Essentials of Medical Genomics

Stuart M. Brown
NYU School of Medicine
New York, NY

with Contributions by

John G. Hay and Harry Ostrer

WILEY-LISS

A John Wiley & Sons, Inc., Publication

Essentials of Medical Genomics

Essentials of Medical Genomics

Stuart M. Brown
NYU School of Medicine
New York, NY

with Contributions by

John G. Hay and Harry Ostrer

WILEY-LISS

A John Wiley & Sons, Inc., Publication

Published by Wiley-Liss, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Brown, Stuart M., 1962-
Essentials of medical genomics / Stuart M. Brown ; with contributions by John G. Hay and Harry Ostrer.
p. cm.
Includes bibliographical references and index.
ISBN 0-471-21003-X (cloth : alk. paper)
1. Medical genetics. 2. Genomics. I. Hay, John G. II. Ostrer, Harry. III. Title.
[DNLM: 1. Genetics, Medical. 2. Genome, Human. 3. Genomics. QZ 50 B879e 2003]
RB155.B674 2003
616′.042–dc21
2002011163

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Contents

Preface

This is a book about medical genomics, a new field that is attempting to combine knowledge generated from the Human Genome Project (HGP) and analytic methods from bioinformatics with the practice of medicine. From my perspective as a research molecular biologist, genomics has emerged as a result of automated high-throughput technologies entering the molecular biology laboratory and of bioinformatics being used to process the data. However, from the perspective of the medical doctor, medical genomics can be understood as an expanded form of medical genetics that deals with lots of genes at once, rather than just one gene at a time. This book is relevant to all medical professionals because *all* disease has a genetic component when hereditary factors are taken into account, such as susceptibility and resistance, severity of symptoms, and reaction to drugs. The National Institutes of Health (NIH) defines medical genetics to include molecular medicine (genetic testing and gene therapy), inherited disorders, and the ethical legal and social implications of the use of genetics technologies in medicine.

> The ultimate goal of genetic medicine is to learn how to prevent disease or to treat it with gene therapy or a drug developed specifically for the underlying defect. Other applications include pharmacogenomics and patient counseling about individual health risks, which

> will be facilitated by new DNA chip technology. Concerns include how to integrate genetic technology into clinical practice and how to prevent genetic-based discrimination.
>
> *Collins 1999*

Before a coherent discussion of genomics is possible, it is necessary to define what is meant by a genome. A genome is the total set of genetic information present in an organism. Generally, every cell in an organism has a complete and identical copy of the genome, but there are many exceptions to this rule. Genomes come in different shapes and sizes for different types of organisms, although there is not always a simple and obvious connection between the size and complexity of an organism and its genome.

An operational definition of genomics might be: The application of high-throughput automated technologies to molecular biology. For the purposes of this book, genomics is defined broadly to include a variety of technologies, such as genome sequencing, DNA diagnostic testing, measurements of genetic variation and polymorphism, microarray gene expression, proteomics (measurements of all protein present in a cell or tissues), pharmacogenomics (genetic predictions of drug reactions), gene therapy, and other forms of DNA drugs. A philosophical definition of genomics might be: A holistic or systems approach to information flow within the cell.

Biology is complex. In fact, complexity is the hallmark of biological systems from cells to organisms to ecosystems. Rules have exceptions. Information tends to flow in branching feedback loops rather than in neat chains of cause and effect. Biological systems are not organized according to design principles that necessarily make sense to humans. Redundancy and seemingly unnecessary levels of interlocking dynamic regulation are common. Molecular biology is a profoundly reductionist discipline—complex biological systems are dissected by forcing them into a framework so that a single experimental variable is

isolated. Genomics must embrace biological complexity and resist the human tendency to look for simple solutions and clear rules. Genomic medicine will not find a single gene for every disease. To successfully modify a complex dynamic system that has become unbalanced in a disease state will require a much greater subtlety of understanding than is typical in modern medicine.

The HGP was funded by the United States and other national governments for the express purpose of improving medicine. Now that the initial goals of the project have largely been met, the burden has shifted from DNA sequencing technologists to biomedical researchers and clinicians who can use this wealth of information to bring improved medicine to the patients—medical genomics. The initial results produced by these genome-enabled researchers give every indication that the promises made by those who initially proposed the genome project will be kept.

