DENNIS W. ROSS

INTRODUCTION TO Molecular Medicine

SECOND EDITION

Springer

Dennis W. Ross

Introduction to Molecular Medicine

Second Edition

With 60 Illustrations

Springer

Dennis W. Ross, M.D., Ph.D.
The School of Medicine
Department of Pathology
University of North Carolina at
Chapel Hill
Chapel Hill, NC 27599, USA
and
Department of Pathology
Forsyth Memorial Hospital
3333 Silas Creek Parkway
Winston-Salem, NC 27104, USA

Library of Congress Cataloging-in-Publication Data
Ross, D. W. (Dennis W.)
Introduction to molecular medicine / Dennis W. Ross. – 2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 0-387-94468-0 (softcover)
1. Medical genetics. 2. Molecular biology. 3. Recombinant DNA.
4. Pathology, Molecular. I. Title.
[DNLM: 1. Genetics, Medical. 2. Molecular Biology. 3. Hereditary
Diseases – genetics. 4. Neoplasms – genetics. QZ 50 R823i 1996]
RB155.R78 1996
616'.042 – dc20
DNLM/DLC 95-36653

Typeset by Bytheway Typesetting Services, Inc., Norwich, NY.

ISBN 0-387-94468-0 Springer-Verlag New York Berlin Heidelberg SPIN 10495100

Preface to the Second Edition

In the four years since the first edition of this book was published, the molecular revolution has continued. DNA has been named by *Time* magazine as the Molecule of the Year, a Nobel Prize has been awarded to a young man for the invention of the polymerase chain reaction, and television viewers have learned of the DNA fingerprint. Molecular technology in medicine is increasing. The availability of DNA probes for cancer susceptibility is stressing our system of insurance, testing our ideas about medical ethics, and teaching us new things about cancer. In this edition, I have added a number of new sections, as well as a new chapter. New examples of molecular medicine have been added to demonstrate current applications of this technology. The basic concepts of molecular biology remain the basis for the first three chapters of the book. The excitement surrounding molecular medicine that I mentioned in the preface to the first edition continues. It is now tinged with a touch of awe and a little bit of fear at the changes that recombinant DNA technology has brought to our society.

Preface to the First Edition

This book describes the discoveries that have created a field called molecular medicine. The use of recombinant DNA technology in medical research and most recently in medical practice constitutes a revolutionary tool in our study of disease. Probing the human genome is rapidly becoming as routine as looking at cells under a microscope. The cloning of a new gene is now a common occurrence, newspapers report. Recombinant DNA technology, like the invention of the microscope, shows us a world of detail richer than we might have imagined.

This book presents the discoveries, basic scientific concepts, and sense of excitement that surround the revolution in molecular medicine. The scientific basis of molecular medicine is explained in a simple and direct way. The level of technical detail, however, is sufficient for the reader to appreciate the power of recombinant DNA technology. This book is clinically oriented throughout. All of the examples and applications are related to medical discoveries and new methods of diagnosis and therapy. A few subjects within molecular medicine are examined in more detail to allow the reader to become aware of the strengths and shortcomings of a molecular approach to disease. I do not hide the incomplete understanding that still surrounds many of the recent discoveries in molecular medicine.

I intend to demonstrate the concepts of molecular medicine in this book by showing examples from all branches of medicine. I include, for instance, infectious diseases, genetic disorders, and cancer. However, I am not trying to be comprehensive in examining all areas of molecular medicine. So many discoveries are made each week in this field that it is not yet possible to draw them together in a comprehensive volume. My goal is to help the reader understand what the future may hold as well as the most important current applications.

This book is not a treatise, but an informal guide to a new field. As a guide, I try to communicate excitement, because this is the predominant feeling among people working in the field of molecular medicine.

Winston-Salem, NC D.W.R.

