An Engineering Introduction to Biotechnology

J. Patrick Fitch

An Engineering Introduction to Biotechnology

Tutorial Texts Series

- *An Engineering Introduction to Biotechnology,* J. Patrick Fitch, Vol. TT55
- *Image Performance in CRT Displays,* Kenneth Compton, Vol. TT54
- *Introduction to Laser Diode-Pumped Solid State Lasers,* Richard Scheps, Vol. TT53
- *Modulation Transfer Function in Optical and Electro-Optical Systems*, Glenn D. Boreman, Vol. TT52
- *Uncooled Thermal Imaging Arrays, Systems, and Applications*, Paul W. Kruse, Vol. TT51
- *Fundamentals of Antennas*, Christos G. Christodoulou and Parveen Wahid, Vol. TT50
- *Basics of Spectroscopy*, David W. Ball, Vol. TT49
- *Optical Design Fundamentals for Infrared Systems*, *Second Edition,* Max J. Riedl, Vol. TT48
- *Resolution Enhancement Techniques in Optical Lithography*, Alfred Kwok-Kit Wong, Vol. TT47
- *Copper Interconnect Technology*, Christoph Steinbrüchel and Barry L. Chin, Vol. TT46
- *Optical Design for Visual Systems*, Bruce H. Walker, Vol. TT45
- *Fundamentals of Contamination Control*, Alan C. Tribble, Vol. TT44
- *Evolutionary Computation: Principles and Practice for Signal Processing*, David Fogel, Vol. TT43
- *Infrared Optics and Zoom Lenses,* Allen Mann, Vol. TT42
- *Introduction to Adaptive Optics,* Robert K. Tyson, Vol. TT41
- *Fractal and Wavelet Image Compression Techniques,* Stephen Welstead, Vol. TT40
- *Analysis of Sampled Imaging Systems,* R. H. Vollmerhausen and R. G. Driggers, Vol. TT39
- *Tissue Optics: Light Scattering Methods and Instruments for Medical Diagnosis*, Valery Tuchin, Vol. TT38
- *Fundamentos de Electro-Óptica para Ingenieros,* Glenn D. Boreman, translated by Javier Alda, Vol. TT37
- *Infrared Design Examples,* William L. Wolfe, Vol. TT36
- *Sensor and Data Fusion Concepts and Applications, Second Edition*, L. A. Klein, Vol. TT35
- *Practical Applications of Infrared Thermal Sensing and Imaging Equipment, Second Edition*, Herbert Kaplan, Vol. TT34
- *Fundamentals of Machine Vision*, Harley R. Myler, Vol. TT33
- *Design and Mounting of Prisms and Small Mirrors in Optical Instruments*, Paul R. Yoder, Jr., Vol. TT32
- *Basic Electro-Optics for Electrical Engineers,* Glenn D. Boreman, Vol. TT31
- *Optical Engineering Fundamentals,* Bruce H. Walker, Vol. TT30
- *Introduction to Radiometry,* William L. Wolfe, Vol. TT29
- *Lithography Process Control*, Harry J. Levinson, Vol. TT28
- *An Introduction to Interpretation of Graphic Images,* Sergey Ablameyko, Vol. TT27
- *Thermal Infrared Characterization of Ground Targets and Backgrounds,* P. Jacobs, Vol. TT26
- *Introduction to Imaging Spectrometers,* William L. Wolfe, Vol. TT25
- *Introduction to Infrared System Design,* William L. Wolfe, Vol. TT24
- *Introduction to Computer-based Imaging Systems,* D. Sinha, E. R. Dougherty, Vol. TT23
- *Optical Communication Receiver Design,* Stephen B. Alexander, Vol. TT22
- *Mounting Lenses in Optical Instruments,* Paul R. Yoder, Jr., Vol. TT21
- *Optical Design Fundamentals for Infrared Systems,* Max J. Riedl, Vol. TT20
- *An Introduction to Real-Time Imaging,* Edward R. Dougherty, Phillip A. Laplante, Vol. TT19
- *Introduction to Wavefront Sensors,* Joseph M. Geary, Vol. TT18
- *Integration of Lasers and Fiber Optics into Robotic Systems,* J. A. Marszalec, E. A. Marszalec, Vol. TT17
- *An Introduction to Nonlinear Image Processing,* E. R. Dougherty, J. Astola, Vol. TT16
- *Introduction to Optical Testing,* Joseph M. Geary, Vol. TT15
- *Image Formation in Low-Voltage Scanning Electron Microscopy,* L. Reimer, Vol. TT12
- *Diazonaphthoquinone-based Resists,* Ralph Dammel, Vol. TT11
- *Infrared Window and Dome Materials,* Daniel C. Harris, Vol. TT10
- *An Introduction to Morphological Image Processing,* Edward R. Dougherty, Vol. TT9
- *An Introduction to Optics in Computers,* Henri H. Arsenault, Yunlong Sheng, Vol. TT8

An Engineering Introduction to Biotechnology

J. Patrick Fitch

Tutorial Texts in Optical Engineering
Volume TT55

Arthur R. Weeks, Jr., Series Editor
Invivo Research Inc. and University of Central Florida

SPIE PRESS
A Publication of SPIE—The International Society for Optical Engineering
Bellingham, Washington USA

Library of Congress Cataloging-in-Publication Data

Fitch, J. Patrick
An engineering introduction to biotechnology / by J. Patrick Fitch
p. cm. — (Tutorial texts in optical engineering; v. TT55)
Includes bibliographical references and index.
ISBN 0-8194-4497-9
1. Biotechnology. 2. Genetic engineering. I. Title. II. Series.
TP248.2 .F55 2002
660.6—dc21 2001060203
CIP

Published by

SPIE—The International Society for Optical Engineering
P.O. Box 10
Bellingham, Washington 98227-0010
Phone: 360/676-3290
Fax: 360/647-1445
Email: spie@spie.org
www.spie.org

Printed in the United States of America.

Cover art:

Model of a clamp protein (navy blue) sliding along a stretch of DNA. All cells depend on interactions among proteins and DNA. Sliding-clamp proteins are involved in tethering other proteins to DNA for its replication and repair. Bases are in pink (A), sky blue (T), green (C), and yellow (G) colors and the backbone sugars are grey and silver on opposite strands and phosphates are red and gold clusters.

(Courtesy of Daniel Barsky, Lawrence Livermore National Laboratory)

Introduction to the Series

The Tutorial Texts series was initiated in 1989 as a way to make the material presented in SPIE short courses available to those who couldn't attend and to provide a reference book for those who could. Typically, short course notes are developed with the thought in mind that supporting material will be presented verbally to complement the notes, which are generally written in summary form, highlight key technical topics, and are not intended as stand-alone documents. Additionally, the figures, tables, and other graphically formatted information included with the notes require further explanation given in the instructor's lecture. As stand-alone documents, short course notes do not generally serve the student or reader well.

Many of the Tutorial Texts have thus started as short course notes subsequently expanded into books. The goal of the series is to provide readers with books that cover focused technical interest areas in a tutorial fashion. What separates the books in this series from other technical monographs and textbooks is the way in which the material is presented. Keeping in mind the tutorial nature of the series, many of the topics presented in these texts are followed by detailed examples that further explain the concepts presented. Many pictures and illustrations are included with each text, and where appropriate tabular reference data are also included.

To date, the texts published in this series have encompassed a wide range of topics, from geometrical optics to optical detectors to image processing. Each proposal is evaluated to determine the relevance of the proposed topic. This initial reviewing process has been very helpful to authors in identifying, early in the writing process, the need for additional material or other changes in approach that serve to strengthen the text. Once a manuscript is completed, it is peer reviewed to ensure that chapters communicate accurately the essential ingredients of the processes and technologies under discussion.

The Tutorial Text series, which now numbers more than fifty titles, has expanded to include not only texts developed by short course instructors but also those written by other topic experts. It is my goal to maintain the style and quality of books in the series, and to further expand the topic areas to include emerging as well as mature subjects in optics, photonics, and imaging.

Arthur R. Weeks, Jr.
Invivo Research Inc. and University of Central Florida

To Kathy, Christine, Stephanie, and Alex with love

Contents

Preface

Biological discovery has accelerated tremendously in the decades preceding the twenty-first century and promises to continue. As the solid-state transistor enabled so much of the information revolution, a few fundamental contributions were critical to the life science revolution. Technologies such as polymerase chain reactions, restriction enzymes, and recombinant DNA have opened new scientific possibilities. A new vocabulary has been brought to life—genomics, proteomics, physiomics, etc. What may be missing from the public view is the dependence of these new biological approaches on principles and technologies from engineering and the physical sciences.

There have been several significant breakthroughs in biotechnology that require mentioning the scientists and inventors. This is not meant to be an exhaustive list, but rather a mini who's who of biotechnologists.

1865 Gregor Mendel is the parent of classic genetics. His experiments and observations of hybridizing pea plants were reported in 1865.

1944 Oswald Avery, Colin MacLeod, and Maclyn McCarty in the 1940s put DNA and inheritance together and called it the transforming principle. Before their research, proteins were believed to be the mechanisms for transference of inherited properties.

1953 James Watson and Francis Crick measured the structural form and other properties of the DNA double helix.

1958 Frederick Sanger developed the Sanger-sequencing approach and determined the structure of insulin, for which he was awarded the Nobel Prize in chemistry in 1958.

1973 Stanley Cohen and Herbert Boyer created ways to engineer the recombination of DNA in living organisms. The first results in 1971 used calcium chloride to make *escherichia coli* more permeable so it would accept a small circular ring of DNA known as a plasmid. Later results were used to create Genentech with businessman Robert Swanson (co-founder with Boyer), one of the first successful biotechnology companies.

1975 M. E. (Ed) Southern developed DNA fragment separation on an agarose gel, followed by blotting onto a membrane where sequence specific probes can be used for identification. This invention led to numerous other assays based on binding or hybridization.

1993 Kary Mullis invented the polymerase chain reaction, which can copy specific regions of DNA, and received a Nobel Prize in 1993.

Biology is usually presented differently than engineering and physical science. The physical sciences strongly promote a reductive approach that decomposes complex phenomena into simpler subsystems that can be incorporated into models. As an example, consider the high-energy physics community's pursuit of subatomic particles. The expectation is that models increase in accuracy with the goal of becoming predictive of the phenomena. Biology has been an observational science. The quantification and reduction of the complex phenomena observed in biology has not usually allowed a reductive approach. This can be a source of frustration for nonbiologists, who often conclude that the science is only a memorization activity with vocabulary and experimental anecdotes in place of models and predictive theories. One of the goals of this book is to present introductory biotechnology from the perspective of someone trained as a physical scientist.

This book is for technical professionals—engineers, physical scientists, and technical managers and marketers. The goal is to create the opportunity for these professionals to determine if their technologies and organizations have relevant application in the life sciences. The plan is to introduce the basic concepts of biology, emphasizing "omic" or "whole mass" approaches, describe large-scale applications such as DNA sequencing, and illustrate technical successes with a few case studies of bioinstrumentation. The subtitle might be "omic" technologies for "ohmic" engineers. We do not attempt to present the detailed biochemistry, safety, or ethical issues that are also important components of biotechnology. It is hoped that this approach will appeal to the reader, will facilitate new discoveries through the interaction of disciplines, and will enable follow-on reading of existing molecular biotechnology texts. The canonical reader of the book is an engineer who has not had biology or chemistry for a while. This book is offered as a "prep class" for more detailed books on biology and biotechnology. I hope this book fills the gap and makes texts like *Molecular Biotechnology* by Glick and Pasternak reachable.

There are many ways to define biotechnology. Sometimes biotechnology is considered to be the use of engineering principles in biology. Two examples are producing new enzymes for laundry detergent and brewing beer, constrained by safety, consumer appeal, and business considerations. Sometimes biotechnology is used to describe the technical or engineering part of a life science program. This might include instrumentation and software for applications that include drug discovery and DNA sequencing. One common interpretation is that biotechnology is synonymous with genetic engineering, where functions are added or removed by modifying the nucleic acids in an organism. In this book, biotechnology is any technique, technology or application that depends on or benefits from information obtained through the ability to extract, copy, modify, or reintroduce the nucleic acids of an organism.

Many people have enabled this writing endeavor and deserve acknowledgments. I am grateful to my family for their loving support and donating many nights and weekends required for this project. Thanks to the biologists at Lawrence Livermore National Laboratory who have encouraged me,

especially T. Carrano, L. Ashworth, J. Felton, P. McCready, and L. Stubbs. I have enjoyed collaborating with many exceptional engineers, computer scientists, and physical scientists, including T. Slezak, J. Balch, and C. Davidson. Thanks to the staff at the SPIE for encouraging the short course on an introduction to genomics and then this writing endeavor.

A special thanks to Bahrad Sokhansanj, a Ph.D. candidate at the time of this writing, for his many contributions and discussions and for his enthusiastic pursuit of our joint projects. Our joint paper, "Genomic engineering: moving beyond DNA sequence to function" (*Proceedings of IEEE*, vol. 88, no. 12, pp. 1949–1971, Dec. 2000) was an excellent starting point for collecting the thoughts presented in this book and for revising the SPIE short course.

There were also several people who provided significant input on drafts of the manuscript—in particular, Janine Garnham, Beth Vitalis, and Tom Kuczmarski. A special thank you to Kathy Fitch for helping review every version of the manuscript.

I would also like to acknowledge the Public Health Image Library PHIL™ provided at the Centers for Disease Control (CDC) web site http://phil.cdc.gov/Phil. The CDC provided in the public domain the PHIL™ images used in this book. I have included the PHIL™ identification number and the source acknowledgement (organization and scientist) when available. Some of this work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

An additional dedication to the heroic people aboard UA93 on September 11, 2001 is appropriate. As one of many people in the Washington, DC area that day, thank you. May the knowledge derived from and the activities that follow this book honor your memory through applications beneficial to humanity.

J. Patrick Fitch
November 2001

An Engineering Introduction to Biotechnology

Part I Introduction to Biology

1

Basic Biology

1.1 Life

Life is the property that makes it possible to reproduce, to adapt to surroundings, and to take in food and create energy from it. These three attributes are also known as reproduction, self-regulation, and metabolism. Biologists have devised several methods for categorizing living organisms. Despite centuries of evaluation, the categories for labeling life continue to be updated as organisms are reclassified, new organisms are discovered, and new approaches to classification are proposed. Approaches to classification include perceived evolution, observed attributes, and nucleic acid similarities. Of the many potential categories, we will focus on what is essential for discussing biotechnology.

Biotechnology applications often require the use of bacteria, yeast, insect, and mammalian cells. Applications of biotechnology to human health may require animal models and human immune response mechanisms. Biotechnology often aspires to exploit the best mechanisms found in nature, and these mechanisms may need to be extracted from the cells of different organisms. These are a few of the reasons that background information on a broad range of organisms is needed for biotechnology.

Unfortunately, the vocabulary for describing and working with cellular and subcellular mechanisms is largely historic and observational. The absence of analytic and predictive models is striking to many physical scientists and engineers. At this point it is essential to introduce the vocabulary needed to discuss living organisms. We have focused the discussion on the topics needed for working in the biotechnology field.

Since the 1930s the living world has been divided into two different domains of organisms called eukaryotes and prokaryotes. Eukaryote translates into “true kernel” or “with nucleus.” Prokaryote translates into “without kernel or nucleus.” We will interchange common use and misuse of the Latin roots and conjugations—e.g., the popular term *prokaryotes* and the proper Latin plural *prokarya*. Traditionally, after the domains Eukarya and Prokarya, the next levels

of separation of living organisms are kingdoms, followed by phyla, classes, orders, families, genera, species, and strains.

The Eukarya include four kingdoms: Protista, Plantae, Fungi, and Animalia. The Protista are single-celled organisms, such as the protozoa. The Plantae are multicellular organisms that manufacture their food, such as ferns and trees. The Fungi are single or multicellular organisms that absorb food from the environment, such as yeasts and molds. The Animalia are multicellular organisms that must capture food and digest it internally, such as dogs, birds, and fish. Eukaryotes are the organisms that most people would observe in their daily routine and include people, animals, plants, and fungi. The presence of a nucleus and other cellular organelles is specific to Eukarya and complicates understanding how these cells work.

Prokarya were traditionally defined as the Monera kingdom and included both bacteria and archaebacteria. In the past two decades, Carl Woese of the University of Illinois and others have proposed the archaea as a separate domain of life—distinct from bacteria. Archaea are often extremophiles—living, for instance, in high salt (halophilic archaea) or high temperature (thermophilic archaea) environments. Using nucleic acids extracted from ribosomes (the ribosome is a ubiquitous molecule in living organisms) as an evolutionary measure, eukarya, archaea, and bacteria have been categorized. Archaea are similar to bacteria when compared using a microscope and functional tests, including the types of proteins produced. However, archaea are actually more closely related to eukarya when they are compared using the ribosomal evolutionary measure and the cellular mechanisms for protein production. In this book we adopt the three domains of life taxonomy proposed by Woese: Archaea, Eukarya, and Bacteria. Table 1.1 is a brief list of examples of the three domains of life.

Table 1.1. Examples of organisms in the three domains of life: Archaea, Eukarya, and Bacteria. These three domains are defined through ribosomal RNA typing.

Archaea	*Aquifex, Archaeoglobus, Pyrococcus, Methanus coccus, Thermococcus*
Eukarya	Plants, Animals, Fungi, Protista
Bacteria	*Chlamydia, Escherichia coli, Legionella, Staphylococcus, Yersinia*

It is usual to specify an organism by just the genus and species. Examples of bacteria include *Escherichia coli (E. coli)* and *Salmonella enteritidis* serotype *typhimurium*. The serotype *typhimurium* provides additional specificity beyond strain when needed. As an example of the full taxonomy, *E. coli K12-MG1655* is an organism in the phylum Proteobacteria (or purple non-sulfur bacteria), class gamma subdivision (or gamma proteobacteria or g-proteobacteria), family Enterobacteriaceae, genus Escherichia, species coli, and strain *K12-MG1655*. New taxonomies mixed with old labels can be confusing, so sometimes the organism is referred to as

Kingdom	Bacteria
Intermediate Rank 1:	Proteobacteria
Intermediate Rank 2:	gamma subdivision
Intermediate Rank 3:	Enterobacteriaceae
Genus:	Escherichia
Species:	coli
Strain:	*K12-MG1655*

Many bacteria remain to be discovered. As a measure of the potential magnitude of what we do not yet know, *Prochlorococcus marinus* was not discovered until the 1980s. Now the *Prochlorococcus* group is speculated to be the most abundant photosynthetic organism on the planet and is responsible for a significant fraction of photosynthesis in the world's oceans including, perhaps, 50% of the chlorophyll in the subtropical Pacific Ocean.

Microscopic organisms or microorganisms have an important role in the earth's biosphere as well as a significant role in evolution. The planet's microorganisms far outnumber the population of multicellular organisms. Microorganisms include bacteria, archaea, fungi, and protozoa. As we will discuss later, viruses are not technically alive. Therefore, viruses are usually included in microbes, but not in microorganisms. Most of the biomass on earth is microbial. The range of microorganisms extends from deep in the earth's crust to the atmosphere. Humans depend on many microorganisms, but most microorganisms would not need animals to survive or to sustain a biosphere. Finally, many microorganisms remain undetected and uncultured. There are even unculturable bacteria that are known to cause human disease. Our understanding of the diversity and interdependence of life on earth will continue to grow with knowledge from all three domains—Archaea, Eukarya, and Bacteria.

1.2 Cells—Archaea, Eukarya, and Bacteria

A typical cell, like the one represented in Fig. 1.1, is 80% water by volume. Of the dry mass, 5% consists of minerals and nucleic acids and less than 20% consists of carbohydrates, and lipids (fats and fatlike substances). The reason nucleic acids receive so much attention is that the deoxyribonucleic acid (DNA) controls the production of proteins, and proteins make up 75% of the dry mass.

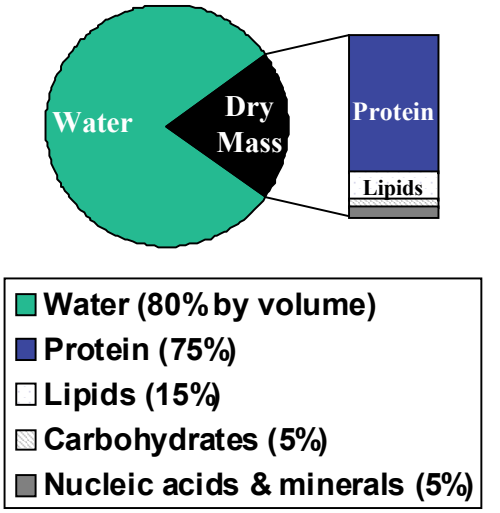


Fig. 1.1. Cells are mostly water, with the dry mass dominated by chains of amino acids known as proteins. Lipids are compounds composed mostly of carbon, hydrogen, and oxygen that are insoluble in water. Carbohydrates are of the chemical form $C_x(H_2O)_y$, i.e., hydrates of carbon.

The information contained in the DNA influences when and how cells respond to environmental conditions through the production of proteins. DNA stores the "parts list" for assembling proteins, and it dynamically interacts with proteins to regulate the timing and amount of their production. Therefore, DNA is a logical place to begin decoding how cellular function works at the molecular level.

DNA contains codes that specify the proteins to be assembled. Proteins contribute significantly to the structures and biochemical reactions within a cell. Cells maintain an internal environment by controlling the permeability of an outer layer known as the cell membrane or plasma membrane. The cell membrane controls flow of materials into and out of the cell. Plants, fungi, and bacteria often have an additional extracellular barrier known as the cell wall. Animal cells rarely have a cell wall, and a few bacteria even lack a full cell membrane.

A typical cell membrane is a lipid bilayer. The bilayer is composed of two layers of phospholipids. A phospholipid is a molecule with a molecular mass of 650 daltons made from a negatively charged phosphate group attached to two hydrophobic hydrocarbon chain tails. As shown in Fig. 1.2, the hydrocarbon tails are on the "inside" of the bilayer and the phospate group faces the hydrophilic environment outside the membrane. A typical membrane is about 5 nm thick and is impermeable to large molecules. The ability of low molecular weight molecules and water to pass through the membrane depends on traversing the hydrophobic region. Real biological membranes are not pure lipids, but include

cholesterol, protein, and other molecules mixed among the lipid "matrix." There are proteins embedded in the membrane (intrinsic proteins) and proteins that are on the surfaces (extrinsic proteins). These proteins may act to help transport molecules across the membrane or to recognize molecules at the surface (receptors).

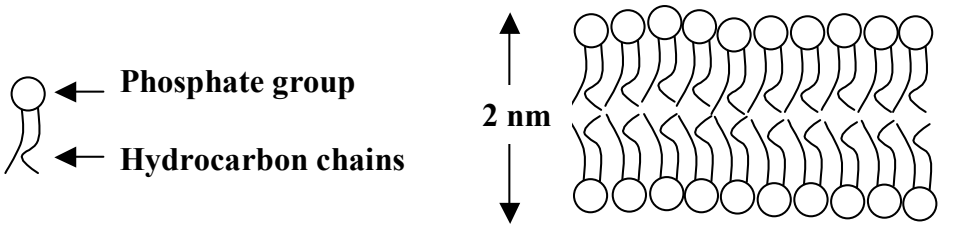


Fig. 1.2. Biological membranes are often lipid bilayers. The phospholipid building block is composed of a negatively charged phosphate group "head" and two hydrophobic hydrocarbon tails. The bilayer is about 5 nm thick and the molecular mass of the phospholipid is about 650 daltons.

Within a cell, most organisms utilize the information coded in the DNA in a very similar manner. Many cells also have common structures and subunits, including ribosomes and vacuoles. Ribosomes are involved in translating nucleic acid (genetic) information into proteins (see next section). As mentioned before, the nucleic acids that define specific ribosomes contain sufficient organism-specific information to define a taxonomy of life based on ribosomal ribonucleic acid (RNA) variation. Vacuoles are cavities in a cell's protoplasm. Cells, however, can have significantly different structures and subunits. In this section on cells we highlight some of the important similarities as well as differences among cells commonly used in biotechnology.

1.2.1 From DNA to protein

Genes produce messages like packets on an electronic network. The messages, carried on messenger RNA or mRNA, direct operations at ribosomes. The conversion of DNA in a gene to mRNA is known as transcription. Ribosomes exist in slightly different forms in eukaryotes, prokaryotes, and even cellular organelles, such as mitochondria. The ribosomes translate mRNA information into proteins (see Fig. 1.3). The parts of the DNA that contribute to the protein code are known as exons. The DNA that does not contribute directly to the protein code is known as introns. Introns can be scattered among exons within a gene as well as between genes.

Introns were thought of as noise sources, unknown signals, or messages that are no longer used. Recent research is showing that some introns do have important functions. As an example, it is now well established that some contiguous regions of DNA are alternatively spliced into the protein coding

messages. That is, intron coding helps determine dynamically which DNA segments serve as introns and exons during a specific transcription event. This is an important result in terms of the complexity of genes and their interactions. What was once labeled a gene and assumed to be a single function or single specific protein is now realized to be a dynamic program that can result in many different proteins, depending on the DNA sequence (both introns and exons) and the biochemical environment of the cell at the time of transcription.

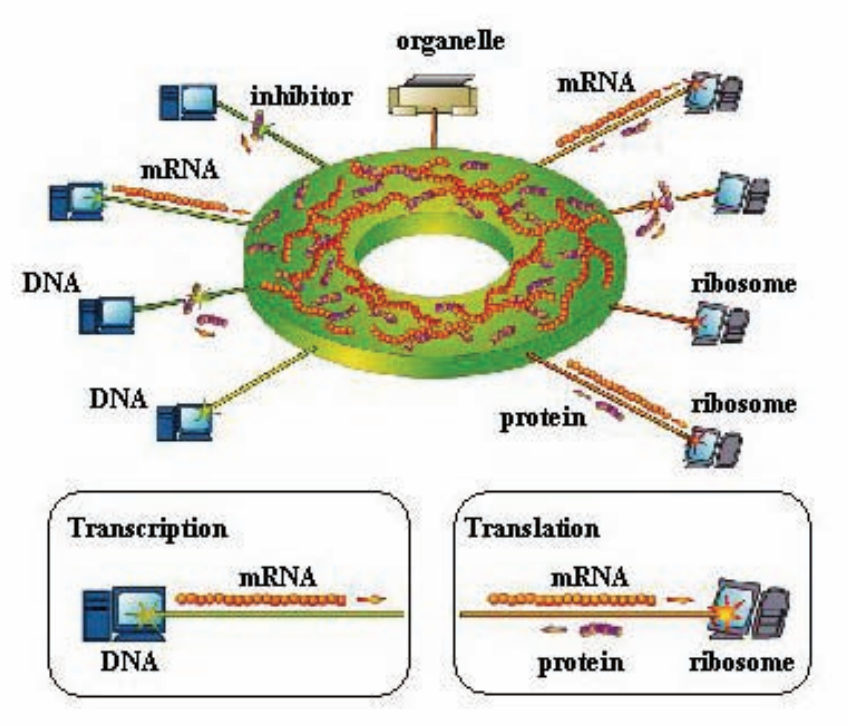


Fig. 1.3. Transcription (DNA to mRNA) and translation (mRNA to protein on the ribosome) represented as a computer network. Proteins also influence the function of the network by promoting and inhibiting the important processes of transcription and translation.

Some genes code for regulatory proteins that can inhibit or enhance the ability of other genes to put information on the "network." This may be accomplished by binding to the region of DNA that is needed to describe the protein and preventing messages from originating there. Regulatory proteins blocking the translation of specific gene messages at the ribosomes can also inhibit genes. Proteins can also be produced that bind to other proteins to inhibit or enhance function. These are only a few of the mechanisms that exist in cells to inhibit functional protein production. Regulatory proteins can also promote protein production by biochemically removing inhibitory proteins from DNA and making the gene available to the "network." Promotion of a gene is also possible

by several proteins forming a complex that creates binding sites for transcription enzymes on DNA. Spatial organization within the cell can play a regulatory role as well, since mRNA and protein have to be transported to different locations in the cell to execute their functions. The inhibitors and promoters compete and complement each other in a complex feedback system that regulates protein production.

In summary, the information stored in the DNA sequence of a gene is transcribed into a message of single-stranded RNA. The messenger RNA moves through the cytoplasm from the DNA toward a ribosome. On a ribosome in the cytoplasm, the mRNA is translated into a string of amino acids to form a protein. Transcription and translation are defined in the flow chart in Fig. 1.4. Conversion of the genetic DNA to messenger RNA is known as transcription. Translation is the process of converting mRNA into protein.

Sometimes biologists refer to the central dogma of biology. This dogma is actually the combination of two ideas. The first idea is that DNA directs the synthesis of protein (the combined steps of transcription and translation). The second idea is that DNA is able to synthesize a complementary strand (cDNA) and therefore to make accurate copies of itself. We will add details to the transcription/translation schematic as we progress. It is important to note that environmental factors contribute to the changing protein expression levels in a cell.

Eukarya, or "true kernel/nucleus" life forms, range from single-celled protozoa to multicellular organisms such as fungi, plants, and animals. A typical eukaryotic cell (Fig. 1.5) contains several subunits and membranes. The nucleus is enveloped by a nuclear membrane that is actually a pair of thin membranes known as the inner and outer membranes. The membrane pair (a bilipid) mediates passage of specific molecules, including RNA. Inside the nucleus are the DNA-rich chromosomes and nucleoli (the singular is nucleolus). Nucleoli produce ribosomal RNA and protein. Mitochondria are energy-producing organelles that contain their own RNA and DNA and can independently replicate and produce some proteins.

The 24 human chromosomes (1–22, X, and Y) range in size from the smallest chromosome 21 at 34 Mb to chromosome 1, with more than 260 Mb. Table 1.2 gives current size estimates for each of the chromosomes. The current estimate for the human genome is 3,200 Mb. Typical genome sizes for eukarya are given in Table 1.3 and range from a 12-Mb yeast to plants with genomes four times the size of the human genome.

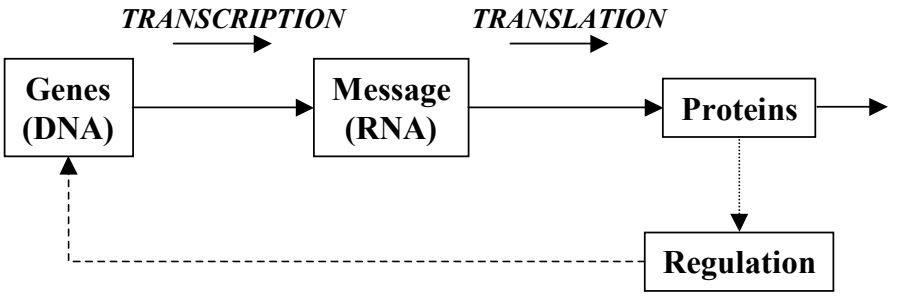


Fig. 1.4. Schematic of DNA coding through to protein synthesis. Conversion of the genetic DNA to messenger RNA (mRNA) is known as transcription. Translation is the process of converting mRNA into protein.

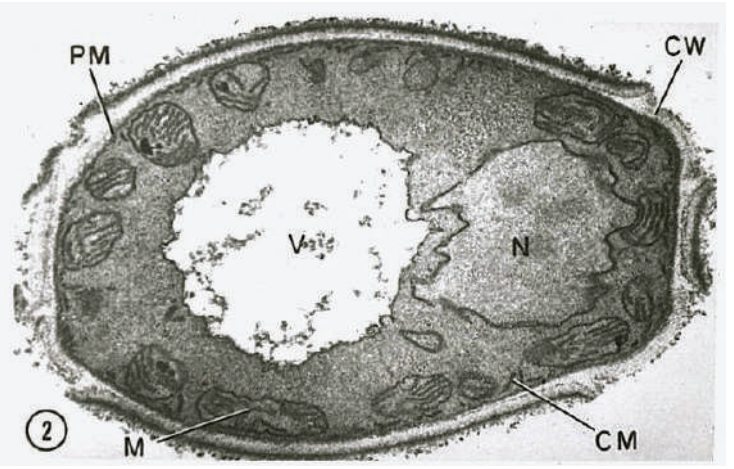


Fig. 1.5. A transmission electron micrograph (TEM) image of the eukaryotic cell *Candida* (a genus of fungi that typically grows to 2–3 μm). PM, plasma membrane; M, mitochondrion, CW, cell wall; N, nucleus; and V, vacuole. Of the 250,000 species of fungi, only about 150 have been shown to cause human disease. (Source: PHIL™296)

The bacteria and archaea share many proteins and cellular structures. The archaea have very interesting properties and a significant role in evolution. However, for this book, we focus on the bacteria. (Historically, the rickettsiae and chlamydiae were treated separately, but now each is recognized as a genus of bacteria.) Similar to Eukarya, bacteria have ribosomes. Chromosomes contain DNA required for sustaining the life cycle. Eukaryotes have linear chromosomes that are packaged with protein. Most bacteria have only a single chromosome comprised of bare rings of double-stranded DNA. Because these cells do not have a membrane-bound nucleus enclosing the chromosome, the DNA is often found near or attached to the cell membrane. Bacteria also do not have functional

organelles like mitochondria, endoplasmic reticulum, a true nucleus, or chloroplasts found in eukaryotes. Although rickettsiae and chlamydiae are both small, the smallest bacteria are called mycoplasmas. The mycoplasmas are unique because they do not have a true cell wall, but rather are bounded by a triple-layered membrane.

Table 1.2. Size in millions of base pairs (Mb) for the 24 human chromosomes.

Chromosome	(Mb)	Chromosome	(Mb)
1	263	14	93
2	255	15	89
3	214	16	98
4	203	17	92
5	194	18	85
6	183	19	67
7	171	20	72
8	155	21	34
9	145	22	34
10	144	X	164
11	144	Y	35
12	143	Total euchromatic	3175
13	98		

Table 1.3. A Sampling of eukaryotic genome sizes.

16,500 Mb	*Triticum aestivum* or wheat (A, B & D genomes)
11,400 Mb	*Avena sativa* or oat
3,200 Mb	*Homo sapiens* or humans
3,088 Mb	*Mus musculus* or mouse
430 Mb	*Gramineae Oryza sativa* or Asian rice
400 Mb	*Fugu rubripes* or puffer fish
137 Mb	*Drosophila melanogaster* or fruit fly
121 Mb	*Arabidopsis thaliana* or thale cress
97 Mb	*Caenorhabditis elegans* or nematode (worm)
25 Mb	*Plasmodium falciparum* or human malaria parasite
14 Mb	*Schizosaccharomyces pombe* or fission yeast (a fungus)
12 Mb	*Saccharomyces cerevisiae* or bakers' budding yeast

Table 1.4. A sampling of bacterial genome sizes.

6.26 Mb	*Pseudomonas aeruginosa PA01*, a major agent of secondary infections in debilitated patients
5.50 Mb	*Bacillus anthracis* or anthrax bacterium
4.65 Mb	*Yersinia pestis CO-92* or plague bacterium
4.63 Mb	*Escherichia coli K12*, found in the intestines of humans and animals and one of the most studied organisms
4.21 Mb	*Bacillus subtilis*, a bacterial fungicide
3.28 Mb	*Deinococcus radiodurans R1* or radioresistant bacterium.
1.83 Mb	*Haemophilus influenzae Rd* has humans as its only natural host and was mistakenly identified as the cause of an influenza pandemic in 1890—and therefore is inaccurately named.
1.64 Mb	*Cambylobacter jejuni*, a common foodborne pathogen
1.11 Mb	*Rickettsia prowazekii Madrid E*, the cause of typhus
1.04 Mb	*Chlamydia trachomatis*, cause of a sexually transmitted disease
0.58 Mb	*Mycoplasma genitalium*, a urogenital tract parasite

Bacterial genomes are usually efficient, with most DNA regions coding for proteins or involved in regulation. In other words, there are very few introns. Bacterial genomes range from half a million bases (0.5 Mb for *Mycoplasma genitalium*) to over 5 Mb. Table 1.4 shows a comparison of genome size for several organisms that are sequenced or drafted.

Short loops of DNA, known as plasmids, complement the chromosomal DNA in bacteria. Plasmids also occur in some eukaryotic cells. Some plasmids are specific to the bacterium and others are not an essential part of the bacterium's genomic definition. Plasmids that provide no obvious benefit to the host are called cryptic. Sometimes plasmids can be transferred into bacteria from external sources and may confer new functions on the organism. For instance, there are plasmids that confer antibiotic resistance on some bacteria. The number of instances that a particular plasmid occurs in a cell is called the plasmid copy number. A low copy number plasmid has one to four copies per cell. A high copy number plasmid may have ten to over one hundred copies in a single cell. Plasmid copy number is usually regulated to prevent harm to the cell from diverting too many resources to plasmid activities.

There are three common forms of bacteria—cocci, bacilli, and spirilla, with examples shown in Fig. 1.6. Cocci are spherical or egg shaped, bacilli are rod shaped, and spirilla are curved walled or helical strands. Examples of cocci, bacilli, and spirilla include *Streptococcus pyogenes, Bacillus anthracis*, and *Treponema pallidum*, the causative agents for "strep" throat, anthrax, and

syphilis, respectively. In each of these examples we provided the genus—e.g., *Streptococcus*, and the species—e.g., *pyogenes*.

