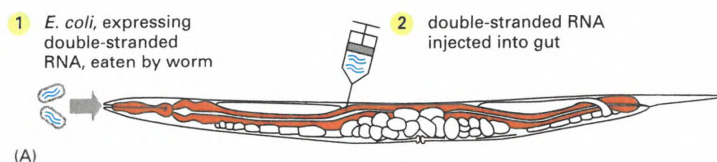
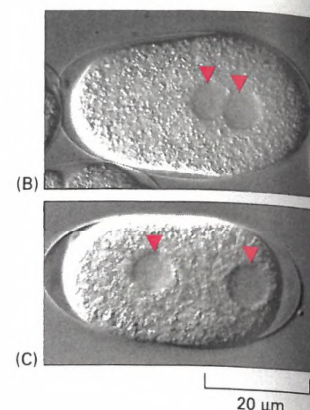


Figure 8-66 Dominant negative mutations created by RNA interference. (A) Double-stranded RNA (dsRNA) can be introduced into *C. elegans* (1) by feeding the worms with *E. coli* expressing the dsRNA or (2) by injecting dsRNA directly into the gut. (B) Wild-type worm embryo. (C) Worm embryo in which a gene involved in cell division has been inactivated by RNAi. The embryo shows abnormal migration of the two unfused nuclei of the egg and sperm. (B, C, from P. Gönczy et al., *Nature* 408:331–336, 2000. © Macmillan Magazines Ltd.)



injected nucleic acid, a stable synthetic RNA analog, called morpholino-RNA, is often used instead of ordinary RNA.

As investigators continued to explore the antisense RNA strategy, they made an interesting discovery. An antisense RNA strand can block gene expression, but a preparation of double-stranded RNA (dsRNA), containing both the sense and antisense strands of a target gene, inhibit the activity of target genes even more effectively (see Figure 7-107). This phenomenon, dubbed *RNA interference* (RNAi), has now been exploited for examining gene function in several organisms.

The RNAi technique has been widely used to study gene function in the nematode *C. elegans*. When working with worms, introducing the dsRNA is quite simple: RNA can be injected directly into the intestine of the animal, or the worm can be fed with *E. coli* expressing the target gene dsRNA (Figure 8-66A). The RNA is distributed throughout the body of the worm and is found to inhibit expression of the target gene in different tissue types. Further, as explained in Figure 7-107, the interference is frequently inherited by the progeny of the injected animal. Because the entire genome of *C. elegans* has been sequenced, RNAi is being used to help in assigning functions to the entire complement of worm genes. In one study, researchers were able to inhibit 96% of the approximately 2300 predicted genes on *C. elegans* chromosome III. In this way, they identified 133 genes involved in cell division in *C. elegans* embryos (Figure 8-66C). Of these, only 11 had been previously ascribed a function by direct experimentation.

For unknown reasons, RNA interference does not efficiently inactivate all genes. And interference can sometimes suppress the activity of a target gene in one tissue and not another. An alternative way to produce a dominant negative mutation takes advantage of the fact that most proteins function as part of a larger protein complex. Such complexes can often be inactivated by the inclusion of just one nonfunctional component. Therefore, by designing a gene that produces large quantities of a mutant protein that is inactive but still able to assemble into the complex, it is often possible to produce a cell in which all the complexes are inactivated despite the presence of the normal protein (Figure 8-67).

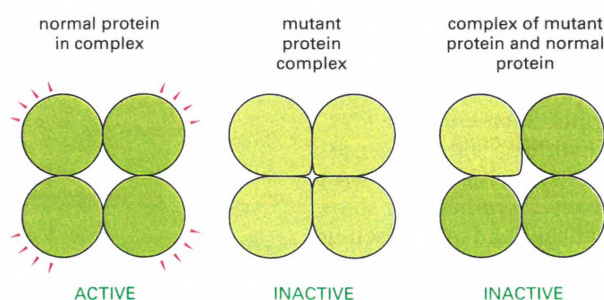


Figure 8-67 A dominant negative effect of a protein. Here a gene is engineered to produce a mutant protein that prevents the normal copies of the same protein from performing their function. In this simple example, the normal protein must form a multisubunit complex to be active, and the mutant protein blocks function by forming a mixed complex that is inactive. In this way a single copy of a mutant gene located anywhere in the genome can inactivate the normal products produced by other gene copies.