Acknowledgments

Short summaries of some of the material in this book have appeared previously in my column "Advances in Science and Pathology," which appears in the *Archives of Pathology and Laboratory Medicine*. I would like to thank the editorial staff of the *Archives* and the American Medical Association for giving me the opportunity to develop my skills in science writing. A number of people have given me valuable suggestions after reviewing draft chapters of this book. I wish to thank them for their time and thought: Susan Atwater, Lanier Ayscue, Phil Carl, Joseph Dudley, Margaret Gulley, Charles Hassell, Edward Highsmith, Roy Hopfer, and William Kaufmann. I must also especially thank Daniel Sinclair, my research technician, for his considerable help on many features of the manuscript including preparing the computer graphics used as illustrations. I also want to thank Sarah Kielar, my secretary, for careful attention to the preparation of the typescript. I also wish to acknowledge most gratefully funding from The Blood Cell Fund, which has supported my work in cancer research and education.

Contents

II. Molecular Approach to Disease

PART I BASICS OF MOLECULAR BIOLOGY

CHAPTER 1

Human Genome

The Genetic Message

The human genome consists of 6 billion nucleotides in a double-stranded helical deoxyribonucleic acid (DNA) molecule. This genetic message codes for the building and operation of the human body. The message is written in an alphabet that uses only four letters: A, C, G, and T. Each of the letters represents one of the four bases, which are the chemical building blocks of DNA: A—adenine, T—thymine, C—cytosine, and G—guanine. The nucleotides that spell out the genetic message are arrayed in a linear sequence in the double-stranded helical DNA molecule. The two strands of the DNA molecule are complementary copies of each other. The nucleotides on one strand pair with the complementary nucleotide on the other strand, as demonstrated in the top portion of Figure 1.1. A pairs with T; G pairs with C.

The genetic message is read not as single letters; rather, it is grouped into three-letter words. Each three-letter word is called a *codon*. A codon specifies one of the 20 possible amino acids that are the building blocks for all proteins. Figure 1.1 shows the organization of the genetic message at the level of DNA.

What has been called the fundamental paradigm of molecular biology is best stated as "one gene equals one enzyme." A gene encodes a specific linear sequence of amino acids assembled on the polyribosomes of the cell. The final form of the protein is achieved after spontaneous folding of the linear chain into a three-dimensional structure as demonstrated in Figure 1.2. It is this three-dimensional structure that gives the protein the ability to carry out its function. The sequence of amino acids in the linear chain predetermines the final folded structure of the protein. We have seen in Figure 1.1 that the nucleotide sequences in DNA serve as a template for messenger ribonucleic acid (mRNA), which directs the translation into a linear amino acid sequence. The last step, folding of the protein into a three-dimensional structure with specific enzymatic activity, is a spontaneous step. The protein structure is dependent on the gene, which is encoded

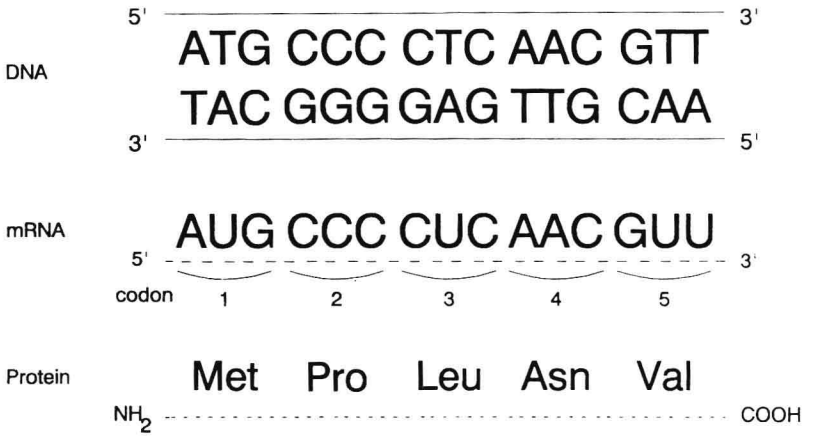

FIGURE 1.1. A genetic message begins as a double-stranded DNA molecule, which serves as the template for messenger RNA. The mRNA, in groups of three nucleotides to a codon, directs the order of amino acids in protein.