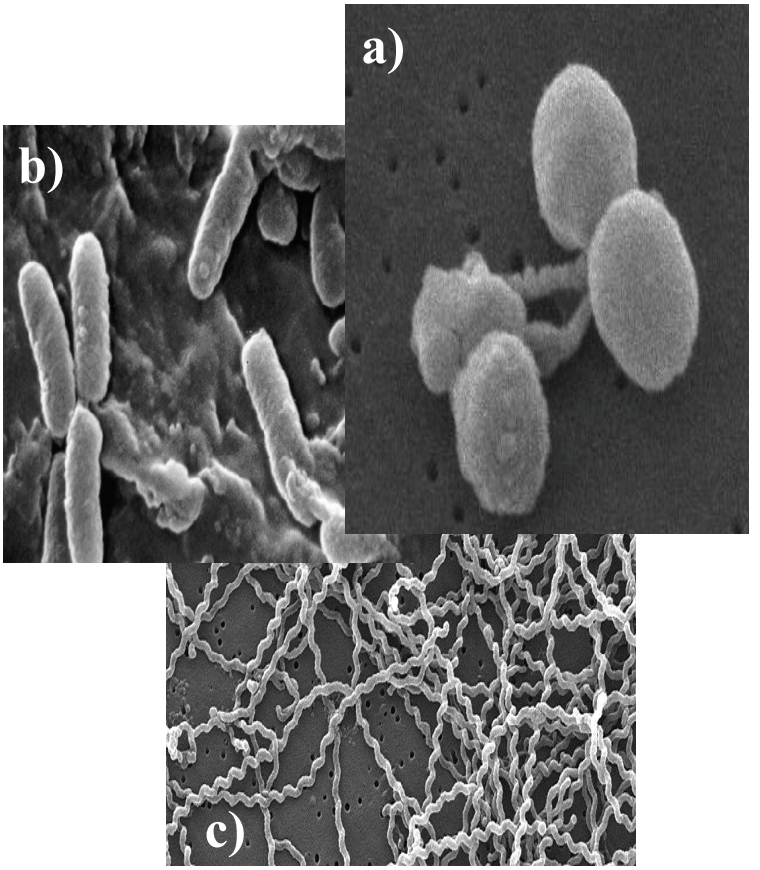


Fig. 1.6. Examples of cocci, bacilli, and spirilla bacteria. (a) *Streptococcus pneumoniae* (PHIL™ photo 265 courtesy of Richard Facklam, Centers for Disease Control), (b) *Pseudomonas aeruginosa* (PHIL™ photo 232 courtesy of Janice Carr, Centers for Disease Control), and (c) *Leptospira* (PHIL™ photo 138 courtesy of Rob Weyant, Centers for Disease Control). Bacteria range in size from 0.2 to 60 μm.

Bacterial species are defined as cell populations with similar characteristics. Definitions of plant and animal species are usually based on the ability to breed. The next level of phylogenetic classification is strain followed by "Vars," such as bioVar or seroVar for physiological or antigenic (serological) assay discrimination, respectively. The ability to ferment glycerol would be an example of physiological discrimination. Some organisms in a species may be able to ferment glycerol and other members of the same species may not. The two organisms would be placed in different bioVars based on their glycerol fermentation. When working with a specific organism, it may be necessary to use all of these designations.

Bacteria are often characterized by motility and Gram stain response. A specific bacterium is motile if it has the capacity for spontaneous movement; otherwise it is referred to as nonmotile. Motility may be enabled by the presence of a tail-like filament known as the flagellum. See Fig. 1.7 for a transmission electron micrograph (TEM) of *E. coli* showing flagella. The Gram stain is a multistep process of staining, rinsing, and counterstaining named after the Danish physician Hans Christian Joachim Gram. The Gram stain is associated with differences in cell wall structure. Organisms that retain the violet stain are called Gram-positive and usually have cell walls 10 to 15 times thicker than organisms that lose the stain and are Gram-negative. Gram-positive organisms retain a purple-black color in the cell walls and Gram-negative organisms retain a pink color. A comparison of Gram-positive and Gram-negative stains is shown in Fig. 1.8. Even after 100 years of use, the specific mechanism of the stain remains unknown. Figure 1.9 is a cartoon of the rough structural differences between the membranes of Gram-positive and Gram-negative organisms.

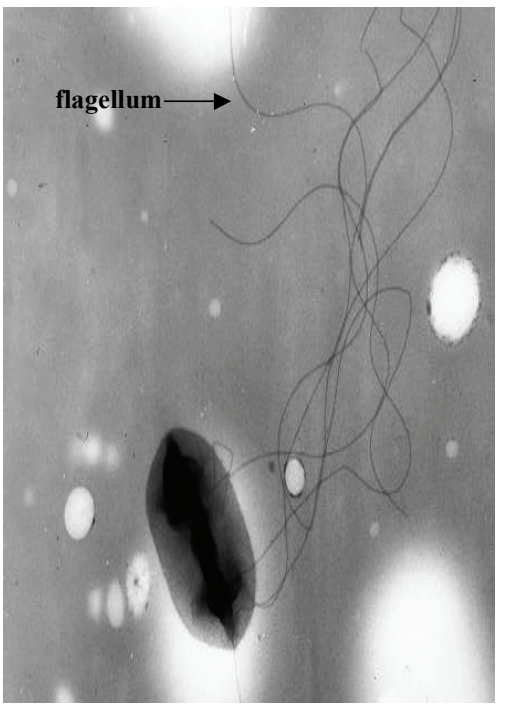


Fig. 1.7. Transmission electron micrograph (TEM) of Gram-negative, motile, *Escherichia coli* O157:H7 showing flagella. PHIL™ photo 188 courtesy of Peggy S. Hayes, Centers for Disease Control. Typical *E. coli* cells are 1 to 2 μm long and 0.5 μm in radius.

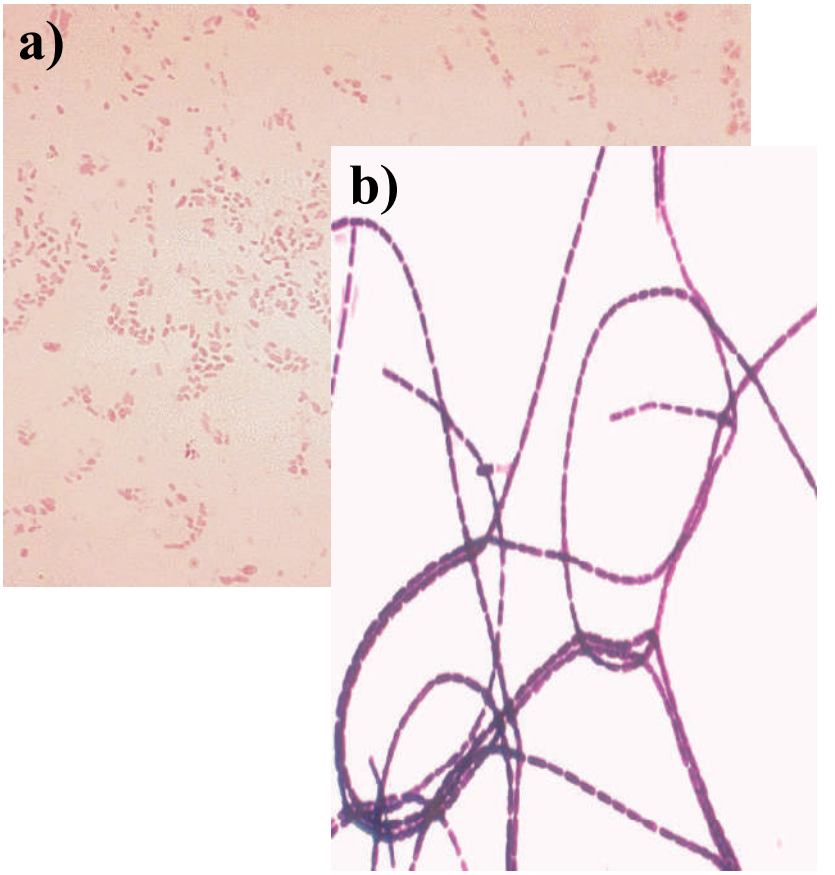


Fig. 1.8. Comparison of (a) Gram-negative stained *Agrobacterium radiobacter* (PHIL™ photo 1254 courtesy of Dr. W. A. Clark, Centers for Disease Control) and (b) Gram-positive stained *Bacillus anthracis* bacteria (PHIL™ photo 1163 courtesy of James Feeley, Centers for Disease Control). Typical *B. anthracis* rods are 1–1.5 μm by 4–10 μm. In 1877, Robert Koch showed that this spore-forming bacterium caused disease. This was the first demonstration of bacteria causing disease.

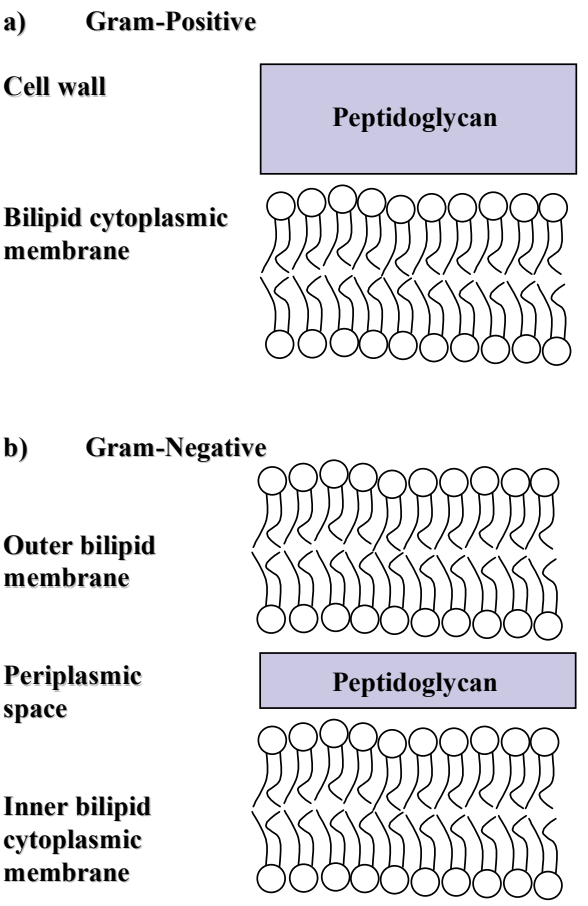


Fig. 1.9. Cartoon comparing the structural differences between (a) Gram-positive and (b) Gram-negative organisms. The Gram-positive membrane uses a thick peptidoglycan layer and other molecules as a cell wall, and the Gram-negative membrane uses a thinner peptidoglycan layer in the periplasmic space. The different attachment molecules between the peptidoglycan layers and the neighboring bilipid membrane(s) are not shown.

A bacterium that produces disease is called a pathogen. Bacteria that cause human diseases are known as human pathogens. There are also plant, animal, and other pathogens. *S. pyogenes* and *B. anthracis* bacteria are nonmotile, Gram-positive human pathogens. *T. pallidum* bacteria are motile, Gram-negative human pathogens. The nature of the disease may also be included in its name. For instance, *Legionella pneumophila* is a motile, bacilliform (rod-shaped), Gram-negative bacterium that causes lung infections, including Legionnaire disease. The genus name *Legionella* was selected because the organism was first isolated at an American Legion meeting. The species name, *pneumophila*, associates the

bacterium with the pneumonia-like symptoms. Figure 1.10 shows a thin section of lung tissue infected with *Legionella pneumophila*, including many copies of the infecting bacteria.

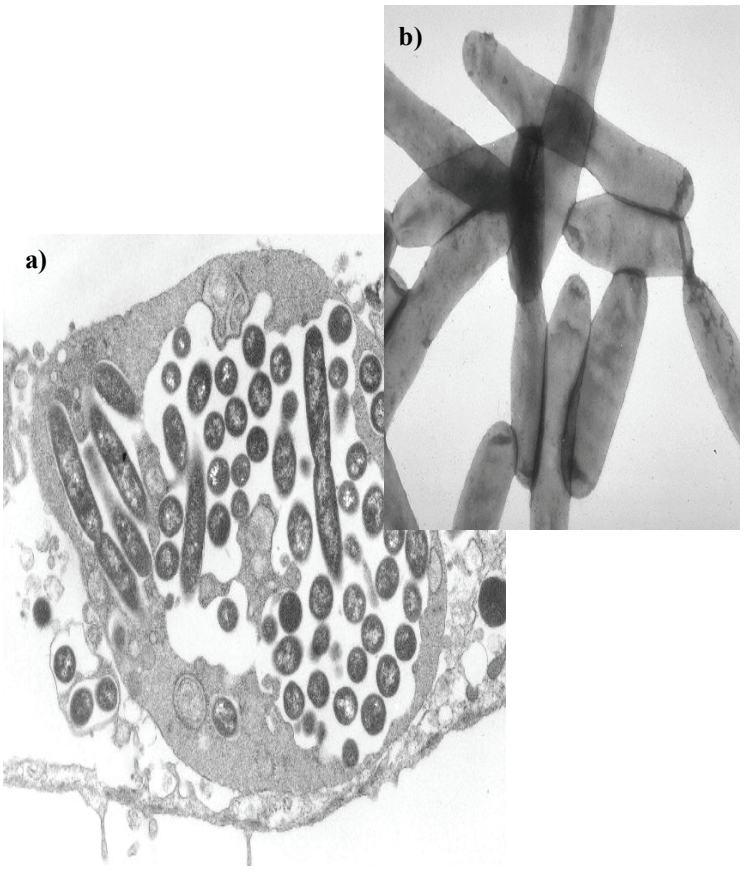


Fig. 1.10. (a) *Legionella pneumophila* multiplying inside a cultured human lung fibroblast (PHIL™ photo 934 courtesy of Edwin P. Ewing, Jr., Centers for Disease Control) and (b) expanded TEM of individual *Legionella* bacillus 0.3 to 0.9 μm in diameter by 2 to 20 μm in length (PHIL™ photo 1187 courtesy of Centers for Disease Control).

1.3 Viruses, Viroids, and Prions

By most definitions, viruses, viroids, and prions are not alive. A virus is a small (15 to 300 nm) infectious agent that is a complex combination of proteins and nucleic acids. A virus replicates only within the cells of a living host and does not self-regulate or metabolize. A viroid is a single-stranded RNA infectious agent that is smaller than a virus. Viroids lack the protein coat of viruses and do not code for specific proteins. Viroids replicate in the nucleus of higher plants, such as potatoes. A prion is a self-replicating protein that exploits a living host. Prions lack nucleic acids and cause slow infections, including scrapie and

Creutzfeld-Jakob disease. Since prions appear to violate the central dogma of biology, we will briefly describe them. Most of this section, however, focuses on viruses and their use in biotechnology.

Prions (pronounced "pree-ons") is short for "proteinaceous infectious particles" and is a term coined by Stanley Prusiner in 1982. Prion disease, often called transmissible spongiform encephalopathy (TSE), can occur through infective, inherited, and spontaneous mechanisms. All currently known prion diseases are fatal. In addition, all known prion disorders develop from a biochemical modification of a protein that is a normal constituent of all mammalian cells. While it is not completely understood, it is believed that prion disease is caused by changes in the shape of an otherwise normal protein. The self-replication attribute derives from the ability of a prion to assist or chaperone an unmodified protein into the prion shape. Understanding the mechanisms used to reshape proteins is an interesting challenge in biology that extends beyond prion applications.

A typical virus is an RNA or DNA (but not both types of nucleic acids) protected by a protein coat known as the capsid. Viruses are often categorized by their host—bacteria, plants, or animals. Other categorizations for viruses include place of discovery, mode of transmission, and manifestations of infection. For instance, Marburg virus was discovered in Marburg and Frankfurt, Germany. It produces hemorrhagic fever symptoms similar to the Ebola virus, first discovered in the Sudan, Africa. The influenza virus is well known for causing the flu and opening the immune system to secondary infections, including bacterial infections from *Haemophilus influenzae* (a bacterium once thought to be the cause of epidemic flu in humans), leading to pneumonia. Bacteriophages are viruses that infect bacteria. Many of these "phages" have a specific bacterial host target.

Upon infecting a cell, viruses may produce a lysin that dissolves or destroys cells (lysis). This process of lysin production and cell destruction is known as the lytic cycle. Alternatively, upon infection the viral nucleic acids may integrate into the DNA of the host cell and replicate with the host cell for generations. This is known as the lysogenic cycle, and bacteriophages that have the option of setting up this state of coexistence with the host are called temperate phages. The lysogenic cycle may last for thousands of cell divisions or more until a lytic cycle with infectious viral particle production and cell lysis occurs. The viral particles released are metabolically inert and called virions. As the virions infect other cells, the process begins anew.

By modifying viruses that have a prolonged lysogenic cycle, DNA can be introduced into a cell and integrated into the cell's genome—the host chromosome or extrachromosomal elements. Because viruses can transfer genetic material between different species of host, they are used extensively in genetic engineering. Viruses also carry out natural "genetic engineering." A virus may incorporate some genetic material from its host as it is replicating and transfer this information to a new host, even to a host unrelated to the previous

one. This is known as transduction, and in some cases it may serve as a means of evolutionary change.

Bacteriophages are important to biotechnology and have significant potential medical applications. Examples include the T4 phage that infects only the intestinal bacterium *E. coli* and the bacteriophage lambda (λ) used for storing and copying large DNA libraries of about 20 kilobases (kb) into *E. coli*. The ability to insert foreign DNA into a temperate phage and then use the lysogenic cycle to introduce the DNA into a host is a fundamental procedure that has enabled many discoveries in biotechnology.

Phages usually have a protein capsule head that surrounds the DNA strand that defines the phage. Sometimes a tail is appended to the capsule, with legs that are used to attach to a specific species of bacteria and inject the DNA payload into the bacterial cell. Figure 1.11 shows a T4 phage and illustrates why tailed phages are often described as tadpoles or "lunar landers." Physically, phages are 40 to 500 times smaller than their bacterial hosts. Phage genomes range in size from 20 kb to 650 kb. Once inside the bacterial cell, the phage DNA may direct production of over 100 new phages in less than half an hour. The bacterium swells and releases the phages, completing the lytic cycle. Having discussed phages as one possible mechanism for introducing foreign DNA into a host bacterium, we return to the properties of DNA and how we might exploit bacteriophages in biotechnology.

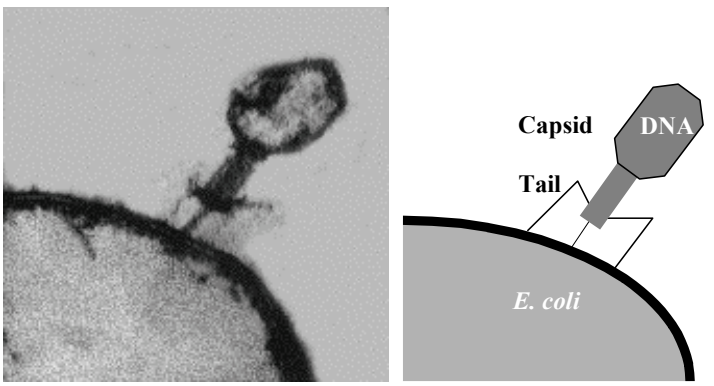


Fig. 1.11. The T4 bacteriophage infecting *E. coli* and a phage schematic. (Phage TEM from Elizabeth Kutter, Evergreen State College, Olympia, WA).

2

Nucleic Acids as the Blueprint

2.1 Genetics to Genomics

It has been more than 135 years since Gregor Mendel observed that several distinct traits of peas were inherited at statistical rates predicted by the traits of the parents. However it was not until 1944 that inherited traits and deoxyribonucleic acid were linked. DNA contains the biochemical codes for the inheritance that Mendel observed. The DNA associated with a specific trait or function is known as a gene. The entire set of information represented in the DNA is known as the genome. This combines the word "gene" with the suffix "ome" for mass.

DNA is a macromolecule built from repeating subunits (see Fig. 2.1). Each of the subunits contains one of four bases. The "size" of a genome is usually expressed as the number of base pairs (bp) of double-stranded DNA (dsDNA) in an organism. Because the base pairs of dsDNA can be generated from the bases of either of the complementary pieces of single-stranded DNA (ssDNA), the "size" of the genome may also be expressed as the number of bases (b). For instance, the human genome contains about 3 billion base pairs of DNA. Convenient units are thousands of bases (kb) and millions of bases (Mb). Ribonucleic acid is a macromolecule similar to DNA that is also measured in units of bases.

In humans, our DNA is packaged in 24 linear macromolecules of double-stranded DNA and protein known as chromosomes. The chromosomes are usually distributed spatially in the nucleus, but are often referred to as "pairs" because they physically arrange in pairs during cell division. Each parent contributes to one of the chromosomes in each pair of the child's genome. The different human chromosomes are designated by 1, 2, 3, …, 21, 22, X, and Y. Table 1.2 lists the size in DNA base pairs of the 24 human chromosomes. In normal cells, there are two copies of the numbered chromosomes and these are called the autosomes. The autosomes pair by indices—a chromosome 19 from the mother pairs with a chromosome 19 from the father. The X- and Y-chromosomes determine sex with the pairs X-to-X and X-to-Y, resulting in female and male progeny, respectively. Since the female parent can only

contribute an X chromosome, the male parent's contribution determines the gender of the progeny. Collectively, the chromosomes determine inheritance, including gender and many other traits.

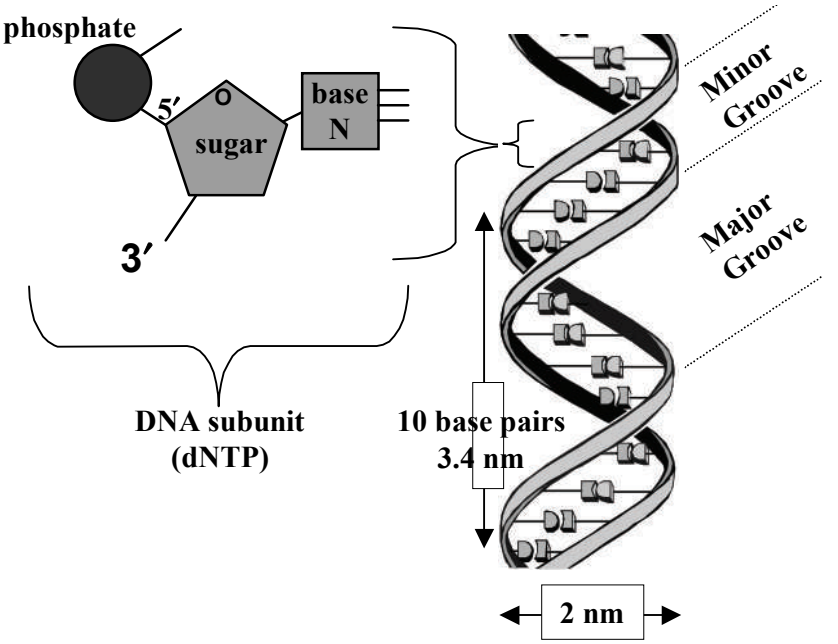


Fig. 2.1. DNA is composed of subunits called deoxyribonucleotide triphosphates (dNTPs). Each dNTP is composed of a sugar, a phospate group, and a nitrogenous base (A, C, G, or T). The bases bind in pairs (A-to-T or C-to-G) to form a double-stranded piece of DNA. Note the major and minor groove in the helix. The series of bases on a strand define the DNA sequence and provide a metric for DNA length. Each base is about 0.34 nm long and the structure of DNA repeats every 10 bases.

Figure 2.2 shows an ideogram of a chromosome. The ends of the chromosomes are called telomeres. During cell division, chromosome pairs are joined at a constricted region called the centromere. The long arm from the centromere is called the q-arm and the short arm is called the petite arm or p-arm. Bands and enumerated regions on the ideogram designate where different markers stain, bind, or cleave the chromosome.

Deviations of DNA from "normal" are known as mutations and may be inherited and/or derived from interactions with the environment. Some mutations affect health. As an example of an environmental mutation, exposure of skin to ultraviolet light can cause sunburn as well as damage to DNA that may lead to skin cancer. As an example of an inherited mutation, in the 1960s it was determined that the presence of an extra chromosome 21 causes Down syndrome. Curiously, Dr. Langdon Down first described this syndrome in 1866—the same year Mendel reported his famous observation.

A brief review of Mendelian inheritance is appropriate. Each chromosome comes from one parent. The genes on any specific chromosome may be different

than the genes on the corresponding chromosome contributed by the other parent. Genetic differences may be due to errors in the DNA sequence or the gene may code for a different trait. A trait is said to be dominant over another trait if it is the trait observed (phenotype) when genes for both traits are present in the genome. For example, Mendel observed that the trait of tallness in peas was dominant over the shortness trait. This means that Mendel's peas were tall if one or both of the chromosomes had the tallness gene. Only if both chromosomes had the gene for the recessive trait of shortness were the peas short.

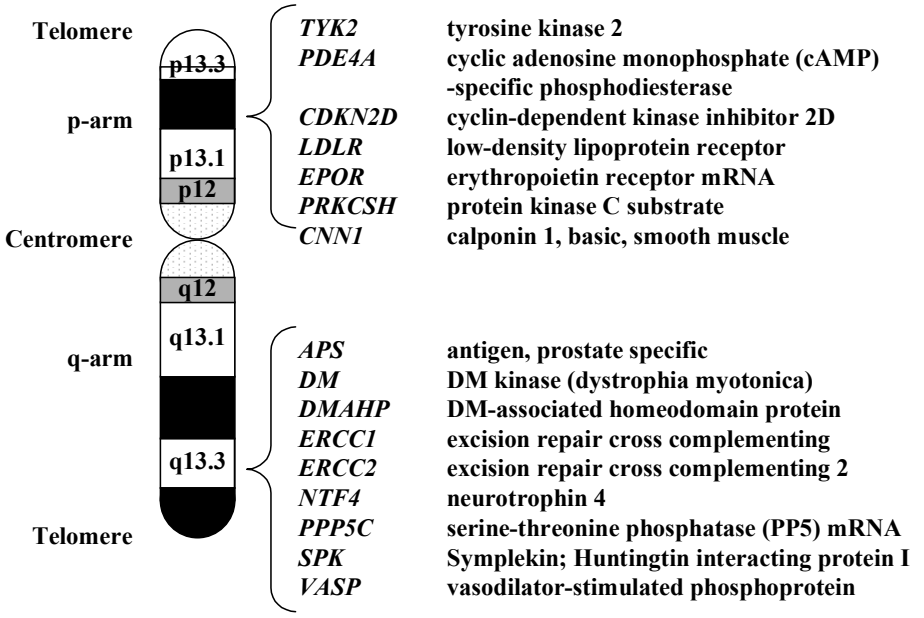


Fig. 2.2. Ideogram of human chromosome 19 with a partial list of genes for p13.2 and q13.3 on the right.

Different forms of a gene, such as tallness and shortness, are known as alleles. When the genes for a trait are the same, that individual is called homozygous for the trait. When the genes differ, the individual is called heterozygous for the trait. Because it is a dominant trait, Mendel's pea plants that are heterozygous in the height gene will be tall. Pea plants that are homozygous for the shortness gene will be short. Of course, Mendel made his observations in the context of breeding patterns and not chromosomes. An understanding of chromosomes and DNA has further enhanced Mendel's genetics.

Autosomal dominant disorders express the trait if either one or both of the chromosomes contain the disease-related genes. For an autosomal dominant disorder, the affected child must have an affected parent. With one heterozygous affected parent, the probability of affected offspring is 50%. Verify this by looking at Fig. 2.3 and noting that traits A, B, C, and D each appear in two of the four children. If the affected parent is homozygous, the disease-related genes

occur on both chromosomes (A and B, for instance) and all children have the disease. Marfan syndrome, myotonic dystrophy, Huntington disease, and familial hypercholestrolemia (FH) are examples of autosomal dominant disorders. FH has been traced to a mutation of the low-density lipoprotein receptor (LDLR) gene on region p13.2 of human chromosome 19. A partial list of 19p13.2 and 19q13.3 genes, including LDLR, is given next to the ideogram of Fig. 2.2.

An autosomal recessive disorder requires the disease-related genes on both chromosomes in a pair. Tay-Sachs disease is an example of a recessive disorder. With two heterozygous recessive parents, there is a 25% chance of a normal genotype (the disease-related gene is not on either chromosome producing a normal phenotype), a 50% chance of a heterozygous genotype (the disease-related gene is on one chromosome and not on the other chromosome in the pair producing a normal phenotype), and a 25% chance of disease (the disease-related gene is on both chromosomes in the pair producing the diseased phenotype). These numbers can also be verified using Fig. 2.3 and selecting B and D as the recessive disease genes. Note that AC is normal (one in four or 25%), AD and BC are recessive heterozygous genotypes (two in four or 50%), and BD is the disease genotype (one in four or 25%). Examples of single gene disorders that are autosomal recessive include Tay-Sachs disease, phenylketonuria (PKU), cystic fibrosis (CF), and sickle cell anemia.

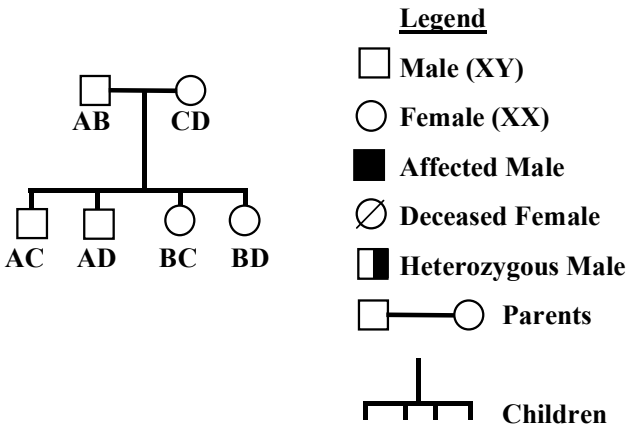


Fig. 2.3. Graphic description of mendelian inheritance of genetic traits A, B, C, and D in progeny. Using the different designations, the reader might develop a family tree for an individual with an autosomal recessive disorder. How might grandparents and siblings appear on the tree?

Mendel's approach, now called classic genetics, predicts many inheritance rates accurately. Some caution is needed, however. The contribution from one parent does not have to come entirely from one strand of the parent's chromosome pair—pieces from both chromosomes in a pair can recombine into a new chromosome for the child. Therefore, the chromosome 19 pair in a child is

not an exact copy of one chromosome 19 from each parent. The two 19 chromosomes in each parent can recombine into a different chromosome that ultimately becomes one of the strands in the child. This is why children have genes on each chromosome pair from all four grandparents.

It is estimated that there are about one trillion cells in each person. Developmental processes differentiate the single-celled fertilized egg into a complex network of organs and tissues that work together. Even though every cell in a human shares the same genetic code that originated in the fertilized egg, the shapes and functions of cells may be very different. For instance, nerve cells and white blood cells have radically different shapes and functions. Monozygotic identical twins arise from the same fertilized egg and share the same genetic code. Despite sharing many physical characteristics, however, monozygotic twins are not truly identical. For example, twins have similar but noticeably different fingerprints. More generally, events will shape twins differently. One twin may get an infection that results in an immune disorder such as multiple sclerosis. The other twin may eat carcinogens in grilled meat and develop cancer. Knowing the genetic code tells us what *might* happen, but it does not tell us what *will* happen.

DNA is responsible for far more than passing static information on inherited traits from parents to children. The DNA has a significant role in the biochemical dynamics of every cell. The DNA contains the parts list and assembly instructions for cell activities that include metabolism, growth, and reproduction. Every cell in a human has the same DNA sequence in its chromosomes. Even cells with very different structures and functions, such as brain and liver cells, have the same DNA sequence. Developmental processes differentiate the cells, changing which genes are "on" and which are "off". For bacteria that cause human disease, a few different genes in the bacterial DNA can determine if the organism causes sickness rather than death. The underlying motivation for biology is improving human health through a better understanding of cellular and subcellular mechanisms.

2.2 DNA

Deoxyribonucleic acid is a macromolecule built from repeating subunits. The subunits are composed of a nitrogenous base, a sugar, and a phosphate group generically denoted dNTP for deoxyribonucleotide triphosphate. The nitrogenous base is one of adenine (A), cytosine (C), guanine (G), or thymine (T), with the associated deoxynucleotides denoted dATP, dCTP, dGTP, and dTTP. The dNTPs can be joined along a sugar-phosphate backbone to form a single strand of DNA, with the bases occurring in any order.

The "N" in dNTP is used as a placeholder for any nucleotide—A, C, G, or T. "N" is the most common placeholder. There are several other shorthand designations used for patterns of bases, including

N = A or C or G or T,
S = A or T,
W = C or G,
B = Not A = C or G or T,
D = Not C = A or G or T,
H = Not G = A or C or T, and
V = Not T = A or C or G.

The dNTPs and the strand have an orientation based on the orientation of the five carbon atoms in the sugar. One end of the strand is designated five prime and the other three prime, 5′ and 3′, respectively. The list of bases in a strand of DNA is known as the DNA sequence and might appear as

5′-CGCGCTCCCTGAACC-3′.

Single-stranded DNA is somewhat fragile and DNA usually occurs as a double strand, with each nitrogenous base attached via hydrogen bonds to a complementary base on the opposite strand. The base pairs in double-stranded DNA must occur as A-to-T or C-to-G. The strands are also antiparallel—e.g., the strand

3′-GCGCGAGGGACTTGG-5′,

is the complement of the earlier example. The two strands tend to twist into the familiar double helix shape associated with DNA and shown in Fig. 2.4. The helical structure repeats every 10 base pairs (roughly 3.4 nm) and is roughly 2 nm in diameter.

The C-to-G attraction is from three hydrogen bonds. The A-to-T attraction is from two hydrogen bonds. The distribution of bases in double-stranded DNA is 50% small pyrimidine (C and T) bases and 50% large purine (A and G) bases. The pyrimidine bases have a single chemical ring structure and the larger purine bases have a double-ring chemical structure. The 50/50 distribution of purine/pyrimidine is due to the complementary binding of bases. Because of the biochemical similarity, the pyrimidine and purine bases also have shorthand designations for patterns of bases:

R = A or G (purine) and
Y = C or T (pyrimidine).

In humans, each DNA molecule folds among various proteins (mostly histones) into a compact package called a chromosome. Histone proteins are typically very rich in the two amino acids arginine and lysine. Different classifications of histones are described by differing concentrations of arginine and lysine. Chromosomes consist of roughly equal portions of protein and DNA, with several hundred base pairs of DNA wrapped around eight histone molecules

and their linkers. This DNA/eight-histone unit is referred to as a nucleosome and has a diameter of about 30 nm (see Fig. 2.5). Chromosomes are made of repeating nucleosome subunits.

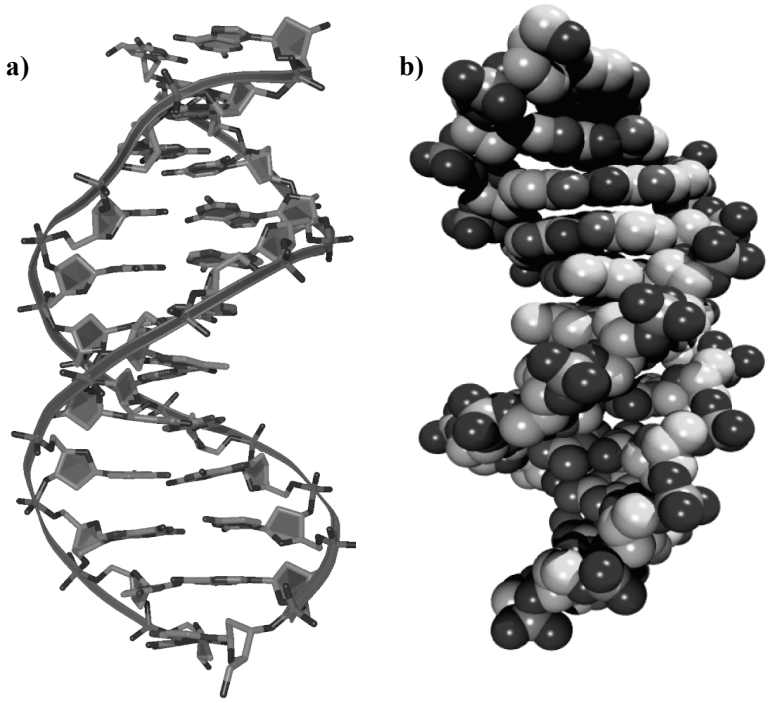


Fig. 2.4. Familiar double helix structure of DNA or twisted ladder with a sugar-phosphate backbone and nitrogenous base rungs. The a) "licorice and ribbons" and b) space-filled views of the Drew-Dickerson dodecamer, the first high-resolution measured crystal structure of B-DNA. B-DNA is the dominant form of DNA under physiological conditions. [Photo courtesy of Daniel Barsky, Lawrence Livermore National Laboratory, using the VMD program (Humphrey) and with Rayshade 4.0 (Kolb and Bogart) and Raster3D (Merritt and Bacon).]

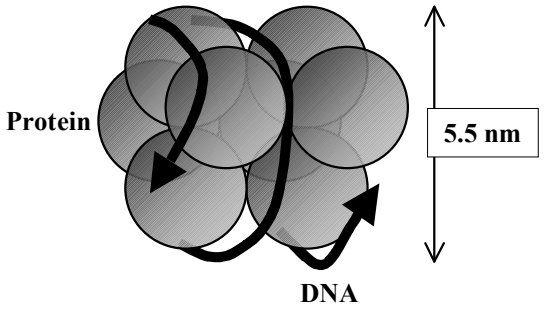


Fig. 2.5. The DNA-histone package that makes up chromosomes is called a nucleosome. Chromosomes are about 50% protein (usually histone) and 50% DNA.

2.3 RNA

As with DNA, RNA is a macromolecule built from repeating subunits. Deoxyribonucleic acid has one less oxygen atom (deoxy) in the ribose sugar than Ribonucleic Acid (RNA). Specifically, RNA has a hydroxyl (OH) attached to the 2′ carbon on the ribose sugar and DNA has a hydrogen atom (H) attached at that site on the sugar. In RNA, the pyrimidine base uracil (U) replaces thymine. As with the pyrimidine thymine, uracil complements adenine (A). These subunits are comprised of a nitrogenous base, a ribose sugar, and a phosphate group generically denoted NTP (note dNTP was used for DNA) for ribonucleotide triphospate. The NTPs are ATP, CTP, GTP, and UTP.

