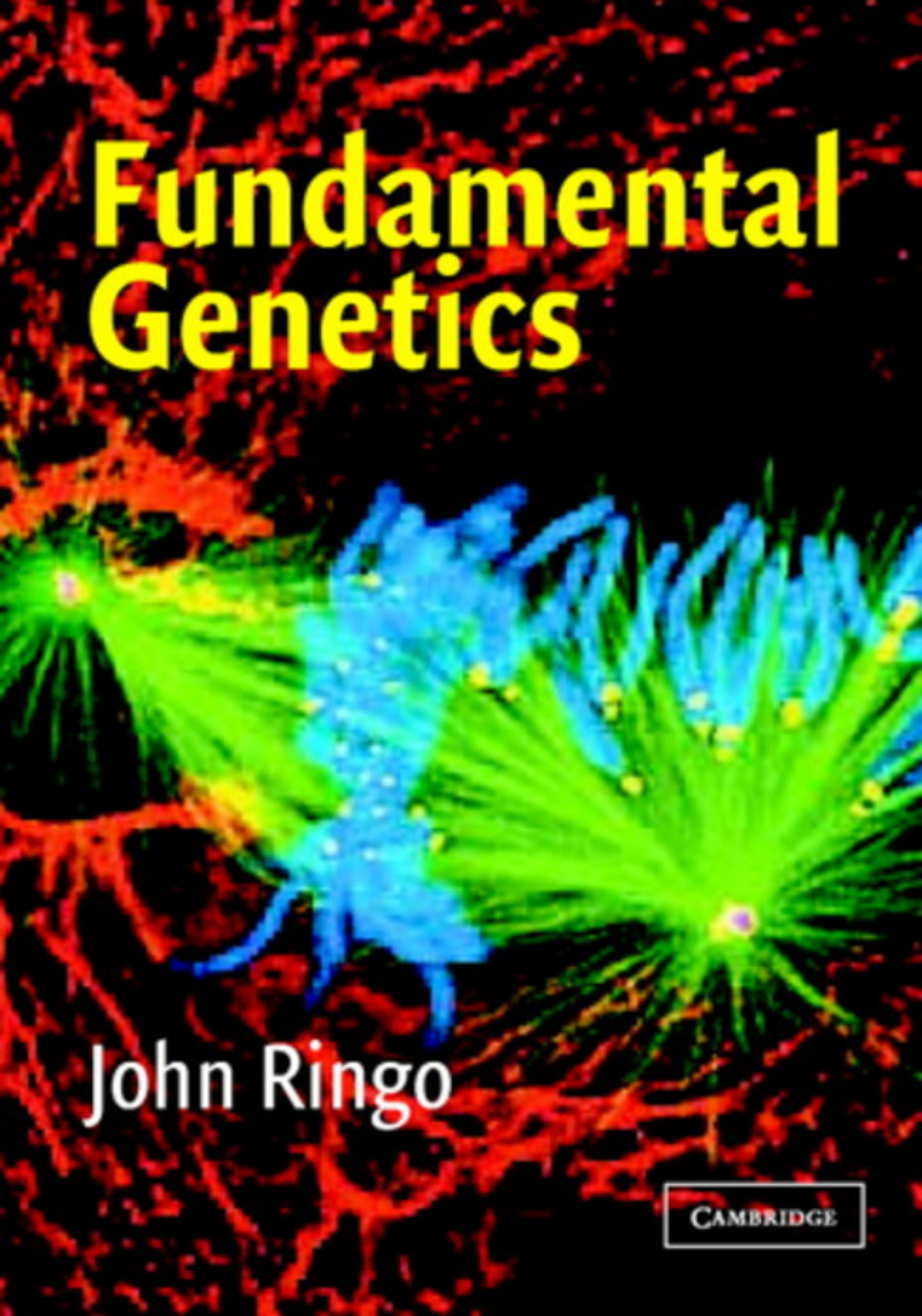
Fundamental Genetics



John Ringo

CAMBRIDGE

Fundamental Genetics

Fundamental Genetics is a concise, nontraditional textbook that explains major topics of modern genetics in 42 mini-chapters. It is designed as a textbook for an introductory general genetics course. *Fundamental Genetics* is also a useful reference or refresher on basic genetics for professionals and students in health sciences and biological sciences. It is organized for ease of learning, beginning with molecular structures and progressing through molecular processes to population genetics and evolution. Students will find the short, focused chapters approachable and more easily digested than the long, more complex chapters of traditional genetics textbooks. Each chapter concentrates on one topic, so that teachers and students can readily tailor the book to their needs by choosing a subset of chapters. The book is extensively illustrated throughout with clear and uncluttered diagrams that are simple enough to be re-drawn by students. This unique textbook provides a compact alternative for introductory genetics courses.

John Ringo is a Professor of Biology at the University of Maine where his research focuses on *Drosophila* genetics, behavior, and evolution. He has carried out laboratory and field studies in the United States and Israel. A member of the Genetics Society of America and the Society for the Study of Evolution, he has published works in various scholarly journals.

Fundamental Genetics

John Ringo
University of Maine

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2004

Printed in the United States of America

Typeface Swift 10/14 pt. and Gill Sans *System* LaTeX 2$_\varepsilon$ [TB]

A catalog record for this book is available from the British Library.

Library of Congress Cataloging in Publication Data
Ringo, John, 1943-
Fundamental genetics / John Ringo.
p. cm.
Includes bibliographical references and index.
ISBN 0-521-80934-7 (hb) – ISBN 0-521-00633-3 (pbk.)
1. Genetics. I. Title.
QH430.R55 2003
576.5–dc21 2003048463
ISBN 0 521 80934 7 hardback
ISBN 0 521 00633 3 paperback

This book is dedicated to the memory of my early teachers:
My parents Mary and Gene
My grandmothers Grace and Margaret
My second grade teacher, Myrtle McCullough, who saved me

Contents

Preface

I wrote this book both for the student learning genetics for the first time and for the biologist or health professional looking for background information. *Fundamental Genetics* is a brief account of the basic facts, theories, and experimental approaches of genetics. The book is unconventional. The organizing principle is to progress from structure to function, from simple to complex, and from molecular events to epiphenomena, such as how genes are inherited from parent to child. The history of genetics is scrupulously avoided, mainly to save space but also to avoid the trap of genuflecting to famous geneticists rather than discussing their experiments in the detail necessary for a full understanding – such a discussion, properly done, would fill several large volumes. This is a short book, and each chapter is focused.

A friend of mine says, "It's not how long life is that matters, but how thick." The chapters of *Fundamental Genetics* are short but thick, so if you are a student using the book as a text, it is probably best to read one and only one chapter at a sitting – any less, and you will have trouble getting the whole picture of the chapter; any more, and you will have mental indigestion. If you are using the book as a reference, the glossary may often be a helpful starting point. The teacher who chooses *Fundamental Genetics* as a textbook can skip chapters that are not appropriate for the course; it will be easy to use the chapters in a different order. The chapters are modules.

I have tried hard to be accurate and correct in writing this book, but it will contain mistakes. If you find a mistake, dear reader, or if you disagree with something in the book, or if some point seems opaque to you, contact me. I guarantee that I will answer any legitimate email inquiry.

Acknowledgments

I could not have written the book without the enthusiastic, unflagging support and encouragement of my wife, Ada Zohar, who kept me on track through the writing. Also, her many valuable comments on the early stages of the manuscript made for substantive improvements, especially in the introductions to each chapter.

Many colleagues helped me by explaining things and giving me source material; these include Danny Segal, Dusty Dowse, Bill Glanz, Irv Kornfield, Becky Van Beneden, Rondi Butler, Keith Hutchison, Malcolm Shick, Sara Orgad, Sarit Cohen, and Knud Nierhaus. Larry Beauregard, head of the Genetics Department, Affiliated Laboratories, Bangor, Maine, generously supplied micrographs of human chromosomes. Andrew Trevor, Tamar Schlick, Juergen Suehnel of IMB Jena Image Library, and the Protein Data Bank graciously consented to my using their images. I thank Alexei Khodjakov, and Nikon Corporation, for allowing me to use his beautiful micrograph for the cover.

I thank anonymous reviewers for their comments at an early stage in the writing. I am very grateful for the helpful, scholarly comments of Jeff Hall and Irv Kornfield, made late in the development of the book. Jeff's comments, insightful and spirited, were voluminous enough to fill a small book.

Katrina Halliday is my excellent editor. I thank her for her wisdom and understanding as much as for her help, as she shepherded the book along its way. Eleanor Umali, senior project manager at Techbooks, has been nothing short of astounding. She is a wizard of organization and efficiency, and I was lucky indeed to have had her on the team.

Chapter 1

Life Forms and Their Origins

Overview

This chapter introduces several basic genetic concepts, without going into detail about any of them. These genetic concepts are as follows:

- life form
- nucleic acid
- gene
- chromosome
- organism
- virus
- semiautonomous organelle

The origin of life and the evolution of the three domains of life are described briefly.

Life Forms Are Genetic Systems

Two essential components of every life form are proteins and nucleic acids. Nucleic acids (DNA and RNA) are thread-like coding molecules, the building material of genes and chromosomes. Genetics is about genes and chromosomes – their structure and function, their behavior and misbehavior, their evolution, and methods of studying them. Because genes are the coding molecules of life, they are complicated and varied. It is difficult to pin down the term "gene" in a simple definition, but, to a first approximation, a gene is a segment of nucleic acid whose immediate function is to encode a piece of RNA (Figure 1.1). The key concepts here are **replication** (copying) of genes and coding. The replication of genes and their coding properties are described in detail in later chapters.

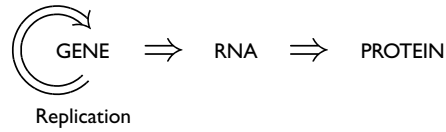


Fig 1.1 Genes replicate and code for RNA. For most genes, the final gene product is protein.

