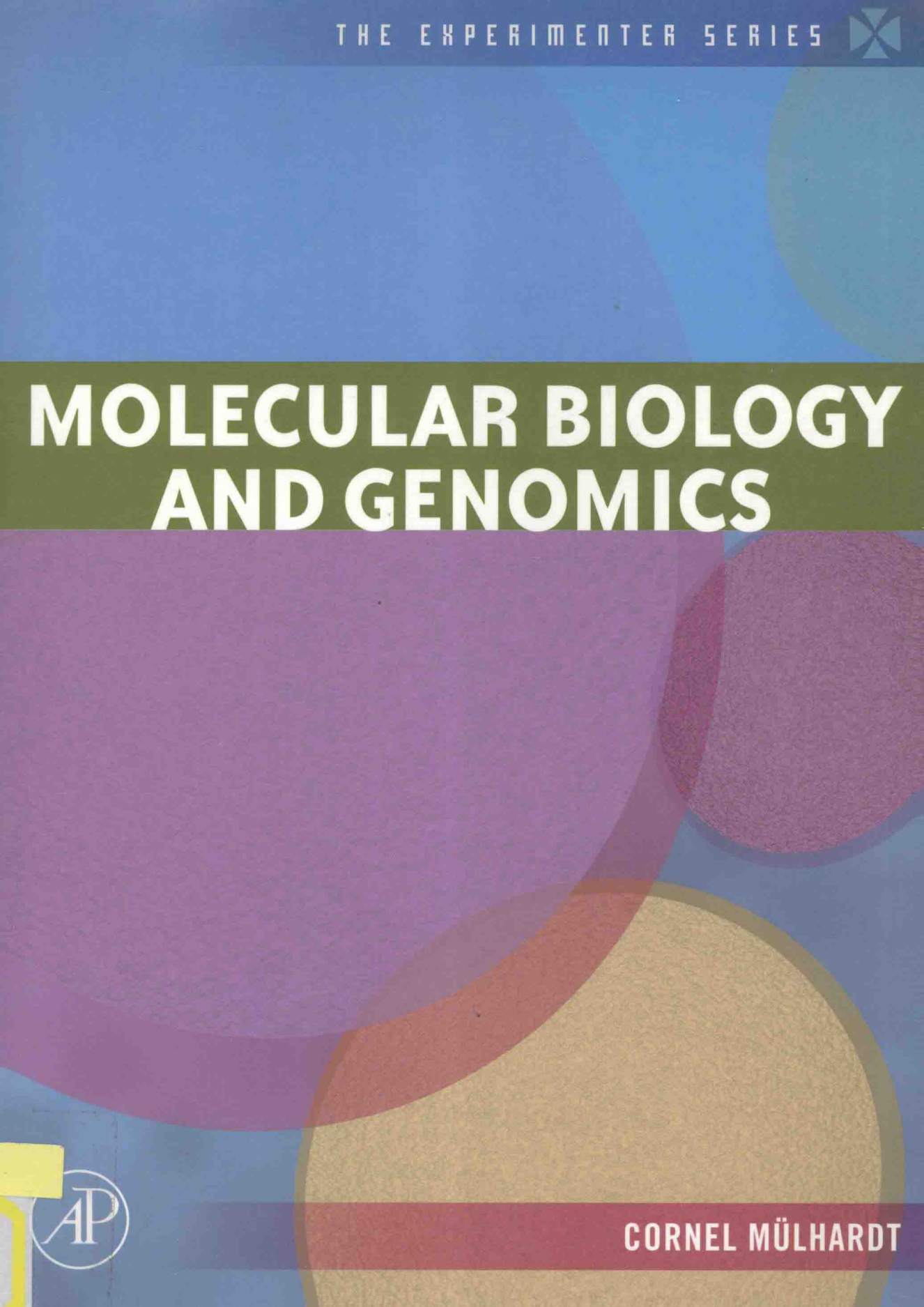
THE EXPERIMENTER SERIES

MOLECULAR BIOLOGY AND GENOMICS



CORNEL MÜLHARDT

Molecular Biology and Genomics

CORNEL MÜLHARDT

Translated by
E.W. Beese, M.D.
Munich, Germany

The Experimenter Series

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
84 Theobald's Road, London WC1X 8RR, UK

This book is printed on acid-free paper.