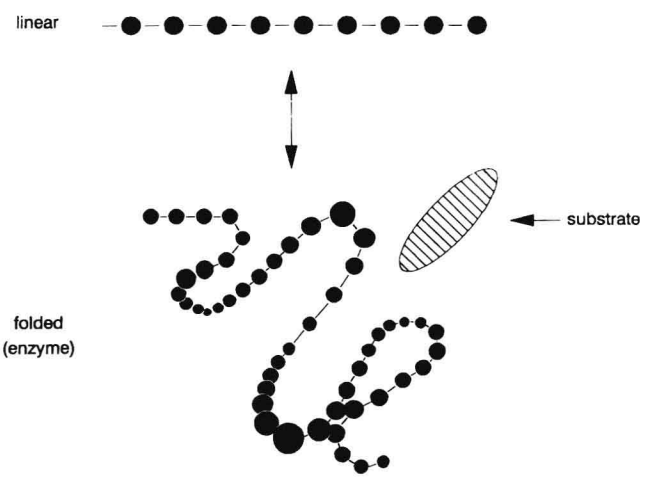

FIGURE 1.2. A chain of amino acids encoded by a gene and assembled on a polyribosome undergoes spontaneous folding to the final form of the protein. The form gives the protein its function, such as binding to a specific substrate in an enzyme-mediated reaction.

in the DNA. The events that control the expression of the genetic message control the function of the cells in the body by modulating the synthesis of proteins.

Shortly we will look at the details of how the genetic message specifies the synthesis of proteins and how proteins in turn regulate cell function. However, it is important first to get some idea of the amount of information contained in the human genome. A useful analogy compares the human genome to a library. The human genome is 6 billion letters long; there are only four letters in the genetic alphabet. The genome is written only with three-letter words. The message of the human genome is fragmented into 46 pieces, which we call chromosomes.

The library system of the University of North Carolina (UNC) has about the same number of letters in its library as the human genome—6 billion or a little bit more. The UNC library is written mostly in the English language, which has an alphabet of 26 letters plus ten numbers and some punctuation. The message in the library is organized into words, sentences, and books. The books are housed in 14 different library buildings around campus. We do not know the reason for the 14 different buildings nor do we know the reason for 46 human chromosomes. Maybe the number is not important.

We know a lot about the way information is organized in the library and used, and how it directs life on a university campus. Let's draw analogies to the human genome. To locate a book in the library, you go to the master card catalog (which is now computerized at UNC). You can look up a book by title, author, or subject. The card catalog tells you the Library of Congress call letters. A map of the library tells you where each call number section is physically located. For example, books about cancer with call letters "RC" are located on the eighth floor of the Davis Library. You can go the eighth floor and find the book on the shelf. If the book is a popular one, there may be several copies or several editions available. Taking the book you want down from the shelf, you use the index at the back of the book to find a particular subject of special interest. You might wish to photocopy a paragraph or two or to return to the circulation desk and check the book out. After you have consulted several books, you have the information you need. You may decide to write a report summarizing your research.

Using information from the human genome to synthesize a specific protein necessary for cell function is similar to consulting a library. We do not yet understand all the details of how the information is handled. To begin, some signal must tell the cell that it needs a certain protein. We call that event an inducer signal. Somehow the inducer signal causes the specific gene for the required protein to be located.

Finding the correct gene is like locating the necessary information in one paragraph of a specific book. We do not know how the cell does it. However, we do know that when the correct gene is found, a copy of the necessary information is transcribed into mRNA. This is the cell's method

of photocopying information for use out of the library. The mRNA is then translated into protein. A number of genes may have to be transcribed to make the composite chains of the final protein. Thus the synthesis of the completed protein molecule is equivalent to our written report based on consulting several books.