From a genetic point of view, a life form is an assemblage of large molecules capable of reproducing itself and including at least one chromosome. A chromosome is a long, thin thread made of DNA or, in some cases, RNA and may also contain proteins. To qualify as a chromosome, a nucleic acid molecule must contain one or more genes, be replicated faithfully in a regulated manner, and be transmitted from a life form to its descendants in a reproductive cycle. Not every molecule of nucleic acid is a chromosome, even if it contains genes. The nucleic acid part of a life form's set of chromosomes is its **genome**.

All life forms arise from preexisting life forms via a reproductive cycle during which chromosomes are copied and the copies are passed on from parent to progeny (Figure 1.2). According to this broad, genetically based definition, life forms include organisms (cellular forms), viruses, mitochondria, and chloroplasts. This book concentrates on the genetics of organisms.

Organisms

Organisms are made of cells, membrane-bound structures capable of reproduction, growth, and metabolism. The genome of a cell encodes all the proteins required for that cell's survival. Every cell has at least one chromosome, which is made of DNA and proteins. Cells also have many ribosomes, micromachines for synthesizing proteins. A membrane surrounds every cell. In some organisms, the cell envelope includes a cell wall and one or more additional membranes. An organism can be a single cell or many cells joined together.

Organisms comprise three major divisions or domains: Bacteria, Archaea, and Eukarya (Figure 1.3). The compelling genetic evidence for this broad taxonomic division comes from DNA sequences of slowly evolving genes. Despite their genetic and biochemical differences, bacterial and archaeal cells are morphologically similar: they lack nuclei, and they reproduce asexually, by simple cell division. Little is known about the genetics of archaea.

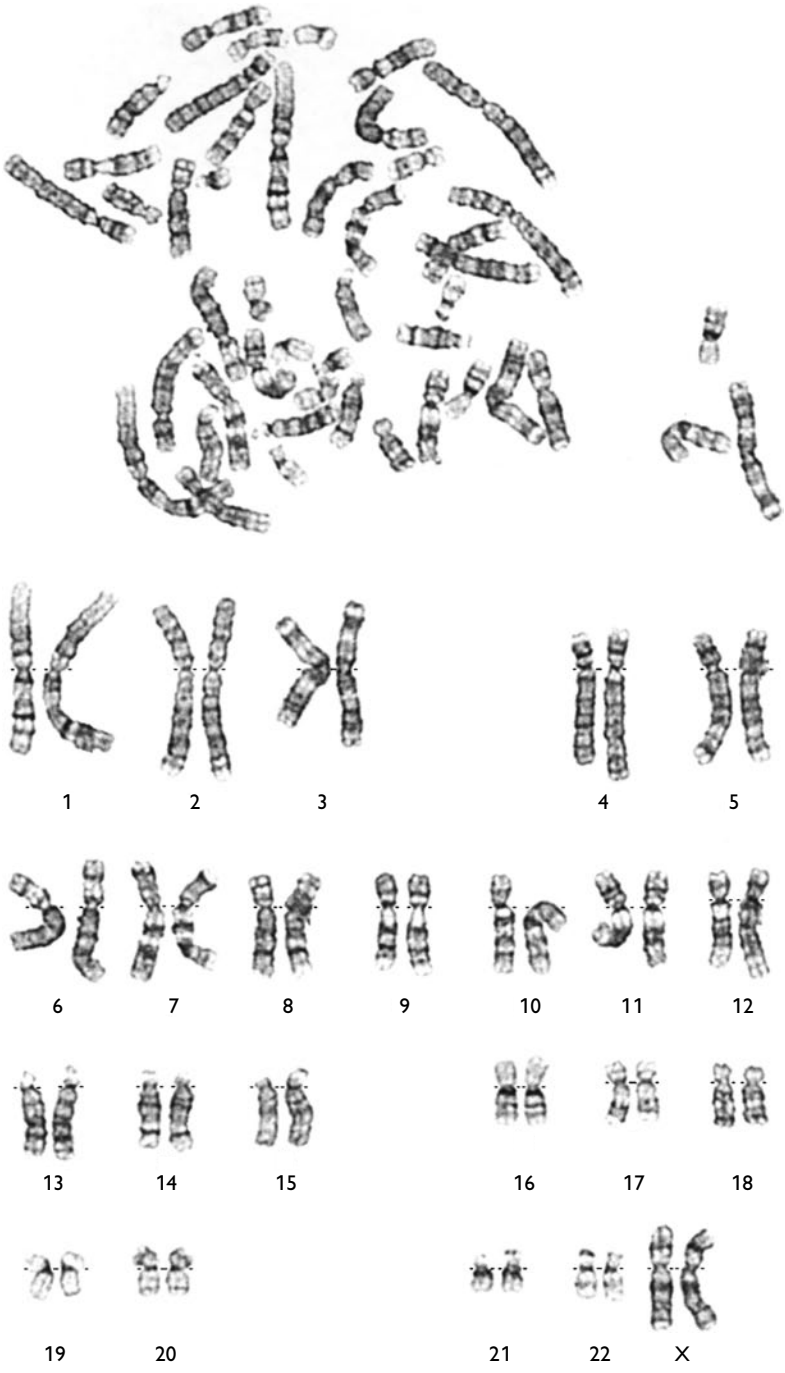


Fig 1.2 Above: Stained chromosomes of a normal female human, from a cell nearing division. Below: The same 46 chromosomes rearranged into numbered pairs, the karyotype. Photograph by Dr. Laurent Beauregard, Genetics Department, Affiliated Laboratory, Inc., Bangor, Maine.

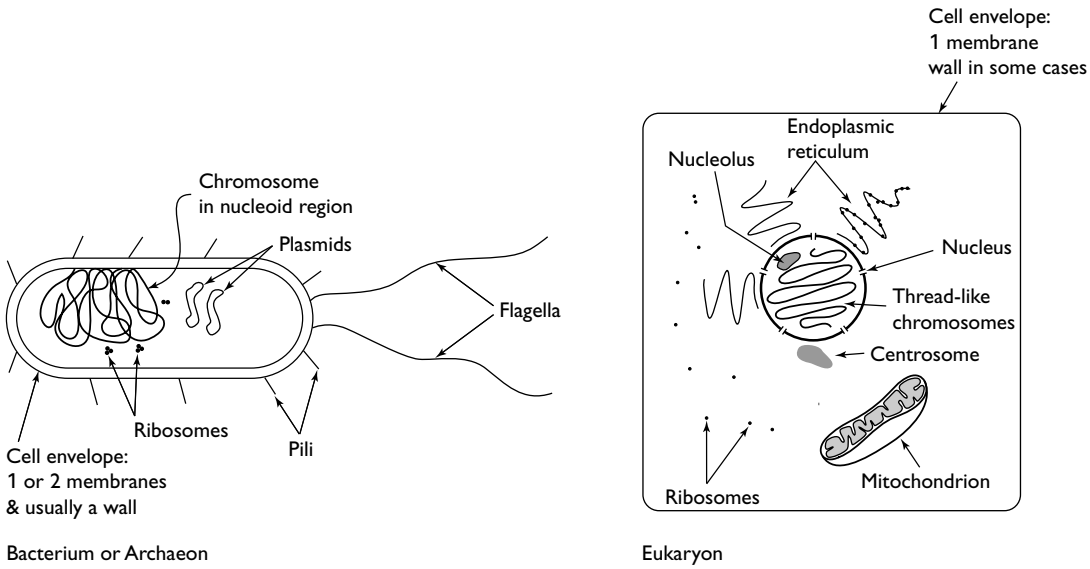


Fig 1.3 A quick look at two kinds of cell.

In contrast to bacteria and archaea, eukaryal cells possess membrane-bound nuclei, an internal system of membranes, and a cytoskeleton made of microtubules.

The Origin of Life

From the time the earth began to form, 4.6×10^9 years (4600 Ma = megaanum, or million years) ago, until it cooled sufficiently for liquid water to exist on its surface, 4400 to 4200 Ma ago, the temperature was too high for life to exist. Meteorites bombarded early earth, and some geophysicists believe these ocean-vaporizing impacts likely did not abate sufficiently for life to emerge until 4200 to 4000 Ma ago (Figure 1.4). The ^{12}C:^{13}C ratio of organic carbon is higher than that of inorganic carbon. This isotopic ratio in fossils of the most ancient sediments known suggests that life was abundant 3900 Ma ago or a bit earlier; sedimentary apatite

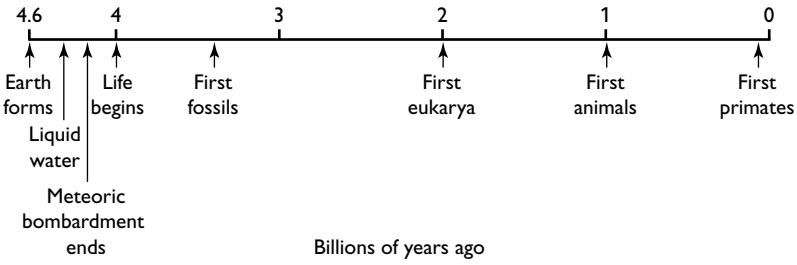


Fig 1.4 Time line of earth's history. The times given for biological "firsts" are very approximate.

(a mineral consisting of calcium phosphates) is a biomarker first appearing in large amounts 3800 Ma ago. The implications are remarkable: life emerged from non-life during a period lasting only 100 to 300 Ma.

Fundamental organic molecules required for life (e.g., amino acids and nucleotides) are thought to have originated through natural chemical reactions starting with simple molecules such as methane, ammonium, phosphate, and water, with the energy for the reactions being heat and electrical discharges in the atmosphere. Modern experiments have shown these reactions to be feasible. Also, under realistic conditions not involving enzymes, amino acids polymerize into polypeptides and nucleotides polymerize into nucleic acids.

The First Organisms: RNA-Based?