The initial sequencing of the 3.2 billion base pairs of the human genome is now essentially complete. A lot of fancy phrases have been used to tout the enormous significance of this achievement. Francis Collins, director of the National Human Genome Research Institute called it ''a bold research program to characterize in ultimate detail the complete set of genetic instructions of the human being.'' President Clinton declared it ''a milestone for humanity.''

This book goes light on the hyperbole and the offering of rosy long-term predictions. Instead, it focuses on the most likely short-term changes that will be experienced in the practice of medicine. The time horizon here is 5 years into the future for technologies that are currently under intensive development and 10 years for those that I consider extremely likely to be implemented on a broad scale. In 5 years' time, you will need to throw this book away and get a new one to remain abreast of the new technologies coming over the horizon.

This book is an outgrowth of a medical genomics course that I developed in 2000 and 2001 as an elective course for medical students at the New York University School of Medicine. Based on this experience, I can predict with confidence that medical genomics will become an essential and required part of the medical school curriculum in 5 years or less. I also learned that medical students (and physicians in general) need to learn to integrate an immense amount of information, so they tend to focus on the essentials and they ask to be taught ''only what I really need to know.''

It is difficult to boil down medical genomics to a few hours' worth of bullet points on *PowerPoint* slides. There is a *lot* of background material that the student must keep in mind to understand the new developments fully. Medical genomics relies heavily on biochemistry, molecular biology, probability and statistics, and most of all on classical genetics.

My specialty is in the relatively new field of bioinformatics, which has recently come in from the extreme reaches of theoretical biology to suddenly play a key role in the interpretation of the human genome sequence for biomedical research. Bioinformatics is the use of computers to analyze biological information—primarily DNA and protein sequences. This is a useful perspective from which to observe and discuss the emerging field of medical genomics, which is based on the analysis and interpretation of biological information derived from DNA sequences. Two chapters were written by colleagues who are deep in the trenches of the battle to integrate genome technologies into the day-to-day practice of medicine in a busy hospital. Harry Ostrer is the director of the Human Genetics Program at the New York University Medical Center, where he overseas hundreds of weekly genetic tests of newborns, fetuses, and prospective parents. John Hay is co-director of the molecular biology core lab for the New York University General Clinical Research Center and the principle

investigator of numerous projects to develop and test gene therapy methods.

Stuart M. Brown

REFERENCE

Collins F., Geriatrics 1999; 54: 41–47

To Kim, who encourages me to write
and to Justin and Emma, who make me proud

Acknowledgments

This book grew out of a course that I taught to medical students at NYU School of Medicine in 2000 and 2001 as part of an interdisciplinary effort called the ''Master Scholars Program.'' Joe Sanger, the Society Master for Informatics and Biotechnology, cajoled, coaxed, and guilt-tripped me into teaching the course. I also thank Ross Smith for hiring me as the *Molecular Biology Consultant* to the Academic Computing unit at NYU School of Medicine. He created a work environment where I could freely organize my time between teaching, consulting, maintaining the core computing systems, and writing. I must also thank my System Managers Tirza Doniger and Guoneng Zhong for picking up the slack for maintaining the UNIX systems and handling the tech support questions so that I could have time for writing.

In a larger context, I must thank my wife Kim for encouraging me to write something less technical that would appeal to wider audience, and for frequently suggesting that I take ''writing days'' to finish up the manuscript. She also provided some clutch help on several of the figures.

At Wiley, I thank Luna Han for having interest and faith in my concept for this book, and Kristin Cooke Fasano for shepherding me through all of the details that are required to make a manuscript into a book.

Finally, I must give credit to Apple Computer for the wonderful and light iBook that allowed me to do a great deal of the writing on the Long Island Railroad.