If a protein is required for the survival of the cell (or the organism), a dominant negative mutant dies, making it impossible to test the function of the protein. To avoid this problem, one can couple the mutant gene to control sequences that have been engineered to produce the gene product only on command—for example, in response to an increase in temperature or to the presence of a specific signaling molecule. Cells or organisms containing such a dominant mutant gene under the control of an *inducible promoter* can be deprived of a specific protein at a particular time, and the effect can then be followed. Inducible promoters also allow genes to be switched on or off in specific tissues, allowing one to examine the effect of the mutant gene in selected parts of the organism. In the future, techniques for producing dominant negative mutations to inactivate specific genes are likely to be widely used to determine the functions of proteins in higher organisms.

Gain-of-Function Mutations Provide Clues to the Role Genes Play in a Cell or Organism

In the same way that cells can be engineered to express a dominant negative version of a protein, resulting in a loss-of-function phenotype, they can also be engineered to display a novel phenotype through a *gain-of-function* mutation. Such mutations may confer a novel activity on a particular protein, or they may cause a protein with normal activity to be expressed at an inappropriate time or in the wrong tissue in an animal. Regardless of the mechanism, gain-of-function mutations can produce a new phenotype in a cell, tissue, or organism.

Often, gain-of-function mutants are generated by expressing a gene at a much higher level than normal in cells. Such overexpression can be achieved by coupling a gene to a powerful promoter sequence and placing it on a multicopy plasmid—or integrating it in multiple copies in the genome. In either case, the gene is present in many copies and each copy directs the transcription of unusually large numbers of mRNA molecules. Although the effect that such overexpression has on the phenotype of an organism must be interpreted with caution, this approach has provided invaluable insights into the activity of many genes. In an alternate type of gain-of-function mutation, the mutant protein is made in normal amounts, but is much more active than its normal counterpart. Such proteins are frequently found in tumors, and they have been exploited to study signal transduction pathways in cells (discussed in Chapter 15).

Genes can also be expressed at the wrong time or in the wrong place in an organism—often with striking results (Figure 8–68). Such misexpression is most often accomplished by re-engineering the genes themselves, thereby supplying them with the regulatory sequences needed to alter their expression.

Genes Can Be Redesigned to Produce Proteins of Any Desired Sequence

In studying the action of a gene and the protein it encodes, one does not always wish to make drastic changes—flooding cells with huge quantities of hyperactive protein or eliminating a gene product entirely. It is sometimes useful to make slight changes in a protein's structure so that one can begin to dissect which portions of a protein are important for its function. The activity of an enzyme, for example, can be studied by changing a single amino acid in its active site. Special techniques are required to alter genes, and their protein products, in such subtle ways. The first step is often the chemical synthesis of a short DNA molecule containing the desired altered portion of the gene's nucleotide sequence. This synthetic DNA oligonucleotide is hybridized with single-stranded plasmid DNA that contains the DNA sequence to be altered, using conditions that allow imperfectly matched DNA strands to pair (Figure 8–69). The synthetic oligonucleotide will now serve as a primer for DNA synthesis by DNA polymerase, thereby generating a DNA double helix that incorporates the altered sequence into one of its two strands. After transfection, plasmids that carry the fully modified gene sequence are obtained. The appropriate DNA is

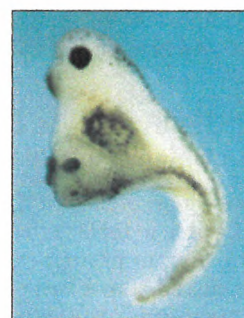


Figure 8–68 Ectopic misexpression of Wnt, a signaling protein that affects development of the body axis in the early *Xenopus* embryo. In this experiment, mRNA coding for Wnt was injected into the ventral vegetal blastomere, inducing a second body axis (discussed in Chapter 21). (From S. Sokol et al., *Cell* 67:741–752, 1992. © Elsevier.)