All living organisms have genes made of DNA, which code for RNA. RNA molecules are intermediate coding molecules in the synthesis of proteins, which make important structures of the cell and carry out virtually all the metabolic functions (Figure 1.5). According to one theory, the original life forms used RNA for coding and for metabolic functions. Some RNAs act as enzymes; these are ribozymes. Biochemists are finding many chemical reactions that are catalyzed by RNAs. If ancient proto-organisms possessed RNAs capable of directing the synthesis of more copies of RNA molecules, then both genes and enzymes could have been made of RNA in those ancient times, perhaps between 4000 and 3500 Ma ago. In this "RNA world" RNA served double duty: genes and enzymes.

The ancestral cells or protocells had evolved into bacteria-like cells by 3500 Ma ago; fossil cells that resemble bacteria were very abundant by then. The split between archaea and bacteria occurred between 3500 and 1900 Ma ago, and the eukarya-archaea split probably occurred between 1900 and 1500 Ma ago

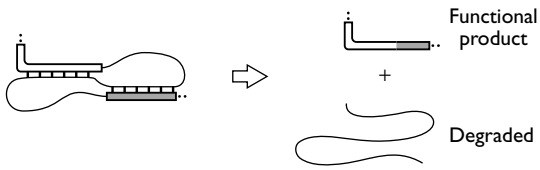


Fig 1.5 Some RNAs can both cut themselves and ligate the pieces.

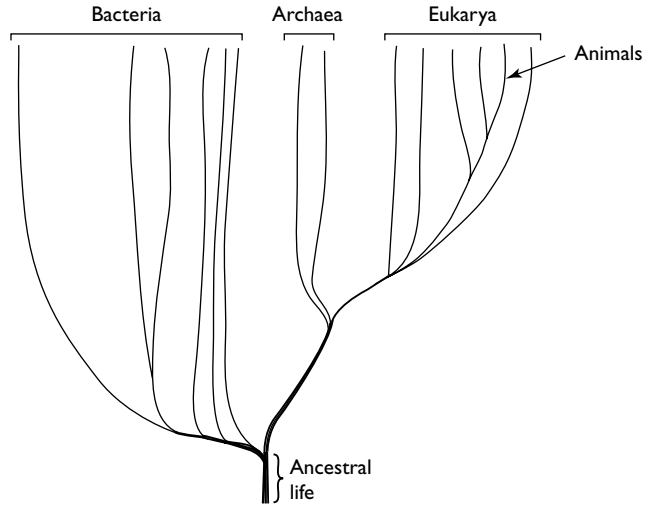


Fig 1.6 Evolutionary tree of three domains of life. Each line is a major taxonomic lineage.

(Figure 1.6). The eukarya are enormously diverse; taxonomists classify them into a stupefyingly detailed and complex hierarchy of taxa.

Mitochondria and Chloroplasts: Semiautonomous Organelles

Eukaryal cells contain several organelles or "mini-organs" inside the cell. Two important membrane-bound organelles found in many eukarya are the mitochondrion [pl., mitochondria] and the chloroplast. Their main functions are oxidative metabolism (mitochondria) and photosynthesis (chloroplasts). They evolved from purple bacteria and cyanobacteria, respectively (Figure 1.7). The ancestral bacteria became mutualistic endosymbionts in eukaryal

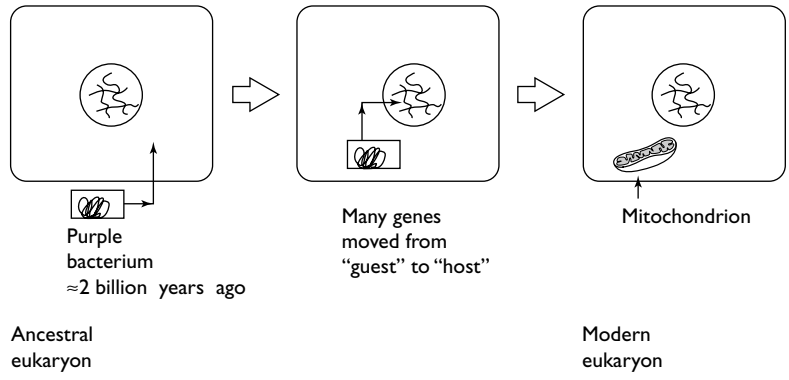
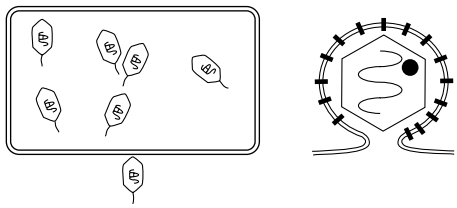


Fig 1.7 Evolution of mitochondria.

cells, meaning that both the host cell and the life form that lives inside it benefited. Mutualistic endosymbiosis is not rare. However, in these cases, the now-organelles have clearly lost their status as bacterial cells, for genes in the "host" eukaryon's nucleus encode many proteins of these organelles. Mitochondria and chloroplasts are therefore genetically parasitic. On the other hand, every mitochondrion and every chloroplast has its own chromosome and its own protein-coding machinery. Furthermore, mitochondria and plastids (chloroplasts and related organelles) are unlike any other membrane-bound organelle, such as the nucleus: mitochondria and chloroplasts are never disassembled and reassembled; instead, they reproduce by division, as did their ancestral bacteria.

Viruses, the Completely Acellular Life Forms

Viruses are acellular life forms: obligate intracellular parasites possessing one or more chromosomes. During the infectious stage of a virus's life cycle, the virus is a virion – the viral genome encapsulated in a structure made of protein. Sometimes the virion includes a membrane envelope stolen from the host's cytoplasmic membrane (Figure 1.8). Most viruses are genetically parasitic, relying on their host for enzymes used in genetic processes. Only viruses with relatively large genomes code for many of the proteins required for their own reproduction. Viruses infect all three domains of life and may be classified according to host, genetic material, or phylogeny. An overall phylogeny of viruses is not appropriate, for viruses appear to have evolved independently many times. Any phylogeny of viruses, except for closely related ones, is difficult to establish, owing to rapid evolution and the tendency of viruses to acquire cellular genes. The going theory is



Bacterial viruses

HIV budding off host cell membrane

Fig 1.8 Some infectious virus particles.

that virus genomes evolved from bits and pieces of their hosts' genomes.

A Few Odd Forms

Plasmids are small, nonessential, "extra" chromosomes of cells. Plasmids code for proteins, including proteins useful to the cell (e.g., genes conferring antibiotic resistance). Some plasmids move between cells and may have genes encoding the machinery for intercellular movement. Are plasmids life forms? Some say yes, because plasmids are parasite-like, but I opt for the idea that plasmids are merely small, inessential chromosomes. Another category of DNA molecule with some of the basic properties of a life form is the **transposon**. Transposons are sequences of DNA that can move about the genome, within or between chromosomes; transposons code for proteins that help them to move.

There are nucleic acid molecules that do not qualify as life forms by the definition offered here but that some biologists do consider living, or at least lifelike. These are **viroids** and **virusoids**, small circular RNA molecules that do not code for protein. Viroids are parasites of plants and cause significant economic damage. Virusoids are parasites of viruses.

The strangest of all life-oid things is the **prion**, an infectious protein that can cause the modification of similar proteins in a cell, ultimately leading to the cell's death. Prions cause certain slow, infectious, neurological diseases, including bovine spongiform encephalopathy ("mad cow disease").

Most of Genetics Is Based on a Restricted Sample of Organisms

There are over 10 million species in the three domains of life. Much of what is known about the genetics of cellular organisms has been learned from intensive study of a limited sample of species clustered in a few branches of the evolutionary tree of organisms, most prominently two bacteria (*Escherichia coli* and *Bacillus subtilis*), a few fungi (notably the mold *Neurospora crassa* and the bread yeast *Saccharomyces cervisiae*), two flowering plants (*Zea mays* and *Arabidopsis thaliana*), and four animals (the fruit fly *Drosophila*

melanogaster, the nematode *Caenorhabditis elegans*, the mouse *Mus musculus*, and *Homo sapiens*). Important genetic phenomena have been substantially investigated in hundreds of other species of bacteria, fungi, plants, animals, and ciliates, as well as in many viruses – only viral species, though, that infect bacteria or multicellular/multinucleate eukarya. The most conspicuous and troubling gaps in knowledge of basic genetics occur in the archaea and in the early-branching taxa of eukarya – troubling, because we have no idea how large those gaps may be. Fortunately, straightforward analysis of DNA sequences is beginning to allow us to infer a great deal about the genetics of these organisms.

Further Reading

Gesteland RF, Cech T, Atkins JF. 1999. *The RNA World,* 2nd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Gray MW. 1999. Evolution of organellar genomes. *Curr. Opin. Genet. Dev.* 9:678–687.

Holland HD. 1997. Evidence for life on earth more than 3850 million years ago. *Science* 275:38–39.

Lazcano A, Miller SL. 1996. The origin and early evolution of life. *Cell* 85:793–798.

Levine A. 1991. *Viruses.* WH Freeman, New York.

Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA* 87:4576–4579.

Chapter 2

Nucleic Acids

Overview

To understand genes, one must first consider nucleic acid, for nucleic acid is the stuff that genes are made of. Inasmuch as function follows structure, a clear picture of nucleic acid will illuminate all genetic processes.

The genetic material of all life forms is nucleic acid, either deoxyribonucleic acid (**DNA**) or ribonucleic acid (**RNA**). DNA and RNA are linear chains made of subunits called **nucleotides**. Two strands of DNA typically associate in a two-stranded helix. This chapter describes the structure of nucleotides, the way nucleotides are connected to make chains, and a bit about the shapes of RNA and DNA, including some properties of double helices.

Polymers

Nucleic acids, the coding molecules of life, are linear polymers. A linear polymer is an unbranched chain made of many subunits connected by covalent chemical bonds; short polymers are known as oligomers (Figure 2.1). The subunit of a nucleic acid polymer is the nucleotide. Energy [typically about 400 kJ/mol] is required to make or to break a covalent bond between subunits of a polymer.

In this simplified representation of an RNA segment, the circles and triangles stand for nucleotides and each line stands for a covalent bond joining adjacent nucleotides. A = adenosine, G = guanidine, U = uridine, and C = cytidine.