Library of Congress Cataloging-in-Publication Data
Mülhardt, Cornel.
[Experimentator. English]
Molecular biology and genomics / Cornel Mulhardt ; translated by E.W. Beese.
p. ; cm. – (The experimenter series)
Translation of: Experimentator : Molekularbiologie. 4th ed.
Includes bibliographical references and index.
ISBN-13: 978-0-12-088546-6 (alk. paper)
ISBN-10: 0-12-088546-8 (alk. paper)
1. Molecular biology—Research—Methodology. 2. Genomics—Research—Methodology.
I. Title. II. Series.
[DNLM: 1. Genomics. 2. Molecular Biology. QU 450 M956m 2007a]
QH506.M84713 2007
572.8—dc22

2006025989

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

ISBN 13: 978-0-12-088546-6
ISBN 10: 0-12-088546-8

For information on all Academic Press publications
visit our Web site at www.books.elsevier.com

Printed in the United States of America
06 07 08 09 10 11 9 8 7 6 5 4 3 2 1

Molecular Biology and Genomics

Foreword

It is hard to believe that we are now already producing the fourth edition within such a brief time period of only five years. In this edition we have once again increased its dimensions; here, however, the basics have remained more or less unchanged, although the number of "specialties" covered continues to increase. This time, for instance, minisequencing, pyrosequencing, and RNA-interference have been included. In addition, the time has now come to write a short introduction concerning the planning of careers for younger researchers—after all, one does not live from the nobility of research alone. Numerous, minute corrections have had to be performed as well, since nothing is perfect. This book as well, in spite of the enormous amount of work which has already been carried out in its production, nevertheless continues to demonstrate room for improvement. One can do whatever is possible, but entropy always seems to win in the end! However, there are enough typographical errors still to be found here as well and we are therefore thankful for every piece of advice which we might receive!

Foreword to the First Edition

Do you actually want to read a foreword?

Read the preface and concluding comments in Hubert Rehms' "Experimentator: Proteinbiochemie" ("Protein Biochemistry" in the "Experimentator" series from Spektrum Verlag) and, finally, after getting a taste for them, read those statements in Siegfried Bars' "Forschen auf Deutsch: Der Machiavelli für Forscher—und solche, die es werden wollen" ("How to Do Research in German: The Machiavellian Guideline for Researchers—and for Those Who Would Like to Become Such" from Verlag Harri Deutsch). Even if the beginner is not inclined to believe so, the wind actually blows as has been described here. The most successful researcher, namely, is not to be distinguished by any pioneering accomplishments, but rather through his/her clever actions in scenes involving publications and in his/her large number of friends. The goal is to keep the phase as short as possible between one's time as a student and that during which one is authorized to function as a professor. Once one has finally reached this goal, however, the research is then left to the others; in the same way in which a successful military strategist no longer reaches for his/her own saber, but instead only carries out his/her advances theoretically through the actions of ambitious military leaders. At this stage, one merely has an indirect relationship to research. Should you not have given up after completing this reading material, you may possibly demonstrate the necessary composure to dedicate yourself to research in molecular biology. This composure, namely, will prove to be quite essential in order to avoid being driven crazy from the endless series of failures which can be expected in the daily life which is to be experienced in the laboratory.

"Don't let yourself be discouraged [...]! The others are also working themselves to a frazzle without any success; it is quite normal that no results are to be observed in the beginning. Just keep it up!—Or take up a reasonable occupation."