Some major differences exist between the way we use a library and the way a cell uses the information stored in its genome; nevertheless the analogy is useful. The cell can find any gene it needs within seconds and synthesize a new protein within minutes. For a human to go to the library, research a subject, and write a report takes much longer. When a human writes a report, it is just possible that the report may find its way back into the library as a new book. This never happens with cells. The cell cannot add to its storehouse of information. The genome does not change rapidly, like a human library. The human genome changes by recombination of genes from parents as part of sexual reproduction, and by mutation, which is a rare event.

However, at the end of the 20th century, the rules for the human genome are changing. As you will see in reading this book, it is now possible via genetic engineering to write new information into the genome. Just as I am writing this book, I can also write and introduce new information into human cells. In the future, the human genome may become more like our libraries. For better or worse, the information written there will be, in part, the product of human ideas.

I will use the library analogy occasionally throughout this book. As we explore the applications of molecular biology to medicine, it is important not to lose track of the transfer of information, which is what DNA is all about.

The Genetic Code

Figure 1.3 shows the genetic code for translating the triplet nucleotide codon of mRNA into the amino acid in proteins. For mRNA, uracil (U) replaces thymidine (T), which is used only in DNA, not RNA. Since there are four nucleotides, read in groups of three, 64 possible combinations are available. The number of amino acids used as building blocks for proteins is only 21. This results in significant redundancy in the genetic code, with several different codons specifying the same amino acid. Examination of Figure 1.3 reveals that AUU, AUC, and AUA all code for isoleucine. The redundancy of the genetic code offers some protection against the adverse effects of mutation. If a single base-pair mutation occurs that alters the sequence codon from AUA to AUC, no change in the protein will occur. The genetic code is very old in evolution. Virtually all the living organisms on the planet use the code listed in Figure 1.3.

Anatomy of a Gene

Our knowledge of the structure of a gene, both physically in terms of its representation as a DNA molecule and functionally, is not yet complete,

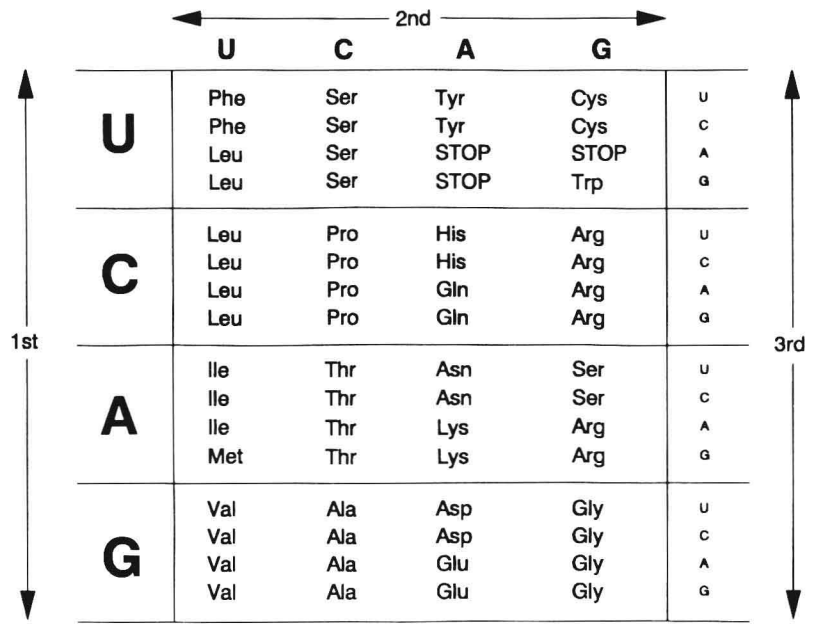

FIGURE 1.3. The genetic code for translation of the triplet nucleotide codons of mRNA into amino acids in proteins.

but much detail is known. A gene is a sequence of nucleotide base pairs that contains the genetic information necessary for directing synthesis of a protein. Figure 1.4 shows the schematic organization of a typical gene, in this instance the c-*myc* growth control oncogene. Figure 1.4 could be a page from an anatomy book of the future. Medical students of the 21st century will read not only gross anatomy and microscopic anatomy but a new text on the anatomy of the human genome.