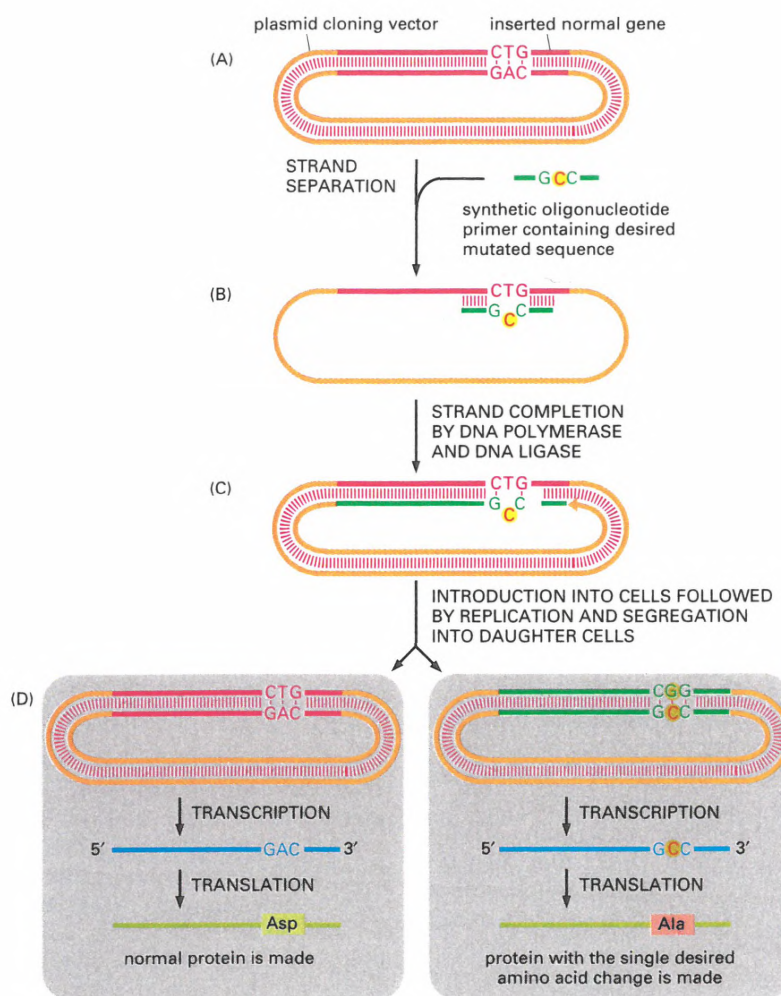


Figure 8-69 The use of a synthetic oligonucleotide to modify the protein-coding region of a gene by site-directed mutagenesis. (A) A recombinant plasmid containing a gene insert is separated into its two DNA strands. A synthetic oligonucleotide primer corresponding to part of the gene sequence but containing a single altered nucleotide at a predetermined point is added to the single-stranded DNA under conditions that permit less than perfect DNA hybridization. (B) The primer hybridizes to the DNA, forming a single mismatched nucleotide pair. (C) The recombinant plasmid is made double-stranded by *in vitro* DNA synthesis starting from the primer and sealed by DNA ligase. (D) The double-stranded DNA is introduced into a cell, where it is replicated. Replication using one strand of the template produces a normal DNA molecule, but replication using the other (the strand that contains the primer) produces a DNA molecule carrying the desired mutation. Only half of the progeny cells will end up with a plasmid that contains the desired mutant gene. However, a progeny cell that contains the mutated gene can be identified, separated from other cells, and cultured to produce a pure population of cells, all of which carry the mutated gene. Only one of the many changes that can be engineered in this way is shown here. With an oligonucleotide of the appropriate sequence, more than one amino acid substitution can be made at a time, or one or more amino acids can be inserted or deleted. Although not shown in this figure, it is also possible to create a site-directed mutation by using the appropriate oligonucleotides and PCR (instead of plasmid replication) to amplify the mutated gene.

then inserted into an expression vector so that the redesigned protein can be produced in the appropriate type of cells for detailed studies of its function. By changing selected amino acids in a protein in this way—a technique called **site-directed mutagenesis**—one can determine exactly which parts of the polypeptide chain are important for such processes as protein folding, interactions with other proteins, and enzymatic catalysis.

Engineered Genes Can Be Easily Inserted into the Germ Line of Many Animals

When engineering an organism that is to express an altered gene, ideally one would like to be able to replace the normal gene with the altered one so that the function of the mutant protein can be analyzed in the absence of the normal protein. As discussed above, this can be readily accomplished in some haploid, single-celled organisms. We shall see in the following section that much more complicated procedures have been developed that allow gene replacements of this type in mice. Foreign DNA can, however, be rather easily integrated into random positions of many animal genomes. In mammals, for example, linear DNA fragments introduced into cells are rapidly ligated end-to-end by intracellular enzymes to form long tandem arrays, which usually become integrated into a chromosome at an apparently random site. Fertilized mammalian eggs behave like other mammalian cells in this respect. A mouse egg injected with 200 copies of a linear DNA molecule often develops into a mouse containing, in many of its cells, a tandem array of copies of the injected gene integrated at a single random



GS Garland Science
Taylor & Francis Group

MOLECULAR BIOLOGY OF CELL
-w/CD
BIOL/361
978-0-8153-3218-3
99990
2900815 332182
USED
101

CELL BIOLOGY
INTERACTIVE
cd-rom inside