Ribonucleic acid usually occurs as a single strand. The single-stranded structure is less stable than the double-stranded DNA structure. For analysis and manipulation, RNA is often copied into a complementary strand of DNA that can be paired into a double strand.

2.4 How DNA Codes for Protein

2.4.1 Transcription

Transcription is the process of creating an mRNA strand that contains genetic information from the DNA. DNA information is transcribed to a single-stranded RNA messenger that delivers the genetic information to a ribosome. This is illustrated in Fig. 2.6 for prokaryotic cells. First, proteins known as helicases unwind a portion of double-stranded DNA. The entire chromosome is not unwound—the hydrogen bonds connecting base pairs are broken locally. The region of interest on the DNA is often preceded by a promoter section of DNA. Specific proteins bind to the promoter region and "promote" attachment of an enzyme known as RNA polymerase. Synthesis of a complementary RNA strand begins near the promoter position and moves along the DNA strand from the 3′ to the 5′ end. Because the RNA is complementary to the DNA, it is synthesized from the 5′ to the 3′ end. In Fig. 2.6 and as a convention for this book, we will represent double-stranded DNA with the top strand going from 5′ on the left to 3′ on the right. Therefore the RNA polymerase is attached to the lower strand and is moving from left to right (3′ to 5′ on the DNA). There are free nucleotides available in solution as ATP, CTP, GTP, and UTP and these polymerize to their complements catalyzed by the RNA polymerase.

As shown in Fig. 2.6, if 5′-AAT-3′ is the top strand, then 3′-TTA-5′ is the bottom strand and the RNA should be 5′-AAU-3′. Recall that uracil in RNA replaces thymine in DNA. This is the mRNA that communicates the genetic information to the ribosomes and it is the same as the "sense" 5′-to-3′ strand (the top strand in the figure with 5′ at the left side) with uracil replacing thymine. The "nonsense" or "antisense" strand is the bottom strand in the figure and is also

known as the template strand because the polymerase enzyme actually moves along that strand (from the 3′ to the 5′ end).

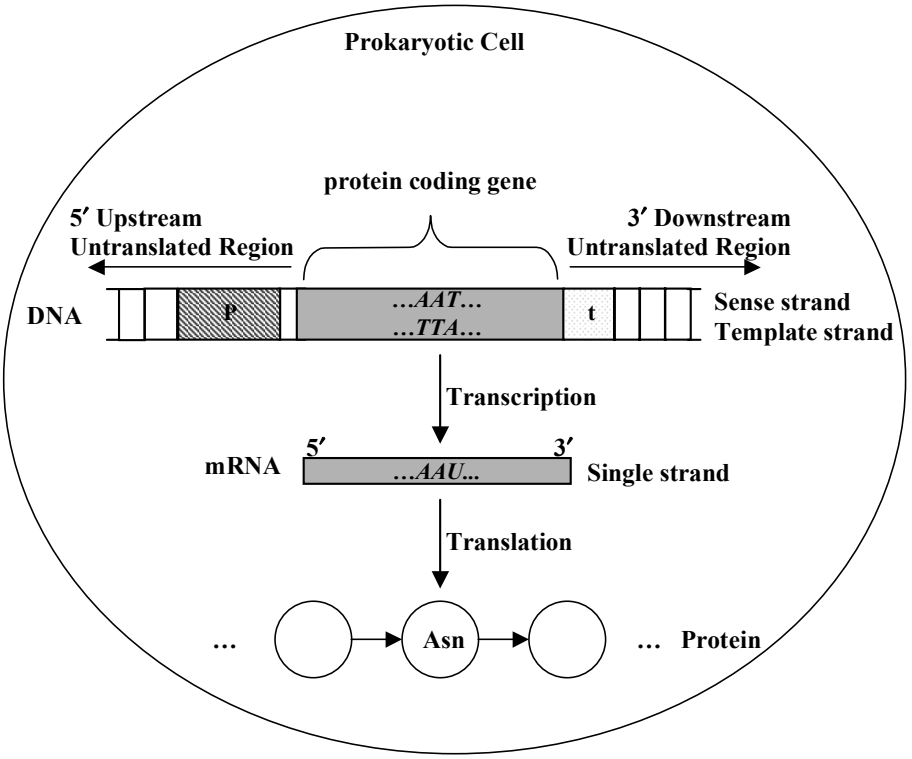
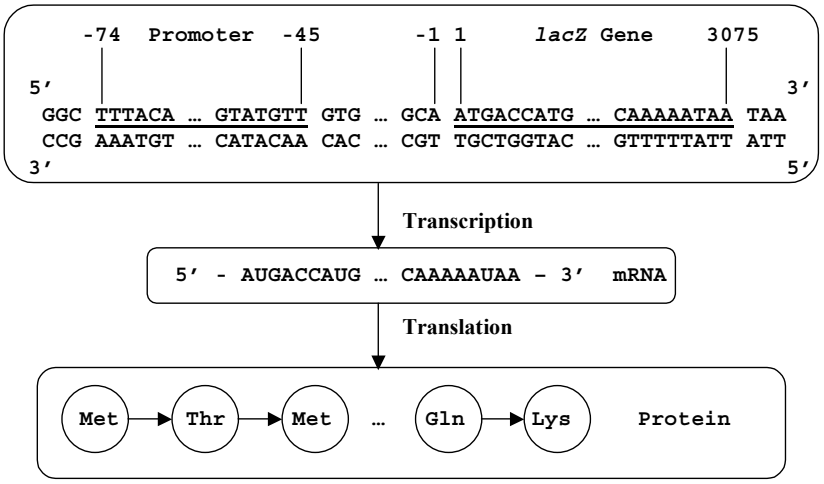


Fig. 2.6. Transcription of a gene requires the conversion of DNA into mRNA by an enzyme that attaches to the template strand near a promoter (P) and moves toward the terminator (t). Groups of three bases (codons) in the mRNA code for specific amino acids that are assembled into chains of amino acids on ribosomes. The process of mRNA directing the formation of amino acid chains (proteins) is known as translation. The codon AAU codes for the amino acid asparagines (Asn).

Some other terms are easily defined with Fig. 2.6. The top strand region to the left is known as the 5′ upstream. The region on the right is the 3′ downstream. By orienting the strands with the RNA polymerase running left to right, nucleic acids of interest (DNA and mRNA) usually run 5′ to 3′ left to right. It is common to only list the top 5′ to 3′ strand since the second strand can be generated by complementing the bases. The sections of DNA that code mRNA are known as protein coding regions. Transcription begins near a promoter site and ends at a terminator site. In prokaryotes, genes tend to be represented by contiguous bases in the DNA, with promoter and other transcription factors nearby. There are short three base DNA sequences that code for start (ATG) and stop (TGA, TAG, TAA). Unfortunately, other factors also mediate this process, making the start and stop codes useful markers but not sufficient to identify genetic regions.

A specific example of transcription in prokaryotic cells is given in Fig. 2.7. The 3,075-base *lacZ* gene in *E. coli* is shown to transcribe into a 1,023-amino acid protein. The 30-base promoter region includes two binding sites (TATGTT and TTTACA). Some of the data available at the National Center for Biotechnology Information (NCBI) online at http://www.ncbi.nlm.nih.gov/ are also listed for the DNA sequence of the promoter region and *lacZ* gene. Note that the NCBI data are for the opposite strand of the DNA—it was labeled online as the "complement." In order to arrange the DNA sequence in Fig. 2.7(a) consistent with the sense strand on top going left to right from 5′ to 3′, the NCBI data must be reversed and complemented. Simply reversing the NCBI data provides the bottom or template strand for the *lacZ* gene.



(a)

```
5581 AGTACATAAT GGATTTCCTT ACGCGAAATA CGGGCAGACA TGGCCTGCCC GGTTATTATT
5641 ATTTTTGACA CCAGACCAAC TGGTAATGGT AGCGACCGGC GCTCAGCTGG AATTCCGCCG
     …
8641 GTTGGGTAAC GCCAGGGTTT TCCCAGTCAC GACGTTGTAA AACGACGGCC AGTGAATCCG
8701 TAATCATGGT CATAGCTGTT TCCTGTGTGA AATTGTTATC CGCTCACAAT TCCACACAAC
8761 ATACGAGCCG GAAGCATAAA GTGTAAAGCC TGGGGTGCCT AATGAGTGAG CTAACTCACA
```

(b)

Fig. 2.7. Transcription and translation demonstrated using genes in the lactose metabolism operon of *E. coli*. The sequence data for this example, partially repeated in (b), are available online at the National Center for Biotechnology Information (NCBI) web site. (a) Specific example of transcription and translation for the *lacZ* gene related to lactose metabolism in the bacterium *E. coli*. The transcribed strand (sense strand) is shown on top with 5′ at the left end. (b) Data from NCBI AE000141 *E.* coli K12 MG1655 for the *lacZ* gene (complement of the 3075 bases 8713 to 5639) and its promoter (complement of the 30 bases 8787 to 8758). There are 44 bases between the promoter and the start codon.

Transcription in eukaryotic cells is more complex and is outlined in Fig. 2.8. In eukaryotic DNA, noncoding regions (introns) interrupt protein-coding regions (exons). Introns range in size from 40 to 10,000 bases. Eukarya have very complex promoter logic, often requiring multiple sites and multiple proteins to promote transcription. Sometimes large protein complexes span several sites on the DNA. Collectively these regulatory proteins are referred to as transcription factors. A first transcript or principal transcript of the DNA strand is made that includes RNA that complements both the exons and the introns. In addition to the bases from the DNA template, there are also bases appended to the ends of the principal transcript. At the 5′ end, a G base is appended and is known as the guanine cap. At the 3′ end, a string of up to 200 adenine bases is appended and is known as the poly(A) tail or polyadenylation. A second RNA strand known as the functional transcript is made by splicing exons together between the G-cap and the poly(A) tail. Occasionally some exons are omitted during splicing and an alternative protein is coded in the functional transcript. The functional transcript then migrates outward from the nucleus toward ribosomes in the cytoplasm.

The information stored in the DNA sequence of a gene is transcribed into a message of single-stranded RNA. On a ribosome in the cytoplasm, the mRNA is translated one codon (three bases) at a time into one of 20 amino acids that are also sometimes referred to as residues. The amino acids designated by the codons are chained together until a full-length protein is formed (see Fig. 2.9). Peptides are short chains of amino acids less than 40 residues long. Most functional proteins are longer than 40 residues. The amino acids have a modular structure built around a central carbon (Cα) flanked by a hydrogen (H) atom, an amino group (NH_2), a carboxyl group (COOH), and a side chain (R) that defines the specific amino acid of the 20 possible. Note that "R" in this case is used for "residue" and not to denote the specific amino acid arginine. Figure 2.10 is a chemical schematic of an amino acid structure. The amino acids are joined together by peptide bonds where the carboxyl group gives up an OH and the amino group donates an H. The bond between the carbon and the nitrogen atoms of the carboxyl and amino groups is known as a peptide bond. As with DNA, amino acid chains can be oriented using the N-terminal group (NH_2) or the C-terminal group (COOH) of the peptide backbone. Proteins are the workhorses of the cell, performing chemical (enzymatic) or structural functions. DNA and the environment control the quantity, timing, and selection of proteins expressed.

The ribosome uses the mRNA and another type of RNA called transfer RNA (tRNA) to construct proteins. As shown in Table 2.1, there are 64 (four-cubed; three-base sets pulled from four possible bases A, C, G, and T) possible codons that redundantly code for the 20 amino acids. An amino acid is attached to the 3′ end of a charged single strand of RNA (the tRNA) with a complementary codon (anticodon) available to bind to the RNA. The pairing of the mRNA with the appropriate series of tRNA collects amino acids on the ribosome so that the formation of peptide bonds can produce a protein.

In many cells there are processes that interfere with the production of protein. In some cases the mRNA is consumed before translation. In eukaryotic cells, the

mRNA may not successfully leave the nucleus to reach a ribosome. It is also possible to have very efficient DNA-to-mRNA to protein processes. Under some conditions, bacterial cells will have translation occurring at a ribosome on an mRNA that has not yet been fully transcribed! Cellular differences, biochemistry, and other factors affect translation efficiency. Because of these dependencies, an increase in mRNA transcription of a particular gene does not guarantee an increase in protein expression in the cell.

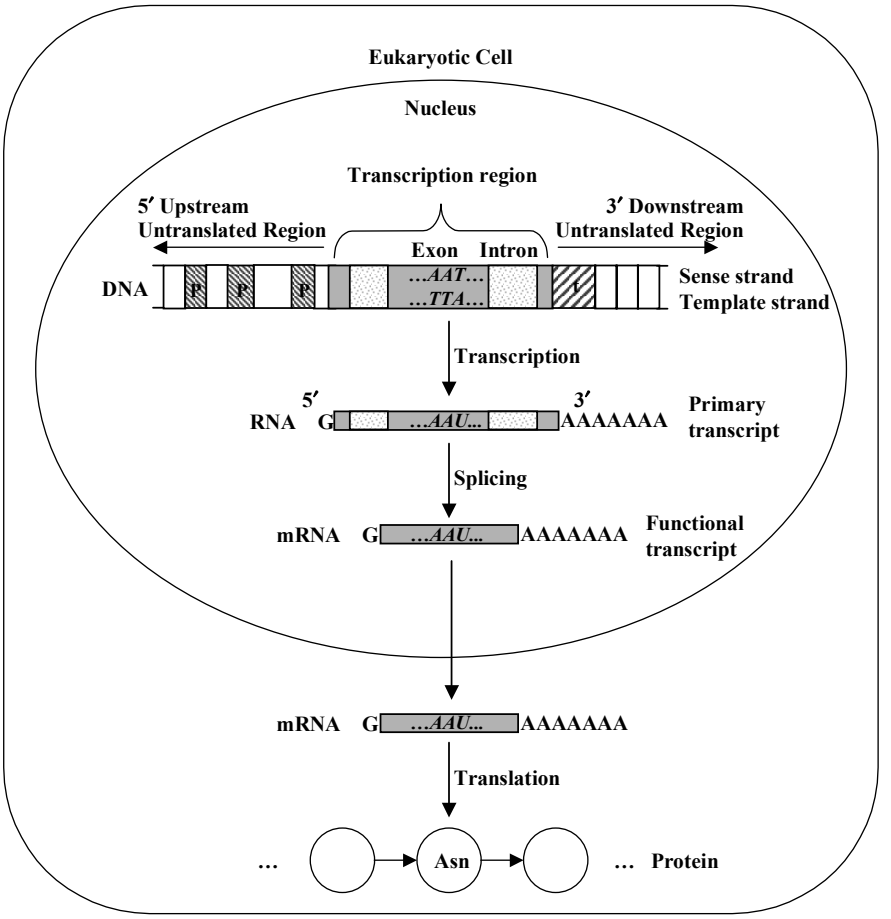


Fig. 2.8. Transcription of a gene in a eukaryotic organism requires the conversion of DNA into a primary RNA transcript as well as the splicing of exons (removal of introns) into the mRNA. The mRNA leaves the nucleus and is translated into protein on a ribosome. Promoters (P) and enhancers can be distributed in the untranslated regions near the gene. Eukaryotic organisms often have a regulatory DNA sequence in more locations and spread over more bases than prokaryotic cells.

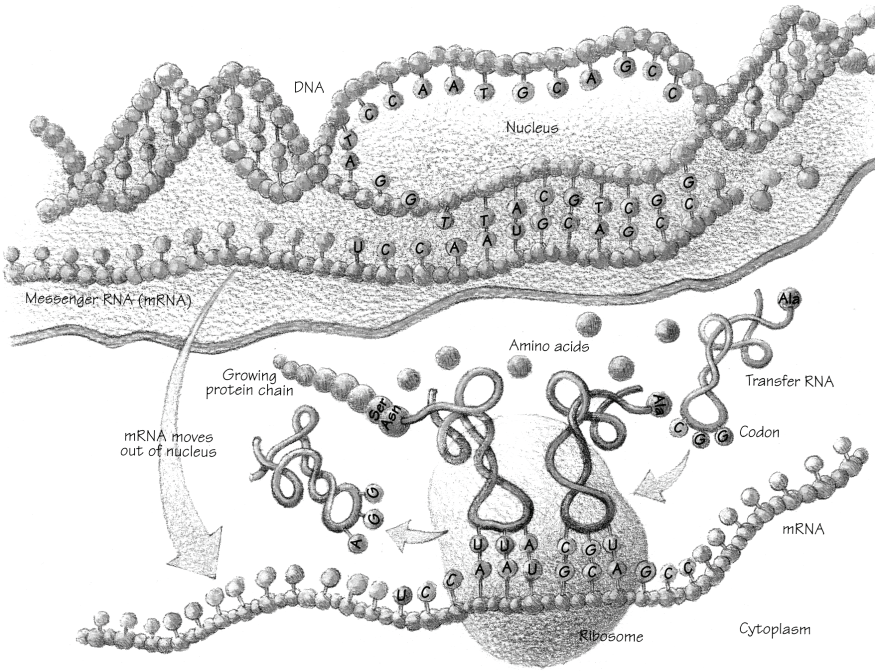
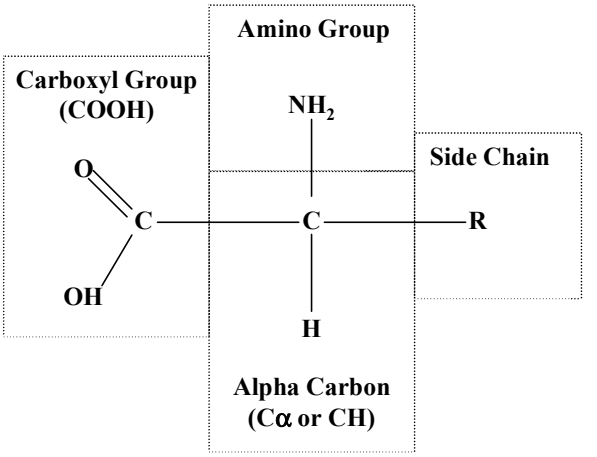


Fig. 2.9. Transcription of DNA into mRNA in the nucleus followed by translation of mRNA at the ribosome into a growing amino acid chain to form a protein. (From *To Know Ourselves*, U.S. Dept. of Energy Rept. PUB-773, July 1996 online at www.lbl.gov/Publications/TKO.)

As an example of working from DNA to protein, consider the double-stranded eukaryotic DNA shown in Fig. 2.11. The top strand is 5′ to 3′ from left to right. RNA is spliced into a primary transcript (b); coding regions (exons) are spliced together as the mRNA is formed (c); and translation of the mRNA into a protein completes the process (d).

Once translation is completed and the protein is away from the ribosome, the amino acid chain assumes a three-dimensional (3-D) shape. Protein structure and function are closely related. There are several levels for describing protein structure. The primary structure is the order of amino acids in the protein. By convention, the amino acids are listed from the amino end of the protein (N-terminal) to the carboxyl end (C-terminal). Recall that the amino acids are joined by peptide bonds. These bonds form with the removal of water (condensation or dehydration synthesis reaction). Protein secondary structures are common repeating structures found in many proteins known as the alpha helix and the beta-sheet. Alpha helices are the most common of the two and occur when hydrogen bonds form between the CO of one amino acid and the NH group of another amino acid four residues away. Beta-sheets or beta-pleated sheets are the other type of secondary structure. A tertiary protein structure is the full three-dimensional structure of the amino acid chain.

a) Generic amino acid chemical structure with the side chain R specifying the 20 amino acids.



b) Peptide bond between two amino acids with side chains R1 and R2.

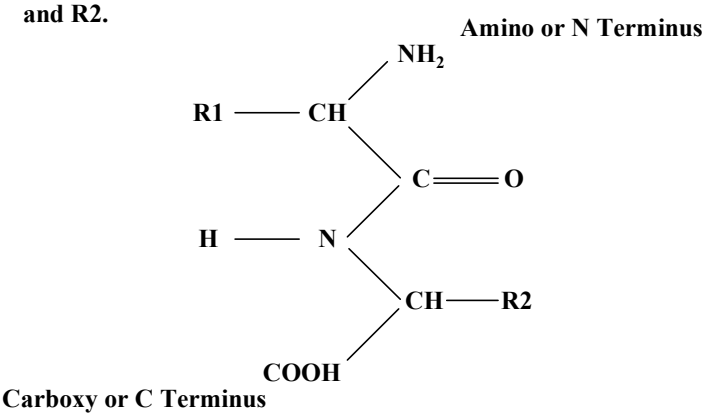


Fig. 2.10. (a) Chemical schematic of an amino acid showing the carboxyl group, amino group, hydrogen atom, alpha carbon, and residue R. (b) Chain of amino acids connected via peptide bonds.

Table 2.1. The 64 three-base codons (5′ to 3′ DNA) with the corresponding mRNA (5′ to 3′ mRNA) and the 20 amino acids. The single letter abbreviations are only used in long lists. Note that the DNA corresponds to the 5′ to 3′ gene and so the mRNA bases are identical with the T-to-U substitution. The mRNA is the complement of the 3′ to 5′ DNA strand that participates in mRNA transcription. ATG codes for the amino acid methionine and also serves as the start codon.

DNA	mRNA	Amino Acid	Abbrev.	DNA	mRNA	Amino Acid	Abbrev.
AAA	AAA	Lysine	K-Lys	GAA	GAA	Glutamic Acid	E-Glu
AAC	AAC	Asparagine	N-Asn	GAC	GAC	Aspartic Acid	D-Asp
AAG	AAG	Lysine	K-Lys	GAG	GAG	Glutamic Acid	E-Glu
AAT	AAU	Asparagine	N-Asn	GAT	GAU	Aspartic Acid	D-Asp
ACA	ACA	Threonine	T-Thr	GCA	GCA	Alanine	A-Ala
ACC	ACC	Threonine	T-Thr	GCC	GCC	Alanine	A-Ala
ACG	ACG	Threonine	T-Thr	GCG	GCG	Alanine	A-Ala
ACT	ACU	Threonine	T-Thr	GCT	GCU	Alanine	A-Ala
AGA	AGA	Arginine	R-Arg	GGA	GGA	Glycine	G-Gly
AGC	AGC	Serine	S-Ser	GGC	GGC	Glycine	G-Gly
AGG	AGG	Arginine	R-Arg	GGG	GGG	Glycine	G-Gly
AGT	AGU	Serine	S-Ser	GGT	GGU	Glycine	G-Gly
ATA	AUA	Isoleucine	I-Ile	GTA	GUA	Valine	V-Val
ATC	AUC	Isoleucine	I-Ile	GTC	GUC	Valine	V-Val
ATG	AUG	Methionine	M-Met	GTG	GUG	Valine	V-Val
ATT	AUU	Isoleucine	I-Ile	GTT	GUU	Valine	V-Val
CAA	CAA	Glutamine	Q-Gln	TAA	UAA	Stop Codon	
CAC	CAC	Histidine	H-His	TAC	UAC	Tyrosine	Y-Tyr
CAG	CAG	Glutamine	Q-Gln	TAG	UAG	Stop Codon	
CAT	CAU	Histidine	H-His	TAT	UAU	Tyrosine	Y-Tyr
CCA	CCA	Proline	P-Pro	TCA	UCA	Serine	S-Ser
CCC	CCC	Proline	P-Pro	TCC	UCC	Serine	
CCG	CCG	Proline	P-Pro	TCG	UCG	Serine	S-Ser
CCT	CCU	Proline	P-Pro	TCT	UCU	Serine	S-Ser
CGA	CGA	Arginine	R-Arg	TGA	UGA	Stop Codon	
CGC	CGC	Arginine	R-Arg	TGC	UGC	Cysteine	C-Cys
CGG	CGG	Arginine	R-Arg	TGG	UGG	Tryptophan	W-Trp
CGT	CGU	Arginine	R-Arg	TGT	UGU	Cysteine	C-Cys
CTA	CUA	Leucine	L-Leu	TTA	UUA	Leucine	L-Leu
CTC	CUC	Leucine	L-Leu	TTC	UUC	Phenylalanine	F-Phe
CTG	CUG	Leucine	L-Leu	TTG	UUG	Leucine	L-Leu
CTT	CUU	Leucine	L-Leu	TTT	UUU	Phenylalanine	F-Phe

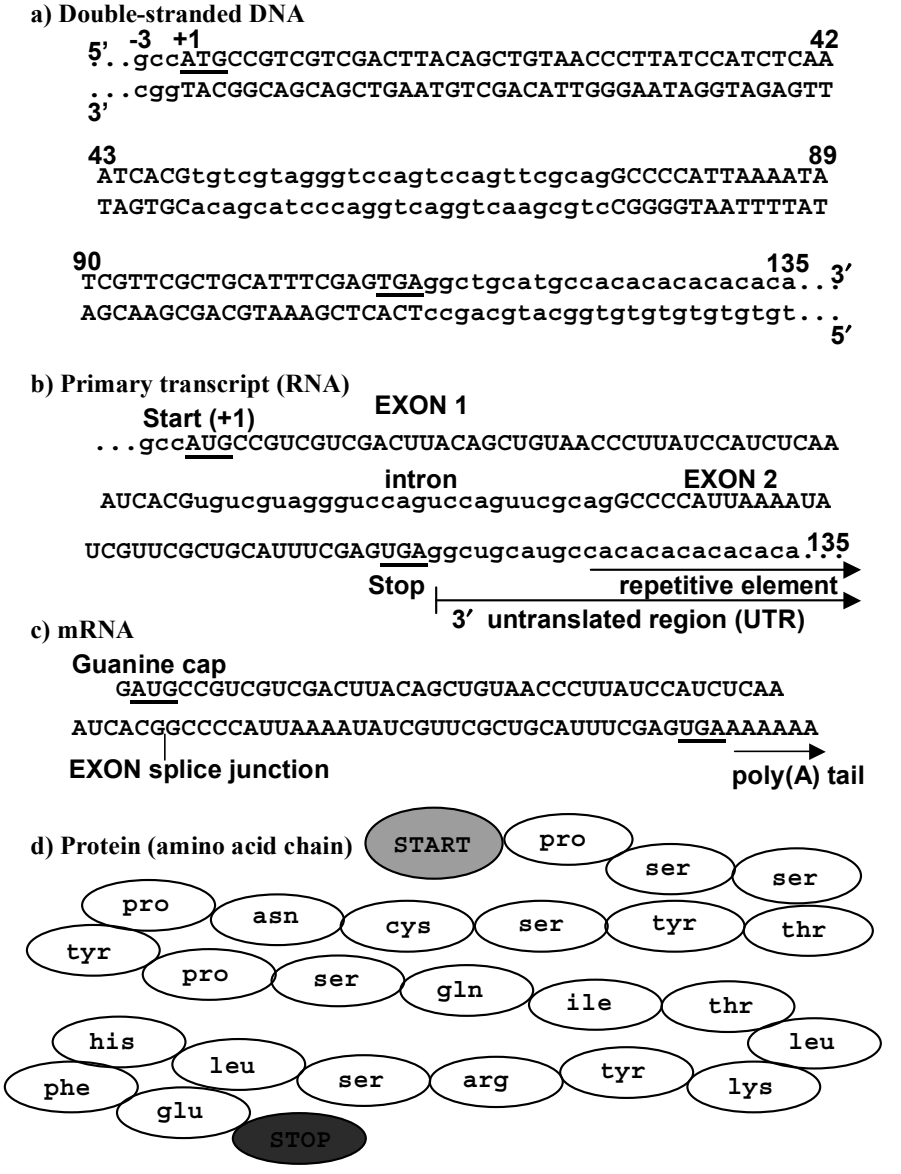


Fig. 2.11. (a) Double-stranded DNA, (b) conversion to primary transcript, (c) mRNA, and (d) amino acid chain. Exons are given in capital letters.

Changes in tertiary structure distinguish prions from their "normal" protein counterpart. Quaternary structure is the interconnection of multiple amino acid chains. Many proteins are made of a single polypeptide and do not have a quaternary structure. Several important proteins, however, are composed of multiple polypeptide chains, including hemoglobin with four polypeptides—two alpha-globins and two beta-globins.

The protein is often modified after translation. In eukaryotes this often includes cleavage of the end amino acid of the protein or cleavage at particular peptide bonds to form a set of smaller proteins. There may also be modification of the protein by addition of phosphate groups or carbohydrates. These "post-translational" modifications can enable or disable functions of the protein.

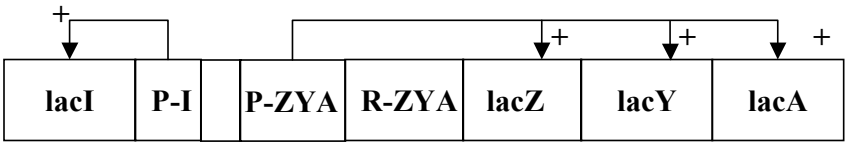
2.5 Genetic Regulation

Our understanding of genetic regulation is improving, and artificial genetic control systems are emerging. For example, a genetic "switch" has been demonstrated in the *E. coli* bacterium (see Gardner, Cantor, and Collins, *Nature*, vol. 403, pp. 339–342, Jan. 2000). One of the biggest obstacles to gene therapy is that it is almost impossible to regulate a dose. The patient has to be given a large quantity of genes through a modified virus (also known as a virus vector), resulting in a high probability of negative immune response and possible death. Thus, the ability to implement stable regulatory pathways for gene expression is critical for genomic approaches to fully affect human health. In this section, we describe a simple example of a biological pathway.

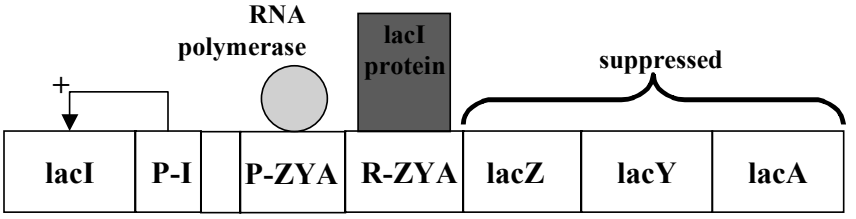
Regulation in biological pathways is often described using the operon model proposed by Jacob and Monod in 1960. The example they first described is the lactose pathway of *E. coli* (the *lac* operon). Figure 2.12 (a) shows how the operon is arranged on the *E. coli* chromosome. Lactose metabolism requires enzymes coded by the genes *lacZ* and *lacY*. For transcription of mRNA to occur, RNA polymerase must bind to a short promoter sequence. In Fig. 2.12, the promoters are labeled P-I and P-ZYA. If the RNA polymerase binds to P-ZYA, the DNA sequence of the *lacZ*, *lacY*, and *lacA* structural genes will be transcribed into mRNA molecules that are then translated on ribosomes to form enzymatic proteins. The enzymes will break down lactose and generate energy for the cell. For simplicity, proteins are referred to by the genes that encode them. For completeness, the *lacI*, *lacZ*, *lacY*, and *lacA* genes encode the lac repressor protein, beta-galactosidase (an enzyme that hydrolizes the bond between the two constituent sugars of lactose: glucose and galactose), lactose permease (an enzyme that brings lactose into the cell through the membrane), and thiogalactoside transacetylase (function unknown), respectively.

If RNA polymerase binds to P-I, the DNA sequence of the *lacI* regulatory gene will be transcribed to an mRNA molecule, which is then translated to form its protein. The arrow from P-I to *lacI* heads in the opposite direction of the other arrow because transcription occurs in the opposite direction. The *lacI* and P-I

a) Promoters and structural genes



b) lacI repressor protein binds to R-ZYA and inhibits polymerase from transcribing lacZ, lacY, and lacA



c) lacI protein is bound to free lactose, reducing repression

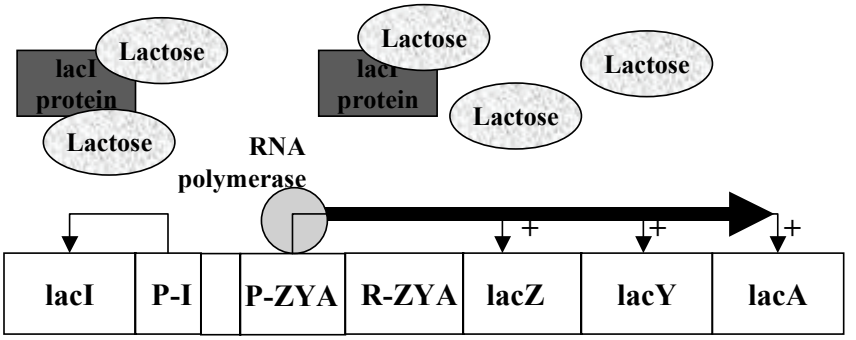


Fig. 2.12. Organization of the *lac* operon in *E. coli* with (a) repressor R-ZYA, promoters P-ZYA and P-I, and genes *lacZ* and *lacY* coding for lactose metabolism enzymes, (b) the repressor protein coded by *lacI* binds to P-ZYA, preventing *lacZ, lacY,* and *lacA* transcription and a corresponding increase in lactose concentration, and (c) lactose binds with *lacI,* removing it from the DNA and allowing RNA polymerase to transcribe the structural genes that in turn digest lactose.

"sense" strand is on the strand opposite to the *lacZ, lacY,* and *lacA*, and P-ZYA "sense" strand. The gene *lacI* produces a repressor protein. The protein binds to R-ZYA, as shown in Fig. 2.12(b), and results in the RNA polymerase being blocked from moving past P-ZYA and therefore inhibiting transcription of the lactose digestive genes, including *lacZ* and *lacY*. Figure 2.12(b) shows the equilibrium case. In fact, the repressor protein from *lacI* is rapidly binding and unbinding the DNA. The binding of RNA polymerase with DNA has a much lower rate constant, so it cannot compete with the repressor and cannot bind to the promoter and transcribe past P-ZYA. But, if the repressor-DNA rate constant decreases, the binding of RNA polymerase to DNA will become favored and the polymerase will transcribe past R-ZYA. The lacI repressor protein binds to lactose with an even higher rate constant than with DNA. In the presence of lactose, the repressor preferentially binds with it and leaves the DNA free. This leaves the path from P-ZYA through R-ZYA available for binding with RNA polymerase, as shown in Fig. 2.12(c), and allows the production of lactose digestive enzymes that break down lactose for energy. When most of the lactose is exhausted, the lacI repressor protein will be free to bind to P-ZYA and suppress further enzyme production, and the system will return to the state of Fig. 2.12(b).

The *lac* operon represents a simple negative feedback system. The system is represented in a schematic in Fig. 2.13. An increase in the lacI protein reduces the amount of lacZ, lacY, and lacA proteins produced—i.e., negative feedback. Note that relative terms like "increased" or "decreased" are used to describe the network of actions and interactions. This is typical in biology where description is often more appropriate than quantitative estimates or reductionistic models. Our group at Lawrence Livermore National Laboratory (LLNL) and several others have been applying fuzzy logic to these types of systems to increase understanding of the biological processes within the constraints of experimental measurement error. Positive feedback loops also exist for some operons, by which a protein binding to DNA promotes the production of downstream genes. An example of this is the *ara* operon for arabidose regulation. In these cases, the protein might change the 3-D structure of the DNA, making it easier for RNA polymerase to attach to the promoter. In general, promotion and repression occur through a variety of DNA–protein interactions (transcription factors).

Although operons are a genetic unit of coordinated expression of both mRNA and proteins, they do not necessarily or ordinarily act in isolation. Operons can be part of a larger coordinated network. The *lac* operon, for instance, is executed in concert with glucose metabolism. Most processes require more complex control than can be encoded in a single operon. Depending on the source of control—single, protein regulator, external stimulus, etc.—the network may have different names such as regulon, stimulon, or modulon. The specifics of biological regulation are beyond the scope of this book.

Soon after the operon idea was introduced, theoretical biologists tried to model it as a Boolean system: a gene was turned on (1) or off (0) by the promoter or repressor protein. This led to important discoveries about the nature of

network dynamics, autocatalytic sets, and how order emerges from the interconnected reactions of a cell. Ultimately, however, Boolean models failed to predict the behavior of many of the biological systems of interest. A full treatment of a complicated biological system is often difficult, so a Boolean representation of the reactions is still used when the pathway size is large. Modern simulations incorporate as much detail as possible: the transcription of DNA to mRNA, the translation of mRNA to protein at the ribosome, binding of the RNA polymerase to the DNA, promoter and repressor kinetics, and the decay of mRNA and protein molecules.

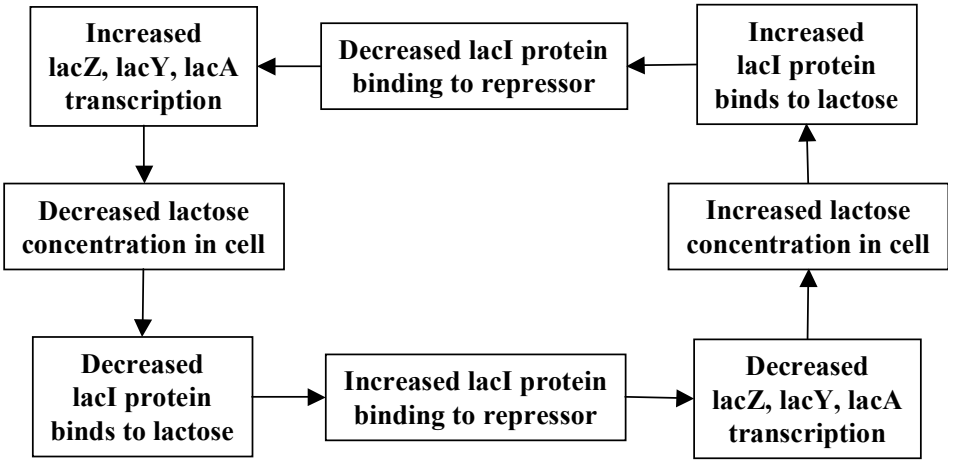


Fig. 2.13. Schematic of the *lac* operon feedback system.