Stuart M. Brown

CHAPTER

1

DECIPHERING THE HUMAN GENOME PROJECT

The Human Genome Project is a bold undertaking to understand, at a fundamental level, all of the genetic information required to build and maintain a human being. The human **genome** is the complete information content of the human cell. This information is encoded in approximately 3.2 billion base pairs of DNA contained on 46 **chromosomes** (22 pairs of autosomes plus 2 sex chromosomes) (Fig. 1-1). The completion in 2001 of the first draft of the human genome sequence is only the first phase of this project (Lander et al., 2001; Venter et al., 2001). *This figure also appears in the Color Insert section.*

To use the metaphor of a book, the draft genome sequence gives biology all of the letters, in the correct order on the pages, but without the ability to recognize words, sentences, punctuation, or even an understanding of the language in which the book is written. The task of making sense of all of this raw biological information falls, at least initially, to **bioinformatics** specialists who make use of computers to find the words and decode the language. The next step is to integrate all of this information into a new form of experimental biology, known as

Essentials of Medical Genomics, Edited by Stuart M. Brown.
ISBN 0-471-21003-X.

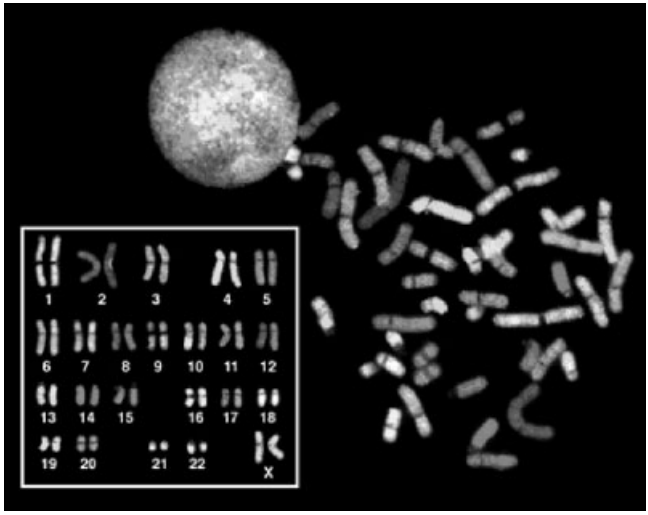

FIGURE 1-1. The human karyotype (SKY image). Figure also appears in Color Figure Section. Reprinted with permission from Thomas Ried National Cancer Institute.

genomics, that can ask meaningful questions about what is happening in complex systems where tens of thousands of different genes and proteins are interacting simultaneously.

The primary justification for the considerable amount of money spent on sequencing the human genome (from both governments and private corporations), is that this information will lead to dramatic medical advances. In fact, the first wave of new drugs and medical technologies derived from genome information is currently making its way through clinical trials and into the health-care system. However, in order for medical professionals to make effective use of these new advances, they need to understand something about genes and genomes. Just as it is important for physicians to understand how to Gram stain and evaluate a culture of bacteria, even if they never actually perform this test themselves in their medical practice, it is important to understand how DNA technologies work in order to appreciate their strengths, weaknesses, and peculiarities.

However, before we can discuss whole genomes and genomic technologies, it is necessary to understand the basics of how

genes function to control biochemical processes within the cell (molecular biology) and how hereditary information is transmitted from one generation to the next (genetics).

THE PRINCIPLES OF INHERITANCE

The principles of genetics were first described by the monk Gregor Mendel in 1866 in his observations of the inheritance of traits in garden peas. Mendel described ''differentiating characters'' (*differierende Merkmale*) that may come in several forms. In his monastery garden, he made crosses between strains of garden peas that had different characters, each with two alternate forms that were easily observable, such as purple or white flower color, yellow or green seed color, smooth or wrinkled seed shape, and tall or short plant height. (These alternate forms are now known as **alleles.**) Then he studied the distribution of these forms in several generations of offspring from his crosses.

Mendel observed the same patterns of inheritance for each of these characters. Each strain, when bred with itself, showed no changes in any of the characters. In a cross between two strains that differ for a single character, such as pink vs. white flowers, the first generation of hybrid offspring (F_1) all looked like one parent—all pink. Mendel called this the **dominant** form of the character. After self-pollinating the F_1 plants, the second-generation plants (F_2) showed a mixture of the two parental forms (Fig. 1-2). This is known as **segregation.** The **recessive** form that was not seen in the F_1 generation (white flowers) was found in one-quarter of the F_2 plants.

Mendel also made crosses between strains of peas that differed for two or more traits. He found that each of the traits was assorted independently in the progeny—there was no connection between whether an F_2 plant had the dominant or recessive form for one character and what form it carried for another character (Fig. 1-3).