Nucleotides

Every nucleotide has three parts: sugar, nitrogenous base, and one to three phosphates; the subunits of nucleic acid are

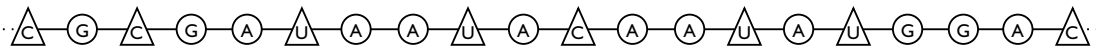


Fig 2.1 An RNA polymer.

monophosphates. A **nucleoside** is a nucleotide minus the phosphate (nucleoside = sugar + nitrogenous base) (Figure 2.2). Ribose is the sugar of RNA (ribonucleic acid) and 2′-deoxyribose is the sugar of DNA (deoxyribonucleic acid).

In ribose, the sugar of ribonucleotides, and 2′-deoxyribose, the sugar of deoxynucleotides, carbon atoms are numbered 1′ ("one-prime") through 5′ (Figure 2.3). Numbers without primes designate the carbon and nitrogen atoms in the nitrogenous bases.

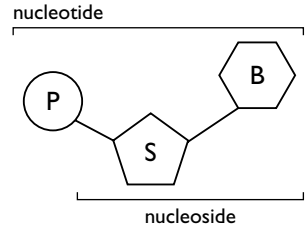


Fig 2.2 Nucleotide and nucleoside. S = sugar, B = base, P = phosphate.

The nitrogenous bases are derivatives of **purine** and **pyrimidine** and are called by those names, even though RNA and DNA do not contain the parent molecules purine and pyrimidine (Figure 2.4). Purine and its derivatives have a double ring structure; pyrimidine and its derivatives have a single ring structure. The most common purine derivatives are guanine and adenine; the most common pyrimidine derivatives are cytosine, uracil, and thymine. Only four nucleotides are used for the synthesis of RNA in organisms — those containing the bases adenine, guanine, cytosine, and uracil. DNA synthesis in organisms is similar, except that thymine is used instead of uracil. In RNA molecules, there are in addition about 50 less common bases, produced by enzymatic modification of bases within an RNA chain. The DNA of some viruses contains uncommon bases. In the DNA of plants and animals, roughly 5% of the cytosines have a methyl group added, at position 5.

The deoxynucleotides are sometimes abbreviated with an initial "d"; thus, deoxyguanidine is sometimes abbreviated dG. Adding a "d" for deoxy is done only when RNA and DNA sequences are given together, to avoid confusion. A single nucleotide may be a monophosphate, diphosphate, or triphosphate. Thus, AMP, ADP, and ATP are the mono-, di-, and triphosphates of adenosine (Figure 2.5).

The names of common bases and corresponding nucleotides are given in Figure 2.6.

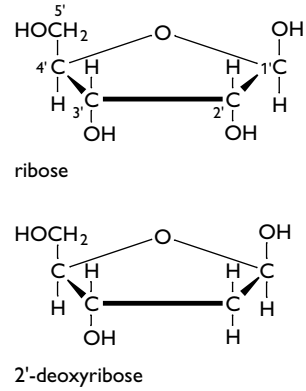
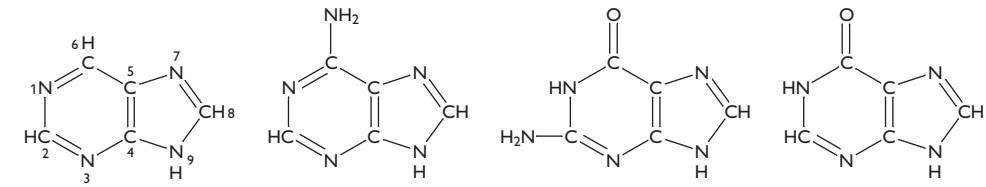


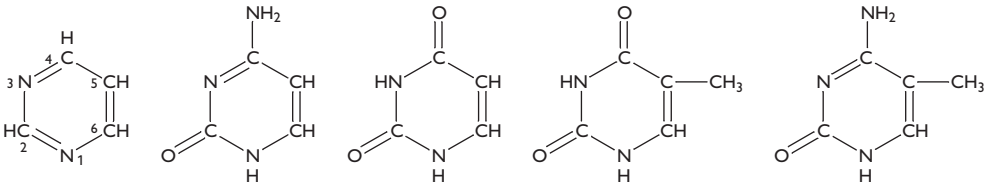
Fig 2.3 Ribose and 2′-deoxyribose

The Phosphodiester Bond between Nucleotides

In all known life forms, nucleotides in a polymer are joined covalently by a phosphodiester bond between the 3′ carbon of one sugar and the 5′ carbon of the next sugar. The string of sugars



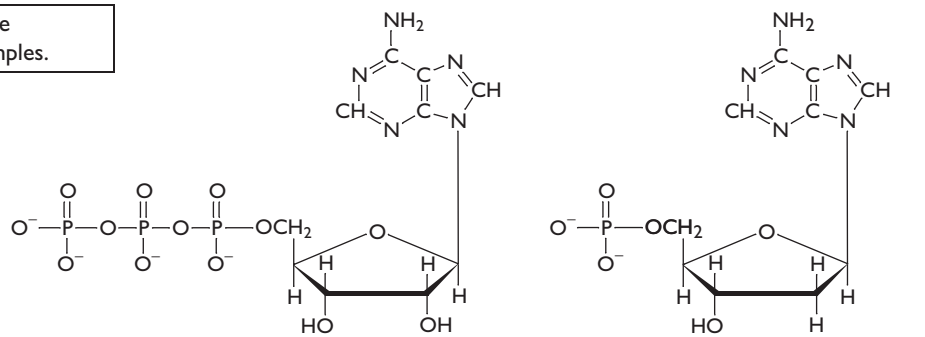
a)



b)

Fig 2.4 Examples of bases found in DNA and RNA.
(**a**) Purine and its derivatives. Adenine and guanine are the commonest purine bases. Hypoxanthine, less common, is found in many tRNAs (small RNAs that carry amino acids into the ribosome during protein synthesis).
(**b**) Pyrimidine and its derivatives. Cytosine, uracil, and thymine are the commonest pyrimidine bases.

Fig 2.5 Nucleotide structure: two examples.



Adenosine triphosphate (ATP)
(MW = 503 Da)

Deoxyadenosine monophosphate
dAMP is a subunit of DNA (MW = 329 Da)

Purine base	Ribonucleotide	Pyrimidine base	Ribonucleotide
adenine	adenosine (A)	cytosine	cytidine (C)
guanine	guanidine (G)	thymine	thymidine (T)
hypoxanthine	inosine (I)*	uracil	uridine (U)

*An exception to the pattern of a nucleotide's name being derived from the name of its base.
unspecified nucleotide (N) unspecified purine nucleotide (Pu) unspecified pyrimidine nucleotide (Py)

Fig 2.6 Names and abbreviations of common bases and the corresponding nucleotides.

and phosphates makes the "backbone" of the polymer, and bases stick out from the sugars (Figure 2.7).

Nucleotides Pair via Hydrogen Bonds

Nucleotides tend to pair via hydrogen bonds between bases. A hydrogen bond is weak relative to a covalent bond and takes little energy to make or break [roughly 5 kJ/mol]. Hydrogen bonds form between adenine and thymine (AT), between adenine and uracil (AU), and between guanine and cytosine (GC). AT and AU pairs have two hydrogen bonds, and GC pairs have three hydrogen bonds (Figure 2.8). Other nucleotide pairings are possible between RNAs in special contexts. Pairing may occur between two strands of DNA, two strands of RNA, or strands of DNA and RNA.

Nucleic Acids Make Double Helices by Base-Pairing

The pairing of nucleic acids promotes formation of a helical duplex. A single strand of nucleic acid may bend around and pair with itself, in which case a short, single-stranded duplex is

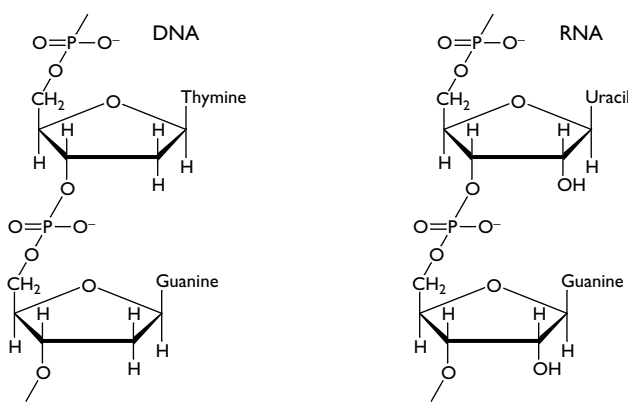
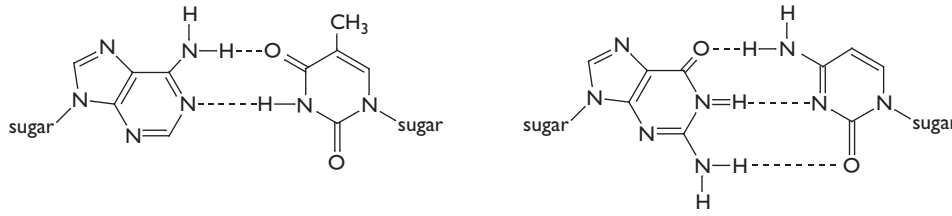


Fig 2.7 Sugar-phosphate backbone of nucleotides.



A:T base pair (MW = 615 Da)

G:C base pair (MW = 616 Da)

Fig 2.8 Normal base-pairing between nucleotides of DNA.

produced. When two separate strands pair they make a long, helical, double-stranded duplex. The chromosomes of all organisms are made of double-stranded DNA. A region of pairing may contain as few as 2 base pairs, or as many as $\sim 10^8$ base pairs — an entire, large chromosome (Figure 2.9).