2.6 Mutations and Disease

Deviations of genes from "normal" can occur as the result of inheritance and/or exposure to environmental factors. For example, sickle cell anemia is caused by a change in a single base of the DNA in an otherwise normal gene. Although it is just one base, the change causes a substitution of a single amino acid (valine for glutamine) in the protein that the gene encodes. The mutation results in abnormally shaped fragile red blood cells or "sickle cells." It is important to note that in many cases, changing only one base does not necessarily change the amino acid sequence of a protein, and changing one amino acid in a protein does not necessarily affect its structure or function.

It is instructive to work through a specific example of how to access DNA sequence data. We selected as an example the inherited disorder known as myotonic dystrophy, which is the most common form of muscular dystrophy that affects adults. Its symptoms range in severity from male-pattern baldness to those that are lethal. The cause of myotonic dystrophy is a set of CTG repeats that occur in the 3′ untranslated region of the dystrophia myotonica (DM) protein

kinase gene on the long arm of chromosome 19 (19q13.2 to 19q13.3; see the ideogram in Fig. 2.2). The CTG pattern repeats only 5 to 20 times in the normal population. Individuals with myotonica dystrophia have from 50 to thousands of CTG repeats, and the symptoms appear stronger with each affected generation. Figure 2.14 shows a list of DNA bases beginning at the 3′ end of the protein-coding region.

```
1751 ...TCCCATC TAGATGGCCC CCCGGCCGTG GCTGTGGGCC AGTGCCCGCT
1801 GGTGGGGCCA GGCCCCATGC ACCGCCGCCA CCTGCTGCTC CCTGCCAGGG
1851 TCCCTAGGCC TGGCCTATCG GAGGCGCTTT CCCTGCTCCT GTTCGCCGTT
1901 GTTCTGTCTC GTGCCGCCGC CCTGGGCTGC ATTGGGTTGG TGGCCCACGC
1951 CGGCCAACTC ACCGCAGTCT GGCGCCGCCC AGGAGCCGCC CGCGCTCCCT
2001 GAaccctaga actgtcttcg actccggggc cccgttggaa gactgagtgc
2051 ccggggccag cacagaagcc gcgcccaccg cctgccagtt cacaaccgct
2101 ccgagcgtgg gtctccgccc agctccagtc ctgtgatccg ggcccgcccc
2151 ctagcggccg gggagggagg ggccgggtcc gcggccggcg aacggggctc
2201 gaagggtcct tgtagccggg aatgctgctg ctgctgctgc tgctgctgct
2251 gctgctgggg ggatcacaga ccatttcttt ctttcggcca ggctgaggcc
2301 ctgacgtgga tgggcaaact gcaggcctgg gaaggcagca agccgggccg
```

Fig. 2.14. The myotonic dystrophy protein kinase gene contains a trinucleotide (CTG) repeat in its 3′ untranslated region. The TGA stop codon for the exon and the CTG repeats are underlined. The normal range of a repeat number is less than 30 and that of a pathological repeat number is from 50 to more than 2,000.

We could find information about myotonic dystrophy in the scientific literature and we could find it in the DNA sequence database. To use the sequence data itself, go to the National Center for Biotechnology Information online at http://www.ncbi.nlm.nih.gov/ and enter "dystrophia myotonica" in the GenBank search window. A list of hyperlinks will be returned. Select the link for accession number NM_004409 "Homo sapiens dystrophia myotonica protein kinase (DMPK), mRNA." If for some reason the search is not working, the accession number should be able to link to the same data. The information available at NM_004409 includes published information related to the gene, the source (*Homo sapiens*, 19q13.3), a list of the 629 amino acids in the expressed protein (MSAEVRL…PGAARAP), a list of the 3,407 bases from the DNA sequence, the position in the sequence that codes the protein (bases 777 to 2,666), and the location of the 3′ untranslated CTG repeat (begins at base location 2890). The last seven amino acids in the protein and the associated DNA sequence are shown in Table 2.2. A GenBank search is just one of many ways to find and compare DNA and protein sequence information. Another approach, the Online Mendelian Inheritance in Man (OMIM™) is also accessible through the NCBI web site. This database links significant scientific information with the gene databases. A search for "myotonic dystrophy" will provide significant detailed information on many muscular disorders and is an excellent launching point for reviewing the scientific literature.

Table 2.2. The last 7 of 629 amino acids in the dystrophia myotonica protein kinase (DMPK) and the associated DNA bases in the gene.

Position	623	624	625	626	627	628	629
Letter	P	G	A	A	R	A	P
Abbrev.	Pro	Gly	Ala	Ala	Arg	Ala	Pro
Acid	Proline	Glycine	Alanine	Alanine	Arginine	Alanine	Proline
mRNA	CCA	GCA	GCC	GCC	CGC	GCU	CCC
DNA	CCA	GGA	GCC	GCC	CGC	GCT	CCC

The gene associated with myotonic dystrophy was discovered as part of research being conducted on the disease. The MD approach of using disease markers to find the associated DNA regions and then sequence those specific regions represents the approach of the 1980s and early 1990s. In contrast, the location and function of genes is largely unknown for the DNA sequence being submitted by the large sequencing centers. Regions of the DNA that are amenable to transcription are called open reading frames (ORFs). Software to find ORFs has been developed and can be customized for a particular organism. For instance, bacterial and human/mouse ORF finders usually use different algorithms. Generic markers such as start and stop codons may often designate the beginning and end, respectively, of an ORF. However, just looking for these codons is generally not sufficient, and algorithm complexity has grown to include such techniques as hidden Markov models. As the DNA sequence for more organisms becomes available, the accuracy of ORF finders can be evaluated and improved. A continuing opportunity exists to use algorithms developed for other applications, such as speech recognition, and customize those techniques to find ORFs. Once the ORF finding problem is addressed and genes are provisionally located and identified, the next major step is to determine the function of the genes.

There are many approaches to assigning function to an ORF. One approach is through similarity. The unknown function of an ORF is sometimes assigned the function of a gene that has a similar amino acid sequence or a similar three-dimensional structure. Basic Local Alignment Search Tool (BLAST®) is a popular software approach to making these types of comparisons computationally.

3

Manipulating Nucleic Acids and Proteins

3.1 Sizing DNA and Proteins

Since nucleic acids have an inherent negative charge, an electric field can be used to apply a force to the molecule and move it through a medium like gel. By selecting a propagation medium that provides appropriate "friction," the nucleic acids can be sorted by size (length). Depending on the application, different gels, electric fields, and instruments are needed. DNA sequencing usually requires single-base resolution for DNA fragments with 20 to 1,000 bases. The details of this will be discussed later. DNA "sizing" may be done in simple gel boxes with large electrodes (see Fig. 3.1). The samples are introduced into the gel at one end near the cathode and propagate toward the anode at the other end of the gel. At a fixed time, DNA molecules of similar size will collocate in a band in the gel. Shorter DNA fragments move more quickly through the gel. These bands can be visualized with dyes, optical labels, or radioactive labels. An example is shown in Fig. 3.2 for a gel using myotonic dystrophy studies. Proteins are more complex, but are approached in a similar manner.

An entire gel can be imaged at one point in time to produce a two-dimensional map of the gel with DNA bands. It is also possible to monitor the gel with a detector at a fixed distance from the cathode or loading well. The bands of DNA move past the detector and produce a one-dimensional time plot. The time series approach is particularly useful for molecules separated by a gel in glass capillaries. Samples are introduced at one end of the capillary and the molecules are detected after being driven down the capillary by an electric field. A synthetic set of data for capillary collection of the gel in Fig. 3.2 is shown in Fig. 3.3.

Proteins (chains of amino acids) are more complex than DNA (chains of nucleic acid). For a chain with N elements, there is a 5^N combinatorial difference for specifying a protein verses a DNA sequence—i.e., 20^N versus 4^N possible chains. The chemical variation among amino acids is also greater than among the four nucleic acid bases. Despite these differences, proteins are sized or separated using very similar techniques. Of the gel-based techniques, two methods are most common: linear agarose gel separation and gradient gel separation. Agarose separation is similar to DNA separation; the "friction" of the gel separates

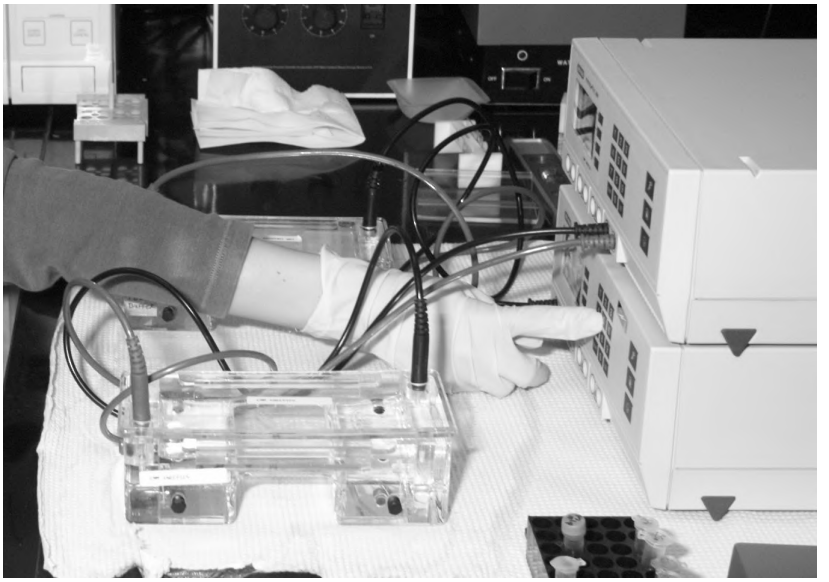


Fig. 3.1. Agarose gel box for electrophoretic separation of negatively charged DNA by size (length).

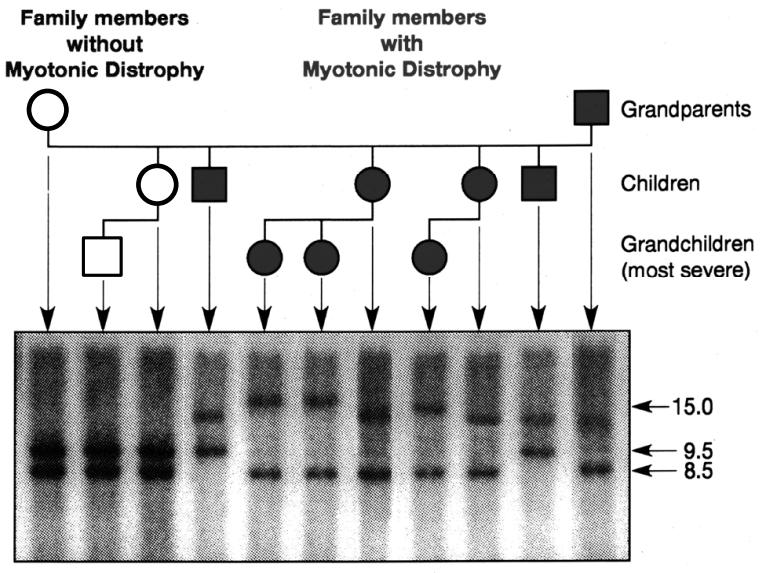


Fig. 3.2. Radioactively labeled gel comparing DNA size for a family that includes members with and without myotonic dystrophy. As the number of trinucleotide (CTG) repeats increases, the severity of the disease increases. The increase is more likely with each affected generation. Circles are women; squares are men. Open circles and squares indicate the absence of the disease.

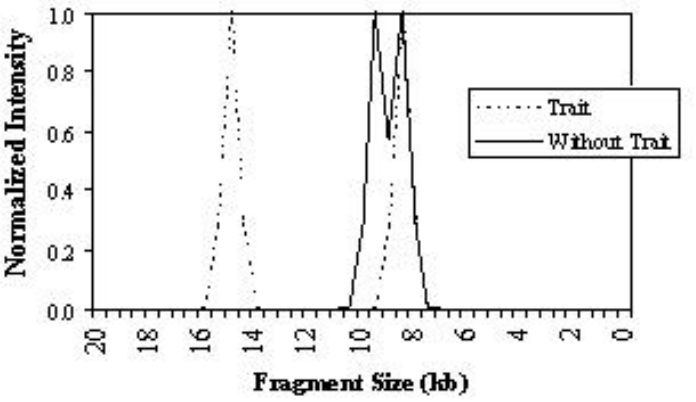


Fig. 3.3. Synthesized capillary electrophoresis signatures for myotonic dystrophy markers (one sample with and one without the trait). Both samples have DNA fragments of about 8 kb. The sample with the trait also has a much larger fragment associated with the trinucleotide repeat at 15 kb. This is 5.5 kb longer than the fragment for the sample without the myotonic dystrophy trait.

molecules roughly by size and an electrophoretic force is used to move the molecules through the gel. Figure 3.4 shows a protein separation using a linear gel. Gradient gels have a range of pH across the gel. As molecules are electrophoretically moved through the pH gradient, they tend to stop when the available charge in the gel neutralizes the charge on the molecule. This is known as isoelectric focusing. It is also possible to use the initial sample buffer to create a gradient that propagates with the molecules through a linear gel. The linear gel separation for molecule size and the gradient gel for determining molecule charge can be combined in one gel experiment known as the 2-D protein gel. Two-dimensional gels are a rapid way to get a fingerprint of protein distribution by size and charge. As 2-D gels become more repeatable and quantitative, their importance as a tool for analyzing proteins is growing. Just as genomics represents approaches to measuring the total genetic complement of an organism, proteomics represents whole-organism protein studies. The tools are evolving for proteomics.

Proteins can also be measured using mass spectrometry. There are many types of mass spectrometry instruments. Briefly, the mass-to-charge ratio is measured for the collection of protein fragments introduced into the spectrometer. The different techniques for introducing the sample greatly affect the fractionation and charge of the protein and are beyond the scope of this discussion. It is interesting that a 1-D isoelectric focusing gel can be used as the front end on a mass spectrometer in order to create a "virtual 2-D gel." The mass spectrometer has the potential to provide higher resolution and more quantification than linear agarose gels.

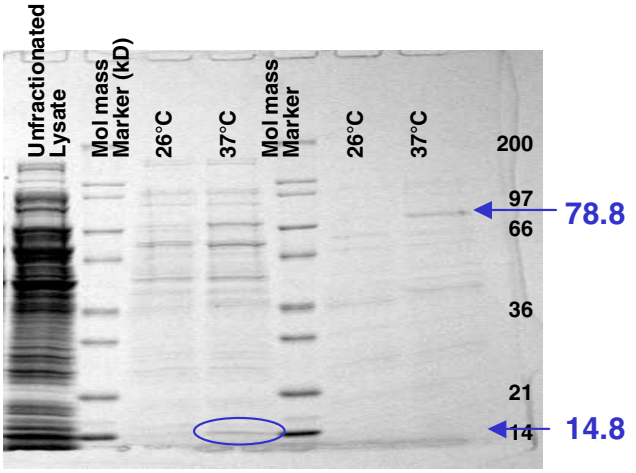


Fig. 3.4. Protein separation using a linear gel. The proteins are from the bacterium *Yersinia pestis*. The scale on the right is approximate mass in kilodaltons. (Gel image courtesy of Sandra McCutchen-Maloney, Lawrence Livermore National Laboratory).

3.2 Blots

After electric field separation, sometimes nucleic acids and proteins are transferred from the bands in the gel to a membrane. This transfer step is known as blotting. There are several useful blotting techniques. The three described here are Southern, Northern, and Western blots and are summarized in Table 3.1.

Table 3.1. Comparison of different blotting techniques.

Blot	Target	Typical gel	Detector
Southern	DNA	Agarose	DNA
Northern	RNA	Agarose	DNA
Western	Protein	Polyacrylamide	Antibody

The Southern blot is named after E. M. Southern and is used to separate and identify DNA. The DNA fragments are separated first on an agarose gel (the electrophoresis previously described) and then blotted onto a nylon or nitrocellulose membrane. Hybridization is the term for joining two separate single strands of DNA into one double-stranded piece of DNA, with each base pair following the A-to-T and G-to-C rules. Specific DNA sequences in a Southern blot are identified by hybridizing the membrane-bound DNA with labeled test DNA. When the strands are complementary (or very close to

complementary), the labeled DNA hybridizes to the spot and its label can be detected.

Northern blots are a similar procedure used to separate and identify RNA fragments. Western blots are also similar to Southern blots but are applied to proteins. For proteins, the electrophoresis is usually done in polyacrylamide. After being blotted onto a nylon or nitrocellulose membrane, the proteins are detected with labeled antibodies.

3.3 Cutting DNA and Protein

There are several ways to fragment chains of nucleic or amino acids. A direct approach still used in DNA sequencing laboratories is mechanical shearing. DNA is forced through a pore and the mechanical stresses cause fragmentation. The velocity and pore size can be adjusted to fragment the macromolecule into different size distributions. Although it is largely random, there is evidence that fragmentation occurs preferentially at the location of some specific sequences.

3.3.1 Restriction enzymes

In addition to mechanical approaches to cutting DNA, there are biochemical approaches. Some enzymes have the property of cleaving specific patterns of nucleic acids and are known as restriction enzymes. See Table 3.2 for a partial list of restriction enzymes and their specificity. The site recognition patterns are often palindromes (or at least partially palindromic) from the top to bottom strand. (Palindromes are words or sentences that spell the same forward and backward, such as "madam I'm Adam.") As an example, the restriction enzyme *Eco*RI has a cut site G:AATTC on the 5′ to 3′ strand and CTTAA:G on the 3′ to 5′ strand. The colon denotes the location of the cut. If the location of the cut is not in the center of the palindrome, one of the strands in the double-stranded DNA is longer after the enzyme cuts the molecule. *Eco*RI leaves a four-base overhang after cleaving the double-stranded DNA.

Restriction enzymes were first discovered in bacteria. Certain bacterial enzymes "restricted" infection by bacterial viruses. The enzymes were found to cleave the DNA of the invading bacteriophage, making incorporation of the phage DNA into the host significantly less likely. More than 3,000 enzymes are now known to cleave DNA. If a restriction enzyme does not reproducibly cleave at the same DNA sequence, it is called a Type I restriction enzyme. Type II restriction enzymes cleave at specific recognition sites on double-stranded DNA. The recognition sites are usually 2 to 20 bases long. The naming convention for restriction enzymes such as *Eco*RI is that the first capital letter is the genus (*E* for *Escherichia*) and the second two letters are the species (*co* for *coli*), followed by a designator and Roman numeral that increases for each enzyme discovered (RI).

The cleavage site may be at a base pair (blunt end) or leave a single-stranded overhang that is usually designated by the orientation and bases. For instance, *Fok*I would leave an overhang designated 5′-NNNN. There are applications

Table 3.2. Sampling of restriction enzymes used to cut DNA at specific sites. The colon denotes regions where the cut occurs, R is purine (A or G), and the double-stranded DNA for the cut sites is oriented with the top strand going 5′ to 3′, left to right. Note that *Hae*III has blunt ends and all of the other enzymes listed have overhangs.

Enzyme	Recognition sequence	Example cut site on double-stranded DNA
*Bam*HI	G:GATCC	G GATCC CCTAG G
*Bst*YI	R:GATC	R GATCY YCTAG R
*Eco*RI	G:AATTC	G AATTC CTTAA G
*Fok*I	GGATG	GGATGNNNNNNNNN NNNNN CCTACNNNNNNNNNNNNN N
*Hae*III	GG:CC	GG CC CC GG
Hin dIII	A:AGCTT	A AGCTT TTCGA A

where overhangs are desired and there are also applications where blunt ends are preferred. If the same restriction enzyme is used to cut two different strands of DNA, the palindrome property can be exploited to recombine the DNA using the "sticky ends" of the DNA. This is demonstrated for the restriction enzyme *Bam*HI in Fig. 3.5.

Restriction enzymes scan along double-stranded DNA at very rapid rates. A few *E. coli* enzymes have been measured scanning at over one million base pairs per second. The first Type I restriction enzymes (*Eco*B) were discovered in 1968 through studies of *E. coli*. The first Type II restriction enzymes (site-specific) were described in 1970. Enzymes that cleave the DNA at sites internal to the strand (single or double) are known as endonucleases. Enzymes that trim bases from the ends of DNA are known as exonucleases and may be specific to individual bases as well as the 3′ or 5′ end.

Restriction enzymes have had an important role in the Human Genome Project. Even before single-base resolution DNA sequencing became cost effective, a series of restriction enzyme digests combined with DNA sizing by gel electrophoresis could be used to identify where specific markers occurred in a genome. The fragment patterns could be reassembled into a coarse physical map of the larger source piece of DNA. This allowed production of physical maps of human chromosomes, such as the abbreviated one in Fig. 2.2.

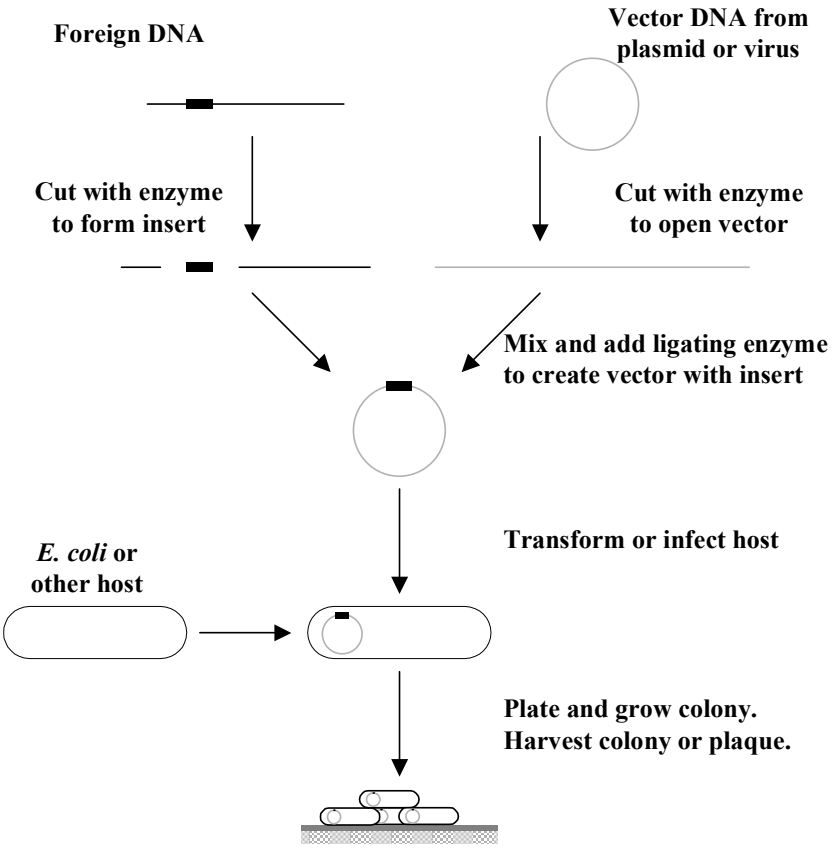


Fig. 3.5. Example of the operation of the restriction enzyme *Bam* HI to cut two different samples of DNA. The two cleaved samples can be recombined at the sticky ends to produce a new strand of DNA.

It is instructive to work through an example of restriction mapping. Beginning with a large DNA fragment, smaller fragments produced by restriction enzymes are labeled and sized. The fragment patterns can then be used to infer a physical ordering of the clones on the larger input DNA. Figure 3.6 shows a synthesized restriction map in which one enzyme (A) produces three fragments and a second enzyme (B) produces two fragments. As in real applications, it is not possible to uniquely assign physical positions to every fragment with the results of (A), (B), and (A+B). Additional information would be needed. The example indicates how much information can be extracted from restriction enzymes. The reader might consider the value of sequencing the ends of the fragments or application of a third enzyme (C) to help deduce the order of the fragments.

It is also possible to use restriction mapping to identify genetic markers, such as the presence of repeat sequences. As an example, consider the CTG repeat

associated with myotonic dystrophy. If a restriction enzyme or a series of restriction enzymes can be used to excise a DNA fragment that includes the 3′ untranslated region of the DMPK gene on chromosome 19, the size of the CTG repeat can be estimated. After restriction enzyme isolation of the fragment, the number of CTG repeats can be estimated using DNA sizing by gel electrophoresis. This was the approach used to deduce the number of repeats for expression of the myotonic dystrophy trait. The use of restriction enzymes and agarose gel electrophoresis to identify differences in DNA sequence is known as restriction fragment length polymorphism or RFLP. As the number of repeats increases or other differences in DNA sequence becomes subtler, it is necessary to more closely bracket the region of interest. A technology known as a polymerase chain reaction (PCR) can be applied to short tandem repeats (STRs) and other small DNA differences, including single-nucleotide polymorphisms (SNPs). PCR is discussed in the next section.

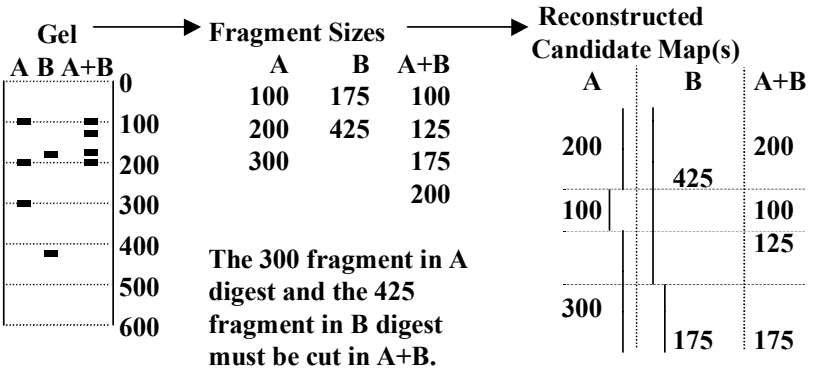


Fig. 3.6. Synthesized restriction map of a 600-base DNA fragment using the enzymes A and B. Note that the fragments cannot be uniquely placed on the map—the 100- and 200-base fragments can be interchanged consistent with the data.

3.3.2 Protein digests

There are a variety of enzymes that are used to digest proteins into fragments. The digested protein can then be separated by size, perhaps using a mass spectrometer, and the resulting fragment pattern can be used to identify the protein. Several of the proteolytic enzymes used today were isolated from the stomach and intestines of humans and animals. In nature these enzymes break down proteins as part of food digestion. There are also proteolytic enzymes isolated from plants including the pineapple. A common enzyme used is trypsin. This enzyme is very specific; it cleaves at arginine-X and lysine-X bonds unless X is proline. Table 3.3 shows several enzymes used to digest proteins.

Table 3.3. Several protein digests and their cleavage sites. See Table 2.1 for amino acid letter designations.

Enzyme	Recognition sequence	
Trypsin	R:X and K:X	Unless X is P
Chymotrypsin	F, Y, W, L	Cleavage C-terminal side
	M, I, S, T, V, H, G, A	Some cleavage C-terminal side
Lys-C	K:X	Cleavage (K=lysine)
	N:X	Some activity
Arg-C	R:X	Cleaves C-terminal side, but not for all X (R=arginine)
	K:X	Also observed
Asp-N	X:D	Cleavage N-terminal side (D=aspartic acid). Some other cleavage reported.
Glu-C	E-X and D-X	In phosphate buffers unless X is P.
	E-X	In some other buffers unless X is P.

3.4 Copying DNA

One approach to making copies of DNA in a laboratory is to utilize living cells. This is known as cloning and will be discussed later. We begin our discussion of how to make copies of DNA with a powerful biochemical technique that allows specific regions of DNA to be amplified (copied) from a template.

PCR is a thermal cycling method that denatures DNA (separates the two strands of DNA) at about 95°C, synthesizes a complementary strand within a region flanked by two primers at about 75°C, and renatures (anneals complementary strands) at about 55°C. PCR starts with a sample that may have very little DNA. In some extreme applications, this may be a single piece of DNA.

The process is begun by denaturing the double-stranded template in a deoxyribonucleotide triphosphate-rich environment. The primers attach to the single-stranded DNA and initiate an enzyme (the polymerase), assembling the dNTPs into a complementary strand growing in one direction from the 3′-hydroxyl end of the primer toward the opposite end of the template. Several enzymes have been used for synthesis, including one extracted from *Thermus aquaticus*. Enzymes that run in the opposite direction are known as reverse transcription enzymes. There are new chemistries that perform similar DNA replication, but do not require thermal cycling.

Assuming the process starts with a single template of DNA, the first denaturing step results in two strands (the complementary strands of the template). After synthesis and renaturing, the second denaturing results in four single strands: the two original template strands and two strands that begin at a primer and run from the 3′-hydroxyl end of the primer toward the opposite end of the template. Primers hybridize to the four strands, synthesize, and renature. The third denaturing cycle results in eight strands: the two original template strands, four strands that begin at a primer and extend to the end of the template, and two strands that begin at one primer and end at the other. The fourth denaturing results in sixteen strands: the two original template strands, six strands that start at a primer and run to the end of the template, and eight strands that begin at one strand and end at the other. As the number of cycles continues, the number of short primer-to-primer strands doubles every cycle; the longer primer-to-template end strands grow two strands per cycle; and the two original template strands remain. In eleven cycles the primer-to-primer strand has been amplified to more than 1,000 copies. The first three cycles of PCR amplification are shown in Fig. 3.7. The Nth cycle would result in

2	Original template strands (T3-T5),
$2N$-2	Partially amplified strands, including templates with one end clipped by one primer on one strand (T3-RP5 and FP3-T5) and templates clipped on one end annealed to an amplified strand that is clipped on both ends (AS3-RP5 and FP3-AS5), and
2^N-$2N$	Strands amplified between both primers (AS3-AS5).

This simple thermal cycling process allows biochemical amplification of DNA.

Primer design allows specific regions of DNA to be extracted from the template. Because primers (even those a few hundred bases long) can be synthesized, the PCR technology can also be applied to detection or in a manner similar to restriction enzymes. PCR amplification followed by electrophoretic gel separation allows PCR to be used for detection as well as amplification. By incorporating fluorescent labels into the PCR process, it is possible to do optical detection in the PCR thermal cycler. One approach is to use a fluorescent dye attached to a primer that is quenched when it is incorporated into one of the short synthesized strands. The presence of the DNA can be deduced by monitoring the reduction of fluorescence as the sample is thermally cycled. Figure 3.8 shows a portable PCR system that incorporates both amplification and detection. These systems use smaller volumes and faster cycling times than conventional bench top laboratory PCR systems. In one system, detection was reported in less than 7 minutes using 25 PCR cycles with a 1-second denature, 1-second anneal/extend, and 17 seconds to switch the sample temperature.

```
5′                        TEMPLATE                        3′
CCGATTAGCCTCACCGATTCAGGCTAAGCTGGCTCGAATGAACAACCC (T3)
GGCTAATCGGAGTGGCTAAGTCCGATTCGACCGAGCTTACTTGTTGGG (T5)
3′                                                        5′

                      FIRST CYCLE
CCGATTAGCCTCACCGATTCAGGCTAAGCTGGCTCGAATGAACAACCC (T3)
GGCTAATCGGAGTGGCTAAGTCCGATTCGACcgagcttacttgtt     (RP5)

     tagcctcaccgattCAGGCTAAGCTGGCTCGAATGAACAACCC (FP3)
GGCTAATCGGAGTGGCTAAGTCCGATTCGACCGAGCTTACTTGTTGGG (T5)


                    SECOND CYCLE
             (T3-RP5, FP3-T5 not repeated)
     tagcctcaccgattCAGGCTAAGCTGGCTCGAATGAACAA     (AS3)
GGCTAATCGGAGTGGCTAAGTCCGATTCGACcgagcttacttgtt     (RP5)

     tagcctcaccgattCAGGCTAAGCTGGCTCGAATGAACAACCC (FP3)
     ATCGGAGTGGCTAAGTCCGATTCGACcgagcttacttgtt     (AS5)


                     THIRD CYCLE
   (T3-RP5, FP3-T5, AS3-RP5, FP3-AS5 not repeated)
     tagcctcaccgattCAGGCTAAGCTGGCTCGAATGAACAA     (AS3)
     ATCGGAGTGGCTAAGTCCGATTCGACcgagcttacttgtt     (AS5)
```

Fig. 3.7. A polymerase chain reaction (PCR) is a biochemical method for copying or detecting DNA. The orientation of the right end of the strand is given by the 3 or 5 in the strand name. The primers are given in lower case, T is for template, RP is for reverse primer, FP is for forward primer, and AS is for amplified strand.

Fig. 3.8. Miniaturized PCR-based detection system invented at Lawrence Livermore National Laboratory. Samples are introduced in plastic tubes at the right of the instrument.

3.5 Genetic Engineering

For centuries farmers have used selective breeding and seed selection to engineer some of the genes in plants. Chemical mutations of plants, animals, and bacteria have also been introduced during breeding to select progeny with preferred traits. In his *Origin of Species* Charles Darwin compared the power of natural selection with man's selection. Darwin states, "Variability is not actually caused by man; he only unintentionally exposes organic beings to new conditions of life, and then nature acts on the organism and causes it to vary."

Today we face a major contradiction to Darwin's statement. Genetic engineering is a significant technical achievement that allows the deliberate modification of an organism's genome. For many applications, our current limited understanding of genetic regulatory systems can lead to unintended consequences from genetic manipulation. As our understanding grows, genetic engineering will contribute significantly to human health and agriculture.

In order to engineer the modification of the DNA of an organism, we need to be able to collect or synthesize DNA, make copies and store DNA for use, insert the DNA into an organism, and have the organism accept or incorporate the new DNA into its genome. Short strands of DNA can be synthesized chemically. Although methods for chemical synthesis are improving, most genetic applications use sequences that are extracted from other organisms. We have already mentioned restriction enzymes and PCR as methods for cutting DNA at specific sites. In this section we introduce technologies that allow the insertion of DNA into a new host organism. A simplified outline of the basic process for inserting foreign DNA into a host (cloning) is given in Fig. 3.5. A summary of how much foreign DNA can be accommodated in several of the typical cloning systems is given in Table 3.4. Figure 3.9 shows hundreds of *E. coli* colonies growing with inserts of foreign DNA.

Table 3.4. Summary of cloning systems.

System	Approximate DNA insert size
Phage	8–20 kb
Plasmid	15–20 kb
Cosmid	35–45 kb
BAC	100–150 kb
YAC	200–400 kb

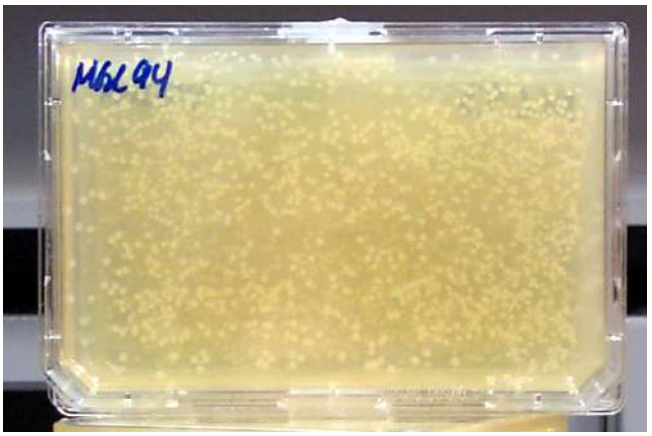


Fig. 3.9. Plate with hundreds of bacterial colonies with each colony containing inserted DNA.

3.5.1 Transformation

Transformation is the genetic integration of a bare, nonviral piece of foreign DNA such as a plasmid into a bacterial cell. The typical target is *E. coli*. Transformation is important to genetic engineering. The analogous process for eukaryotic cells is usually called transfection, and yeast cells are typical targets. The generic term for transfer of nonviral DNA to a cell is transduction. There are three general methods for transformation: heat transformation, electroporation, and conjugation.

Heat transformation. Cells are initially treated in an ice bath with $CaCl_2$. The cold treatment is followed by 42°C exposure for 1 to 2 minutes. Typical transformation success rates are 1 cell per 1,000 cells treated.

Electroporation is short for electric field-mediated membrane permeabilization. A brief electric shock of a few milliseconds and kilovolts is used to drive DNA through the cell membrane. It is not known if the pores are formed during the process or if existing pores are made available.

Conjugation. Some plasmids have a natural ability to replicate and induce cell-to-cell junctions that allow transfer of plasmids. At this point, conjugation is not typically used in transformation.

Transformation efficiencies depend heavily on the size of the DNA insert, the target host, and the conditions used. There are commercially available protocols for transformation of a variety of cells. In addition to these direct approaches, it is possible to infect a target organism with a genetically engineered virus.

3.5.2 Bacteriophage cloning systems

As indicated earlier, bacteriophages are viruses that infect bacteria. The bacteriophage lambda (λ) infects *E. coli* and can be engineered to deliver DNA from other organisms into *E. coli*. Recall that some phages, including bacteriophage λ, can integrate its DNA into the host and continue nearly indefinitely as part of the host (lysogeny). In addition, bacteriophage λ DNA can induce excision from *E. coli* DNA with about 100 replications of its 50-kb DNA made by the host followed by host cell lysis. There is a 20-kb region of the 50-kb bacteriophage λ DNA that is required for these integration and excision events. This is called the I/E region. By replacing the I/E region of phage λ with different DNA, we can use *E. coli* to make many copies.

The replication of phage λ DNA by *E. coli* during the lytic cycle is exploited in this process. In the first step of the lytic cycle, the phage DNA is copied in a repeating linear series. There may be 100 copies of the phage genome in the DNA strand. Next, a protein head forms around the first phage DNA. The DNA is cleaved by a restriction enzyme and if the DNA is roughly 50 kb in length (48–52 kb), a tail is added. The combination of the head, 50-kb strand, and tail makes a complete λ-phage. If the DNA is the wrong length, the phage is not completed—i.e., the DNA is cleaved but the head or tail of the phage is not completed or it is not infective.