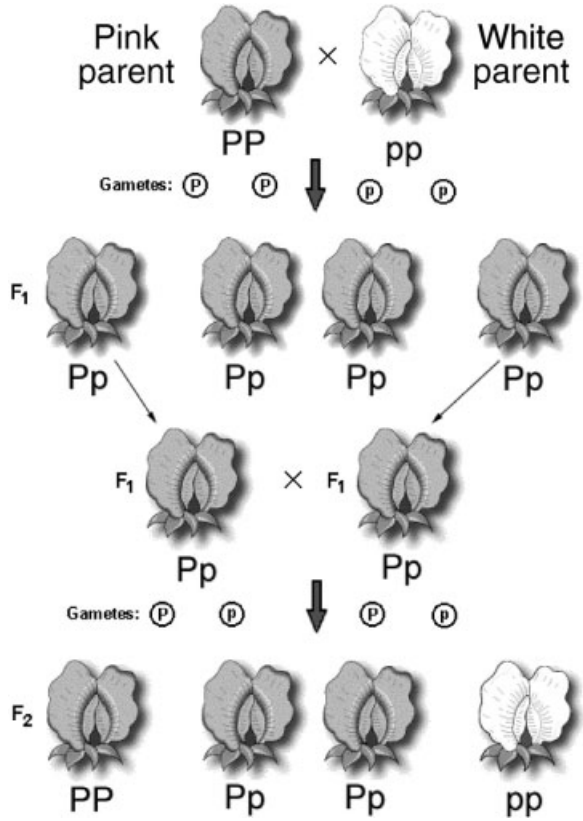

FIGURE 1-2. Mendel observed a single trait segregating over two generations.

Mendel created a theoretical model (now known as Mendel's Laws of Genetics) to explain his results. He proposed that each individual has two copies of the hereditary material for each character, which may determine different forms of that character. These two copies separate and are subjected to independent assortment during the formation of gametes (sex cells). When a new individual is created by the fusion of two sex cells, the two copies from the two parents combine to produce a visible trait, depending on which form is dominant and which is recessive. Mendel did not propose any physical explanation for

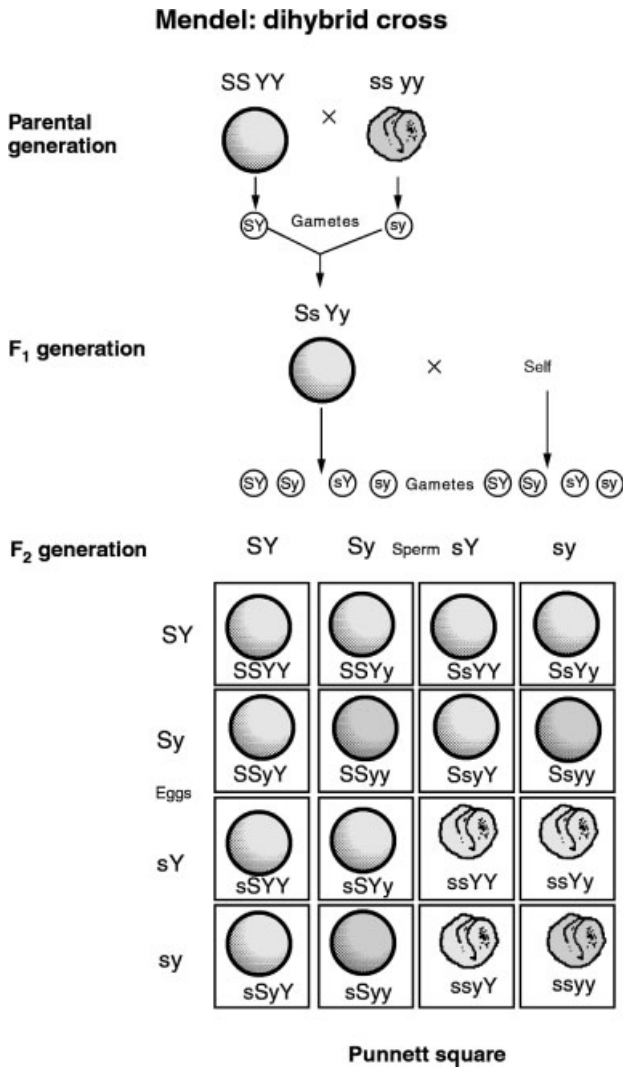

FIGURE 1-3. A cross in which two independent traits segregate.

how these traits were passed from parent to progeny; his characters were purely abstract units of heredity.