With this in regard ...
Cornel Mülhardt

Acknowledgments

I have always considered the effusive acknowledgments at the beginning of a book by the author directed to friends and relatives to represent the boring performance of one's duties. Today, I am finally aware that it truly is a matter near one's heart to express one's thanks for having survived a difficult phase of life. For their contribution toward the success of this book, I would like to thank Marc Bedoucha, Igor Bendik, Hans-Georg Breitinger, Dorothee Foernzler, Christophe Grundschober, Andreas Humeny, Frank Kirchhoff, Nina Meier, Nicoletta Milani, and Ruthild Weber. My special thanks go to Cord-Michael Becker, who has continuously supported me over the years and earned my warmest thanks.

Finally, there is also Spektrum Verlag and the project planner Ulrich Moltmann, who came up with the idea for such a book and thereby enabled me to fulfill a very old dream, Jutta Hoffmann, without whose occasional, friendly pressure I would certainly have given up before reaching the end, and Bettina Saglio, who must always struggle with my newest versions. If God does exist, may He/She bless them with more punctual authors.

It is known that nothing is so good that it cannot be improved upon. That is especially the case for such a book as this, a reason that would make me very happy to receive any constructive criticism.

Contents

1 What Is Molecular Biology?

Gib nach dem löblichen
von vorn Schöpfung anzufangen!
Zu raschem Wirken sei bereit!
Da regst du dich nach ewigen Normen
durch tausend, abertausend Formen,
und bis zum Menschen hast du Zeit.

Yield to the noble inspiration
To try each process of creation,
And don't be scared if things move fast;
Thus growing by eternal norms,
You'll pass through many a thousand forms,
Emerging as a man at last.

–Goethe, *Faust*[1]

Pursuits in the field of molecular biology involve genetic engineering and techniques such as cloning, and they may be interpreted as being quite adventuresome and almost divine. As a molecular biologist, declaring how you spend your days may garner boundless admiration from some people and disapproval from others. Consequently, you must consider carefully with whom you are speaking before describing your activities. You should not mention how many problems and how much frustration you must endure daily, because one group will be disillusioned, and the other will inevitably ask why you continue to work in this field.

I confess that I like the debates about whether we should clone humans, although the discussion often is not engaged at a scientific level. One individual makes a stupid suggestion, and one half of the media and the nation explain why the cloning of humans may not be justified by any means. It seems much ado about nothing, and I cannot comprehend what advantage it would be for me to clone myself. Why should I wish to invest $75,000 for a small crybaby, whose only common ground with me is that it looks like me as I did 30 years ago, if I could instead come to a similar result in a classic manner for the price of a bouquet of flowers for my wife and a television-free evening?

This example reveals the problems that are associated with molecular biology. Incited by spectacular reports in the media, everybody has his or her own, usually quite extreme, ideas on the benefits or disasters related to this topic. In reality, the picture is deeply distressing. Molecular biology, also known as the *molecular world*, generally deals with minute quantities of chiefly clear, colorless solutions. There are no signs of ecstatic scholars who have gone wild in the cinematic setting of flickering, steaming, gaudy-colored liquids. The molecular world deals with molecules, evidence for the existence of which has been attempted in many textbooks, although there is generally little more than a fluorescing spot to be seen on the agarose gel. Every procedure seems to take 3 days, and no Nobel Prize is associated with any of them.

---

[1] This and the following quotations are from Goethe's *Faust* (3rd ed. dtv-Gesamtausgabe, München 1966) and from Goethe's *Faust* (John Shawcross, translator. Allen Wingate Publishers, 1959).

Molecular biology is primarily a process of voodoo—sometimes everything works, but usually nothing works. Very unusual parameters seem to play a role in the result of an experiment, which should represent the subject of the research itself—the last, great taboo of modern science. Based on the empirical data, I have come to the realization that the results of an experiment can be calculated as the quotient of air pressure and the remaining number of scribbled note pages in the drawer raised to the power of grandmother's dog. Until I have been able to verify this experimentally, however, I will continue to limit myself to the classic explanations based on mathematics, which have admittedly enabled this profession to progress quite far.