The cleavage sites are important in correctly packaging the phage. Phage λ has a 12-base, single-stranded extension at each 5′ end of its linear 50 kb of DNA. These extensions are usually called the cohesive ends or cos ends. There is an enzyme that recognizes the cos sequence and therefore can cleave the replicated linear strand at those sites. Because the cos ends are also complementary, upon initial injection into *E. coli*, the λ DNA forms a loop (plasmid). The linear strand with the 100 repeats of λ DNA described in the previous paragraph is generated using the circular plasmid as a template.

Because the phage λ replication process includes lysis of the *E. coli* host, the cloned DNA is not collected in living colonies. Instead, copies of phage λ are harvested where the *E. coli* has been lysed. These regions are referred to as plaques because they are clear areas on the agar plates where the *E. coli* cells have been lysed.

There are multiple ways to engineer a DNA sequence into phage λ. One common technique is to use restriction enzymes. An enzyme, *Bam*HI for instance, is used to cleave both the phage λ and the source DNA. The source DNA must be fractionated by size in order to have inserts that are roughly 20 kb long. The λ DNA will be missing some part of the I/E region. The CTAG sticky ends left by *Bam*HI (see Table 3.2) are used to recombine source and λ DNA in the presence of a DNA ligase, such as T4 ligase. Finally, special *E. coli* cells that do not allow bacteriophage λ with intact I/E regions to replicate are transformed. Only the genetically modified phage λ will perpetuate and can be harvested as plaques.

3.5.3 Cosmid cloning systems

Cosmids are a combination of plasmid and bacteriophage λ cloning systems. The λ phage is used as a vehicle for infection and the cos ends are used to form a stable plasmid in *E. coli*. The λ DNA is modified to include an antibiotic resistance (usually tetracycline resistance) gene as well as an origin of replication (ori) site for *E. coli* recognition as a perpetuating plasmid. There are also usually several restriction enzyme sites used to open the plasmid and insert or remove the DNA of interest. T4 DNA ligase is also used to join the source and vector DNA.

The total size of the cosmid is 50 kb because it must be incorporated into the λ phage for infection. However, since the λ phage will not be replicated (the goal is a stable plasmid, not phage copies), most of the λ DNA can be overwritten. The cosmid cloning system can carry about 40 kb of inserted DNA. If copies of *E. coli* are grown on an antibiotic-treated agar plate, only the *E. coli* with successful inserts of the antibiotic resistance gene will replicate.

3.5.4 Artificial chromosome cloning systems

An approach to storing and copying DNA in a stable manner is to insert the DNA into a chromosome. Several chromosome systems have been engineered to enhance this capability. The principal distinguishing factor is the size of the piece of DNA that is being stored and cloned.

Yeast artificial chromosomes (YACs) are used for cloning very large DNA inserts ranging from 200 to 400 kb in length. Bacterial artificial chromosomes (BACs) are used for cloning large DNA inserts ranging from 100 to 150 kb. The BAC system was derived from a plasmid in *E. coli* known as the F-factor plasmid. Another system is the P1 artificial chromosome (PAC) for cloning 80- to 100-kb DNA inserts. The PAC system is based on the bacteriophage P1 and an *E. coli* plasmid.

3.6 Protein Expression

Nucleic acids are polymers composed of four possible nucleotides connected by phosphodiester bonds between the 3′ and 5′ carbons of the backbone. We have presented methods to detect DNA sequences (hybridization), cut DNA (restriction enzymes), make copies of DNA (cloning, PCR), and insert DNA segments into an organism (electroporation, virus infection).

Many proteins can be manipulated using methods similar to DNA manipulation. Proteins are polymers composed of 20 possible amino acids connected by peptide bonds between the carboxyl group (COOH) and the amino group (NH_2) of the next amino acid. There are 20 different chemical side chains or R groups—a specific one for each amino acid. As with single-stranded DNA, the side chains are chemically active. DNA side chains (bases) hybridize in a relatively predictable manner to complementary bases. The binding of an amino acid chain, however, is more complicated than hybridization of DNA. The available binding sites of a protein depend significantly on its three-dimensional

structure. There are enzymes that cut proteins at specific sites and because we can engineer DNA and insert it into a cell, it is possible to use cells to make foreign proteins. For proteins there currently is no chemical equivalent of making copies like PCR does for DNA. In short, the best current method for making protein is to manipulate the DNA and use cells to produce the protein. There are also emerging techniques for expressing protein *in vitro*.

In order to produce foreign proteins using cells, the DNA sequence that codes for the protein must be introduced into the cell. In addition, a proper promoter and terminator are needed to induce transcription, and the transcription factors must be compatible with the RNA polymerase. It may also be necessary to place the protein-coding DNA inside a regulated region of DNA so that the number of transcript copies can be controlled. If the protein is to be harvested, it may also be appropriate to engineer a tag into the protein that allows efficient purification of the protein.

3.6.1 Using cells to express proteins

In many applications, bacterial cells such as *E. coli* are used to express proteins. The desired gene is cloned into a vector with a promoter and terminator that are compatible with the target host bacterium. The vector must also be compatible with the RNA polymerase. If transcription is to be controlled, as is often the case if large quantities of protein are needed, the vector is tied to a regulatory pathway or even introduced into the chromosomal DNA of the host. Once the insert vector has been engineered to induce transcription of the correct messenger RNA, a ribosome-binding site may also need to be inserted in front of the translation initiation site. It may also be necessary to add a translation termination signal to the gene insert. Sometimes host cell enzymes attack a protein formation. To make the foreign protein more stable and able to resist host cell proteolytic enzymes, additional amino acids can be engineered onto the N-terminal of the foreign protein by adding codons to the cloned gene. The method for harvesting the protein also affects the genetic design. For instance, cell lysis approaches will need different expression timing than proteins that are excreted by the cell.

Every task the cell has to perform to create the foreign protein is an additional cost, known as the metabolic load, to its own "normal" activities. To overcome the metabolic load of foreign proteins, host cells may be selected with special attributes, including the absence of specific proteolytic enzymes, specific oxygen or other metabolic processing attributes, growth rates, and promoter control mechanisms. In some cases, these attributes are also genetically engineered into the bacterial host.

Because bacterial cells cannot always produce functional eukaryotic proteins, host systems and vectors for eukaryotic cells have also been developed. The same transcription and translation issues face eukaryotic expression systems. There are enzymes in eukaryotic cells that help stabilize many proteins as well as enzymes that cleave proteins to improve the functionality of the protein. In addition, eukaryotic proteins may also undergo specific post-translational modifications, including the addition of sugar residues to the protein (known as

glycosylation) and/or the addition of chemicals to specific amino acids within the protein (known as acetylation, phosphorylation, sulfation, etc.). To address many of these issues, protein expression systems for fungi (*Saccharomyces cerevisiae*), insect cells (*Spodoptera frugiperda*, fall armyworm), and other eukaryotic cells are used. The insect cells are usually infected with the *Autographa californica* (alfalfa looper) virus. The protein expression system for the viruses that infect only invertebrates is known as a baculovirus system and has successfully expressed many functional mammalian proteins.

The details of protein engineering are beyond the scope of this book. In addition, the field is changing rapidly as new methods are developed and significant experience is gained.

3.6.2 Natural optical signatures

One very useful technique for tracking the expression of genes and proteins is the use of fluorescent proteins. The gene for a protein with a known optical signature is included next to a gene of interest. As long as the gene is in the same regulatory region of the genome, the fluorescent protein will be transcribed and translated with the neighboring gene and protein. This approach is used to track both the expression of engineered proteins (inserts) and the expression and location of native proteins.

There are now several colors available (green, yellow, cyan, blue, and red) in fluorescent proteins. Historically luciferase and green fluorescent protein led the way. Luciferase is a 62-kDa protein isolated from the firefly *Photinus pyralis* that emits yellow-green light at 560 nm. Green fluorescent protein (GFP) is a small protein (27 kDa) isolated from the jellyfish *Aequorea victoria*. GFP is excited at about 405 nm and emits in the green roughly between 500 and 550 nm. One limitation of GFP is the time delay between protein expression and fluorescence. The delay is due to post translation oxidative modification of GFP that is critical to fluorescence. This limits GFP use for real time *in vivo* expression studies.

3.6.3 *In vitro* protein production

The conventional approach for generating proteins is to take over the cellular machinery for transcription and translation. For many applications it is desirable to mimic these reactions *in vitro*. This approach is also referred to as "cell free" protein expression. The enzymes for transcription and translation from several organisms have been isolated, modified, and tested to induce protein production *in vitro*. As the techniques become validated through comparisons with native proteins, this approach offers tremendous promise for biochemical synthesis and engineering of new proteins. *In vitro* protein production is especially important for proteins that are toxic to the traditional host cells used for protein expression such as *E. coli*. Cancer, infectious disease, and other human health applications require the expression of proteins that are toxic to *E. coli*.

4

An Integrated Approach for Biological Discovery

Just as it was possible to take a whole-organism (all of the genes) approach to human DNA sequencing, it is now possible to consider a whole-organism approach to determining the mechanisms that control the biochemistry in a cell. An understanding of these mechanisms is the basis of disease prevention and treatment. High-throughput whole-organism approaches include genomics; proteomics; functional genomics; and structural genomics for the study of genes, proteins, gene function, and three-dimensional protein structures, respectively. These approaches contrast with and complement the hypothesis-driven research tradition in biology of studying a single isolated phenomenon. The next generation of hypotheses will address an entire complex activity such as metabolism, which requires information about multiple protein–DNA interactions of the cell's regulatory mechanisms.

As shown in Fig. 4.1, an integrated approach should have the potential to go from genes (DNA) to function. This chapter presents approaches to collecting gene, protein, and regulatory data and the challenges of integrating these data into an information system model of the biology. We include in this chapter a brief description of how computer modeling and simulation might facilitate data interpretation and an understanding of complex biochemical pathways and mechanisms.

There are several levels of data abstraction in biology. The basic unit is the one-dimensional gene that is composed of the four building blocks A, C, G, and T. The next level of abstraction is a protein. Proteins are three-dimensional combinations of the 20 amino acids, with structure strongly correlated with function. Pathways are the next level, with many pathways often combining for a robust activity such as metabolism. Systems-level biology is the new frontier and involves the pursuit of mathematics, computer algorithms, and biological data to show that complex networks of organisms (even ecosystems!) may be amenable to modeling and measurement. These are new ways of viewing biological research. Biology is becoming an information-based science.

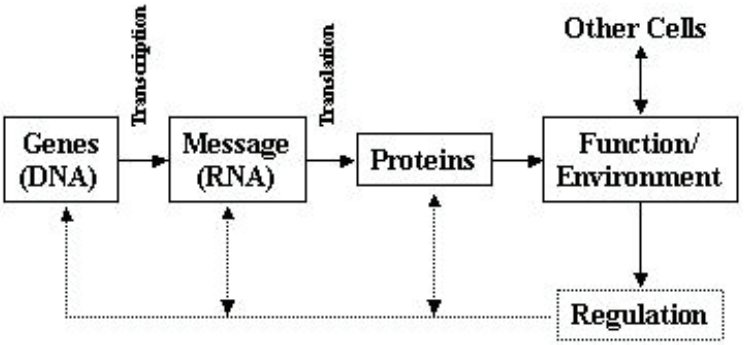


Fig. 4.1. Living organisms are a complex, nonlinear feedback system that derives many traits from inheritance (DNA) and is continuously influenced by the environment, including neighboring cells.

The potential impact of appropriate modeling tools in biology parallels the historic impact of circuit simulation in electrical engineering. Genomic engineers need simulations of complex biochemical networks that can reduce the amount of experimentation needed to understand changes in the networks or to introduce deliberate changes that influence function. Predictive modeling will profoundly influence our ability to safely engineer new crops, medicines, and genetic treatments.

Computer engineering was revolutionized with the introduction of modular subsystems and standards such as lambda design rules for integrated circuits and software for implementing the rules. Electronic circuits are often viewed as analogous to biological circuits, and researchers endeavor to discover the design rules for a biological circuit simulator or bio-SPICE (Simulation Program for Integrated Circuits Emphasis). Electronic and biological circuits both have components, connectors, and power supplies. Both types of circuits can have feedback and feed-forward loops that drive nonlinear circuit behavior. An important aspect of this analogy that is often overlooked is that the resistors, capacitors, inductors, and wires in an electronic circuit (a computer, for example) must be simulated and integrated with thermal and other environmental conditions. Supercomputer designs that do not take into account heat and packaging issues will fail. The environment is also influenced by the existence of other computers that communicate and influence changes in the environment (network). Just as many computers do not operate in isolation but are part of a complex network of other devices, the simulation of a biological system requires at least comparable and probably greater integration than simulation of a computer network. Biological applications are benefiting from several approaches at different scales.

The “parts list” for a biological circuit includes genes and proteins. New technologies are making it possible to measure the wiring connections, including regulators among the many genes and proteins in a circuit or pathway. Unlike

a typical electronic circuit, the values of the components are not well quantified. For instance, even if the DNA sequence of a gene is identified, the gene may not be transcribed without the presence of an appropriate promoter protein. Even if the right promoter is available, intron DNA patterns may modulate transcription, producing an alternatively spliced mRNA and therefore a different protein. The functional value of these biomolecules may change with a variety of factors. This can lead to counterintuitive and mathematically complex scenarios, such as increases in the concentration of a specific mRNA without an increase in the concentration of the associated protein. An additional complication of the system is that these processes are highly dependent on environmental factors, such as pH and temperature.

Is modeling cellular pathways and mechanisms an unrealistic goal? Based on accelerated data collection of genetic and protein expression, new techniques for unraveling regulatory mechanisms, and initial modeling successes at several scales, it appears that modeling is already contributing to our understanding and will grow in importance. At the most fundamental level, biological modeling is chemical modeling—amenable to the existing tools of quantum chemistry, molecular dynamics, stochastic simulation, and differential equations. The difficulty arises from the paucity of quantitative data for the models and from the scale of simulation needed. The scale is large because relevant biochemical reactions take place in solution and require the simulation of water and other molecules present in the environment.

The data and other limitations of conventional simulation techniques have led to a network approach for modeling biology at a systems level. The systems approaches have included binary models (gene on/off, protein on/off, etc.), finite-state (e.g., Markov) models, and fuzzy models. The range of modeling approaches is diagrammed in Fig. 4.2. Our research group at LLNL has utilized differential equations, stochastic simulation, and fuzzy modeling. We are currently focusing on fuzzy models for integration of genomic and proteomic data because they allow a direct conversion of biological hypotheses into mathematical constructs and computer models.

The ability to pursue systems-level biology depends on mathematical constructs, computer models, and biological data. The ability of modern biotechnology to generate genomic and proteomic data is significant. The amount of DNA sequence data available is growing at a significant rate. DNA sequence data for the human genome and many other organisms are available (see Tables 1.3 and 1.4). With the throughput of the large DNA sequencing centers, it is now possible to draft the sequence of an entire bacterial genome in less than a day. The availability of these data is changing the approach to many biological research questions.

In principle, knowing every gene in an organism provides the sequence of every protein that organism can produce. A nerve cell and a white blood cell in a human are distinguished because a different subset of genes is expressed to produce RNA messages for protein synthesis. Expression patterns also change with time and environment. When a nerve cell receives a signal from another cell

across the synapse, there is a change in the genes that are expressed. The changed expression results in protein products that signal the next cell in the brain's neural network.

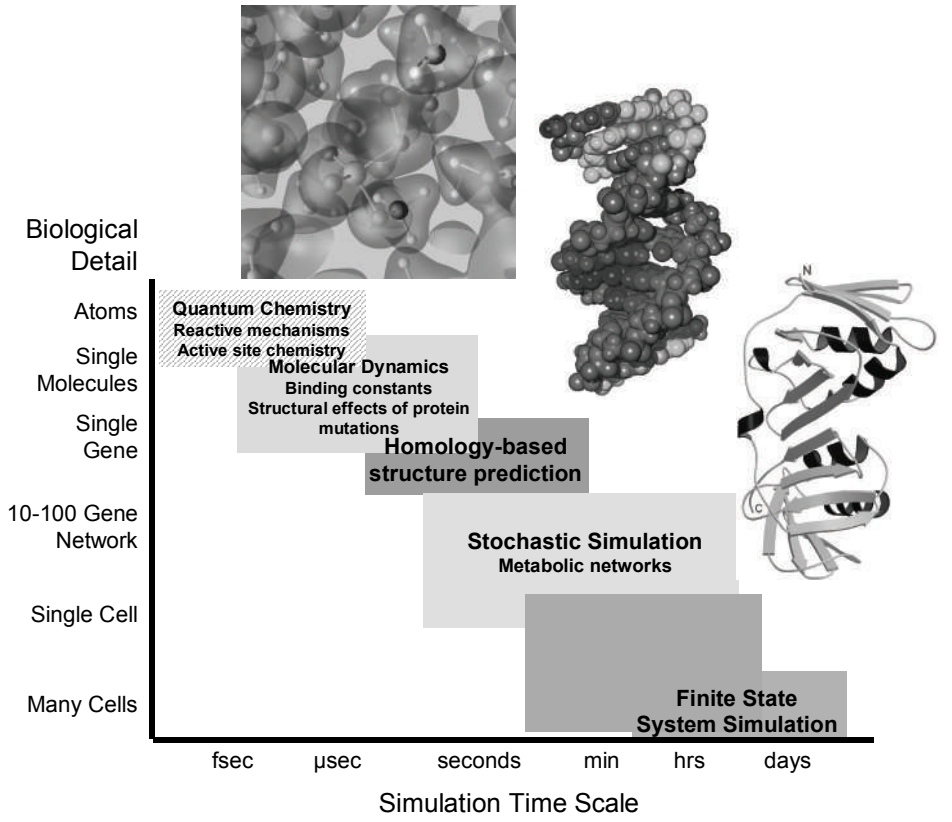


Fig. 4.2. A variety of simulation tools are needed to address the many time and spatial scales that are relevant in biology. Small spatial simulations are dominated by chemical reactions and larger simulations are similar to complex network analysis.

We define a pathway as a series of reactions that perform some function for the cell. For example, the breaking down of starch to produce energy is a pathway. An enzyme catalyzes each step of the pathway. The amount of the enzyme is controlled by the expression of its structural gene, which is in turn controlled by the regulatory genes associated with the pathway. A pathway does not have to be sequential. Reactions can occur simultaneously and there can be branching. Pathways are generally self-regulated by feedback loops. A pathway to digest a particular molecule typically turns itself off when that molecule is no longer present. Some pathways turn on if they sense a different external temperature or chemical concentration.

Consider *Yersinia pestis,* the bacterium that causes plague in humans. As shown in Fig. 4.3, the virulence mechanism of *Y. pestis* is not activated when the bacteria are living in fleas at 26°C. When this bacterium enters human hosts through a flea bite, its temperature increases to 37°C, the calcium concentration falls, and the bacterium begins to produce proteins associated with virulence. Figure 4.1 shows the interaction of genes (DNA), message (RNA), proteins, function outcome, and environment in a pathway. Pathways are not independent; they often share enzymes and can stimulate or suppress each other, and they are not necessarily confined to a single cell.

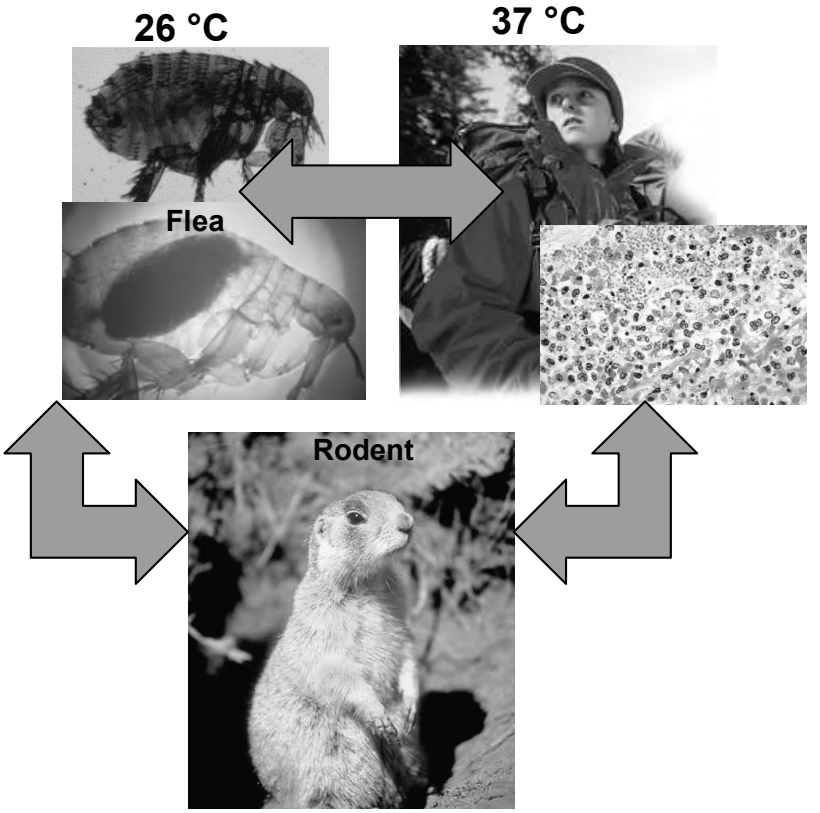


Fig. 4.3. Natural life cycle of *Yersinia pestis,* the bacterium that causes plague in humans. Prairie dogs or rats may have the infection. The *Y. pestis* bacterium is transferred to other rodents via insects (typically the flea). When humans are infected, the disease can take several forms, including bubonic plague (inflammatory swelling and discoloration of lymph nodes known as buboes) and pneumonic plague (infection of the lungs—a highly transmissible form of the disease among humans). A section of infected lung tissue is shown. PHIL™ photo 741 (histopathology of lung in fatal human plague) courtesy of Marshall Fox, Centers for Disease Control. PHIL™ photo 969 (prairie dog) courtesy of Centers for Disease Control.

Our group at LLNL has begun to apply fuzzy logic to several aspects of our genomic approach to understanding virulence in *Y. pestis*. Fuzzy logic was selected because it is a framework that can take into account the nature of biological science: a history of linguistic and graphic models and numerically imprecise data, especially for living organisms. Boolean logic is an insufficient representation, but fuzzy logic can mathematically represent the problem in a context biologists can understand and appreciate—i.e., the user interface and utility of the approach are quickly appreciated. A biologist might describe the level of gene expression as low, medium, or high. Fuzzy logic provides a methodology for converting to and from numbers and linguistic values such as "low" and "high." An example lookup table for a gene expression array is given in Fig. 4.4. The problem with fuzzy logic implementations has been the exponential increase in computation needed with the number of inputs and states. Our group is utilizing a more restricted subset of the logic that grows linearly with an increased number of inputs and states. This approach is known as the union rule configuration (URC). It is similar to doing image processing with a subset of morphological operators. Other nonmorphological operators can be approximated by a series of morphological operations.

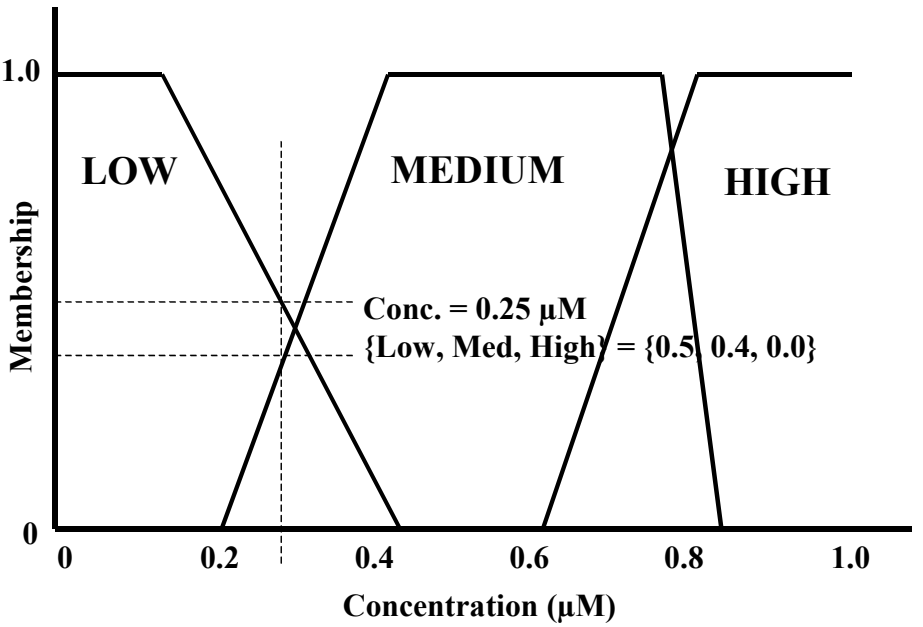


Fig. 4.4. Quantitative values are fuzzified and incorporated into linguistic models with values such as "low" and "high." The linguistic values can be manipulated using fuzzy logic. The linguistic output can be converted to numerical values.

Traditionally, a pathway is experimentally studied by a series of knockout experiments. In each experiment, a single structural or regulatory gene is mutated or removed from the genome. In a simple example, if a pathway is responsible

for digesting fructose, its failure means the cell no longer has access to fructose as a source of energy and cannot grow if it is fed only fructose. So, if a gene is mutated and the cell continues to survive, the mutation did not affect its fructose metabolism pathway. What biologists have found repeatedly is that different combinations of genes may lead to different results. For obvious evolutionary reasons, pathways often have redundant branches. The loss of one gene may reduce efficiency or have no effect at all. So while losing either gene A or gene B might produce no observable changes, losing both genes A and B would result in cell death. In the case of regulatory genes, the situation is more complex. Experiments have shown that there are cases where losing either gene A or B may kill a cell, but losing both regulatory genes will not! So, fully understanding a pathway requires testing every possible case of gene expression under all environmental conditions. Even a very small pathway contains about 10 genes or 2^{10} possible gene deletion experiments if every combination of genes is deleted. Some human pathways result from the interaction of hundreds of different genes, and each cell contains hundreds of interconnected pathways.

Even given the whole genetic code, it is obvious that traditional molecular biology would take centuries to tackle even the 470 genes of the smallest known genome of any free-living organism (the bacterium *Mycoplasma genitalium*). Functional genomics is a group of massively parallel, high-throughput experimental and computational techniques used to study the function of every gene in an organism. This includes measuring the mRNA concentration for every gene, determining the function and structure of every protein, and finally being able to model the interconnected regulatory network of the whole cell. While the individual subsystems will be refined as technologies improve, the overarching approach shown in Fig. 4.5 will likely persist for the foreseeable future. The specific methods and instrumentation of this approach are presented in the second part of this book.

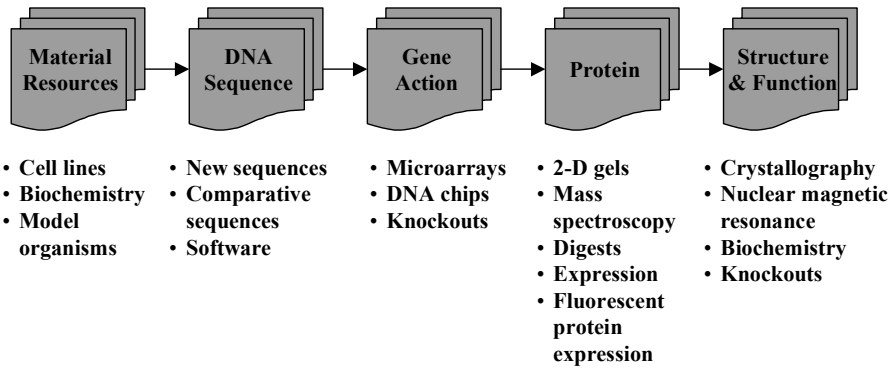


Fig. 4.5. Platform for genomic study of organisms. The technologies for achieving each piece of the platform are changing rapidly. The collective system approach is likely to remain unchanged for years.

The high-throughput approach begins with DNA sequence information. If an entire genome or section of a genome is accurately sequenced, the data can be used to predict the location of open reading frames. An ORF is a likely location for transcription and therefore a putative gene. Recall that some codons (three adjacent DNA bases that code for an amino acid) are start and stop locations for transcription enzymes. Unfortunately, locating start-to-stop regions is not sufficient to identify an ORF. Start and stop codons are useful checkpoints, but most ORF-finding software tracks the statistics of intron DNA as well as representative patterns for parts of proteins. Multiple scenarios of potential genes and proteins can be predicted using Markov state models. Selection may be done to maximize the probability a specific ORF is correct, given statistics on DNA sequence, amino acid chain data compared with known proteins, and comparisons with known genes in other organisms.

The ORF data can be used to design arrays that can detect which mRNA molecules are generated from transcription. The ORFs selected via computer software are often cut from the genome and copied using PCR. Gene expression is monitored using the PCR-generated DNA to capture the anticipated mRNA via hybridization experiments. Levels of mRNA do not always correlate with the amount of protein produced, so a variety of techniques are used to determine the proteins being produced. Two very common techniques for determining protein expression are mass spectrometry and two-dimensional polyacrylamide gel electrophoresis (2-D PAGE). Genes and proteins are usually not measured in an absolute context like concentration but, instead, are usually measured differentially in comparison with a reference. The comparison of gene expression level for a *Yersinia pestis* virulence gene with a reference culture at a lower temperature is an example of differential gene comparison. Once the genes and proteins are at least known relative to a reference, the protein must be correlated with a function. For *Y. pestis*, the gene expression differences based on temperature often correlate with gene expression differences in the flea vector and the human host. Protein complexes often perform a specific function. Therefore, the relations among many proteins are needed to completely understand a pathway.

After differential screening of gene and protein expression, there are a variety of steps that can be used to determine function. The specific methods are heavily influenced by the nature of the gene and protein being studied. For instance, knockout mutants and biochemical assays must be designed for the specific gene and protein activity. Our group has had some initial success in using computer models to assist in prioritizing genes and proteins to be studied. For instance, one of our initial virulence factor screens resulted in the discovery of five genes that had not been previously implicated in virulence. We used standard software tools to compare the DNA and amino acid sequences of our putative virulence genes and proteins with those from other organisms. In addition, an in-house code developed for structural alignment of the protein with segments of other well-studied proteins provided clues about function. Figure 4.6 shows one of the structural models from our computational studies of genes and

proteins associated with virulence. The function clues allow us to design biochemical and knockout experiments to directly test the hypothesis.

An integrating theme is the computer tools that link the different elements of the system. We are entering an era where the volume of biological data already posted on the web will require decades to analyze and integrate. As the information tools improve, we will find the informatics driving both discovery-based and hypothesis-driven biological experiments. The interested reader may want to track the progress of several groups, including the Institute of Systems Biology (http://www.systemsbiology.org/) founded by Leroy Hood.

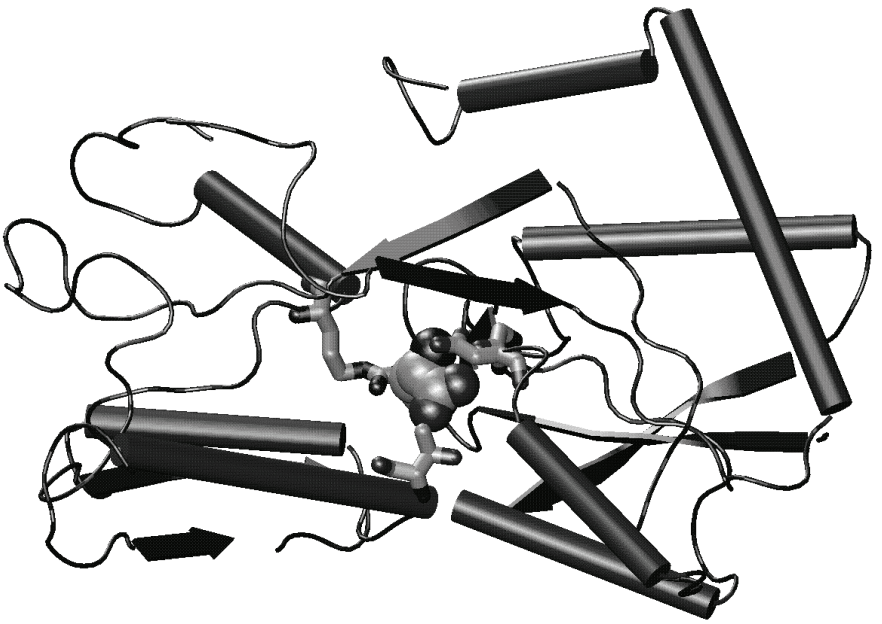


Fig. 4.6. Model of one of the proteins encoded by a newly discovered thermally regulated putative virulence gene from *Yersinia pestis*. Coils, arrows, and cylinders represent coils, beta strands, and helices, respectively. Photo courtesy of Daniel Barsky, Adam Zemla, and Krzysztof Fidelis, Lawrence Livermore National Laboratory.

5

DNA Sequencing

DNA sequencing has become a high-throughput process for determining the ordered base pairs in a strand of DNA. Manufacturing techniques, including statistical process control, are now routine. An example throughput metric is the number of DNA bases sequenced per day per dollar. Some sequencing centers report their daily and monthly production online. Commercial organizations such as Incyte Genomics (http://incyte.com) and Celera Genomics (http://celera.com) have significant DNA sequencing capacity. Public sequencing activities for the Human Genome Project were mostly conducted in five large centers: the Sanger Center, the Washington University Genome Sequencing Center, the Department of Energy/University of California Laboratory Joint Genome Institute, the Whitehead Institute for Biomedical Research at Massachusetts Institute of Technology, and the Baylor College of Medicine.

There are several sites on the worldwide web that summarize DNA sequencing progress. The European Bionformatics Institute Genome Monitoring Table (MOT) page at http://www.ebi.ac.uk/genomes/mot/ is updated daily with sequencing progress for several eukaryotes. The Institute for Genome Research (TIGR) maintains a list of published microbial genomes and chromosomes at http://www.tigr.org/tdb/mdb/mdbcomplete.html. The NCBI also has a list of microbial DNA sequence, including completed microbial genomes from both archaea and bacteria.

5.1 Sequencing Approaches

In most DNA sequencing approaches, the DNA sequence is assembled from many shorter, overlapping subsequences. The sequence of a strand less than a couple of thousand base pairs in length is measured using four-color electrophoresis. Our description of DNA sequencing here begins with a simplified description of sequencing chemistry, followed by discussions of electrophoresis and computerized base calling and assembly. The next chapter has an overview of the automation used to increase throughput and reduce the cost of DNA sequencing as well as many other biological applications. Submission of the sequence to the public DNA sequence database is the final step for most publicly funded sequencing projects.

Beginning with a purified template of single-stranded DNA, the second complementary strand is generated using an enzyme known as DNA polymerase. Deoxynucleotides (dNTPs) for each of the bases are provided in solution so that the polymerase enzyme can assemble them along the template. First, a small chain of dNTPs, called a primer, is annealed to the template DNA. The primer is designed to be at a unique reference position on the template. Assembly of the complementary strand begins at the primer site and continues toward the 5′ end of the template. As each dNTP is added, a 3′-hydroxyl group is left available for the next dNTP in the growing complementary second strand. The clever modification of a dNTP so that no 3′-hydroxyl group is available for chain extension provides a means to terminate strand assembly. These synthesized molecules, known as dideoxynucleotides or ddNTPs, can also be labeled with a fluorescent dye specific to the base. By balancing the concentration of dNTPs and ddNTPs, an ensemble of DNA strands beginning at the same position on the template DNA but of different lengths can be generated. These strands are also terminated with a fluorescent label specific to the final base in the chain. The two strands are separated using temperature and/or biochemical techniques. Once the "chain termination" or "Sanger sequencing" chemistry has been completed, the DNA sequence can be obtained by ordering the new strands by size and fluorescent label, as seen in Fig. 5.1.

One popular alternative to dye terminator chemistry uses fluorescent labels at the primer site and is known as the dye primer method. The template DNA is separated into four aliquots and a fluorescent label is incorporated as part of the primer. (Aliquots are quantitative fractions of a solution.) Chain extension using dNTPs is performed as described earlier. However, chain termination differs in that ddNTPs for only a single base are used in each aliquot and the ddNTPs are not labeled. After chain extension and termination in the dye primer method, each aliquot has fragments labeled at the primer site and terminated at the same base. The DNA sequence can be obtained by electrophoretic sizing of the new strands separately for each of the four aliquots (DNA bases). The four electropherograms are then computationally combined to determine the DNA sequence. If a different color of primer label is used in each of the four aliquots, the aliquots can be pooled before DNA sizing, and the process continues as with dye terminator chemistry. One of the principal advantages of dye terminator chemistry is the ability to do the chain extension in a single aliquot.