Modern genetics has completely embraced Mendel's model with some additional detail. There may be more than two different alleles for a gene in a population, but each individual

has only two, which may be the same (**homozygous**) or different (**heterozygous**). In some cases, two different alleles combine to produce an intermediate form in heterozygous individuals; for example, a red flower allele and a white flower allele may combine to produce a pink flower; and in humans, a type A allele and a type B allele for red blood cell antigens combine to produce the AB blood type.

GENES ARE ON CHROMOSOMES

In 1902, Walter Sutton, a microscopist, proposed that Mendel's heritable characters resided on the chromosomes that he observed inside the cell nucleus (Fig. 1-4). Sutton noted that ''the association of paternal and maternal chromosomes in pairs and their subsequent separation during cell division . . . may constitute the physical basis of the Mendelian law of heredity'' (Sutton, 1903).

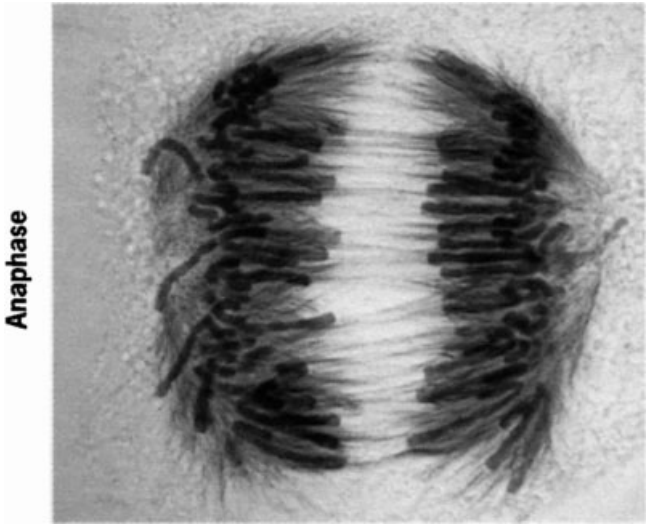

FIGURE 1-4. Chromosomes during anaphase in a lily cell.

In 1909, the Danish botanist Wilhelm Johanssen coined the term *gene* to describe Mendel's heritable characters. In 1910, Thomas Hunt Morgan (1910) found that a trait for white eye color was located on the X chromosome of the fruit fly and was inherited together with a factor that determines sex. A number of subsequent studies by Morgan and others showed that each gene for a particular trait was located at a specific spot, or **locus,** on a chromosome in all individuals of a species. The chromosome was a linear organization of genes, like beads on a string. Throughout the early part of the twentieth century, a gene was considered to be a single, fundamental, indivisible unit of heredity, in much the same way as an atom was considered to be the fundamental unit of matter.

Each individual has two copies of each chromosome, having received one copy from each parent. In sexual cell division (**meiosis**), the two copies of each chromosome in the parent are separated and randomly assorted among the sex cells (sperm or egg) in a process called segregation. When a sperm and an egg cell combine, a new individual is created with new combinations of alleles. It is possible to observe the segregation of chromosomes during meiosis using only a moderately powerful microscope. It is an aesthetically satisfying triumph of biology that this observed segregation of chromosomes in cells exactly corresponds to the segregation of traits that Mendel observed in his peas.

Recombination and Linkage

In the early part of the twentieth century, Mendel's concepts of inherited characters were broadly adopted both by practical plant and animal breeders as well as by experimental geneticists. It rapidly became clear that Mendel's experiments represented an oversimplified view of inheritance. He must have intentionally chosen characters in his peas that were inherited

independently. In breeding experiments in which many traits differ between parents, it is commonly observed that progeny inherit pairs or groups of traits together from one parent far more frequently than would be expected by chance alone. This observation fit nicely into the chromosome model of inheritance—if two genes are located on the same chromosome, then they will be inherited together when that chromosome segregates into a gamete and that gamete becomes part of a new individual.

However, it was also observed that ''linked'' genes do occasionally separate. A theory of **recombination** was developed to explain these events. It was proposed that during the process of meiosis the homologous chromosome pairs line up and exchange segments in a process called crossing-over. This theory was supported by microscopic evidence of X-shaped structures called chiasmata forming between paired homologous chromosomes in meiotic cells (Fig. 1-5).