1.1 The Substrate of Molecular Biology: The Molecular World for Beginners

War es ein Gott, der diese Zeichen schrieb,
die mir das innre Toben stillen,
das arme Herz mit Freude füllen
und mit geheimnisvollem Trieb
die Kräfte der Natur rings um mich her enthüllen?

Was it a God, who traced this sign,
Which calms my raging breast anew,
Brings joy to this poor heart of mine,
And by some impulse secret and divine
Unveils great Nature's labors to my view!

Molecular biology is by no means the biology of molecules, which is better characterized by biochemistry. Molecular biology is the study of life as reflected in DNA. It is a small, closely knit world that differentiates itself from all other related fields, including zoology, botany, and protein biochemistry. Only a few individuals in the field of molecular biology are biologists, and those who are biologists would be most happy to deny it. If you are not a biologist, be assured that you now find yourself in the best of company and that the small amount of material needed to master this work can quickly be understood.

Molecular biology deals with nucleic acids, which come in two forms: deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). The chemical differences between the two substances are minimal. They are both polymers that are made of four building blocks each, deoxynucleotides in DNA and nucleotides in RNA.

The nucleotides of RNA are composed of a base constituent (adenine, cytosine, guanine, or uracil), a glucose component (ribose), and a phosphoryl residue, whereby two nucleotides are connected to each other through phosphoglucose bonds. In this way, one nucleotide can be connected to the next, forming a long chain known as a *polynucleotide*, which is designated as RNA. DNA is formed in almost the same manner. The deoxynucleotides of DNA are composed of a base constituent (adenine, cytosine, guanine, or thymine), 2′-deoxyribose (instead of ribose), and a phosphoryl residue.

More details can be found in textbooks of biochemistry, cell biology, or genetics. I limit my comments on these topics to a few aspects that are significant for laboratory practice.

Deoxynucleotide is a tongue twister, which is avoided whenever possible. Researchers instead speak of *nucleotides* when using these substances in the laboratory, although they generally refer to the deoxy variant.

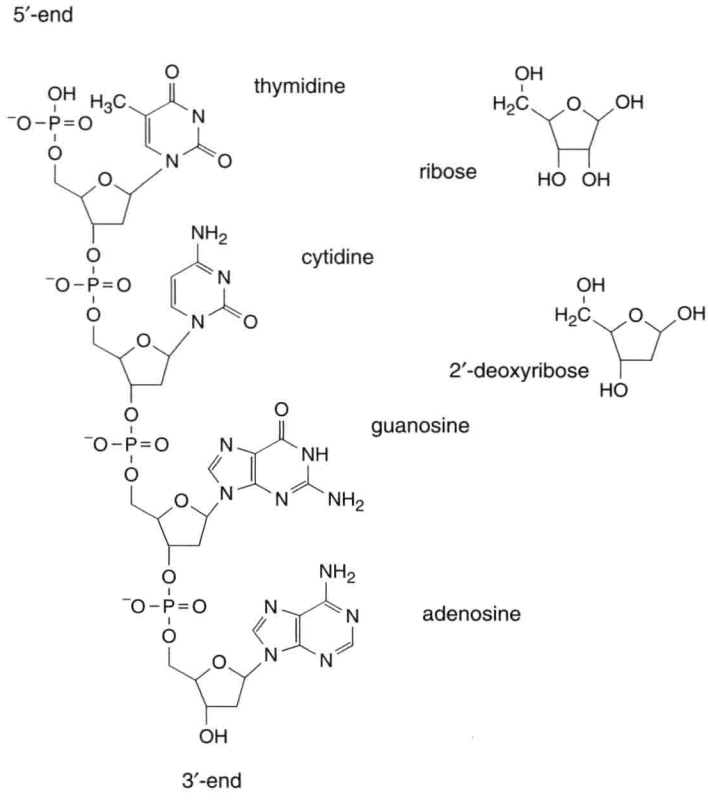


Figure 1-1. The four nucleotides and the way in which a DNA helix typically is portrayed.