5.2 Instruments

The method of choice for determining the size of the DNA strands is four-color electrophoresis. Electrophoresis to separate biomolecules began with the Nobel Prize-winning work by Tiselius on proteins in 1937. In DNA sequencing, electrophoresis uses the force from an applied electric field to move the negatively charged single-stranded DNA molecules through a separation medium. DNA has a roughly constant charge-to-mass ratio. The sieving medium

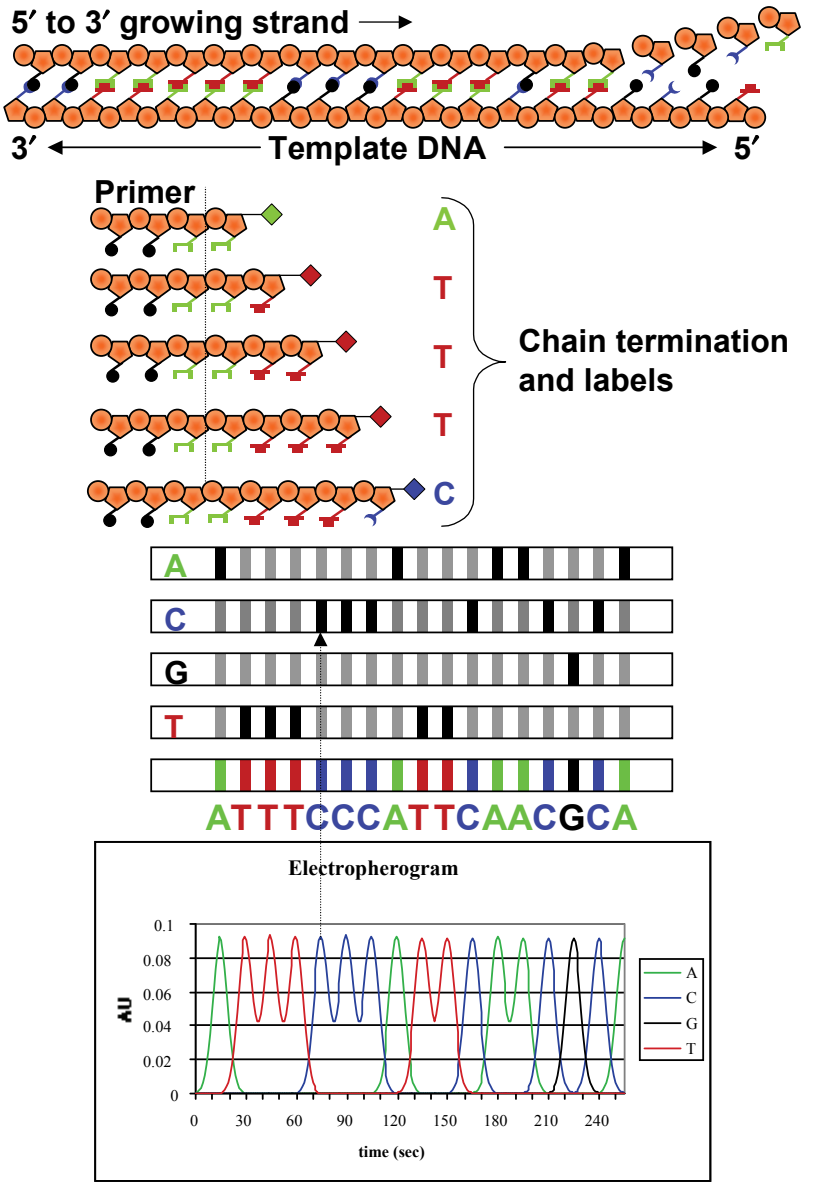


Fig. 5.1. DNA sequencing using four-color electrophoresis and the Sanger chain termination chemistry. A template of DNA is copied into many random-length pieces of DNA that start at the same primer and terminate with an optical label specific to the last base in the chain. The strands are separated by length using electrophoresis, allowing the DNA base sequence to be deduced.

and the electric field are engineered to produce differential drift velocities that are proportional to the length of the DNA and usually for DNA are less than a thousand bases long. Limitations arising from diffusion and convection led to the use of polyacrylamide or agarose sieving media in many instruments. High-throughput instruments have utilized several approaches, including gels spread thinly across slabs of glass and gels injected into glass capillary arrays, glass microchannels, and plastic microarrays (Fig. 5.2). Each of these types of instruments is in use today, with instruments using arrays of glass capillaries currently dominating sequencing in the large centers. We describe here the electrophoresis process, with a bias toward the capillary systems. Figure 5.3 is a schematic of a generic capillary electrophoresis DNA sequencing instrument.

Before DNA can be loaded into an electrophoresis instrument, a gel is pumped from the data collection end of the system into the capillaries using a syringe-type pump or compressed gas at over 1,000 psi. Usually several capillary volumes are pumped through the system. The excess gel is aspirated from the sample end of the capillaries. The loading well around each capillary entrance is then filled with a buffer solution. The sample (usually a few microliters) is introduced into the loading buffer. The goal in DNA loading is to create a thin stack of DNA in the gel. If the stack spreads out before electrophoresis, the resolution of the system degrades and fewer DNA bases can be deduced from the run. In electrokinetic injection, an electric field (with the anode at the detection end of the system) is applied and the negatively charged DNA in the sample is moved into the gel in the capillary. The loading buffer is then aspirated out of the system and a running buffer is introduced to promote migration of the DNA down the capillary. Commercially available sequencing instruments now require very little operator intervention. In one of the commercial systems, the Applied Biosystems 3700 DNA analyzer shown in Fig. 5.4, a robotic arm performs sample loading and some of the aspiration operations. The input sources for all high-throughput commercial systems are standard laboratory 96- or 384-well plastic microtiter plates.

An innovative alternative to injecting the sample into the gel electrically is to create a narrow cross-channel (see Fig. 5.5) that moves DNA across the gel in the sequencing channel. After loading, the stack of DNA in the channel is roughly the width of the cross-channel. After the cross-channel is isolated electrically, electrophoresis begins in the sequencing channel. Although electric fields have been used to move the DNA for this loading scheme, the geometry has mechanically isolated the DNA for electrophoresis. The cross-channel loading has been implemented in several systems using microelectromechanical systems (MEMS) techniques, including lithographic patterning of the channel and cross-channel. Compared with electrokinetic sample loading, cross-channel loading requires additional electrical circuitry. The voltage and current must be controlled in both the main and the crossing channels to minimize diffusion of the sample into the gel. Although it has had several promising demonstrations, this method is currently not available in any commercial system.

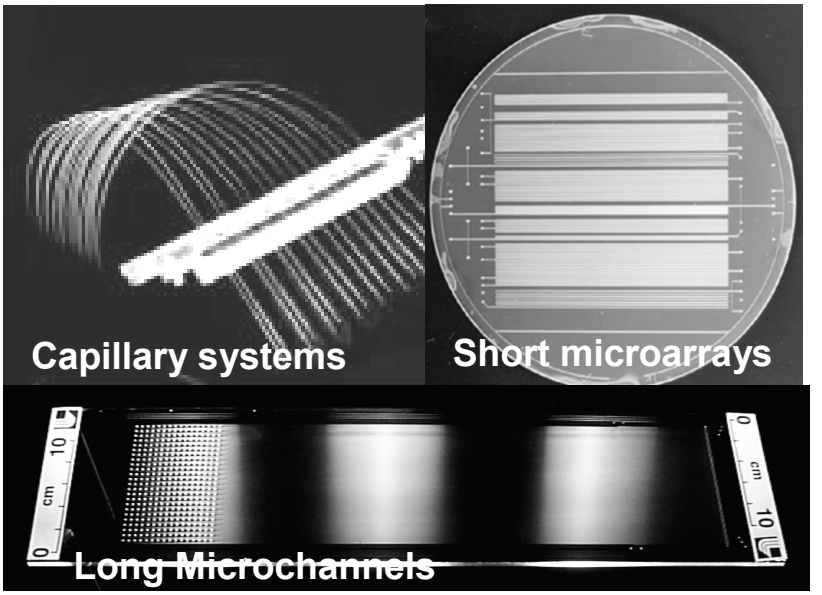


Fig. 5.2. Three types of sequencing channels: glass capillaries, short plastic microarray (with cross-channel loading), and long (50 cm) glass microchannel plates.

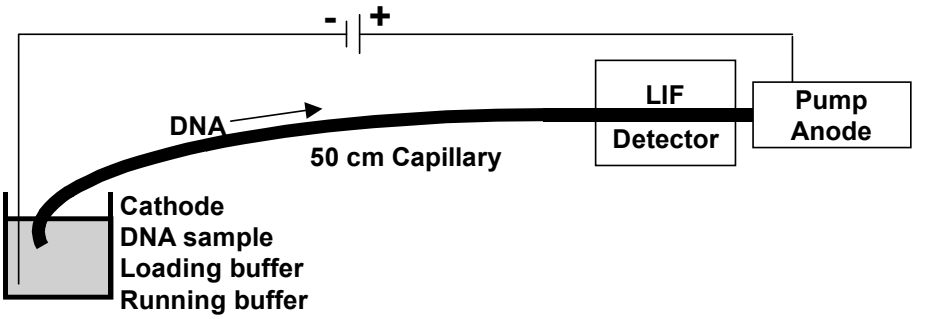


Fig. 5.3. Capillary-based electrophoresis DNA sequencing instrument with a laser-induced fluorescence (LIF) detection system. The separation medium is pumped from the anode toward the cathode. Excess medium is aspirated from the cathode well. A sample of DNA is introduced near the capillary at the cathode. The negatively charged DNA is moved through the capillary by electrophoresis and detected by LIF.

Fig. 5.4. Applied Biosystems 3700 96-channel capillary electrophoresis DNA sequencer with robotic loader detail shown. To minimize evaporation and well-to-well contamination, the 384-well microtiter plates are sealed. The two aspiration/dispense tips puncture the seal when a capillary is available for sample introduction.

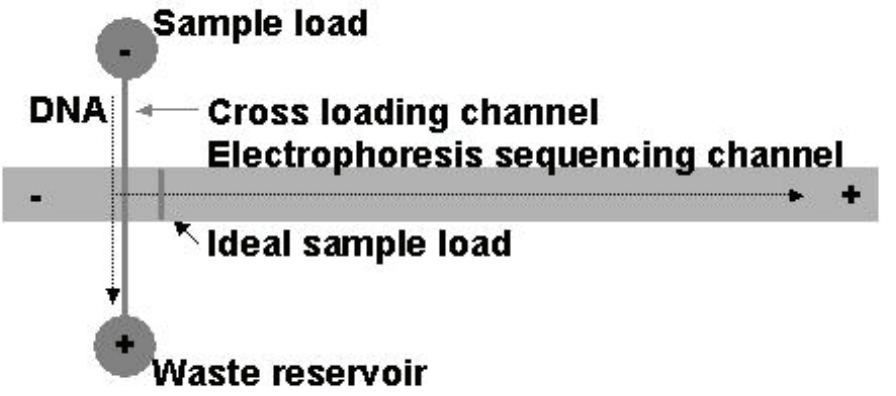


Fig. 5.5. Cross-channel loading for microchannel electrophoresis. A typical electrophoresis or "running" channel is 2,500 μm² in cross-sectional area (similar to commercial glass capillaries). The width and depth are usually optimized for the detection method. The cross-channel is much thinner and may be as narrow as 1 μm.

Once the DNA sample is loaded into the gel in the capillary, an electric field of 100 to 200 V/cm is applied to move the DNA through the gel toward the detector. The shorter fragments and surplus primer from the enzymatic reaction arrive first at the detector. Surplus template DNA without attached fluorescent labels can contaminate a run by arriving at the detector at the same time as shorter DNA fragments that have an attached label that slows migration. It is also possible for the single-stranded DNA to fold on itself and cause poor electrophoretic separations. As with aliasing in an analog-to-digital sampling system, it is not possible to determine from the electrophoresis data alone if a peak is due to a short fragment or a longer fragment that has folded on itself and migrates faster than it would without the fold. To reduce some of these noise sources, the samples are often purified before loading; the gel and buffer chemistries are engineered to keep single-stranded DNA from hybridizing to complementary DNA strands; and the temperature and running conditions are optimized for electrophoretic resolution.

As an example, the Applied Biosystems 3700 DNA analyzer shown in Fig. 5.4 often loads 2 µl from a 25-µl source in microtiter format with a 30-second electrokinetic load at 1 kV. This represents approximately 20 ng of DNA loaded onto the column. The run voltage is often 6.5 kV. The 50-cm long and 50-µm inner diameter capillaries are filled with a polydimethylacrylamide (PDMA) sieving gel. The run duration is about 2 hours for 500 bases with single-base resolution. Similar run parameters are used for other glass capillary systems, including the MegaBACE 1000 DNA sequencer shown in Fig. 5.6.

Fig. 5.6. MegaBACE 1000 DNA sequencer with 96 glass capillaries. New 384-capillary machines are now available.

A standard measure of resolution of DNA sequence is the ratio of the peak width to the peak spacing. The peak width is usually taken as the full width at half the maximum value. The number of bases at which this resolution is unity is known as the crossover point for the system (this is similar to the Rayleigh diffraction limit of an optical system). The signal-to-noise ratio and other parameters influence performance, but the crossover point is a good measure of the inherent capability of an instrument to resolve DNA fragments differing in length by one base. When the DNA fragments are resolvable, the fluorescent labels on the terminating ddNTP will indicate the last base on the fragment of interest and there will be many copies of that label (one for each DNA fragment), allowing for optical detection.

For most of the commercially available ddNTP labels, an argon-ion laser (488 and 514 nm wavelengths)-induced fluorescence (LIF) system is used for detection. The optics of the two most common detection systems are a scanning confocal microscope with photomultiplier tubes and a fixed charge-coupled device (CCD) imaging system that collects multiple wavelengths simultaneously through a prism. The detection system can operate through the glass capillary or through a liquid that creates a "sheath flow" around the end of the capillary. Although mechanically complex, the sheath flow eliminates refraction through the capillaries and allows the laser to illuminate many channels simultaneously without scanning. Numerous other detection methods have been proposed, including electrochemical ones, but they have not been adopted in commercially available sequencers.

The fluorescent labels used in most DNA sequencing instruments have emission spectra that overlap. Example spectra are presented in Fig. 5.7. Color correction is needed before beginning data analysis to detect DNA bases. Other characteristics of the electrophoretic separation that must be corrected include length-dependent changes in velocity of the DNA fragments and velocity differences that are due to the four different fluorescent labels. For a fixed detector system, the DNA that arrives later appears to have a broader distribution. This is due to the slower velocity and is not necessarily a more spatially dispersed ensemble of DNA fragments. Corrections for the velocity-dependent artifacts are referred to collectively as mobility correction.

In general, there are two approaches for color and mobility correction. The first approach is system calibration with known samples using the same loading and running parameters that will be used with the unknown samples. The second approach is system compensation by estimating the parameters of a correction model dynamically from the data. Because of the high-throughput nature of DNA sequencing, the system parameters remain fixed for many runs, making the first approach (the use of calibrated test runs) preferred. Ideally, the traces from each of the DNA fragments would have the same shape, samples would be evenly distributed by size in the electropherogram; and the fluorescent labels would not overlap spectrally. Unfortunately, the electropherogram has many distortions, and the signal environment is similar to a digital communication system with fading channels and crosstalk. When the biochemistry or temperature is

suboptimal, the folding of the single strand of DNA can also change the electropherogram as if the signal had multipath artifacts.

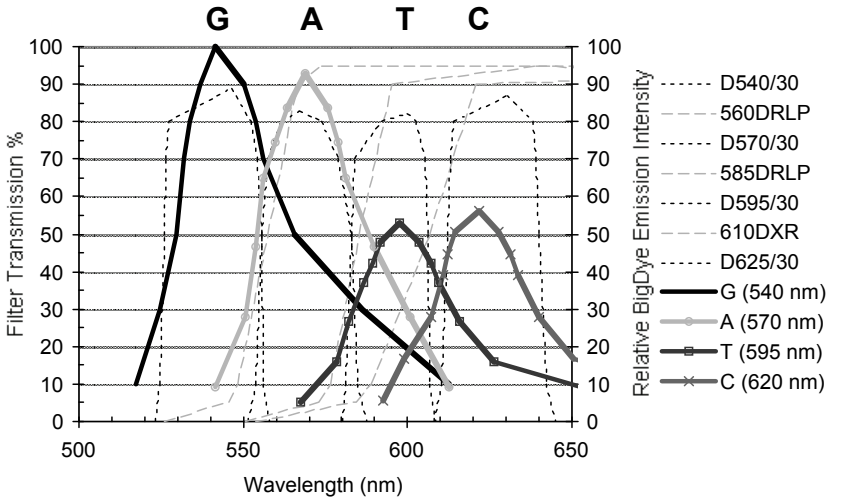


Fig. 5.7. Example spectra for four-color fluorescent DNA labels and the optical transmission characteristics of the Lawrence Livermore National Laboratory microchannel DNA sequencer. The dichroic mirrors are designated xxxD, where xxx is the cutoff wavelength in nanometers. The optical bandpass filters are designated Dxxx/xx, where xxx is the center wavelength and xx is the bandwidth in nanometers. The fluorescent terminator labels were sold under the trademark PE Big Dye.

Figure 5.8 shows an electropherogram before and after color correction, background subtraction, and filtering for mobility and shape correction. The ultimate metric to compare against is not the homogeneity of the electropherogram, but rather the accuracy of the assignment of DNA bases. This process is known as base calling and the state of the art has been defined by the early work in industry for supporting the ABI 373 sequencing instrument and more recently by Phil Green's group at the University of Washington with codes named Phred and Phrap. The Phred and Phrap codes are probably the gold standards for base calling, assessing the quality of a base call, and assembling sequence data from many DNA fragments into an estimate of a longer contiguous DNA sequence. The code performs all of the necessary filtering mentioned earlier, does peak tracking to identify potential locations of bases in the electropherogram, and then performs a model fit to call a base and to look up a probability of error in a calibrated table for the particular instrument. "Phred 20 bases" has become the industry standard for identifying the number of bases that have roughly a 1 in 10,000 probability of error.

Electrophoresis allows determination of the sequence of template DNA of lengths of about 1,000 bases. How are entire genomes with over a billion bases

sequenced? The sequence of longer segments of DNA is assembled from many overlapping shorter sections of DNA. The average number of times each base appears in a different DNA fragment is called coverage. Sequencing projects range from "draft" quality with 3× to 5× coverage to "finished" quality, with about 10× coverage and on average less than one incorrect base call per 10,000.

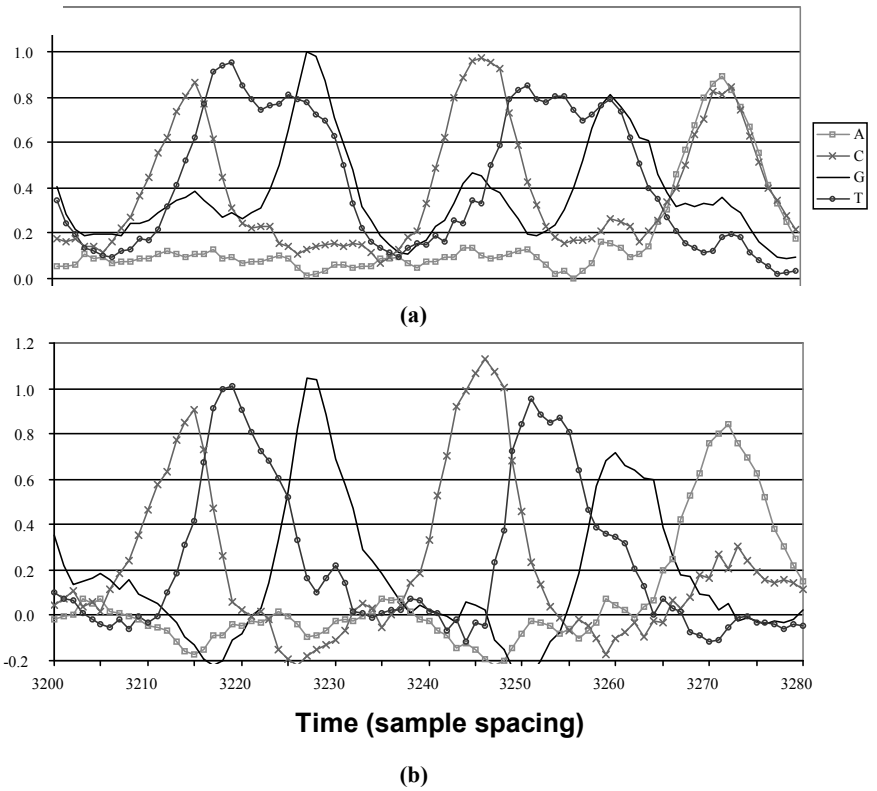


Fig. 5.8. Electropherograms before and after the color correction step. The sample spacing is roughly 1 second. Note that the width of the peaks and peak-to-peak spacings are not uniform. (a) Normalized, smoothed electropherogram, (b) color-corrected electropherogram.

The overlapping, but not identical, DNA fragments are the input for the sequencing chemistry. These fragments are usually generated mechanically or biochemically. In the mechanical approach, a purified source of DNA is sheared or fragmented into many random-length subsequences, often by forcing the DNA through a pore. In the biochemical approach, multiple restriction enzymes that digest DNA at fixed sites are used to generate different fragments, depending on the order in which the enzymes are applied. In the Human Genome Project, both techniques have been used.

There are two different strategies on how to obtain a 1,000-base template DNA from multimegabase chromosomes. In the mapping approach, mechanical shearing is used on the chromosomes, with fragments selected to be about 150,000 base pairs. The fragments are coarsely assembled into a rough map that identifies the location of the particular DNA fragment on the chromosome. Sometimes it is necessary to "end sequence" these large fragments to facilitate assembling the map. The size of the fragment and the limited sequence information are used to identify a tiling pattern that covers the section of interest on the chromosome. Once the 150-kb clones are mapped, the selected clones in the tiling pattern are fragmented randomly into DNA short enough (less than 2 kb and known as subclones) to use as templates for electrophoretic sequencing. In sequencing smaller genomes, such as microbes, and in the Perkin Elmer(PE)/Celera "shotgun" approach to human genome sequencing, the mapping step is omitted and the whole genome is fragmented. Eliminating mapping saves time and effort, but assembly of the significantly larger number of DNA fragments is more difficult.

For most publicly funded projects, the final step for sequence data is submission to the public database, GenBank. GenBank is managed by the National Center for Biotechnology Information at the National Library of Medicine (NLM) of the National Institutes of Health (NIH). There are a variety of ways to submit data. In the end, a unique accession number is assigned to each submission so that the data may be appropriately referenced. High-throughput sequencing centers can also submit data at different levels of completion. The finished sequence is assigned phase 3 status, and draft data are either phase 0, 1, or 2, depending on the base call quality and the number of gaps in the data.

5.2.1 Optical detection subsystem example

A photomultiplier tube (PMT) system that was used in an in-house sequencing instrument is shown in Fig. 5.9. A confocal microscope with focal length (F) of 1 cm is scanned across a glass plate with 96 separate 50-cm-long microchannels. The sample is excited with an argon-ion laser bandpassed at 488 nm and imaged through a 1-mm pinhole. Dichroics and a 488-nm block filter keep laser light out of the four PMT detectors. The image shown in Fig. 5.10 is obtained by stacking a single PMT's output for consecutive scans across the glass plate.

5.3 Automation

Automation has been applied to many steps of the DNA sequencing process. Parallel aspiration out of and dispensing into microtiter format plates with 96, 384, or 1,536 wells allows a large number of samples to be processed simultaneously. Examples of microtiter plates are shown in Fig. 5.11. The plastic plates have been standardized at 12.9 by 8.6 cm. Wells are indexed alphabetically along the short dimension and numerically along the long dimension. A 96-well

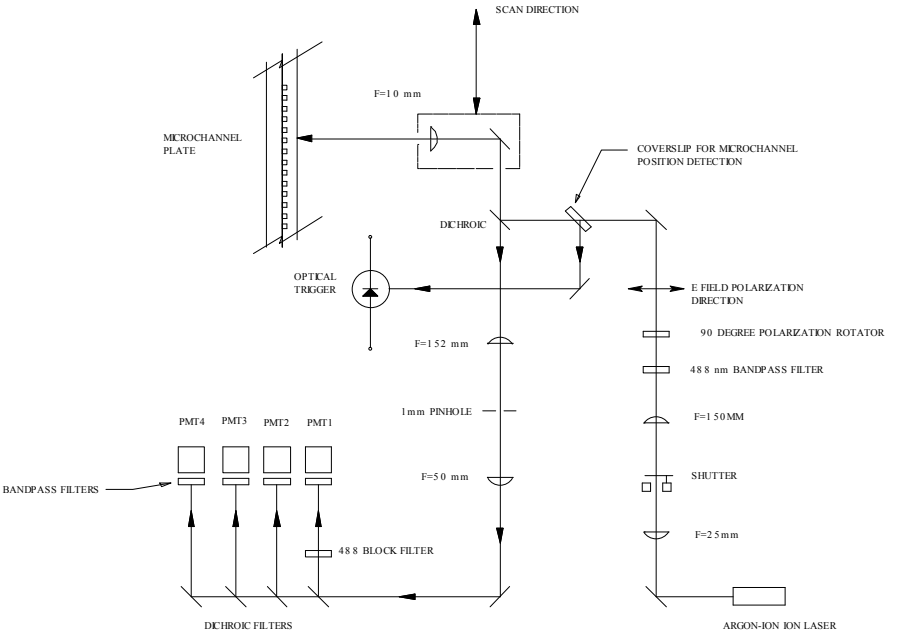


Fig. 5.9. Optical subsystem of the Lawrence Livermore National Laboratory microchannel DNA sequencer. The optical characteristics of the filters are given in the legend for Fig. 5.7.

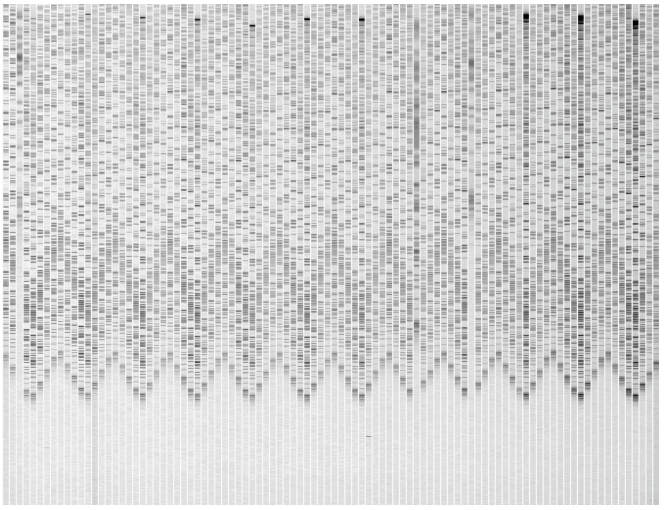


Fig. 5.10. Output of a single PMT in the Lawrence Livermore National Laboratory microchannel DNA sequencer. Each column represents a different DNA sample, with the staggered output due to the microtiter format of the input wells. The first (fastest) DNA to arrive includes short DNA fragments and template DNA. Larger fragments arrive later (toward the top of the image).

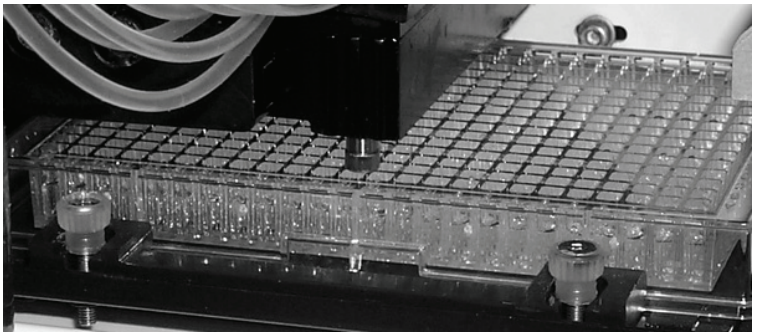


Fig. 5.11. Microtiter plate example of a 384-well plate. Standards for the plate dimensions and well shapes have allowed automation instruments and biochemical protocols to be developed for higher throughput.

plate therefore uses A, B, …, H and 1, 2, …, 12 as the indices. The "first" well is A1 and the "last" well at the other corner is H12. Similarly, a 384-well plate ranges from A1 to P24. The plates are available in different depths to help accommodate the effects of volume and surface area on the biochemistry. As the number of wells increases in each plate, the reduced cross-section of each well requires more and more accuracy from the automation systems. A 384-well plate, for example, has 4.5-mm center-to-center spacing of the wells. Flexible tips are often used so that slight misalignments do not damage the plates or the robots.

Automatic sample aspiration and dispensing can be done for entire plates in one step using arrays of machine-driven pipettes. Sometimes fabricated as gangs of piston-driven syringes, parallel dispensing robots are very useful for setting up reactions. See Fig. 5.12 for a 96-channel aspirate/dispense robot. It is tempting to consider constructing in-house dispensing systems. The implementation of many parallel channels is often limited by the ability to provide consistent pressure and volumes across the entire array.

Since aspirating and dispensing of fluids are not the only tasks that benefit from automation, general-purpose robots with a moving gang of fluid-handling tips represent a useful alternative to highly parallel systems. The Packard Multiprobe shown in Fig. 5.13 has been very useful in LLNL laboratories. When evaluating general laboratory automation systems, consider:

- Tips: Use flexible tips that can be changed for handling different volumes as well as disposable tips (especially if sterilization is not an option and contamination is a concern). There are also many new small-volume tips available, including piezoelectric nozzles.
- Software: It should be simple to specify the aspirate/dispense steps, positions, and other actions that are needed for multiple microtiter plates on the deck. In order to prevent significant mistakes, it may be necessary to add sensors that can monitor dispensing levels; read barcodes to

identify plates; and interface with a computer database to verify protocols, samples, and timing.

- Functionality: Many operations beyond aspirating and dispensing are now possible on the decks of these robots; these include vacuum extraction, PCR, stirring, chilling, and moving the plates to new positions (including transfer systems to other decks).

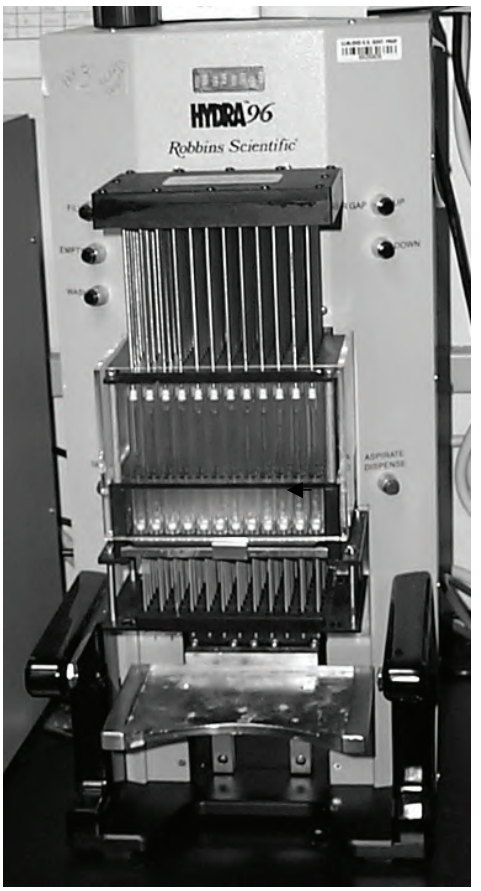


Fig. 5.12. Parallel 96-channel dispensing robot—the Robins Scientific HYDRA™ 96. Each channel is basically a precision syringe that is controlled by gang manipulation of the plunger.

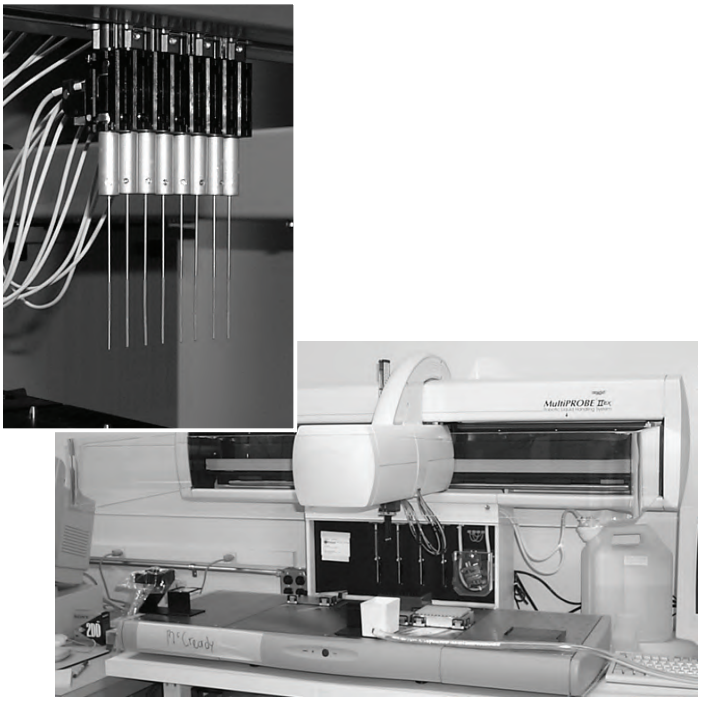


Fig. 5.13. Packard Multiprobe dispensing robot with vacuum extraction, washing, and disposal stations on the deck. Inset shows a view of flexible dispensing tips on a similar Tecan robot system.

For DNA sequencing, samples remain in microtiter plates for many operations, including centrifugation, thermal cycling, template purification, and sequencing reaction set up. Each of these steps can therefore be done to many DNA samples in parallel. Only a few nanograms of DNA would be required for DNA sequencing if it were possible to reliably make and handle volumes that small. For an example of a high-throughput instrument, see the thermal cycling system in Fig 5.14.

With many molecular biology instruments now accommodating microtiter format plates, tracking and moving plates around the laboratory has become important to efficiency. Plate shuttling is often accomplished with a conveyor belt that has stations along the track. Shuttling can also be done with a robotic arm that picks up and places plates on subsystems around the arm. Figure 5.15 shows a small robotic arm that moves plates off and onto a set of plate "hotels." The nearby plate-filling robot dispenses reagents into each plate before the plate is returned to the "hotel." Larger robots can move plates among several stations, including dispensing, reaction set up, and thermal cycling stations.

Fig. 5.14. Thermal cycling instruments that can perform up to 384 PCR reactions simultaneously (four 96-well microtiter plates). There are similar instruments in 384-well format for 1,536 simultaneous reactions.

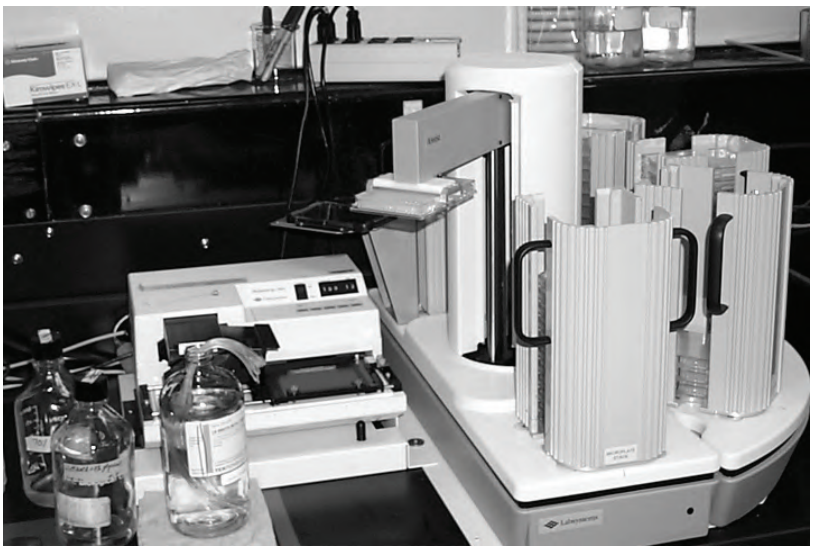


Fig. 5.15. Rotating arm robot that can remove and return plates to four "hotel" stacks as well as a plate filling station.

The first step in many DNA sequencing centers is the growing of many copies of bacteria with an inserted piece of DNA on flat culture plates like those shown in Fig. 5.16. As the bacteria replicate, copies of the inserted DNA are made. "Picking" robots can harvest bacterial plaques and colonies into microtiter plates. This requires sophisticated imaging systems to identify the location of the bacteria to be harvested and high-speed positioning systems that can direct the tip that picks up the bacteria. The picked DNA is then transferred to a microtiter plate. Figure 5.17 shows a robot that uses a rotating picker over translating plates. Figure 5.18 shows an x-y-z robot with multiple picking tips. We have had tremendous success with both types of systems. Given the alternative of picking colonies and plaques by hand using toothpicks, these robots represent a tremendous savings in effort.

Recent biochemical breakthroughs in a technique known as rolling circle amplification (RCA) are allowing some of the sequencing centers to eliminate the culture growth step. As both assays and instruments improve, it is important to keep the two technologies matched. In summary, automation has produced significant time and cost savings, reduced sample volumes, improved protocol consistency, and allowed more accurate sample tracking. There are numerous other automated systems used in biology today.

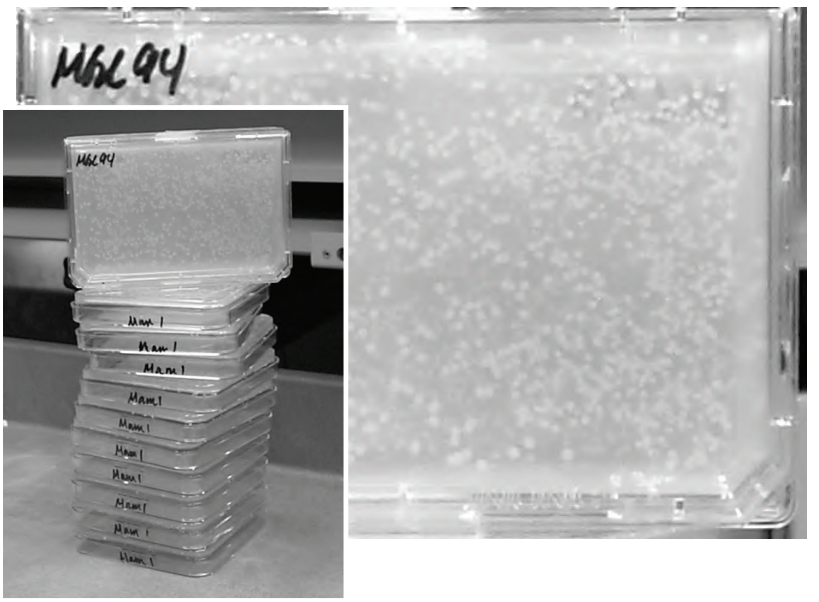


Fig. 5.16. Stack of bacterial gels and an expanded view of a single plate with hundreds of colonies visible. These colonies are identified with automated imaging software and harvested by a "picking" robot. Recent biochemical innovations may obviate the need for this complex, time-consuming, and messy culturing step.