If a parent cell contains two different alleles for two different linked genes, then after the cross-over, the chromosomes in the gametes will contain new combinations of these alleles. For example, if one chromosome contains alleles **A** and **B** for two genes, and the other chromosome contains alleles **a** and **b**, then—without cross-over—all progeny must inherit a chromosome from that parent with either an **A-B** or an **a-b** allele combination. If a cross-over occurs between the two genes, then the resulting chromosomes will contain the **A-b** and **a-B** allele combinations (Fig. 1-6).

FIGURE 1-5. Chiasmata visible in an electron micrograph of a meiotic chromosome pair.

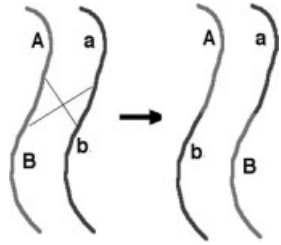

FIGURE 1-6. A single cross-over between a chromosome with **A-B** alleles and a chromosome with **a-b** alleles, forming **A-b** and **a-B** recombinant chromosomes.

Morgan, continuing his work with fruit flies, demonstrated that the chance of a cross-over occurring between any two linked genes is proportional to the distance between them on the chromosome. Therefore, by counting the frequency of cross-overs between the alleles of a number of pairs of genes, it is possible to map those genes on a chromosome. (Morgan was awarded the 1933 Nobel Prize in medicine for this work.) In fact, it is generally observed that on average, there is more than one cross-over between every pair of homologous chromosomes in every meiosis, so that two genes located on opposite ends of a chromosome do not appear to be linked at all. On the other hand, alleles of genes that are located very close together are rarely separated by recombination (Fig. 1-7).

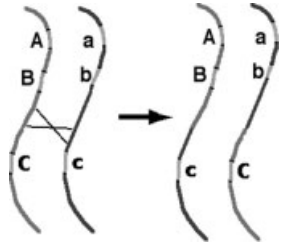

FIGURE 1-7. Genes A and B are tightly linked so that they are not separated by recombination, but gene C is farther away. After recombination occurs in some meiotic cells, gametes are produced with the following allele combinations: **A-B-C**, **a-b-c**, **A-B-c**, and **a-b-C**.

The relationship between the frequency of recombination between alleles and the distance between gene loci on a chromosome has been used to construct genetic maps for many different organisms, including humans. It has been a fundamental assumption of genetics for almost 100 years that recombinations occur randomly along the chromosome at any location, even within genes. However, recent data from DNA sequencing of genes in human populations suggest that there are recombination hot spots and regions where recombination almost never occurs. This creates groups of alleles from neighboring genes on a chromosome, known as **haplotypes,** that remain linked together across hundreds of generations.

Genes Encode Proteins

In 1941, Beadle and Tatum showed that a single mutation, caused by exposing the fungus *Neurospora crassa* to X-rays, destroyed the function of a single enzyme, which in turn interrupted a biochemical pathway at a specific step. This mutation segregated among the progeny exactly as did the traits in Mendel's peas. The X-ray damage to a specific region of one chromosome destroyed the instructions for the synthesis of a specific enzyme. Thus a gene is a spot on a chromosome that codes for a single enzyme. In subsequent years, a number of other researchers broadened this concept by showing that genes code for all types of proteins, not just enzymes, leading to the ''One Gene, One Protein'' model, which is the core of modern molecular biology. (Beadle and Tatum shared the 1958 Nobel Prize in medicine.)

Genes Are Made of DNA

The next step in understanding the nature of the gene was to dissect the chemical structure of the chromosome. Crude

biochemical purification had shown that chromosomes are composed of both protein and DNA. In 1944, Avery, MacLeod, and McCarty conducted their classic experiment on the ''transforming principle.'' They found that DNA purified from a lethal S (smooth) form of *Streptococcus pneumoniae* could transform a harmless R (rough) strain into the S form (Fig. 1-8). Treatment of the DNA with protease to destroy all of the protein had no effect, but treatment with DNA-degrading enzymes blocked the transformation. Therefore, the information that transforms the bacteria from R to S must be contained in the DNA.