The names of the nucleotides are based on the names of their bases: **adenosine** (A), **cytidine** (C), **guanosine** (G), **thymidine** (T), or **uridine** (U). Many confuse these bases with the nucleotides (Figure 1-1), although this rarely makes a difference in practice. The abbreviations A, C, G, T, and U have been used to immortalize all the DNA and RNA sequences known to the world.

The phosphoryl group of the nucleotides is connected to the 5′ carbon atom of the sugar component. The synthesis of nucleic acid begins with a nucleotide at whose 3′-OH group (i.e., the hydroxyl group at the 3′ carbon atom) a phosphodiester bond connects the phosphate group with the next nucleotide. Another nucleotide can be connected to the 3′-OH group of the nucleotide and so on. Synthesis proceeds in the 5′ to 3′ direction, because a 5′ end and a 3′ end exist at all times, and a new nucleotide is always added to the 3′ end. All known sequences are read from the 5′ end to the 3′ end, and attempts to alter this convention can only create chaos.

If no 3′-OH group exists, it is impossible to connect a new nucleotide. In nature, this situation is rarely observed, although it plays a large role in the sequencing performed in the laboratory. In addition to the four 2′-deoxynucleotides, 2′,3′-dideoxynucleotides are used in DNA synthesis. They can be connected to the 3′ end of a polynucleotide in the normal manner, but because a 3′-OH group is lacking, the DNA cannot be lengthened any further. This principle is essential for understanding sequencing, a subject that will be discussed in more detail later.

Each nucleotide or polynucleotide with a phosphoryl group at its 5′ end can be bonded (ligated) to the 3′ end of another polynucleotide with the aid of a DNA ligase. In this way, even

larger DNA fragments can be bonded to one another. Without the phosphoryl group, nothing works, and elimination of a phosphoryl residue with a phosphatase can be used to help inhibit such ligations.

Two nucleotides bonded to one another are known as a *dinucleotide*, and three together are known as a *trinucleotide*. If more nucleotides are bonded together in a group, the structure is called an *oligonucleotide*, and if very many are bonded together, the entire structure is known as a *polynucleotide*. The boundary between oligonucleotides and polynucleotides is not defined precisely, but oligonucleotides usually are considered to contain fewer than 100 nucleotides.

Whether the structure is a mononucleotide, dinucleotide, oligonucleotide, or polynucleotide, the substance always represents a single molecule, because a covalent bond is formed between each of the initial molecules involved. The length of the chain formed in this way plays no role, because even a nucleic acid that is 3 million nucleotides long is a single molecule.

Polynucleotides have a remarkable and important characteristic: Their bases pair specifically with other bases. Cytosine always pairs with guanine, and adenine pairs with thymine (in DNA) or with uracil (in RNA); no other combinations are functional. The more bases that pair with one another, the more stabile this combination becomes, because hydrogen bonds are formed between the bases of each pair, and the power of these bonds is cumulative. A sequence that pairs perfectly with another sequence is said to be *complementary*. The nucleic acid sequences of two complementary nucleic acids are completely different and pair in a mirror image form. The following example clarifies this arrangement:

5′-AGCTAAGACTTGTTC-3′
3′-TCGATTCTGAACAAG-5′

The orientation of both sequences proceeds in an opposite direction (reading from the 5′ to 3′ end), because the two chains must spatially be aligned in opposition for the pairing to function properly. In the normal 5′ to 3′ method of writing, the complementary sequences are read as follows:

5′-AGCTAAGACTTGTTC-3′

and

5′-GAACAAGTCTTAGCT-3′

The sequences appear to be so different that you must look very closely to see that they are complementary to one another, and you can see the value of the application of the 5′ to 3′ reading convention. A beginner can very quickly write a primer strand sequence of

3′-TCGATTCTGAACAAG-5′

as

5′-TCGATTCTGAACAAG-3′

so that the sequence, despite its normal appearance, looks completely different from that of the original sequence.