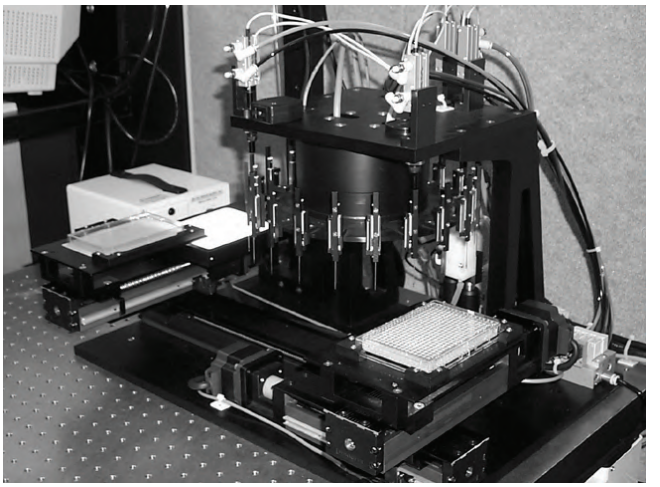


Fig. 5.17. Rotary picker from Norgren Associates with input plate, light source, and imaging system on the left and output microtiter plate on the right. Looking down on the instrument, the rotation is counterclockwise. The sequence is image/locate, pick from plate, place in microtiter, sterilize, and repeat.

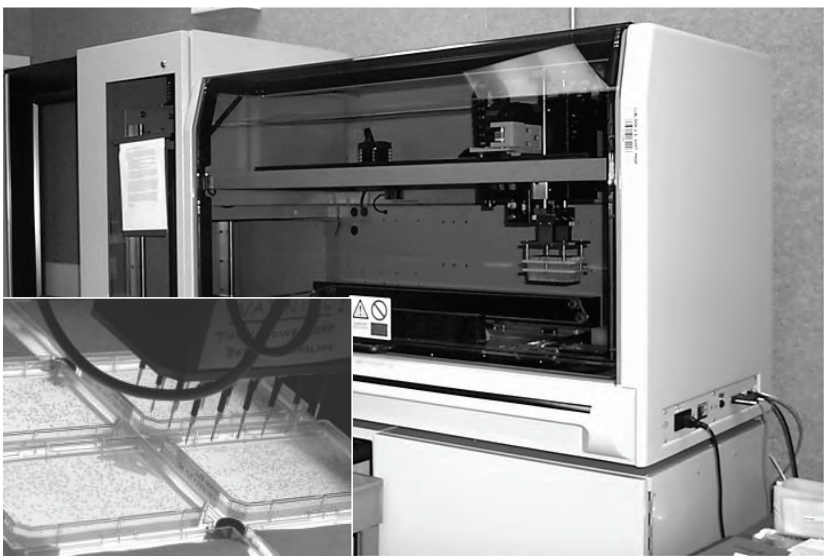


Fig. 5.18. Flexys high-speed colony and plaque picking system using an x-y-z robot and multitip head.

6

Detecting Nucleic Acids

The polymerase chain reaction has already been described as a biochemical method to copy and detect specific sequences of DNA. PCR or restriction enzymes can also be used as the front end to fragment sizing systems such as gel electrophoresis in order to detect DNA. In this chapter we consider three technologies for detecting DNA: DNA chips, DNA microarrays, and affinity capture. All three of these technologies are based on hybridization of reference strands of DNA to the unknown DNA being tested.

DNA chips and DNA microarrays are two-dimensional arrays of reference DNA on glass membranes or microscope slides. Chips are fabricated by synthesizing DNA directly on the substrate. Microarrays are fabricated by printing small volumes of solution containing reference DNA onto the substrate. Affinity capture approaches use beads and other nonplanar surfaces to identify and/or separate hybridized products. Historically, chips used shorter DNA strands than microarrays and affinity capture.

These three hybridization technologies may be applied to detection of DNA in a complex environment, such as detecting a pathogen in an environmental sample. The technologies may also be applied to a specific culture to analyze the genetic response to a stimulus by estimating the change in mRNA levels.

One significant problem with all DNA array experiments is that the hybridization is not perfect. Errors in hybridization become particularly acute at 90% or greater sequence similarity. Applications that include medical diagnostics and regulatory gene expression studies may require discriminating among very similar subsequences. In these cases, redundant probes specific to unique subsequences may be designed that effectively resequence the region. To put this in perspective, a 1% DNA sequence error rate in the human genome is sufficient to map the sequence from human to chimpanzee.

6.1 Environmental Detection Chips

DNA chips are similar to microarrays, but are referred to as "chips" because they are fabricated in a manner similar to computer "chips." DNA chips, which are principally developed by Affymetrix, use oligonucleotide probes: 20- to 30-base sequences. More than 100,000 different sequence probes can be synthesized on a

1.3 cm × 1.3 cm surface (see Fig. 6.1) using photolithographic techniques that originated in semiconductor manufacturing. A series of masks and chemical reactions sequentially add a base to the oligonucleotide probes at specific positions defined by the optical mask.

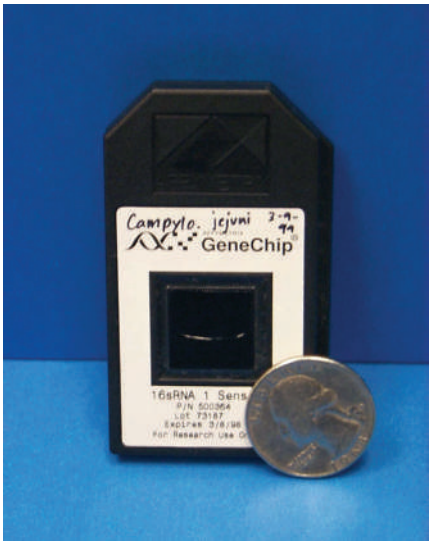


Fig. 6.1. Custom Affymetrix GeneChip™ designed in the laboratory of Gary Andersen, Lawrence Livermore National Laboratory. Current Affymetrix chips can have 250,000 test probes in less than 1 cm².

An expression profiling application is used here to demonstrate DNA chips. For the DNA chip, DNA detection in a complex background is the application. An initial complex sample from blood, tissue, soil, air, or water is collected. The sample must be washed and the DNA (from all sources) separated. Depending on the application, harvesting the DNA may require disrupting the cell membranes to make the DNA available. Once the DNA is separated, the specific DNA of interest can be amplified using PCR. The DNA is labeled with a fluorophore, usually during amplification. Hybridization to the DNA chip results in accumulation of the fluorophore at the location on the chip where a complementary DNA sequence occurs. Knowing the sequence at each position on the chip allows the relative amount of complementary sequence present in the unknown sample to be inferred.

An innovative approach to detecting organisms using DNA chip technology is to mimic the classification used by Woese for the three domains of life: Archaea, Eukarya, and Bacteria. Specifically, the variable and conserved regions of ribosomal RNA are observed to determine placement on the new phylogenetic tree and to detect an organism. Fortunately there are variable regions of RNA

located between conserved regions. For instance, the 1,542-base 16S rRNA contains an 85-base highly variable region that can be used to identify organisms. The Affymetrix GeneChip™ shown in Fig. 6.1 was designed to use this region to identify a variety of bacteria. Obviously the sequence of the variable region is needed to do the chip design.

6.2 Gene Expression Microarrays

Measurement of the levels of gene expression or "transcript profiling" is the application selected for describing DNA microarrays. The variation of cell behavior with changing conditions is a function of differential gene expression. Under a given internal and external state, each gene is copied to mRNA at a particular rate. Thus, in principle, measuring mRNA concentrations under a set of conditions provides a "snapshot" of genetic activity. After the cell is subjected to an external perturbation, the genetic activity changes as some pathways are turned on, some are turned off, others are "tuned" up or down, and many might not change at all. These changes are dynamic, so snapshots after the initial perturbation show continued changes as the first pathways produce intermediate products that stimulate the next wave of pathways. Finally, the cell's response can interact with the environment and other cells. For example, during intense exercise, a human muscle cell runs out of oxygen. The cell responds by activating the much less efficient pathways responsible for anaerobic respiration (energy production without oxygen). Anaerobic respiration produces lactic acid that cannot be broken down fast enough. The accumulation of lactic acid stimulates a signaling pathway that sends a chemical message to a nearby nerve cell, which then sends a "pain" signal to the human brain. When exercise stops because of muscle pain, oxygen becomes available, aerobic respiration resumes, the lactic acid is removed, and the signaling pathway turns off. Obviously, measuring the expression of just a few genes is not enough to characterize these complex changes.

Several technologies have been developed for the simultaneous measurement of the concentration of thousands or more different mRNA sequences. DNA chips and microarrays separate a mixture of mRNA molecules based on knowing their sequences. If the sequence is not known, a method called serial analysis of gene expression (SAGE) can identify an mRNA transcript that did not come from a known gene. Given the rapid acquisition of sequence data discussed earlier, sequence knowledge is typically not a problem. However, it may be difficult to identify what parts of the sequence actually code for genes.

In contrast to DNA chips in which the probe DNA is synthesized, microarrays use cDNA probes copied from an organism's DNA and amplified using a polymerase chain reaction. Each probe can be a section of a gene or the entire gene (about 1,000 bases is typical). Nylon microarrays with radioactive probes have been used for analyzing the simultaneous expression of almost all *E. coli* genes. Glass microarrays with optical fluorescence detection, pioneered by Patrick Brown's lab at Stanford, are now more frequently used because of their

greater sensitivity. In general, microarrays are relatively easy to customize, and public protocols are available on the web. A popular substrate for printing microarrays is a standard microscope slide (approximately 2.5 cm by 8.6 cm with a cover slip area of 22 mm^2).

In a typical DNA chip or microarray experiment, the mRNA is isolated from a sample of cells in the state of interest. The mRNA is then processed by a reverse transcription reaction (5′ to 3′ on a DNA strand), which produces a complementary single strand of DNA. For eukarya, the polyA tail may be used to prime the reverse transcription. For bacteria, random PCR priming may be used to generate a collection of cDNA fragments. A fluorescent marker is attached to the cDNA. Now the cDNA target can bind with a single strand of DNA having the complementary sequence.

DNA chips and arrays have surfaces covered by thousands of spots, and each spot can contain billions of cDNA probes corresponding to a particular known gene. The targets are poured onto the probe array, the targets hybridize with the complementary probes (if present in the array), and the array is washed to remove targets that did not hybridize. The intensity of fluorescence at a spot then indicates how much mRNA with the corresponding sequence was present in the original sample of cells. It is currently not possible to quantitatively determine the original mRNA concentration from this fluorescence signal. Therefore, DNA array experiments usually measure the simultaneous hybridization of mRNA extracted from two samples. A different fluorescent label is attached to the mRNA from each sample [e.g. red cyanine (Cy5) and green fluorescein isothiocyanate (FITC)]. The ratio of fluorescence corresponding to each sample then indicates the relative mRNA concentration between the two samples. Figure 6.2 shows a typical DNA microarray experiment.

Reliable image analysis of microarrays is challenging. Figure 6.3 shows the raw pixel data from the red channel of a microarray made in our lab. Data acquisition considerations are similar to those of other optical imaging systems, including integration time, optical crosstalk of fluorescent labels, and nonuniform illumination. Microarray images often have a very low signal-to-noise ratio. There are generally many more probes than targets, so the spots are only partially fluorescent. Since the technology used to print microarrays cannot form consistent spots, exact *a priori* assignment of spot regions is impossible. Thus, the spots have to be recognized after the experiment from a weak and irregular signal. Also, some targets will bind to the wrong spot, and some will bind to the substrate and fail to be washed off. Stray targets and other sources of fluorescence, including the substrate and coatings, contribute to a significant nonlinear background that must be removed in order to retrieve the signal

A recent innovation that improves spot finding and background estimation and compensation is the inclusion of a third dye, 4′,6′-diamidino-2-phenylindole (blue DAPI), in the microarray experiment. DAPI is a DNA counterstain that binds to the cDNA that failed to hybridize with the target DNA. Thus, the blue channel reveals the shape of each spot, making spot recognition simpler and

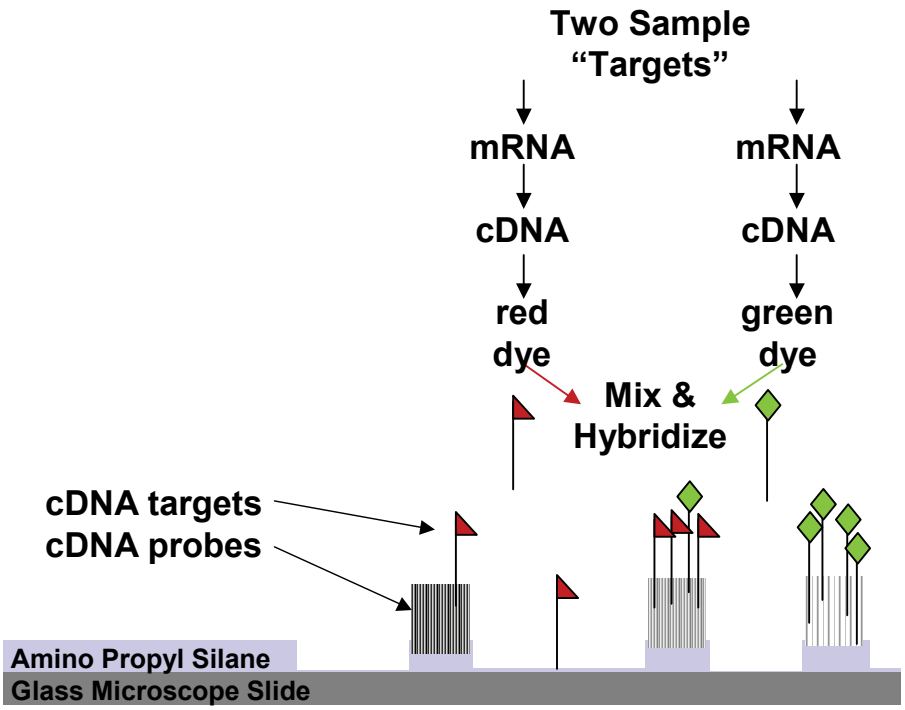


Fig. 6.2. DNA microarray experiment to measure changes in gene expression. Two samples are separately labeled and simultaneously hybridized to an array of complementary DNA on a glass slide.

more accurate. The DAPI stain also helps estimate the background noise from DNA–surface binding. After the signals for each sample are normalized for background and the relative intensity of the fluorescent dyes, the final outcome of a microarray experiment for each probe spot is a ratio of the mRNA concentration in one sample relative to the other. Perhaps the most important step in obtaining useful microarray data is quantifying and reducing the large error inherent in defining this ratio.

Figure 6.4 shows the final processed image corresponding to Fig. 6.3. Along with controls, it contains 85 genes from *Yersinia pestis*. In this picture, mRNA from cells at 25°C is labeled red and mRNA from cells at 37°C is labeled green. The more intense the color, the more mRNA from that sample hybridized to the spot relative to the other sample. Below the microarray data is a table of the gene names corresponding to each spot. Cells colored green are *Y. pestis* genes that are expressed more at 37°C than at 25°C, red cells are genes expressed more at 25°C. Cells colored blue are control spots from mouse and human genes that are absent in *Y. pestis*.

The challenge of visualizing microarray data is hinted at in Fig. 6.4. In particular, a microarray containing every gene in *Y. pestis* would contain more

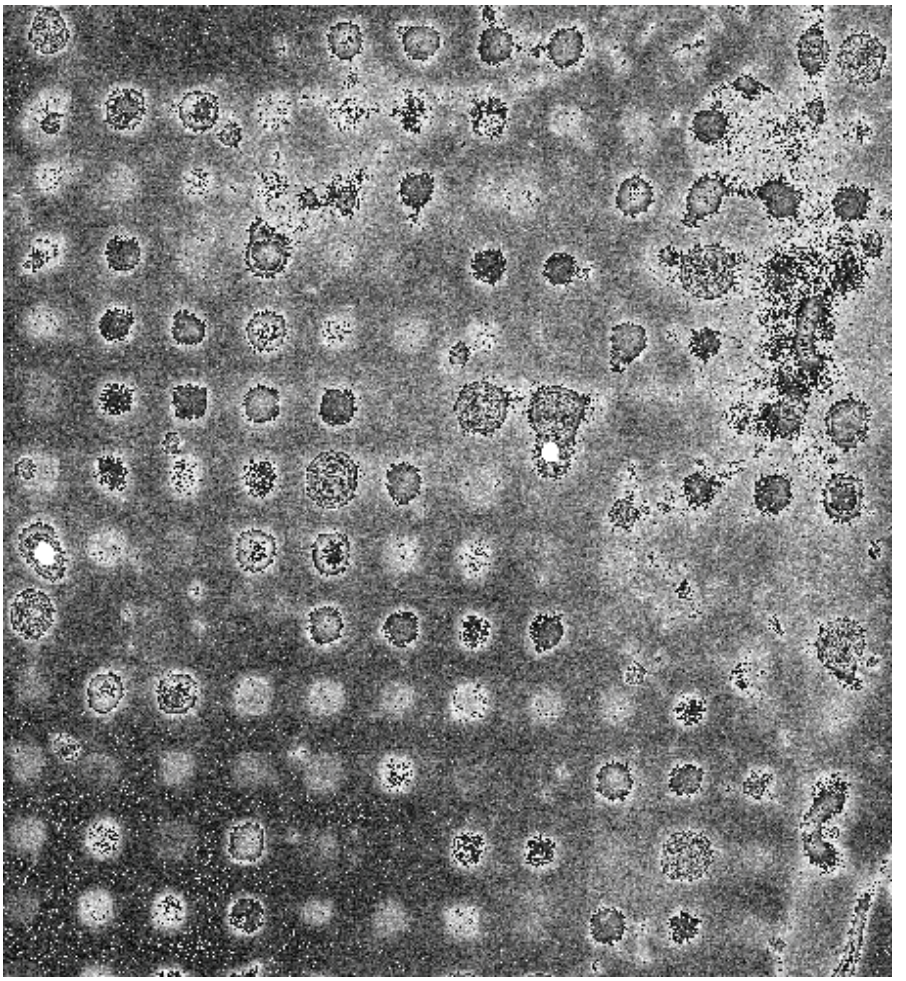
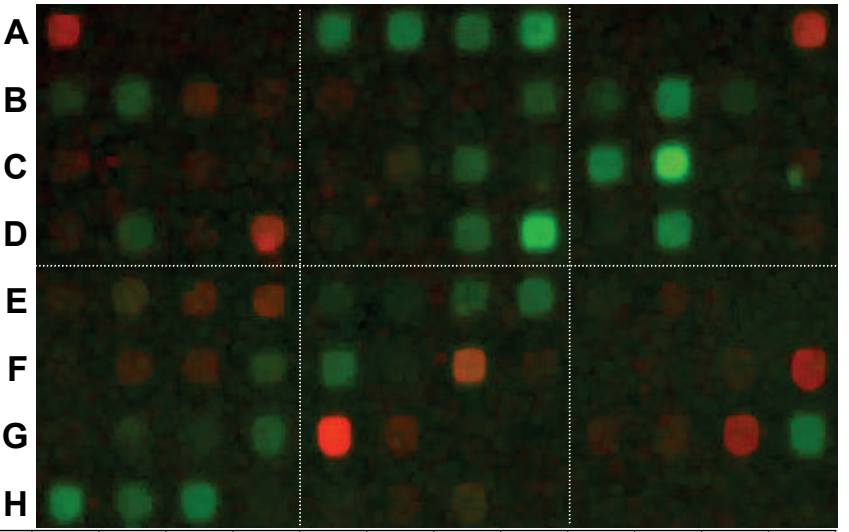


Fig. 6.3. Raw single optical channel microarray image. Notice the nonuniform background, noise spikes in both background and spots, and variable shape and size spots. The average center-to-center spacing is roughly 200 μm. (Courtesy of A.Wyrobek and E. Garcia, Lawrence Livermore National Laboratory.)



A	λ	BRCA	XRCC	PKC	ypkA	yscC	yscO	lcrV	PKC	XRCC	BRCA	λ
B	yopE	caf1A	sodA	envZ	ompR	yenI	rpoE	yopJ	yscB	lcrH	sycE	caf1R
C	tpx	rfaH	phoP	yenR	g6pd	rpoN	yopH	lcrF	lcrE	yopB	sycH	luxS
D	oxyR	psaA	glts	rpoS	lcrQ	yscW	tyeA	yopD	yadA	katY	hemF	toxR
E	rstB	nqrA	dha4	rpoH	yscL	yscU	sycN	yopM	pst	hns	cbl	flhC
F	nhaB	nadB	rpoD	yscH	lcrD	sycT	pla	catA	gcvA	copR	araC	nhaC
G	dam	yscG	lcrR	yopT	ymt	sodC	lysR	ilvY	tsaA	caf1	tuf	yscD
H	yscP	lcrG	yopK	caf1	sodB	crp	nhaR	fliA	h5	RAD	CDC2	Empt
	1	2	3	4	5	6	7	8	9	10	11	12

Fig. 6.4. Microarray data for 85 genes of *Yersinia pestis*. The 96-well microtiter format was used for spotting the array. There are 11 control spots. (Courtesy of E. Garcia and A. Wyrobek, Lawrence Livermore National Laboratory.)

than 4,300 different spots, making an image like Fig. 6.4 difficult to interpret. In general, DNA array experiments generate large, complex data sets. For any one experiment, each gene has three measurements associated with it: the intensities of the two competing samples and the ratio of those intensities. Typically a series of assays are taken over time. Thus, an array of 10,000 genes can be thought of as a set of 10,000 three-dimensional vectors that are changing in time. There is a need for ways to store the data in conveniently accessible, public data warehouses, visualize experimental results, and interpret the relative expression of the genes to identify pathways and common regulatory mechanisms.

To date, most microarray experiments have been published in scientific journals, with the expression data either included or referenced at the uniform resource locator (URL) of the authors' website where the data are posted. There is no current standard for warehousing microarray data, so they are stored in a variety of database formats or even large spreadsheet or text files. Currently, there are different database frameworks under development. One of many

examples is ArrayDB developed at the National Human Genome Research Institute (NHGRI), but there are many other public and private efforts. In the near future, as microarray experiments become increasingly frequent, there will be a need for a central public facility for submitting and accessing experimental data sets, similar to the current NIH GenBank for storing DNA sequences. Coupled to data warehousing, there need to be creative ways for biologists to visualize experimental results. Microarrays with thousands of grid cells can be viewed as similar to multidimensional geographical information, and work is being done to extend tools from that area to be useful in biology.

Even with better visualization and data storage, manually processing tens of thousands of data points is very difficult. Pattern recognition methods developed for imaging can be extended to automatically classify gene expression data. The goal is to divide genes into categories based on expression levels. Classification by expression level can indicate which genes are involved in the same pathway and potentially identify common regulatory mechanisms. The problem is that a 50% increase in expression could be as important to the biology of one gene as a 300% increase in expression in another gene. Simple threshold-based classification approaches erroneously classify these genes in separate categories. Furthermore, experimental error is sufficiently high that differences in relative expression are not very well quantified. Initial approaches to the problem included unsupervised learning methods, hierarchical clustering algorithms, and self-organizing maps. Often the function of at least some genes is known or suspected with a high degree of confidence. A support vector machine (SVM) method has been presented that takes advantage of this prior knowledge. The challenge facing algorithm developers is that even biologists cannot currently distinguish biologically significant features and clusters from artifacts. Progress depends on developing an appropriate biological framework and translating it into a mathematical model.

6.3 Multiaffinity Assays

When more than one binding site is used for detection, the assay can be called multiaffinity. The multiple binding events may be detected with separate dyes or labels. With special linker chemistries, it is sometimes possible to use a single reporter that is present only when both binding events occur. It is also possible to use multiple binding sites on the probe and target so that the probe can be labeled after the hybridization or binding has occurred. To reduce background noise, it is also possible to use one of the binding sites combined with bead or other technologies to isolate the target.

There are numerous multiaffinity assays, and the techniques can be applied to nucleic acids and proteins. For DNA, hybridization of complementary bases provides the discrimination. For proteins, antibodies or ligands are used to bind to the target protein. As examples we describe a hybrid DNA chip and a bead-based flow cytometric method using a sandwich assay. As with DNA chips and

arrays, one method will be used with gene expression and one with medical diagnostic applications.

6.3.1 Hybrid DNA chip

Detection of a test strand of DNA is accomplished by hybridization with a complementary reference strand of length $N=N_a+N_s$. The generalized experiment is shown graphically in Fig. 6.5. The N-mer is composed of an array of N_a-mers attached to a glass substrate in a manner similar to microarrays or DNA chips. The additional N_s bases for the detection are introduced in solution. Using extension enzymes or special linker chemistries, the discrimination power after hybridization is the same as with an N-mer.

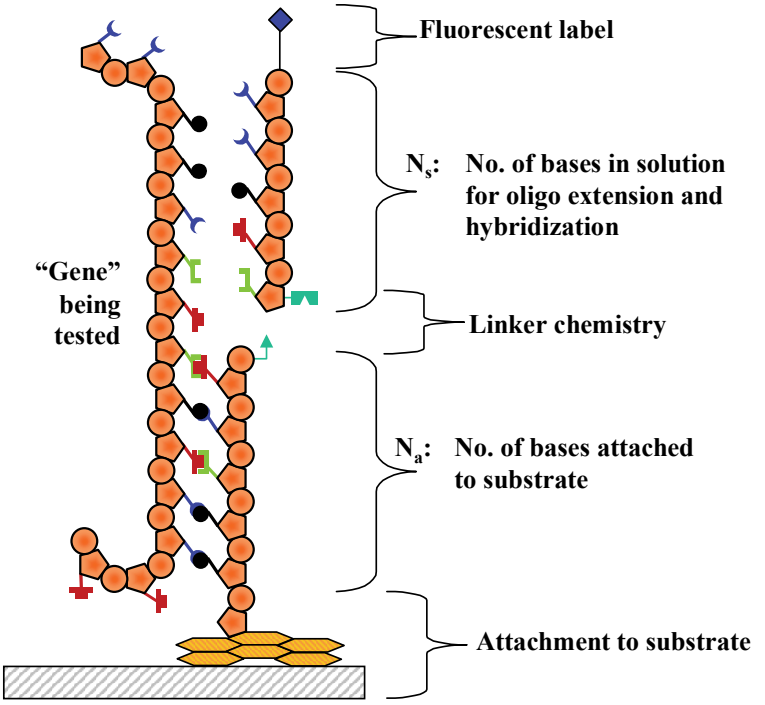


Fig. 6.5. Generalized hybridization experiment with attached oligos and oligos in solution.

To demonstrate this approach, consider the goal of monitoring the messenger RNA level of most genes (over 90%) in a bacterium without custom microarrays. Nothing in the method we describe limits the approach to bacteria. The hybridization, however, is currently experimentally restricted in oligonucleotide length to about 10 bases. As the length increases from 10-mers to 12-mers and longer (see Fig. 6.6), eukaryotic profiling, including human gene profiling, will be possible with the same technique.

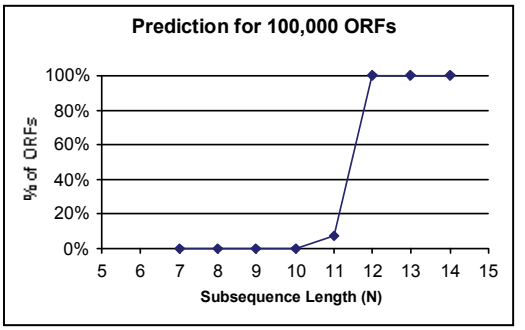


Fig. 6.6. Model prediction that 12 bases are needed to discriminate 100,000 genes (ORFs). This is the basis for predicting that eukaryotic gene expression can be determined using a universal 12-mer array. The typically smaller bacterial genomes would only need an array of 10-mers.

One approach to getting N-mer coverage is to place the competing N_s-mers into separate pools or experiments. Even when multiple N_s-mers bind to the same spot, their source genes can be discriminated. A matrix formulation of the experiment is $\mathbf{M} = \mathbf{P}\,\mathbf{D}\,\mathbf{G}$, where

- $\mathbf{G}$ = gene expression vector to be estimated (g × 1-column vector) representing the level of activity for "g" genes.
- $\mathbf{D}$ = DNA matrix (4^N × g) that maps individual genes into constitutive N-mers. Let $\mathbf{D}(i,j)$ be the value of $\mathbf{D}$ at position (i,j) where $1 \le i \le 4^N$ and $1 \le j \le g$. $\mathbf{D}(i,j)$ is defined as the number of times that the ith N-mer occurs in gene "j."
- N = $N_a + N_s$ is the number of bases in the DNA reference strand.
- N_a = number of bases in the oligos attached to the substrate.
- N_s = number of bases in the oligo extension introduced in solution during hybridization.
- $\mathbf{P}$ = pooling matrix ($p\ 4^{Na} \times 4^N$) where p is the number of experimental pools.
- $\mathbf{M}$ = measurement vector ($p\ 4^{Na}$ × 1-column vector) representing the outcome of an experiment.

With the experiment constructed as a matrix multiply, the same approach as in the classic $\mathbf{Ax}=\mathbf{b}$ matrix formulation can be used. Recall that if $\mathbf{A}\mathbf{A}^T$ is invertible, then the optimal least-squares solution or pseudoinverse is

$$\mathbf{x_{LS}} = \mathbf{A}^T (\mathbf{A}\,\mathbf{A}^T)^{!1}\,\mathbf{b}. \tag{6.1}$$

We have assumed that all possible N_a-mers are attached to the substrate. The matrix formulation is more general than this and any set of oligos (even of varying lengths) could have been used. However, the HyChip™ by HySeq uses all possible 5-mers at this time and so we have restricted the matrix formulation to apply to demonstrate transcript profiling.

To demonstrate the matrix approach, consider an artificial genome with 10 genes of 10 randomly assigned bases. In Fig. 6.7, we show the genes as well as highlighting the unique 3-mers—i.e., a 3-mer that appears in only one gene. For this example, we use a 3-mer reference strand compsed of a single-base oligo attached to a substrate (N_a=1) and a 2-mer in solution (N_s=2). A few of the rows of the **D** matrix generated from the genes in Fig. 6.7 are shown in Fig. 6.8. The first N_a bases in the index column (the first "A" for the rows in Fig. 6.8) correspond to the attached N_a–mers so that the pooling and measurement matrices are easier to interpret. Rows that contain a single nonzero element represent 3-mers that uniquely identify a gene—e.g., AAG, ACC, AGA for genes 7, 2, and 7, respectively. If each gene had many unique identifiers, the detection problem would be simple. For a short oligo detection system, however, there are many shared *N*-mers and so multiple "hits" must be utilized in the algorithm. The design issues are how to implement a least-squares estimate of the level of gene expression while introducing a pooling strategy (the **P** matrix) that minimizes the number of experiments needed.

The experimental constraints result in a **P** matrix that is sparse and with repeating subunits. Denoting each pool by $\mathbf{P_i}$, the pooling matrix is of the form shown in Fig. 6.9. Each $\mathbf{P_i}$ is repeated 4^1=4 times because N_a is 1 in our example. Each $\mathbf{P_i}$ row vector is of length 4^2=16 (N_s is 2) and composed of 1s and 0s. A "1" indicates that a particular N_s-mer is used in the pool. **O** is a zero row vector of length N_s. For this example, we designed the set of p=5 pools shown in Fig. 6.10. Both **D** and **PD** are of matrix rank 10. Therefore the pseudoinverse exists and can be determined using singular-value decomposition or other techniques. The pseudoinverse was verified using MATLAB™. The calculation implements Eq. (6.1) and estimates

$$\mathbf{G_{LS}} = \mathbf{(PD)}^{\mathrm{T}} \left(\mathbf{PD\ (PD)}^{\mathrm{T}}\right)^{-1} \mathbf{M}. \tag{6.2}$$

Applying this approach to a specific bacterium such as *Yersinia pestis* requires the development of several tools and algorithms. *Y. pestis* is roughly a 4.3-Mb genome and has more than 4,000 genes. Using an experimental system based on two 5-mers (N_a=N_s=5), we use the matrix approach to design a pooling strategy. The design algorithm is

1. Determine the occurrence rates of every *N*-mer (0, 1, 2, etc.)
2. Starting with unique occurrences, then doublets, etc. determine which attached N_a-mers are "contaminated" by the proposed *N*-mer and are not already in a pool.
3. Find the largest set of N_s-mers meeting criterion 2 with nonintersecting contamination sets. Place in one pool and then return to step 2.

This approach is basically a Gram-Schmidt orthogonalization. We have calculated the "**D**" matrix for *Y. pestis* and are now designing the pooling strategy. Figure 6.11 shows the percent of ORFs that have a unique *N*-mer for an *N* ranging from 5 to 14. Both model and data are presented and show that a 10-mer (N=10) is sufficient to transcript profile *Y. pestis*.

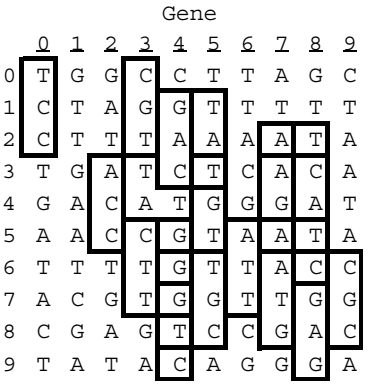


Fig. 6.7. Ten genes of a 100-base length are used as a synthetic genome for demonstrating the technique. Unique 3-mers indicating the discrimination power of the detection are highlighted. Note that there is no 3-mer unique to gene 1.

	Gene									
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
AAA	0	0	0	0	0	0	0	0	0	0
AAC	0	0	0	0	0	0	0	0	0	0
AAG	0	0	0	0	0	0	0	1	0	0
AAT	0	1	0	0	0	0	0	1	0	1
ACA	0	0	0	0	0	0	0	0	0	0
ACC	0	0	1	0	0	0	0	0	0	0
ACG	0	0	0	0	0	0	1	0	0	1
ACT	1	0	0	1	1	0	0	0	0	0
AGA	0	0	0	0	0	0	0	1	0	0
AGC	0	0	0	0	0	0	0	0	0	0

Fig. 6.8. The first 10 rows of the 64 row by 10 column D matrix. Rows with a single nonzero element are unique identifiers of a gene. For instance, AAG is 0 except for the 1 in the gene 7 column. Compare these entries with the highlighted 3-mers in Fig. 6.7.

$$\begin{vmatrix} \mathbf{P}_1 & \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{P}_1 & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{P}_1 & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{P}_1 \\ \mathbf{P}_2 & \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{P}_2 & \mathbf{O} & \mathbf{O} \\ & \cdots & & \\ \mathbf{O} & \mathbf{O} & \mathbf{P}_p & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{P}_p \end{vmatrix}$$

Fig. 6.9. General form of the pooling matrix P.

	P1	P2	P3	P4	P5
AA	0	0	0	0	1
AC	1	0	0	0	0
AG	1	0	0	0	0
AT	0	0	0	0	0
CA	0	0	0	0	0
CC	0	0	1	0	0
CG	0	0	0	0	0
CT	0	0	0	0	0
GA	0	0	0	0	0
GC	0	0	0	0	0
GG	0	0	0	0	0
GT	0	0	0	1	0
TA	0	0	0	0	0
TC	0	0	0	0	1
TG	0	0	0	0	0
TT	0	1	0	0	0

Fig. 6.10. The five pools used for the synthetic genome example. PD has matrix rank 10.

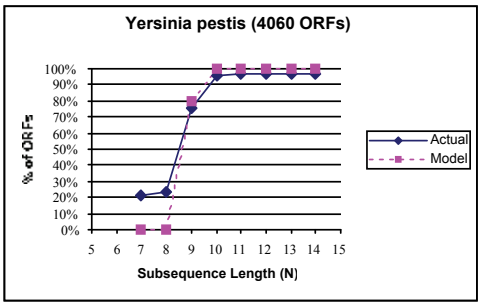


Fig. 6.11. Model and data overlaid showing that over 90% of the *Yersinia pestis* genes have at least one unique 10-mer. A matrix formulation can exploit nonunique hits for a very robust estimate of gene activity.

In addition to making it possible to design the minimal pooling strategy, the matrix formulation facilitates the design of calibration oligos and other strategies for making microarrays more quantitative and repeatable. Design constraints can also be implemented in the pool specifications to accommodate ribosomal and other contaminating signals.

6.3.2 Bead-based flow cytometry for detection

Just as microarrays and DNA chips were used to do many measurements in parallel on a 2-D surface, a variety of other physical separation techniques have been developed. In some cases magnetic beads and attached DNA (or protein,

etc.) are separated from background solutions with magnets. Materials have also been layered as a type of barcode for micromarkers. When larger substrates can be used, radio-frequency (RF) tags and other methods have been used to identify, track, and separate targets from background. The example we present here uses color-coded beads that can be read by a laser-induced fluorescence system. There are a variety of ways to read these beads—a flow cytometric approach is presented here.

In a flow system like the one shown in Fig. 6.12, test particles are suspended in fluid and flowed past a sensor system. A flowing liquid creates a "sheath" to facilitate lining up the sample particles in single file. Laser illumination to detect cells, beads, and/or labeled DNA or protein is usually used in a flow system. The scatter of the laser light can also be used to estimate the relative size and roughness (granularity or complexity) of the particle using forward or side scattering, respectively. Argon-ion lasers (488-nm wavelength) are often used in flow systems. Scatter is usually measured by aligning a detection system at 90º and 180º from the laser illumination. Figure 6.12 shows a different approach for optical detection that minimizes alignment requirements by placing a tapered fiber optic as the detection system in the flow. The fluid acts as a light pipe, with significantly less stringent alignment criteria than using optics outside of the flow.