Hershey and Chase confirmed the role of DNA with their classic 1952 ''blender experiment'' on bacteriophage viruses. The phages were radioactively labeled with either ^{35}S in their proteins or ^{32}P in their DNA. The researchers used a blender to interrupt the process of infection of *Escherichia coli* bacteria by the phages. Then they separated the phages from the infected bacteria by centrifugation and collected the phages and bacteria separately. They observed that the ^{35}S-labeled protein remained with the phage while the ^{32}P-labeled DNA was found inside the infected bacteria (Fig. 1-9). This proved that it is the DNA portion of the virus that enters the bacteria and contains the

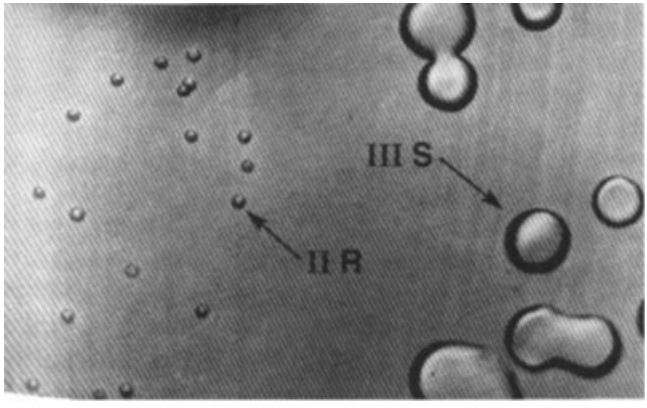

FIGURE 1-8. Rough and smooth *Streptococcus* cells.

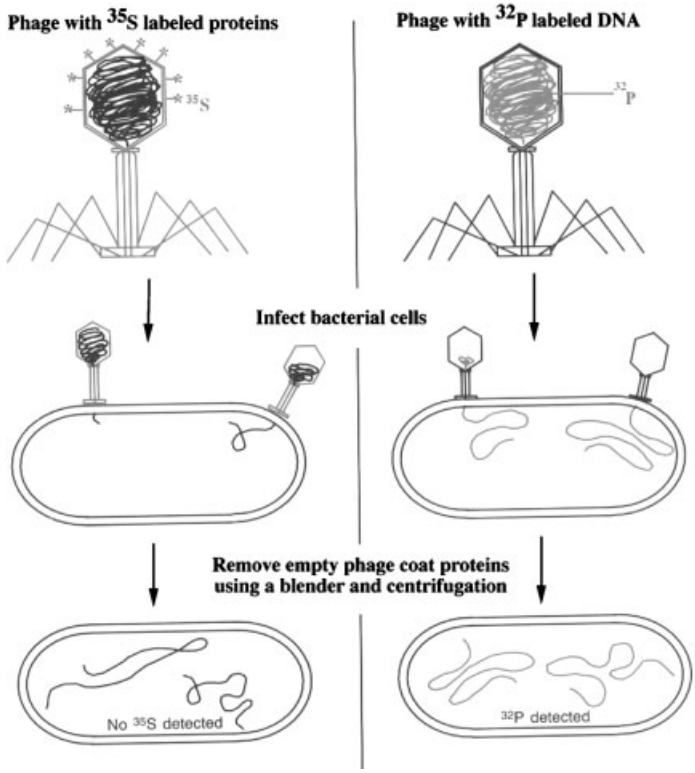

FIGURE 1-9. In the Hershey-Chase blender experiment, *E. coli* bacteria were infected with either ^{35}S-labeled proteins or ^{32}P-labeled DNA. After removing the phages, the ^{32}P-labeled DNA, but not the ^{35}S-labeled protein, was found inside the bacteria. Reprinted with permission from the DNA Science Book, CSHL Press.

genetic instructions for producing new phage, not the proteins, which remain outside. (Hershey was awarded the 1969 Nobel Prize for this work.)

DNA Structure

Now it was clear that genes are made of DNA, but how does this chemically simple molecule contain so much information? DNA is a long polymer molecule that contains a mixture of four

different chemical subunits: adenine (A), cytosine (C), guanosine (G), and thymine (T). These subunits, known as nucleotide bases, have similar two-part chemical structures that contain a deoxyribose sugar and a nitrogen ring (Fig. 1-10), hence the name deoxyribonucleic acid. The real challenge was to understand how the nucleotides fit together in a way that can contain a lot of information.