A gene is organized into segments called *exons*, which are separated by *introns*. The exons contain the DNA sequences, which are transcribed into messenger RNA and then translated into proteins. The base pairs within the introns do not code for protein. There may be no major function associated with introns or the function may as yet may be undiscovered. In bacteria and other simple organisms, the gene sequences are continuous without introns. When a human gene with intron sequences is transcribed into mRNA, the entire DNA sequence from the start to the end of the gene is copied. Since the intron sequences do not contribute to the structure of the protein, these portions must be spliced out of the intermediate RNA transcript to form mRNA.

C-*myc*, as will be presented later in Chapter 8, is an oncogene on chromosome 8 that encodes for a nuclear binding protein that stimulates cell division. The inappropriate expression of c-*myc* protein, usually brought about

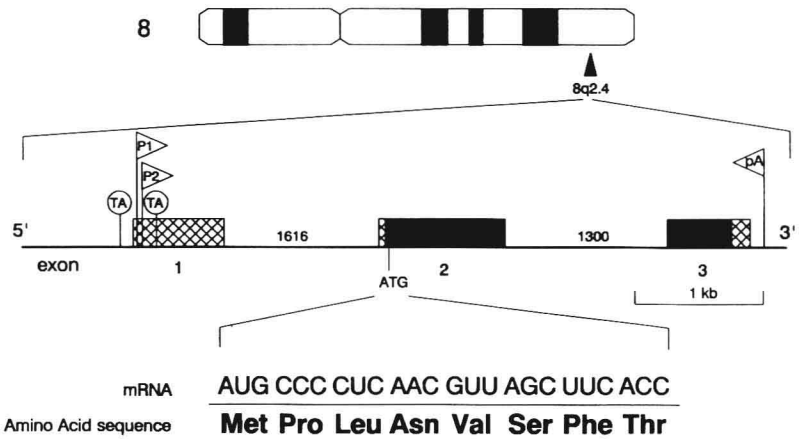

FIGURE 1.4. The anatomy of the c-*myc* oncogene shows its location on chromosome 8 (arrowhead) and its grouping into three exons. The initiation sequence ATG starts the mRNA template coding for the amino acid sequence of the final protein product.

by structural aberrations in the gene's anatomy, is associated with a number of human malignancies. A good example is Burkitt's lymphoma, in which a t(8;14) chromosomal translocation erroneously juxtaposes c-*myc* with the immunoglobulin gene from chromosome 14.

At the top of Figure 1.4 is a schematic diagram of chromosome 8 as it would appear in a Giemsa-stained metaphase karyotype. The light- and dark-staining bands on the short (p) and the long (q) arms of the chromosome define physical areas on the chromosome. The location of c-*myc* on the long arm at 8q2.4 is indicated by an arrowhead.

The center panel of Figure 1.4 shows the genomic structure of c-*myc*. C-*myc* is a small gene, consisting of three exons and spanning only 5,000 base pairs. This is a distance of 0.005 *centiMorgans*. A centiMorgan is defined as a distance along a chromosome that has a 1% probability of undergoing genetic recombination during gamete formation at meiosis. (See also Chapter 6.) As genes get further apart along a chromosome, there is an increasing (though still slight) chance that they will sort independently (as if on separate chromosomes) during sexual reproduction. Genes that are on different chromosomes always sort independently.

The spacing between exons 1 and 2 of the c-*myc* gene is 1,616 base pairs, and between exons 2 and 3 the spacing is about 1,300 base pairs. Important structures within the *myc* gene are indicated in Figure 1.4. P1 and P2 are two promoters that help control the expression of this gene. A promoter is a region of DNA associated with a gene that regulates gene expression. There are other regulatory elements associated with genes besides promot-