When two strands of nucleic acid pair and intertwine they usually make a right-handed double helix. Grasp a pen in your right hand and point your thumb up; your four fingers turn in a right-handed direction, like the strands of DNA. The bases are on the inside of the double helix, and the sugar-phosphate backbones are on the outside. As shown in the figure on the previous page, the two strands of the double helix are oriented in opposite directions: looking down the length of a double helix, the linkages between nucleotides run from 5′ carbon to 3′ carbon in one strand and from 3′ carbon to 5′ carbon in the other strand. The overall appearance is that of a twisted ladder whose rungs are the base pairs. There are grooves in the helix between the pair of backbones. The helix is asymmetrical, with major and minor grooves.

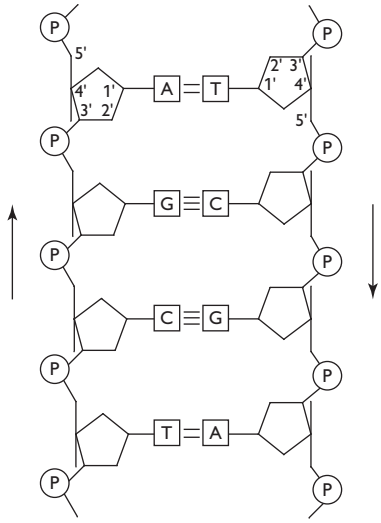


Fig 2.9 Double-stranded nucleic acid. This picture is flattened to simplify the view; the helical shape is shown on the next page. The two strands point in opposite directions.

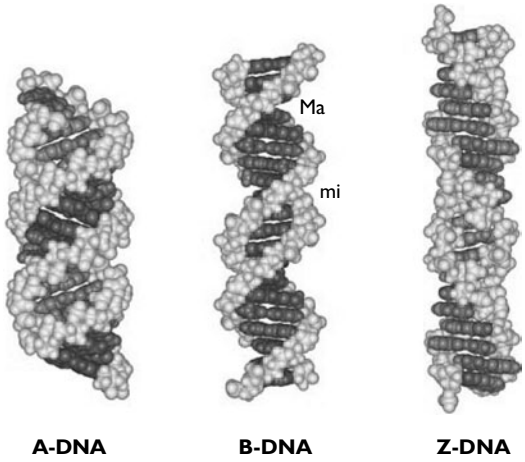


Fig 2.10 Nucleic acid helices. IMB Jena Image Library of Biological Macromolecules, 50 years of double DNA helix structure.

In life, two main forms of nucleic acid duplex are known, A-form and B-form, both right-handed helices. The form of a helix heavily depends on the kind of sugar (ribose or deoxyribose). Ribose contains a bulky OH group on carbon 2′, which influences bond angles in the sugars; these bond angles, in turn, influence the shape of the polymer. RNA and RNA-DNA hybrids are A-form, and DNA is mainly B-form. A rare form of duplex DNA is called Z-form, a skinny, left-handed, zigzag helix that occurs in short regions (Figure 2.10). Short triplex regions very rarely form when one strand loops out of the double helix at one point and adds to a nearby stretch; this is called H-form DNA.

A- and B-form helices differ in several ways (Figs. 2.10 and 2.11). In the B-form helix, the base pairs are centered deep inside and lie flat in a plane perpendicular to the axis of the helix; the phosphates are on the outside; and the sugars lie between bases and phosphates. In the thicker A-form helix, the base pairs sit halfway between the center and the outside of the helix and are tilted 20° from the horizontal; both the phosphates and the sugars are on the outside. Minor, local variations in helix structure are caused by the sequence of nucleotides (base content) and interactions between the nucleic acid and proteins.

Fig 2.11 Some approximate average characteristics of A-, B-, and Z-form nucleic acid. RNA and RNA-DNA hybrids are A-form. Most double-stranded DNA is B-form. The rare Z-form DNA occurs in short sequences.

Form	Mean thickness	Mean number of bases/twist	Length of twist	Direction of twist
A	2.2 nm	11	2.5 nm	righthanded
B	2.0 nm	10	3.3 nm	righthanded
Z	1.8 nm	12	4.6 nm	lefthanded

Other Characteristics of Double-Stranded DNA

The shape of the B-form DNA is neither invariant nor constant. The dimensions of the helix vary slightly from place to place. In addition, base pairs come apart and then re-pair — the double helix "breathes."

Water associates with DNA and influences its structure. In life, the sugar-phosphate backbone of B-form DNA is fully hydrated, and water occupies both the major groove and the minor groove. Water in the minor groove stabilizes the helix.

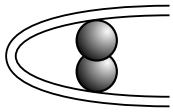


Fig 2.12 Proteins can bend DNA.

Double-stranded DNA as described above is called "relaxed DNA" because it has no extra twists. Relaxed DNA is rather stiff; the smallest circle made of it would be at least 170 nucleotide pairs long. DNA-binding proteins bend DNA, causing it to make tighter turns than it could by itself (Figure 2.12). DNA-bending proteins are commonplace in all organisms. In addition to bending, other variations in the shape of B-form DNA, mostly local variations, occur (Figure 2.13): single-stranded regions, loops and hairpins, and supercoiling.

All DNA chromosomes contain supercoiled regions. Everyday objects become supercoiled – string, rubber band, telephone cord, garden hose. Take a piece of multistrand string or rope in hand

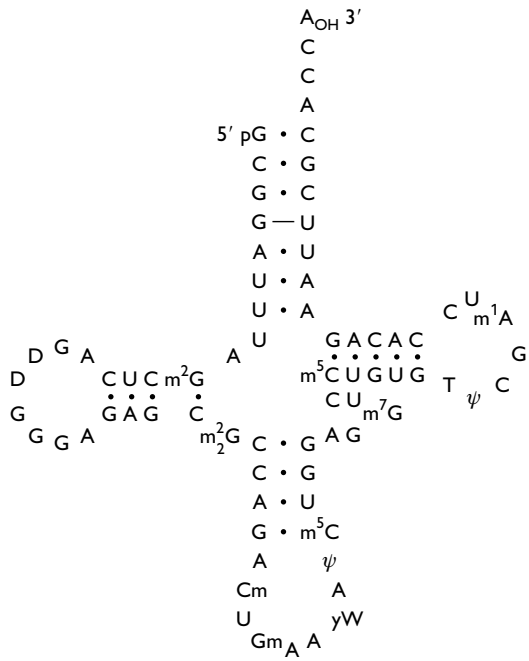


Fig 2.13 Hairpins result from intrastrand base-pairing. In this flattened picture of tRNA, 13 of the 76 nucleotides contain modified bases.

Relaxed chromosome Supercoiled chromosome

Fig 2.14 The appearance of a small, circular chromosome, relaxed and supercoiled, based on electron micrographs.

and twist it in the same direction as the strands twist; the result is overwinding, or positive supercoiling. Take the same piece of multistrand string or rope in hand again and twist it in the direction opposite to the strand twisting; the result this time is underwinding, or negative supercoiling.

DNA in chromosomes can easily become supercoiled. In cells, DNA is slightly underwound through the agency of DNA-binding proteins, so that the mean number of nucleotides per twist is increased over that of fully relaxed DNA by about 6%. Moreover, when duplex DNA is unwound, which must happen every time RNA is synthesized and every time DNA is replicated, positive supercoiling rapidly develops ahead of the unwinding point, which requires the action of special enzymes to relieve this supercoiling (Figure 2.14).

Some Commonly Encountered Abbreviations

Double-stranded DNA is often abbreviated **dsDNA**; single-stranded DNA is often abbreviated **ssDNA**. The size of a nucleic acid polymer is usually given as the number of nucleotides or nucleotide pairs. A standard unit of size is the **kilobase (kb)**. This is applied to DNA and RNA and to both double-stranded and single-stranded nucleic acid. For example, a piece of ssRNA that consists of 3200 nucleotides is 3.2 kb long, and a piece of dsDNA that consists of 3200 nucleotide pairs is 3.2 kb long. One thousand kilobases is a **megabase** (**Mb**).

Further Reading

Adams RLP, *et al.* 1992. *The Biochemistry of the Nucleic Acids,* 11th ed. Chapman and Hall, London.

Dickerson RE, *et al.* 1982. The anatomy of A-, B-, and Z-DNA. *Science* 216: 475–485.

Travers A. 1993. *DNA-Protein Interactions.* Chapman and Hall, London.

Chapter 3

Proteins

Overview

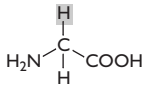
Proteins are linear polymers of **amino acids**. Genes encode all proteins, and proteins perform essential roles in all genetic processes, including the synthesis of DNA, RNA, and proteins. Some proteins bind to DNA and RNA, and some proteins are enzymes that act on nucleic acids. Some proteins bind to specific nucleotide sequences, but others bind equally well to any nucleotide sequence. This chapter describes the main points of protein size and structure.

Amino Acids

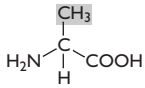
Protein is a generic term for a linear polymer made of amino acids as well as an aggregate of these polymers. An amino acid is a small carboxylic acid with an amino group and a side group that defines it. There are hundreds of different amino acids, but only 22 that are known to be genetically encoded, and two of these – selenocysteine and pyrrolysine – are found in only a handful of proteins. The molecular masses of the 20 common, genetically encoded amino acids range between 75 and 204 Da (Figure 3.1). Notice that amino acids are smaller than nucleotides.