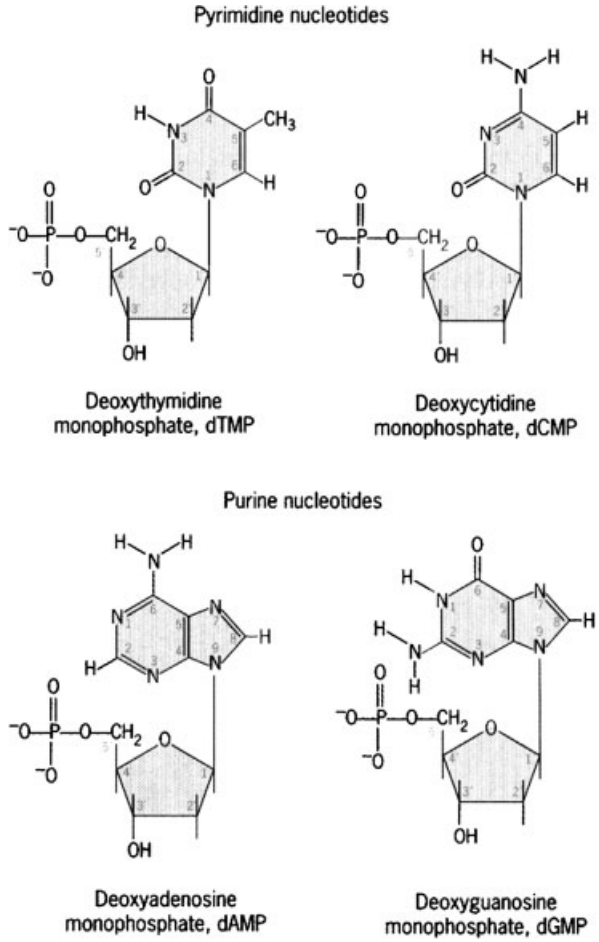

FIGURE 1-10. The DNA bases.

In 1950, Edwin Chargaff discovered that there was a consistent one-to-one ratio of adenine to thymine and of guanine to cytosine in any sample of DNA from any organism. In 1951, Linus Pauling and R. B. Corey described the α-helical structure of a protein. Shortly thereafter, Rosalind Franklin provided X-ray crystallographic images of DNA to James Watson and Francis Crick, which showed many similarities to the α-helix described by Pauling (Fig. 1-11). Watson and Crick's crucial insight was to realize that DNA formed a double helix with complementary bonds between adenine-thymine and guanine-cytosine pairs.

The Wastson-Crick model of the structure of DNA looks like a twisted ladder. The two sides of the ladder are formed by strong covalent bonds between the phosphate on the 5′ carbon of one deoxyribose sugar and the methyl side groups of the

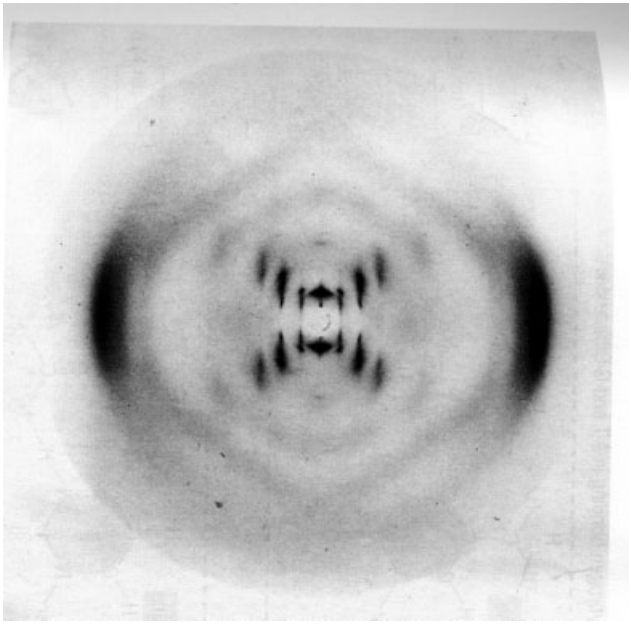

FIGURE 1-11. Franklin's X-ray diffraction picture of DNA.