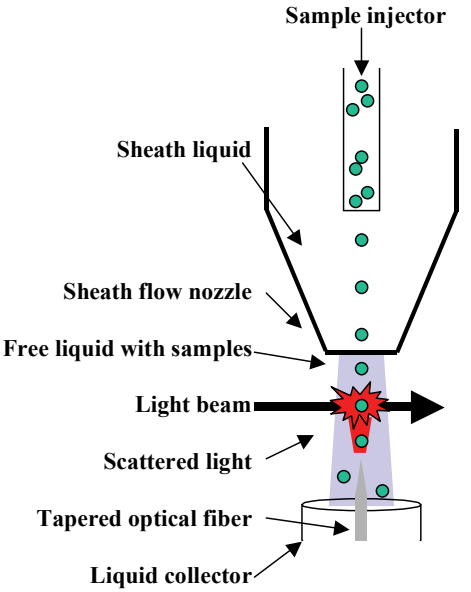


Fig. 6.12. Flow system for detection. A liquid sheath moves quickly out of a flow nozzle, distributing the test sample along the flow path. The test samples are already bound to beads that can then be interrogated individually by laser illumination. The beads and the bound biomolecules may have unique optical signatures, allowing multiple assays to be run simultaneously. The optical fiber connects to a photodetector and the liquid collector takes the beads and fluid.

A sandwich assay is used to detect DNA or protein with the flow system. A sandwich assay captures the target between two probes. Figure 6.13 shows a sandwich assay that uses optical microbeads as the substrate. The probe is attached to the bead using a biotin–avidin link. This bond is the strongest known noncovalent attachment. Biotin is a vitamin (244 Da) and avidin is a protein (68 kDa) obtained from egg whites. Sometimes streptavidin, a protein isolated from the bacterium *Streptomyces avidinii*, is used in place of avidin. A second probe is labeled and binds to a separate site on the DNA or protein.

The final system operates by detecting the color of each bead to determine which probe is being queried. A fluorescent signal indicates that the target is present. By assigning specific probes to a specific color bead, many assays can be done in parallel.

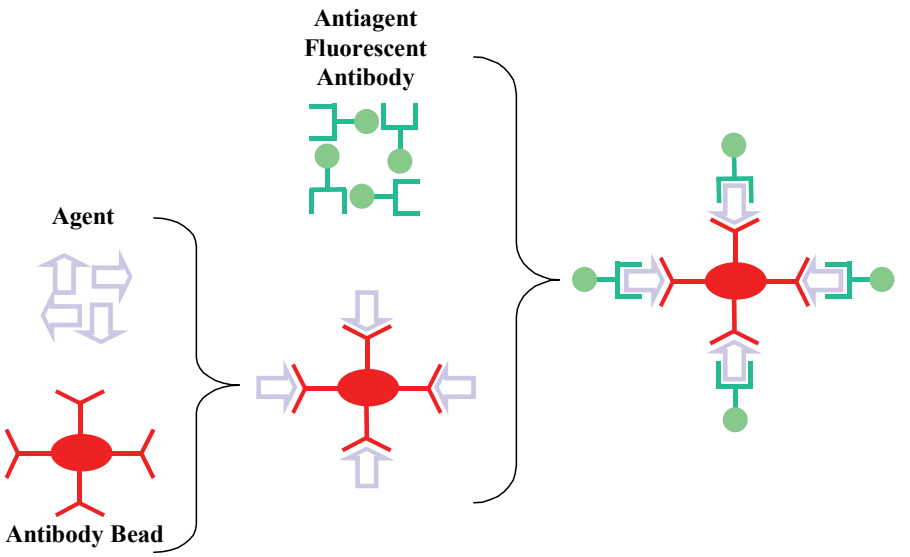


Fig. 6.13. Bead-based sandwich assay for flow system detection. Multiple binding events per bead allow greater signal to noise for detection. By using color-coded beads, multiple assays can be run in parallel.

7

Protein Structure

Protein structure is critically linked to an understanding of protein function. Currently there are three approaches to determining 3-D protein structure: nuclear magnetic resonance (NMR), x-ray diffraction, and computational prediction of structure. NMR examines proteins in solution; x-ray diffraction measures scattering from a solid protein crystal; and computational approaches predict structure by finding similarities in amino acid sequence and substructures between the unknown and known protein structures. At this time, about 80% of the known protein structures were determined using x-ray diffraction.

7.1 Nuclear Magnetic Resonance

Nuclear magnetic resonance measures the coupling of atoms across chemical bonds and short distances under the influence of a magnetic field. An NMR instrument is shown in Fig. 7.1. A typical NMR experiment with a 600-MHz field takes 4 to 5 weeks for data collection and is limited to proteins with fewer than 360 amino acids (40 kDa). Data analysis that once took months can now occur in a single day using new algorithms. New NMR instruments with fields as high as 1 GHz and more sensitive techniques allow faster analysis of larger proteins, but there is still a size limit of a few hundred amino acids.

The principles behind NMR are similar to those used in the medical procedure of magnetic resonance imaging (MRI). In fact, MRI is really a specific application of the more general NMR principles. MRI usually concentrates on hydrogen atoms. Not only does hydrogen have the strongest magnetic response, it is the most common atom in biological systems. NMR spectroscopy includes hydrogen, carbon, nitrogen, and other elements (see Table 7.1). Only certain isotopes of these elements can be detected in NMR systems. Fortunately, these are some of the most common atoms in proteins. If a specific isotope is needed or an increased abundance of one isotope rather than another, organisms can be fed a restricted diet that includes the isotope. For NMR, these isotopes do not need to be radioactive.

Fig. 7.1. Nuclear magnetic resonance instrument.

Table 7.1. Comparison of NMR parameters for several isotopes.

Isotope	Unpaired Protons	Unpaired Neutrons	Net spin	Gyromagnetic ratio (MHz/T)	Natural abundance (%)
^{1}H	1	0	½	42.6	99.99
^{2}H	1	1	1	6.5	0.02
^{13}C	0	1	½	10.7	1.11
^{14}N	1	1	1	3.1	99.63
^{19}F	0	1	½	40.1	100
^{31}P	0	1	½	17.3	100

MRI and NMR use large external magnets to orient molecules and atoms that have an intrinsic magnet polarity. In MRI the magnets are usually between 0.5 and 2 tesla (T). NMR magnets can have much stronger fields, up to 19 T. Many atoms that have an odd number of neutrons, protons, or both have a mechanical spin associated with angular momentum. The nuclear spin quantum number (I) is used to characterize the angular momentum.

$$I = \tfrac{1}{2}\{\text{number of unpaired protons} + \text{number of unpaired neutrons}\}. \quad (7.1)$$

The spin axis defines an atomic-scale bar magnet or magnetic dipole (μ). The naturally occurring "magnets" align with the external field the same way a collection of compasses point north in the earth's magnetic field (see Fig. 7.2). There are two orientations in which an atomic nucleus is aligned to the magnetic field: parallel and antiparallel. The parallel alignment is of slightly lower energy than the antiparallel alignment. The quantum of energy needed to boost a nucleus to the higher energy state is defined as the resonant energy. Because the angular momentum orientation (spin) is changed by 180°, this is referred to as "spin-flip."

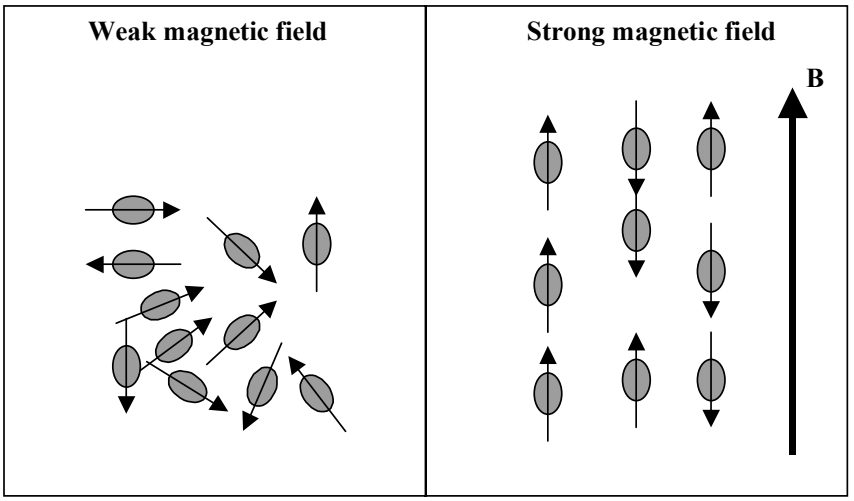


Fig. 7.2. For atoms with appropriate nuclear angular momentum, an external magnetic field can be used to align the mechanical spin vectors. NMR uses the quantum changes in energy absorbed and emitted as atoms go to and from parallel and antiparallel alignments.

The atoms relax back until the magnetic dipole realigns with the external magnetic field, releasing energy at certain characteristic radio (resonance) frequencies. By measuring the reaction of the atoms to different radio pulses, especially the relaxation time for realignment, NMR can be used to estimate the locations of the atoms and often to reconstruct a 3-D protein structure. The

chemical composition is inferred from a shift in the radio-frequency response of specific atoms to the series of RF pulses.

The range of radio frequencies available to an NMR instrument scales with magnetic field strength—the larger the magnetic field, the higher the RF response that can be detected. ^{1}H is often used as the reference isotope. An 11.7 T magnet can measure 500 MHz radio frequencies. Verify this in Table 7.1 with 500 divided by 11.7 being approximately 42.6. The gyromagnetic ratio (γ) is used to relate the RF response and magnetic field for a specific isotope. A more powerful 18.8 T magnet can be used to measure 800-MHz responses for ^{1}H. These magnets are usually built from superconductors that require liquid helium (4 K) and liquid nitrogen (77 K) cooling. Since the magnetic field strength in teslas and the best radio-frequency response are equivalent, it is as common to describe an NMR system in megahertz as in teslas.

For protein studies, the isotopes appear as part of complex molecular structures. The local fields induced by neighboring electrons can change the RF response of specific isotopes, depending on their locations in molecules. NMR must be used as a spectrometer in order to collect data for many resonant frequency responses and shifts. Commonly occurring chemical chains are characterized in detail. The details of NMR spectroscopy are beyond the scope of this text.

NMR instruments are used for protein structure measurements as well as a variety of other applications, including protein–ligand interactions. Because proteins can be introduced in solution, the NMR can be used to interrogate hundreds of possible binding reactions simultaneously. A target protein can be assayed against numerous potential ligands and the spectra of the bound complexes used to determine the ligands that bind.

7.2 X-Ray Crystallography

Unlike NMR that is limited to proteins of less than a few hundred amino acids, x-ray crystallography has no theoretical size limit. A protein crystal diffracts x-rays, generating the Fourier transform of the atomic structure (electron density) of the protein. Since only the intensity of the diffracted x-rays can be directly measured, techniques to create phase modulation of the intensity are needed. These phasing techniques include replacing selected atoms by heavy metals and using different wavelengths. The process is similar to holography. Data analysis used to be slow and complex, but recent algorithms have made it much faster. X-ray crystallography, however, is sharply limited because it is often very difficult and sometimes impossible to obtain a high-quality protein crystal.

X-ray crystallography begins with protein production. *In vitro* approaches to protein production usually insert the appropriate gene (DNA) into *E. coli* or another cell. *In vivo* (also known as "cell-free") approaches have also been used. Protein engineering to replace specific amino acids may be needed to facilitate protein expression. Once a sufficient quantity of protein (usually a few microliters) is produced, the crystallization process begins.

Protein crystals are usually formed by using vapor diffusion. A highly concentrated sample containing many copies of the protein is loaded within a chamber with a separate buffer solution of lower concentration. In many experimental setups, the protein sample is hung on a cover slip above the buffer as diagrammed in Fig. 7.3. The hanging drop can be 1 to 10 μl. The chemicals in the sample diffuse to the buffer in a process similar to evaporation. The remaining drop contains a greater and greater concentration of protein. If conditions are right, the protein molecules align and form a crystal (see Fig. 7.4). A typical crystal is less than a millimeter on a side and contains more than 10^{15} protein molecules. Crystals may form in a couple days or, more typically, they may take a few months. In some cases, crystals do not form. If any crystals form, even ones not sufficient for use in x-ray diffraction, the researcher may vary conditions, including concentration, temperature, or pH, to encourage better crystal formation. For some types of proteins, no one has been able to produce crystals. In order to increase the number of crystal experiments that can be performed, robotic systems like the automatic greasing and fluid-dispensing robot shown in Fig. 7.5 are used. The grease is used to seal the hanging drop over the well. Note that the plastic plates used in this experiment are typically much larger than those used in DNA sequencing. The 24-well plate shown in the figure, unfortunately, does not have the same dimensions as a standard microtiter plate.

Once a crystal has formed, it is mounted, frozen to cryogenic temperatures with liquid nitrogen, and placed in front of an x-ray source. The cryofreezing is used to minimize x-ray damage to the crystal. The x-ray illumination can last from less than a second to several hours. The sample (or equivalently the source/detector) is rotated to collect two-dimensional x-ray scatter data from several angles. The composite data set represents the three-dimensional x-ray scattering of the protein crystal. The positions of atoms within the crystal are determined from the x-ray diffraction data using a Fourier transform. The wavelength of the x-rays is usually between 0.5 and 1.5 angstroms (10^{-10} m).

For simple single crystals, Bragg scattering is a useful description of the mechanism. As shown in Fig. 7.6, constructive interference occurs when the path-length difference ($AB+BC$) is an integer multiple of the wavelength ($n\lambda$). The relationship is given by

$$n\lambda = 2\, d \sin(\theta), \tag{7.2}$$

where θ is the grazing angle of illumination and d is the crystal plane spacing.

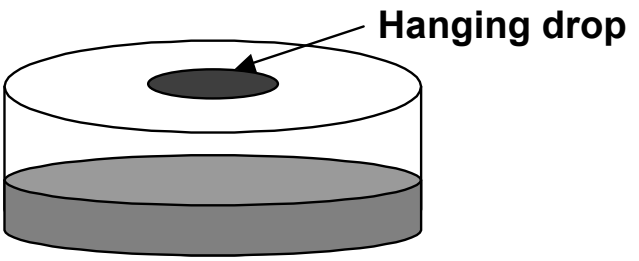


Fig. 7.3. Cartoon of a plastic well with a sealed cover slip. A protein-rich drop is hung above a well filled with buffer solution in order to create a vapor diffusion process to concentrate the protein, it is hoped, into a crystal.

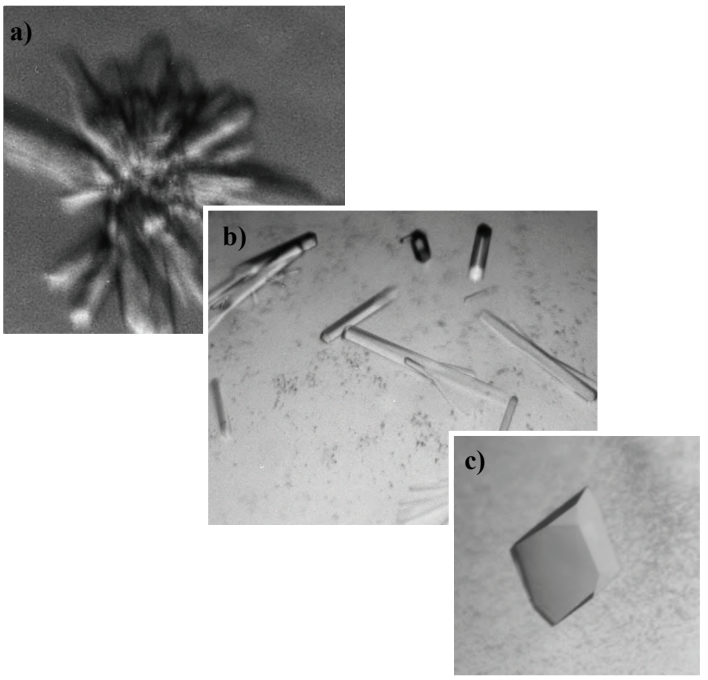


Fig. 7.4. Protein crystals with a variety of expected diffraction qualities. The single, well-formed protein crystal (c) is likely to produce a good diffraction pattern. The other two would encourage modifications to the crystallization recipe and will likely form quality crystals in subsequent trials. (Courtesy of B. Rupp and B. Segelke, Lawrence Livermore National Laboratory.)

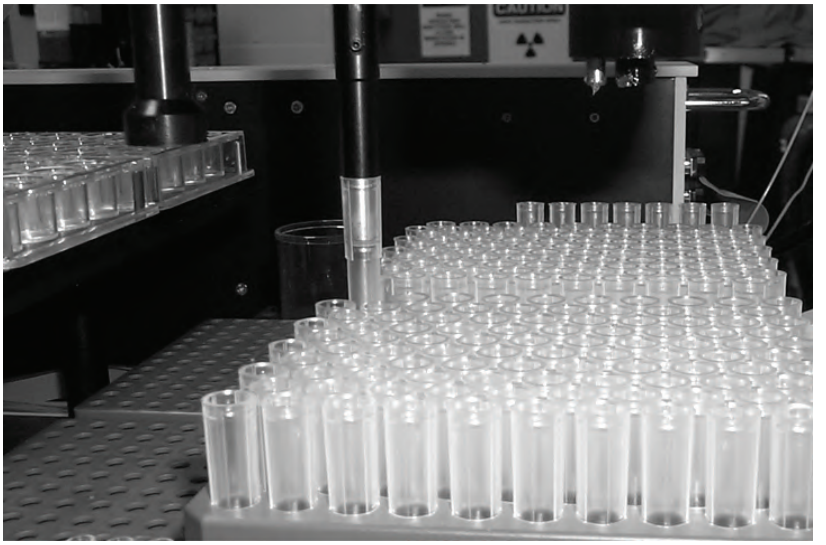


Fig. 7.5. Robot picking up a disposable plastic pipette tip from a microtiter format source plate. An attachment for greasing the lips of the wells is used to seal the cover slips with the protein drop over the wells.

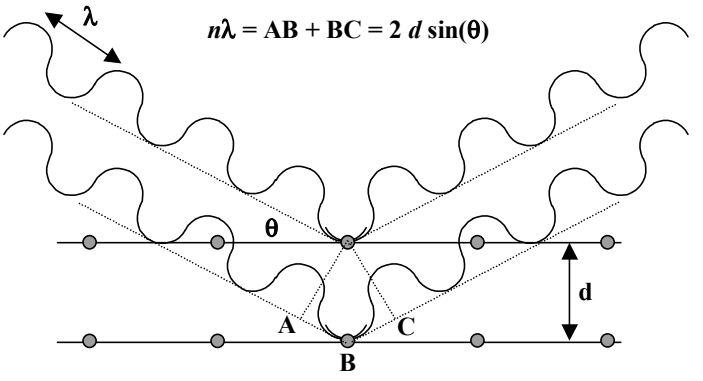


Fig. 7.6. Derivation of Bragg's law relating the wavelength of illumination (λ), crystal plane separation distance (*d*), and grazing angle (θ).

For protein crystals, the x-ray illumination wavelength is roughly the same length as the diameter of the atoms being observed. Therefore, as shown in Fig. 7.7, most of the scatter is forward. A useful mathematical representation of the scatter from a single atom is as the complex-valued phasor

$$f_m\, e^{i\phi m}, \tag{7.3}$$

where ϕ_m depends on path length and the amplitude f_m depends on scatter angle and the atom. With a protein crystal, a collection of different atoms are illuminated, and the scatter at a detector can be expressed as the scatter summed over all atoms:

$$A = \sum_m f_m\, e^{i\phi m}. \tag{7.4}$$

The intensity, I, which would be measured by an x-ray detector, is the amplitude times its complex conjugate $I=AA^*$. Multiplying out the terms in the summation and collecting terms so that single-atom scattering and scattering from pairs of atoms are in separate summations,

$$I = \sum_m |f_m|^2 + \sum\sum_{m\neq n} f_m f_n^*\, e^{i(\phi m ! \phi n)}. \tag{7.5}$$

As the detector is moved (or equivalently, the protein crystal is rotated), the first term varies slowly. The second term is modulated by the phase differences and therefore encodes path-length differences. This is illustrated graphically in Fig. 7.8.

In summary, the intensity of an x-ray diffraction pattern can be related to the electron density map of a protein crystal. The electron density map provides the locations of the constituent atoms of the protein. The relationships among the measurements and the desired parameters are approximated by a Fourier transform similar to Eq. (7.5). The phase of the measurement is the most important component because it strongly encodes the physical characteristics associated with atom position.

There are two techniques that are used to identify landmarks to facilitate reconstruction of the protein structure. In the first approach, protein crystals are doped with x-ray scattering metals such as platinum and the scattering pattern is compared with a pure (undoped) protein crystal. The stronger scattering centers from the metals are correlated to the known binding sites on the protein to provide a coarse map for reconstructing the 3-D protein structure. The second approach, known as multiwavelength anomalous diffraction (MAD), uses a tunable x-ray source. A single protein crystal with selenium atoms replacing the sulfur atoms in the amino acid methionine is illuminated with three or more different wavelength x-rays. One of the wavelengths used is tuned for absorption by selenium atoms and allows a coarse localization similar to the doping method without needing multiple protein crystals.

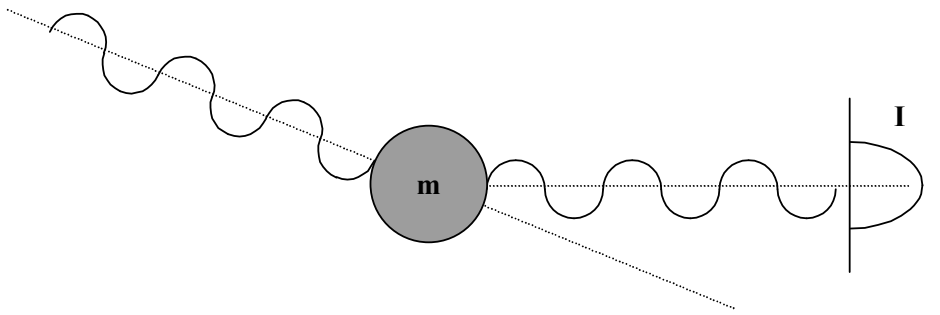


Fig. 7.7. The x-ray wavelength of illumination (λ) is roughly the same size as the diameter of atom "m" (0.5 to 1.5 Å), resulting in mostly forward scatter. The amplitude depends on an angular distribution about some dominant scattering angle and phase modulations due to path length. Measuring the scattered field at a variety of positions would allow estimation of the position of the atom (more specifically, the electron density).

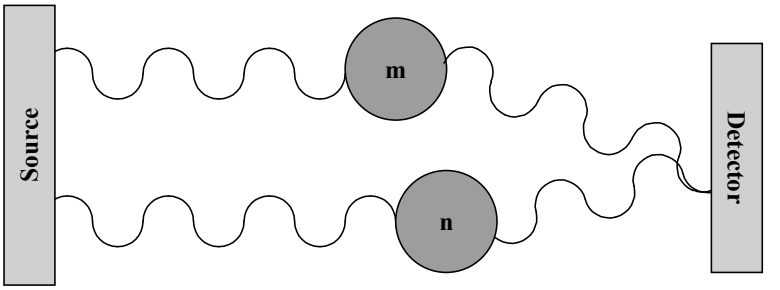


Fig. 7.8. Scattering from a collection of atoms depends on the relative difference in path length between pairs of atoms and the scattering angles.

7.3 Computational Prediction of Structure

Because of the limitations of both measurement techniques, little structural data are known for the important class of membrane proteins, which make up about 30% of the total protein in a cell. Improvements in NMR magnets and methods and protein crystallization protocols are needed.

Even with the implementation of high-throughput techniques, protein structure experiments are not keeping up with the amount of sequence data. Currently, there are 10,675 proteins with known 3-D structures from x-ray crystallography and NMR experiments in the Protein Data Bank (PDB) at http://www.rcsb.org/pdb/. This is a tiny fraction of the millions of gene

sequences in the GenBank from 820 different species. However, while there is a great diversity of sequences, there are certain structural features that arise repeatedly in many different proteins. The 10,000 protein structures in the PDB share about 1,800 folds, or common structural features. It is estimated that fewer than 5,000 folds occur naturally. Thus, with careful selection of experimental targets, perhaps only a few thousand more protein structures are required for every possible structural feature to be found in the database. The current prediction is that by 2003, there will be 35,000 protein sequences in the PDB. In principle, if every fold that can be put together to build a protein is known, it should be possible to compare the sequence of an unknown protein with the sequences corresponding to these folds and predict its final structure. Comparative modeling is a computational approach to doing this.

With computational prediction, a structure is assigned by identifying sequence similarities (homologies) between the unknown protein and proteins with known structures. Generally, a 30% amino acid sequence similarity is thought to be sufficient for accurate structure prediction. Less than 15% similarity is usually considered a random correlation with a low probability of accurately associating structure or function. Sequence similarity is determined by aligning amino acid subsequences to identify similarities and differences.

A significant limitation of this approach is that alignment is an unsolved problem. Algorithms like BLAST can align sequences with many short subsequences and then integrate to compare them as a whole. Protein structure requires aligning different parts of the protein sequence in three dimensions. Figure 7.9 is a simple illustration of how similar subsequences can be located in different regions of the overall sequence, and the subsequences themselves may not be completely identical. In some cases, equally suitable alignments can be found in which every amino acid is at a different position in the predicted structure. Furthermore, most experimental structure data ignore the amino acids at the end of the protein chain, limiting what is available in databases.

in database: -QWRAZWTTWDWHQMMQQQQWWRZHIOPP-
unknown: -HHWRLHIOPQWRRWTTWWWHQL- - -

Fig. 7.9. A simple example of the sequence alignment problem in protein homology assessments.

Moreover, many amino acids share the same basic chemical properties, so exchanging them does not significantly affect structure or function. Therefore two proteins with very different sequences may in fact have the same structure. There are hundreds of known hemoglobin protein sequences in mammals, all of which share a similar structure and the same function. One approach is to consider the evolutionary history of the organism in question and at what point

its sequence diverges from those in the databank. For example, a human gene will have a sequence more similar to that of a chimpanzee gene than to a mouse gene, and will be much less similar than an *E. coli* gene. Another approach is to combine the results of sequence homology and evolutionary history with the results of microarray experiments.

To ensure that algorithms can be applied generally to unknown sequences, algorithm performance is measured in biennual critical assessment of protein structure prediction (CASP) experiments. Before each CASP meeting, the experimental community provides a list of structures that are about to be determined. The sequences are distributed to the computational community, which analyzes them without knowing the structure beforehand. Overall, while the accuracy of predictions is improving, computational prediction of structure is still limited to subsequences of an unknown gene that have high sequence similarity.

The shortcomings of comparative modeling would be avoided if it were possible to predict protein structure *ab initio* based on amino acid chemistry alone. Almost all protein structure is the result of the interaction of amino acids with water in cells. Thus, *ab initio* simulation of the folding of a 1,000-amino acid protein requires a 10,000-body calculation of the interactions of the amino acids and 9,000 surrounding water molecules. Fundamental quantum chemistry calculations are limited to studying the area around a single amino acid or DNA base. Classic molecular dynamics treats amino acids and water molecules as "billiard balls" and can model a subregion of the protein up to about 40–60 amino acids in length. IBM has launched a new effort to produce a computer that is capable of one petaflop, which is about a hundred times faster than the most powerful supercomputers currently under development. However, even if technological obstacles can be overcome, the computer algorithms currently used for *ab initio* predictions of structure do not scale well and will need to be redesigned.

Appendix A

Units and Measures

For engineers and physical scientists, it is often shocking how much biology can be discussed without using quantified physical units and measures. Nucleic acid bases and amino acids could have been used as the principal units, if not the only units, in our introduction to biology. A few brief definitions follow for important units and measures.

Definitions

Biology uses a wide range of scale. Cells may have femtoliter volumes but contain more than a billion copies of a specific type of nucleic acid. The prefixes used to denote powers of ten are

10^{-18} = atto (a)	
10^{-15} = femto (f)	10^{+15} = peta (P)
10^{-12} = pico (p)	10^{+12} = tera (T)
10^{-9} = nano (n)	10^{+9} = giga (G)
10^{-6} = micro (μ)	10^{+6} = mega (M)
10^{-3} = milli (m)	10^{+3} = kilo (k)
10^{-2} = centi (c)	

The liter (l) is a measure of volume of a liquid that equals 0.001 cubic meters or 1,000 cubic centimeters (10 cm on a side cube). A milliliter (ml) is therefore 1 cm^3 (1 cm on a side cube) or 10^{12} cubic microns. A microliter (μl) is 10^{-3} cm^3, 10^9 μm^3, or 1 mm^3 (1 mm on a side cube). A nanoliter (nl) is 10^{-6} cm^3, 10^6 μm^3, or 0.1 mm^3 (0.1 mm on a side cube).

Molecular weight is the sum of the atomic weights of the constituent atoms of a compound. Recall that atomic weight includes the contribution of neutrons in the nucleus of the atom (atomic number is a count of the protons or electrons). The molecular weight of insulin, the protein secreted in the pancreas and whose deficiency is the cause of most diabetes, is 5734.

A dalton (Da) or atomic mass unit (AMU) is $N_g^{-1} = 1.66\times10^{-24}$ gram where N_g is Avogadro's number. This is the same as one-twelfth the mass of carbon-12 (the most abundant isotope of carbon).

There are many ways to express concentration. Because biochemical reactions often depend on the number of molecules available, a count of the number of molecules per unit of volume of the total solution is often used. We first define the molecular weight of a substance in grams as a mole (mol). A mole of water is 18 g. Recall that water has molecular weight of 18 (H=1.0079 + O=15.9994). For comparison, a mole of salt (NaCl) is 58.44 g (Na=22.9898 + Cl=35.453). Molar concentration (M) is defined as the number of moles of a substance per liter of total solution. The molar concentration of any solution can be converted to the number of molecules by multiplying with Avogadro's number (6.02×10^{23}).

The ionization of water is important to many processes. The concentration of H^+ ions and $OH^!$ ions determines acidity and alkalinity, respectively. For water, the distribution of H^+ and $OH^!$ are related and a logarithmic measure of the concentration of H^+ known as pH (for power of hydrogen) is the accepted standard. The pH measure is defined as the negative of the logarithm of the molar concentration of H^+, that is, pH=!log10(H^+). A pH 7.35 (typical of human blood) represents $10^{-7.35}$. Acidic solutions have large concentrations of hydrogen ions and therefore a low pH (approaching 0). For instance, 1 M hydrochloric acid (HCl) is pH 0. A neutral solution would have a pH of 7 because there are 10^{-7} mol of both H^+ and OH^- in water at equilibrium. Above a pH of 7, there is a higher concentration of OH^- and the solution is considered basic. For instance, 1 M NaOH is pH 14.

Isoelectric focusing is applied to a solution of charged molecules by moving them through a gel with a pH gradient. The isoelectric focus point is where the molecules and gradient-pH gel reach a neutral charge and an applied electric field therefore does not mobilize the molecule. The isoelectric focus point is usually abbreviated pI.

Centipoise (cP) is a measure of viscosity in units of grams/meter/second. For example, water at 20°C is 1.00 cP and PDMA typical for pumpable capillary DNA sequencing applications is 14 cP at 4% PDMA.

Order-of-Magnitude Calculations

The following are some "rules of thumb" that can help estimate order of magnitude for parameters.

A typical mass of an amino acid is 120 Da. A 1-kb coding capacity in DNA is equivalent to 333 amino acids and this is approximately a 40-kDa protein.

A single 10-kDa protein (molecular mass) is equivalent to 100 pmol. Since 1 nmol is 10 μg, 100 pmol would be 1 μg or 6×10^{13} molecules (scale by Avogadro's number).

A picomole (pmol) of a 1-kb DNA is roughly 0.66 μg. If there is 1 μg/ml of nucleic acid, there is approximately 3 μM phosphate.

Appendix B

Nonscientific Issues

Safety

Before embarking on R&D in biotechnology, it is important to recognize that many laws and regulations exist to help ensure that the R&D is safe for the researcher, the consumer, the environment, and the general public. We will not attempt to itemize all of the issues, but rather provide some pointers on where to start.

The principal guide for biosafety is *Biosafety in Microbiological and Biomedical Laboratories*, published by the Centers for Disease Control (CDC) (Fig. B.1). This is must reading for everyone in the field and is available a variety of ways, including over the web or in pdf format from http://www.cdc.gov/od/ohs/biosfty/bmbl4/bmbl4toc.htm. Important nomenclature used in the CDC guide includes biosafety levels (BL or BSL). The BSL is used to qualify a laboratory for a certain level of work. The levels range from 1 to 4 and a brief summary follows. Please examine the most recent CDC guide for updates and specific information.

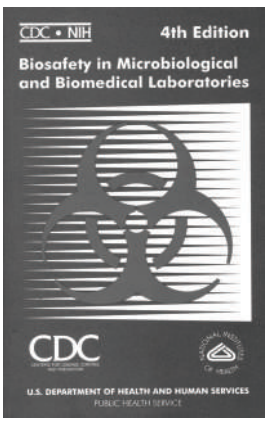


Fig. B.1. The biosafety guide of the Centers for Disease Control showing the universal biohazard symbol on the cover.

A BSL-1 laboratory works with material not known to consistently cause disease in healthy adult humans. Standard microbiology practices are expected and a sink must be available for washing hands.

A BSL-2 laboratory works with human-derived blood, tissues, bodily fluids, or primary human cell lines. The presence of an infectious agent may be unknown. Organisms are not known to transmit via an aerosol route. Splash shields, face protection, gowns, and gloves are required and exposure of percutaneous and mucous membranes must be prevented. Sinks for washing hands and a waste decontamination facility are required.

A BSL-3 laboratory handles organisms that have the potential for respiratory transmission and that may cause serious and potentially lethal infection. BSL-2 capabilities are needed plus additional emphasis on containment to protect contiguous areas and the environment. An example organism is *Mycobacterium tuberculosis*, the causative agent of tuberculosis in primates, including humans.

A BSL-4 laboratory works with organisms for which there is no available vaccine or therapy. There are very few BSL-4 laboratories. An example organism that requires BSL-4 facilities is the Marburg virus.

Other guides that are useful include the NIH guides on recombinant DNA and on Institutional Biosafety Review. See http://www4.od.nih.gov/oba/aboutoba.htm for additional information. Transportation and transfer of many biological substances, including DNA, are regulated and often require registration with the CDC and documentation of handling procedures. There are also Occupational Safety and Health Administration standards for working with blood known as the Blood borne Pathogen Standard. Finally, the Environmental Protection Agency regulates several of the chemicals used in a biotechnology laboratory (such as ethanol). These are issues that everyone faces. Benchmarking nearby laboratories with similar activities is an excellent method for identifying best practices.

Ethical, legal, and social issues

These are some of the questions that professionals working in the biotechnology field may consider.

Genetic testing. Should tests be offered if there is no treatment? Should tests be offered if the prediction of outcome is not definitive?

Insurability. Should people be denied health coverage because of their genes? Should genetically susceptible persons or society be asked to pay a higher insurance rate?

Employment. Should employers be able to deny a job based on a genetic predisposition or susceptibility?

Criminal justice. Should a person be held criminally liable if his/her behavior has a genetic basis?

Agriceuticals. What are the benefits and risks to people and the food supply?

Education. Do we ignore these issues as professionals in the field or do we become proactive?

Recommended Reading

Rather than attempting to acknowledge the first discovery of every scientific and technical accomplishment presented in this book, we have selected a short set of books and reports that may interest the reader. The "best" journals today include *Science*, *Nature*, and *Proceedings of the National Academy of Sciences* (*PNAS*). There are dozens of other journals that focus on medical, engineering, agricultural, and industrial applications. Books that we have found particularly useful in our group include

- *Schaum's Outline of Theory and Problems of Biology*, G. Hademenos and G. Fried, McGraw-Hill, ISBN 0070224056, 1998.
- *Biosafety in Microbiological and Biomedical Laboratories (BMBL)*, CDC and NIH, U.S. Government Printing Office, Fourth Edition, May 1999. Online http://www.cdc.gov/od/ohs/biosfty/bmbl4/bmbl4toc.htm.
- *Molecular Biotechnology*, B. Glick and J. Pasternak, American Society for Microbiology, Washington, DC, 1998.
- *Principles of Molecular Medicine*, J. Larry Jameson (Editor), Francis S. Collins (Editor), 1998, Humana Press; ISBN: 0896035298
- *Dorland's Illustrated Medical Dictionary* (*Standard Version*), W. A. Newman Dorland (Editor), 2000, W B Saunders Co.; ISBN: 0721662544
- *Molecular Biology of the Gene*, James Watson, 1997, Addison-Wesley Publishing; ISBN: 0805348247

Index

G

H

I

L

M

N

O

P

R

J. Patrick Fitch works at the University of California, Lawrence Livermore National Laboratory. For the past decade he has been a division leader with responsibilities that include genomics, bioengineering, and engineering research being conducted by more than 200 scientific and technical staff members. His research interests include bioinformatics, bioinstrumentation (automation, microelectromechanical systems, and photonic), and medical devices. Dr. Fitch received a Ph.D. in electrical engineering from Purdue University, W. Lafayette, Indiana in 1984 and BS degrees in physics and in engineering science from Loyola College, Baltimore, Maryland in 1981. He is a senior member of the IEEE, Fellow of the American Society for Laser Medicine and Surgery, member and short course instructor for SPIE, an editorial board member of *Biomolecular Engineering,* an advisory board member for the Colorado State University College of Engineering, and a former board member of the California State Breast Cancer Research Program. He received an IEEE best paper award in 1988 and national FLC awards for medical devices in 1998 and 1999. Dr. Fitch also successfully developed and marketed a medical device business strategy to venture investors. Prior to working on life science applications, Dr. Fitch was the principal investigator for a variety of imaging and computing projects applied to astronomy, nondestructive evaluation, and national security. In addition to scientific and technical journal articles, conference papers, and U.S. and international patents, he authored *Synthetic Aperture Radar*, published by Springer-Verlag in 1988. He may be reached at fitch@ieee.org