Peptides

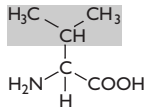
A peptide is a short polymer of amino acids, usually 30 amino acids or fewer. Adjacent amino acids in peptides are held together by a **peptide bond** (Figure 3.2) – a covalent bond between the carboxyl carbon atom of one amino acid and the core amino nitrogen atom of the other. A polypeptide is a large polymer of amino acids, usually 100 amino acids or more. There is no standard



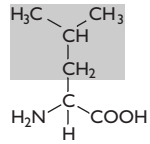
Glycine
Gly, G



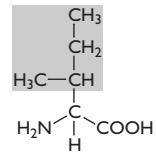
Alanine
Ala, A



Valine
Val, V

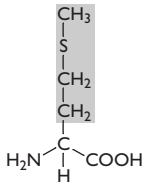


Leucine
Leu, L

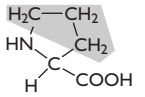


Isoleucine
Ile, I

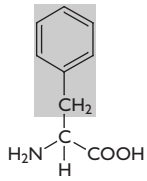
Nonpolar aliphatic (not containing a benzene ring)



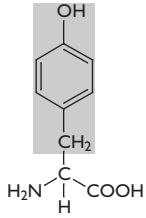
Methionine
Met, M



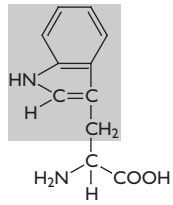
Proline
Pro, P



Phenylalanine
Phe, F

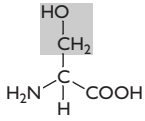


Tyrosine
Tyr, Y

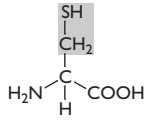


Tryptophan
Trp, W

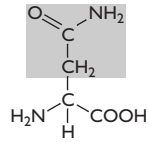
Nonpolar aromatic



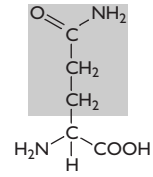
Serine
Ser, S



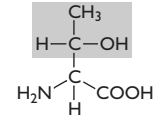
Cysteine
Cys, C



Asparagine
Asn, N

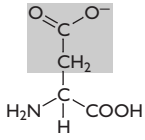


Glutamine
Gln, Q

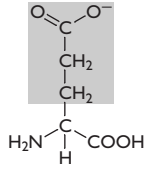


Threonine
Thr, T

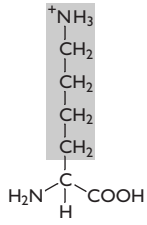
Polar and uncharged



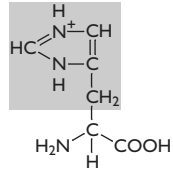
Aspartic acid
Asp, D



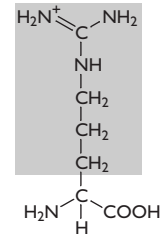
Glutamic acid
Glu, E



Lysine
Lys, K



Histidine
His, H



Arginine
Arg, R

Polar and charged

Fig 3.1 The 20 common, genetically encoded amino acids, with three-letter and one-letter abbreviations. They are arranged according to attributes of the side group. User-friendly three-letter abbreviations of amino acids are being supplanted by less easily remembered one-letter abbreviations.

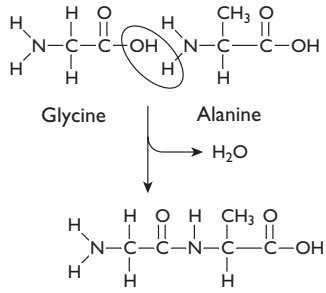


Fig 3.2 The peptide bond.

convention to delineate peptides and polypeptides. A protein may be a single polypeptide, or it may comprise many polypeptides held together by weak bonds such as hydrogen bonds.

Protein structure has several levels of organization: **primary**, **secondary**, **tertiary**, and **quaternary**.

Primary Structure

Primary structure is the sequence of amino acids. The thread of amino acids is considerably thinner than a single strand of RNA or DNA, because amino acids are smaller than nucleotides. The end of the polypeptide with a free amino group is called the N terminus, and the end with a free carboxylic acid group is called the C terminus (Figure 3.3).

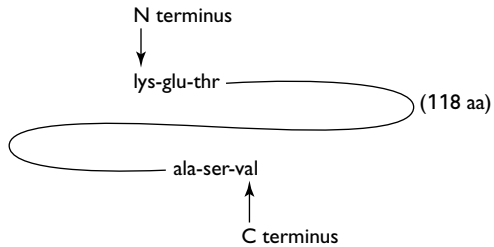


Fig 3.3 Primary structure of bovine ribonuclease A, a 124-amino acid. N terminus at top right 14 kDa protein; C terminus at left center.

Secondary Structure

Secondary structure is the shape assumed spontaneously by each segment of a polypeptide; the most common ones are alpha (α) and beta (β) (Figure 3.4). The α-helix is a right-handed helix about 0.3 nm thick, averaging 3.6 amino acids per turn; the length of one turn is about 0.6 nm. The β-ribbon is a flat section of amino acids; two or more β-ribbons often associate into a β-sheet, formed by hydrogen bonds between two or more of the ribbons. β-sheets impart strength to a protein and are found in proteins with structural roles.

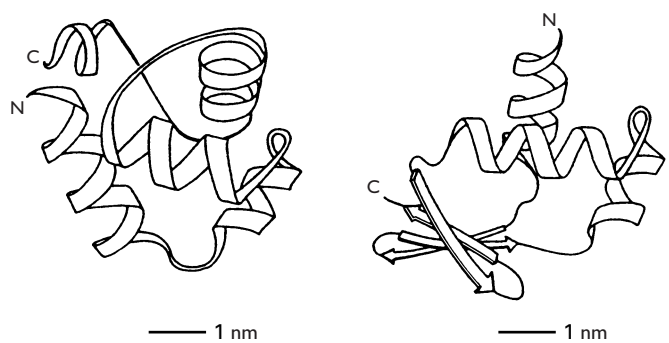


Fig 3.4 (Left) Four α-helices in part of a DNA-binding protein. (Right) β-sheet (arrows) and 3 α-helices in part of a DNA-binding protein. Travers 1993, *DNA-Protein Interactions*, Chapman & Hall.

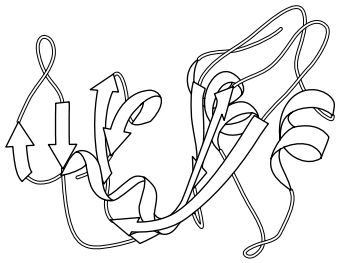


Fig 3.5 Bovine Ribonuclease A, a 124-amino acid, 14-kDa protein; this protein is a dimer, but for simplicity only one of the two identical subunits is shown. Protein Data Bank, Sadasivan et al. 1998.

Tertiary Structure

Tertiary structure is the folding pattern of secondary structures such as α-helices into a three-dimensional conformation (Figure 3.5). Proteins called chaperonins help in the formation of tertiary structure. In diagrams of tertiary structure, the strands of α-helices and β-ribbons are shown as flattened sections, while the connecting regions are depicted as thin threads.

Quaternary Structure

Quaternary structure is the association of two or more polypeptides via weak chemical bonds. Each polypeptide is then a subunit of the whole protein, and subunits are sometimes given Greek letters (Figure 3.6). For example, the fully active RNA polymerase enzyme in the bacterium *Escherichia coli* is a pentamer $= \alpha_2\beta\beta'\sigma$, and the adult human hemoglobin molecule is a tetramer $= \alpha_2\beta_2$.

The number of subunits in a protein or other molecule is designated by a prefix. The following table lists the most-used of these prefixes. Examples: (1) Bacterial RNA polymerase (Chapter 7) is a pentamer (it is an association of 5 polypeptides). (2) A typical

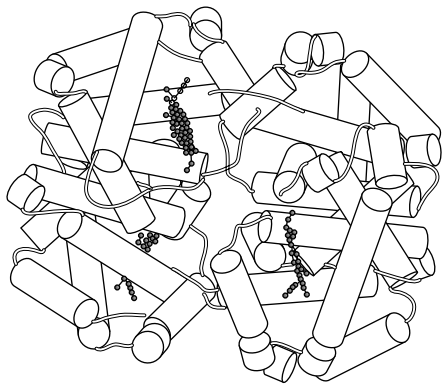
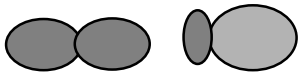


Fig 3.6 Adult human hemoglobin, a 66-kDa tetramer; each α chain has 141 amino acids, and each β chain has 146 amino acids. A heme molecule sits at the heart of each polypeptide. In this figure a cylinder represents an α helix. Protein Data Bank (primary citation not available).

Number of subunits	Name	Number of subunits	Name
1	monomer	11	undecamer
2	dimer	12	dodecamer
3	trimer	13	tridecamer
4	tetramer	14	tetradecamer
5	pentamer	15	pentadecamer
6	hexamer	20	eicomer
7	heptamer	21	uneicomer
8	octamer	22	doeicomer
9	nonamer	30	tricomer
10	decamer		

polymerase chain reaction primer (Chapter 28) is an eicomer (it is a chain of 20 nucleotides).

If the subunits of a protein are identical, use the prefix "homo"; if not, use the prefix "hetero" (Figure 3.7).



a) **b**)

Fig 3.7 Homodimer (**a**) and heterodimer (**b**).

Every polypeptide is encoded by an RNA molecule. After a polypeptide has been synthesized, it may be trimmed or even spliced, and some of its amino acids may be modified enzymatically – for example, by the addition of an acetyl group or a sugar. Peptides are made by cutting a polypeptide into pieces.

Protein Domains

A protein, even a single polypeptide, may have several structurally and functionally distinct regions called **domains** (Figure 3.8). A domain contains several secondary structures (α-helix, β-ribbon). Domains are connected in such a way that they can move separately from each other. A function may be carried out by a single domain, or by two domains working together. On the order of

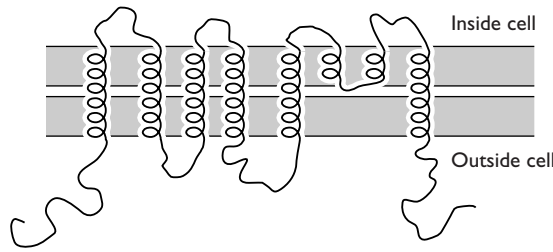


Fig 3.8 Receptor with transmembrane domain. A receptor binds to a small molecule (e.g., estrogen) that acts as a signal to the cell.

10^3 domain motifs associated with specific functions have been found.