DNA typically is made up of two complementary chains, and RNA usually is made of a single chain. When pairs occur in RNA, they are found within the same chain and play a more significant role in the three-dimensional structure of the RNA, whereas double-stranded DNA must be considered a linear molecule.

Have you noticed the error? Double-stranded DNA does not consist of a single molecule; it is made up of a double-stranded structure with a chain and an opposing chain. The two chains can

be separated from one another at any time if sufficient energy is supplied; this process is called *denaturation*. Ten seconds at 95°C (203°F) is sufficient for performing complete separation of both strands. If they join together again, a process known as *hybridization*, they are said to *anneal*. The entire procedure is almost arbitrarily reversible.

DNA is frequently called the *molecule of the life* by the yellow press. Why does such a very long, but quite boring molecule have such a lengthy name? Nature has optimally made use of the characteristics of nucleic acids to create a confusing diversity of life, and this is carried out in the following manner. Because the bases pair, a complementary strand can be synthesized from a single strand of DNA; another strand can be synthesized from this second strand, which is identical in structure to that of the initial strand. In this way, DNA can be duplicated as often as desired while preserving the sequence of its bases. The process, which is carried out by special enzymes called **DNA polymerases**, is known as **replication**.

Because RNA and DNA are similar chemically, a complementary RNA molecule can be synthesized from a DNA strand. This process, which is carried out with the aid of **RNA polymerases**, is known as **transcription**.

With an extremely complicated apparatus involving very many molecules (ribosomes), the sequencing information available in RNA can be used for the synthesis of a protein, a process known as **translation**. Proteins are long polymers, whose synthesis proceeds in much the same way as that of nucleic acids—an amino acid exists to which a second amino acid is added, then a third, and so on, until a polypeptide is finally formed, which is commonly known as a *protein*. Twenty different amino acids are known to exist, but with only four bases, how can the information for a protein be hidden in RNA? The trick is that three bases (codons) together contain the information to specify one amino acid. This code is interpreted in the same manner in almost all living cells on this planet: AAA signifies lysine, CAA represents glutamine, and so forth. Together, there are 64 triplet codons (4^3) that code for 20 different amino acids and 3 stop codons; the system is somewhat redundant, and the genetic code is therefore degenerate. For example, AAA and AAG each code for lysine. If you look through all of the 64 triplet codes, you can see that the first two bases are frequently decisive for the particular amino acid that is to be installed, whereas the third base often proves to be insignificant. This arrangement is very useful for molecular biologists, because they can cause mutations in DNA sequences without altering their coding characteristics.

Not every DNA sequence codes for a protein; in human chromosomes, more than 90% of the available sequences have no recognizable significance. The debate continues about whether these stretches should be considered as junk DNA or as elements with some unknown function, but let us instead concern ourselves with the miserable remnants.

The remnants are the **genes**, transcription units on the long DNA molecule that are made up of regulatory and transcribed regions (Figure 1-2). Their functions are indicated by their names. The regulatory region controls whether the other regions are transcribed or not. The transcribed area can be divided into two subregions, the **exons** and **introns**. This division reflects a peculiarity of transcription in higher organisms (in this case, everything that is not a bacterium): **splicing**. Directly after its synthesis, the transcribed RNA is immediately processed by a complex apparatus, the spliceosome; this process excises parts of the RNA sequence and reattaches the remnants, a complicated procedure that ultimately results in a final RNA with a somewhat shorter length than before. The sequences excised during the splicing are introns on the DNA level, and those that remain are exons. The remaining RNA can again be divided into a region that serves as an information source for the synthesis of a protein, also as a coding area or an **open reading frame** (ORF), and one region each at the 5′ and 3′ termini of the RNA, which are known as **untranslated regions** (UTR) and whose function is still being investigated. The 5′ end of the RNA has a methylated G nucleotide (**5′ cap**), and a **poly A tail** (sequence of 100 to 250 adenosines) is located at the 3′ end. The A residues are not encoded on the DNA but are added by the action of poly A polymerase using ATP as a substrate.