Proteins Assist Genetic Processes

Neither RNA nor DNA works by itself. Proteins participate in every genetic process – for example, synthesis and degradation of DNA, mutation, repair of mutation, recombination, synthesis and degradation of RNA, transport of nucleic acids, protein synthesis, positioning and moving chromosomes, nuclear division, and cell division. Proteins also make up part of the chromosomes of organisms. While carrying out these functions, proteins must bind specifically to other proteins, bind to nucleic acids, and, in some cases, catalyze chemical reactions.

Protein–Protein Binding

Proteins tend to stick to each other. Much of a protein's stickiness is highly specific, promoting the marriage of some proteins but not others, and ensuring that each copy of a multimeric protein is put together in the same way. Proteins are held together primarily by weak chemical bonds (e.g., hydrogen bonds). The association of proteins through many weak bonds is strong enough to withstand random, thermal movement, yet weak enough to allow for disaggregation in response to molecular signals.

Non-Sequence-Specific DNA Binding

Histones are important proteins that bind to DNA in a non-sequence-specific way. DNA wraps around histones to make the nucleosome, an essential molecular structure in eukaryal chromosomes (Chapter 5). Some enzymes that act on DNA bind equally well to all nucleotide sequences (Figure 3.9). How do proteins bind to DNA without regard to sequence? Such proteins recognize DNA by its shape, binding mainly to the DNA's sugar-phosphate backbone.

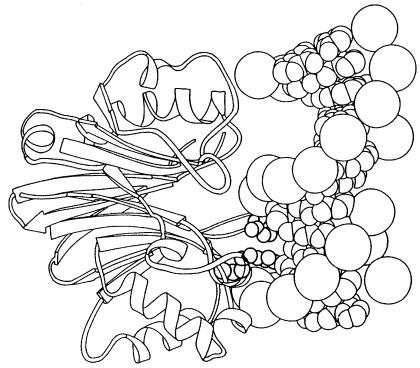


Fig 3.9 DNase I, an enzyme that cuts DNA, binds to the minor groove of DNA. Ma = major groove; protein contacts minor groove. Travers 1993, *DNA-Protein Interactions,* Chapman and Hall.

Sequence-Specific DNA Binding

Proteins that bind to specific sequences of double-stranded DNA include enzymes that synthesize RNA, enzymes that cut DNA, and regulatory proteins (Figure 3.10). Proteins that regulate genetic processes – RNA synthesis, DNA synthesis, recombination, DNA repair – may bind to DNA at specific sequences to do their work. Proteins that bind to specific sequences fit into major or minor grooves, and binding is based on hydrogen bonds between amino acids and specific purine and pyrimidine bases.

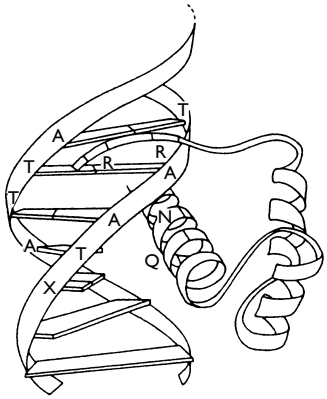


Fig 3.10 A regulatory protein inserts into DNA grooves and binds to specific nucleotides. Travers 1993, *DNA-Protein Interactions,* Chapman and Hall.

RNA-Binding Proteins

Many proteins bind to RNA, including enzymes, ribosomal proteins, and, in eukarya, proteins that carry RNAs from the nucleus to the cytoplasm. RNAs are usually single stranded but form secondary structures such as short A-form duplex regions (usually less than a single twist) and open loops. RNA-binding proteins seem to be of two broad types: groove binders and single-strand binders. Groove binders recognize features of RNA duplex regions and either sit in the broad, shallow minor groove or else insert into the major groove from the ends. Single-strand binders recognize specific bases, often in segments of RNA that lack secondary structure (single-stranded, nonduplex regions). The binding domains of these proteins are often either α-helices or β-ribbons.

Further Reading

Dickerson RE, Geis I. 1983. *Hemoglobin.* Benjamin/Cummings, Menlo Park, CA.

Doolittle RF. 1995. The multiplicity of domains in proteins. *Annu. Rev. Biochem.* 64:287–314.

Draper DE. 1999. Themes in RNA protein recognition. *J. Mol. Biol.* 293:255–270.

Sadasivan C *et al.* 1998. Plasticity, hydration, and accessibility in ribonuclease A. *Acta Cryst. D. Biol. Chryst.* 54 pp. 1343.

Travers A. 1993. *DNA-Protein Interactions.* Chapman and Hall, London.

Chapter 4

Simple Chromosomes

Overview

Every life form's genetic material is packaged in one or more chromosomes. Its **genome** is the nucleic acid in one complete set of chromosomes, excluding nonessential ones. This chapter is about the composition, size, shape, and number of chromosomes in bacteria, archaea, mitochondria, chloroplasts, and viruses – all but the eukarya. The nucleic acid of chromosomes is double-stranded DNA, except for a few single-stranded DNA plasmids and except for some viruses.

Bacteria and Archaea

Bacteria and archaea have two kinds of chromosomes, **essential chromosomes**, which are required for the survival and normal functioning of the cell, and **plasmids**, which are not absolutely necessary for survival and reproduction. Most bacteria and archaea whose genomes have been analyzed have only one essential chromosome, and in nondividing (nonreproducing) cells there is one copy of that chromosome. The number of different types of plasmids per cell varies from zero to several, and the number of copies of each plasmid ranges from 1 to $\sim 10^2$, depending on the plasmid. Usually, essential chromosome is simply referred to as chromosome.

In most bacteria the genome is located in an amorphous region, the **nucleoid** (Figure 4.1), which takes up from a quarter to half the cell's volume. It is not known whether plasmids are also restricted to the nucleoid. The parts of the chromosome where DNA replication begins and ends are attached to the cell's membrane. Unlike the nucleus of a eukaryal cell, a membrane

does not bound the nucleoid region. An exception is the bacterium *Mycoplasma gallisepticum*, whose chromosome sits inside a membrane-bound organelle; perhaps other species of bacteria have this nucleus-like structure.

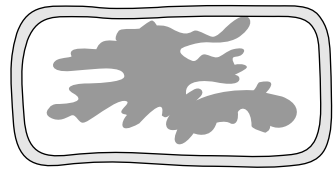


Fig 4.1 Nucleoid region.

Essential Chromosomes

The total genome size of bacteria and archaea ranges from 0.5 to 9.2 megabases (Mb). Most species have one essential chromosome, but some species have two or more (Figure 4.2). A chromosome in bacteria and archaea is made of one molecule of double-stranded DNA plus associated proteins. In most taxa chromosomes are closed circles, but in a few cases chromosomes are linear. Bacteria undergoing cell division cycles have multiple, identical copies of each chromosome.

E. coli has a single essential chromosome, fairly typical in size for bacteria: a 4.6-Mb circle of DNA with a mass of $\approx 3 \times 10^9$ Da, roughly 1% of the cell's total mass. The DNA of *E. coli*'s chromosome, if it were relaxed, would be 1.6 mm long and 2 nm thick – 800,000 times as long as it is thick and about 500 times as long as the 3-μm cell, making chromosome packaging a serious "problem." Proteins bind to DNA to condense it, so that it fits into the nucleoid region. The *E. coli* chromosome makes about 50 loops, about 100 kilobases (kb) per loop, and each loop is

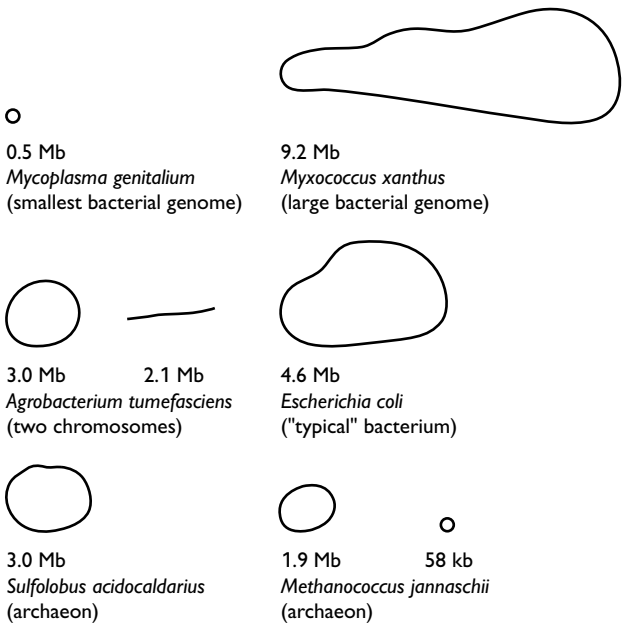


Fig 4.2 Diverse chromosomes in selected bacteria and archaea. These examples were chosen to emphasize diversity in the size and number of essential chromosomes. Chromosomes, approximately to scale, are cartooned as simple circles or rods. The 58-kb chromosome of *M. jannaschii* was reported as an "extrachromosomal element," but it contains genes that code for proteins that are essential to this cell, including two histones and two restriction enzymes.

Fig 4.3 Looped chromosome (**a**) and chromosome loop (**b**).

compacted about 10-fold compared with free, relaxed, B-form DNA (Figure 4.3).

In *E. coli*, the chromosome is 80% to 90% DNA and 10% to 20% protein by mass. Most of its DNA is completely free of protein. A dozen different chromosomal proteins make up the bulk of the protein in the nucleoid region, bound to the B-form DNA with weak chemical bonds. Some of the chromosomal proteins, such as HU, bind with no sequence specificity, but others, such as IHF, bind to specific DNA sequences. Only a minority of the DNA is bound to chromosomal proteins. Overall, the DNA is slightly negatively supercoiled (underwound).

HU is a small, basic heterodimeric protein ($\alpha\beta$, 19 kDa) that binds to DNA with any sequence of nucleotides and bends it. HU is more concentrated at the edges of the nucleoid, where transcription takes place; HU participates in transcription, replication, and recombination. HU and DNA form bead-like structures *in vitro*. IHF, which is similar to HU in amino acid sequences, also makes small heterodimers ($\alpha\beta$, 22 kDa) and also bends DNA. In contrast to HU, IHF binds preferentially to regulatory DNA sequences. Like HU, IHF serves in many genetic processes.

The DNA of archaeal chromosomes is wrapped around small, basic proteins called archaeal histones, homologous to eukaryal histones (Figure 4.4). The archaeon *Methanothermus fervidus* has two histones, HmfA and HmfB, each about 70 amino acids, which make tetramers. Positively supercoiled DNA (90 to 150 base pairs (bp) wraps around a tetramer of HmfA and HmfB with no sequence specificity, to make a particle similar to the nucleosome of eukarya. Native archaeal chromosomes vary from place to

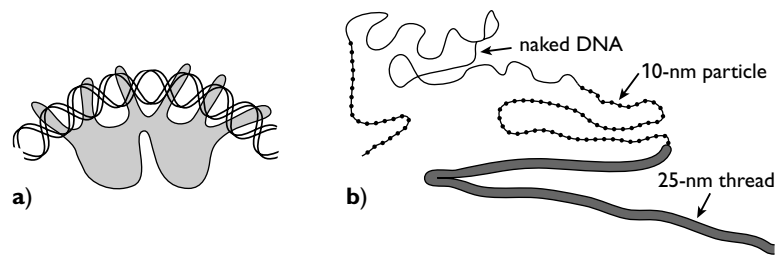


Fig 4.4 Bacterial HU-DNA particle (**a**). Segment of an archaeal chromosome (**b**).

place, including areas of naked DNA, regions of DNA and histones arranged like ≈10-nm beads on a string, and regions of rough ≈25-nm threads.

Plasmids

In addition to its essential chromosome(s), a cell may have one or more plasmids – small, extra chromosomes that are not needed by the cell in a natural environment. The term plasmid is in wide use today, although one sometimes sees the ill-conceived synonym extrachromosomal element. Plasmids are usually double-stranded DNA, rarely single-stranded DNA. They are usually circular but may be linear. Most are in the size range 3 to 100 kb. Plasmids contain functional genes and are replicated, although their replication is independent of replication of the main chromosome(s). Most small chromosomes of bacteria and archaea whose DNA sequences have been analyzed appear not to have essential genes. For this reason, they are likely plasmids.

Some plasmids are **conjugative** – capable of being transferred from a donor cell to a recipient during conjugation, when two cells make physical contact and a passage forms between them. Conjugative plasmids have genes that encode proteins required for conjugation, and nonconjugative plasmids lack them.

Plasmids conferring resistance to an antibiotic are often designated R, followed by letters or numbers; R plasmids in pathogenic bacteria are medically important. Some plasmids encode a bactericidal protein (colicin) as well as an "antitoxin"; these are called Col plasmids. Plasmids appear to be as commonplace in archaea as they are in bacteria.

In abundance, any given plasmid falls into one of three categories: low copy number, about 1 per cell; medium copy number, about 10 per cell; and high copy number, 30 to 100 per cell.

Origin of Replication

All bacterial and archaeal chromosomes, including plasmids, contain a short sequence where chromosome replication begins, termed the origin of replication, or simply **origin**. Proteins of the replication machinery bind to the origin before chromosome replication begins. In *E. coli* the origin, *oriC*, is about 200 bp long.

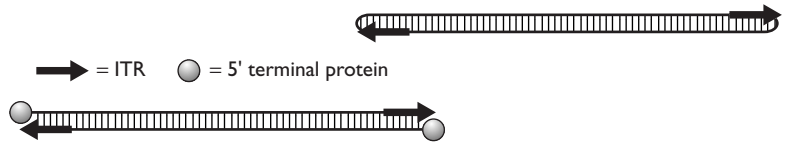


Fig 4.5 Telomeres of linear plasmids contain inverted terminal repeats (ITR). Telomeres of plasmid in *Streptomyces* (left). Hairpin telomeres of plasmid in *Borelia* (right).

Telomeres

An obvious difference between a circular chromosome and a linear chromosome is that the former is endless, whereas the latter has two ends. Constitutively linear chromosomes of bacteria and archaea, including linear plasmids, contain structures at their ends called **telomeres**. These are short terminal sequences of DNA; in some cases protein, part of the telomere, is attached covalently to a terminal nucleotide. Telomeres facilitate replication of the chromosome's ends and may also help protect the linear chromosome from degradation. The main features of bacterial telomeres are (1) inverted terminal repeat (ITR) sequences, and (2) either hairpin loops or 5′-terminal proteins (Figure 4.5). ITRs get their name from the fact that the nucleotide sequence at one terminus is identical to that at the other terminus, except for pointing in the opposite direction (5′→3′ vs. 3′→5′).

Mitochondria

Mitochondria, like bacteria, have essential chromosomes and may also have plasmids, localized in the nucleoid region. The mitochondrial chromosomes are double-stranded (ds) DNA, and plasmids are usually double-stranded DNA but may be single-stranded (ss) DNA. Mitochondrial genes encode the protein-synthesizing machinery (rRNA, tRNA, ribosomal proteins; Chapter 11) and some enzymes of aerobic metabolism. Many proteins in mitochondria are encoded by nuclear genes and synthesized in the cell's cytoplasm – on the endoplasmic reticulum – and then imported into mitochondria.

Most mitochondria have a single essential chromosome, usually present in several copies; one or more plasmids, also in multiple copy number, may be present. Mitochondrial chromosomes are usually circular, but linear mitochondrial chromosomes occur in ciliates, algae, animals, and fungi. Mitochondrial genomes vary in size from 14 to 2500 kb. Linear mitochondrial plasmids of

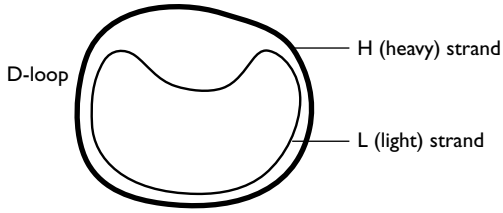


Fig 4.6 Mammalian mitochondrial chromosome.

the inverted repeat type are known: they have terminal inverted repeats and proteins linked covalently to the 5′ ends.

The chromosome of mammalian mitochondria is circular and consists of a purine-rich heavy (H) and pyrimidine-rich light (L) strand (Figure 4.6). During the normal development of a mammal's mitochondrion a structure called the D-loop (displacement loop) forms in the mitochondrial chromosome, where part of a new H strand is synthesized. Each strand has a separate origin of replication, O_H on the heavy strand and O_L on the light strand. Each strand has a single promoter (site where RNA synthesis is initiated). O_H and the two promoters are located in the D-loop. Plant mitochondrial chromosomes have many promoters, and there are large regions of noncoding DNA sequences, interspersed between genes.

Kinetoplasts

The mitochondria of the Protozoan order Kinetoplastida (*Trypanosoma* and *Leishmania*) are bizarre. In addition to a main chromosome in the nucleoid region, the mitochondrion of these protozoans contains a disk-like body, the **kinetoplast**, which contains 40 to 50 large chromosomes (maxicircles) and 5000 to 10,000 small ones (minicircles), all interconnected in a giant network. Maxicircles contain 25 to 35 kb and minicircles contain 0.65 to 2.50 kb of DNA. The maxicircles, which appear to be identical copies of the main chromosome, contain genes typical of mitochondrial genomes. Hundreds of different minicircles are present, each in multiple copies.

Plastids

Plastids (chloroplasts and their relatives) are organelles found in green algae and plants. Chloroplasts descended from cyanobacteria, which became endosymbionts of plants and algae. Other

plastids evolved from chloroplasts; these include chromoplasts (pigment-containing plastids of flowers) and amyloplasts (starch-storing plastids). The evolution of chloroplasts is complex compared with that of mitochondria: there were at least two separate endosymbiotic captures of cyanobacteria, and some chloroplasts evolved by secondary capture – a chloroplast-containing eukaryon became the chloroplast of a new host (e.g., Cryptomonads).

Chloroplasts have one circular essential chromosome, usually present in multiple copies. The chromosome (120 to 200 kb of dsDNA) encodes four rRNAs (RNAs that are part of the ribosome), about 30 tRNAs, and about 100 proteins. Noncoding sequences are interspersed between genes. A chloroplast may also have one or more multiple-copy plasmids.

Viruses

Viral genomes may be RNA or DNA and may be double stranded or single stranded (Figure 4.7). The genome may consist of a single chromosome or several chromosomes. Total genome size ranges from 2 to 670 kb, so that the largest viral genome (of virus G, infecting *Bacillus subtilis*) is bigger than the smallest bacterial genome. Chromosomes may be circular or linear. Linear viral chromosomes lack telomeres. The genomes of some viruses reproduce independently of the host genome. In other viruses, one or more copies of the genome may integrate into the host genome, where they reside for part of the virus's life cycle.

Fig 4.7 Diverse viral genomes. Host in parentheses: B = bacterium, An = animal, P = plant. Genome size in kb, number of chromosomes, number of genes, and chromosome shape are given for each virus; ss = single strand, ds = double strand. For single-strand viruses, (+) = chromosome is equivalent to encoded RNA; (−) = chromosome is complementary to encoded RNA.

Type	Name	Host	Genome	Chromosomes per genome	(+) or (−)
ssRNA	MS2	B	3.6 kb	1 linear	(+)
	polio	An	7 kb	1 linear	(+)
	TMV	P	6.5 kb	1 linear	(+)
	HIV-1	An	8.5 kb	1 linear	(+)
	influenza-A	An	13.6 kb	8 linear	(−)
dsRNA	reovirus	An	30 kb	10 linear	
ssDNA	Ff	B	6.4 kb	1 circular	
	parvovirus	An	2 kb	1 linear	(+) or (−)
ss/dsDNA	hepatitis-B	An	3.3 kb	1 circular	
dsDNA	λ	B	48 kb	1 linear	
	CaMV	P	8 kb	1 circular	
	reovirus	An	30 kb	10 linear	
	HSV-1	An	153 kb	1 linear	
	vaccinia	An	200 kb	1 linear	
	G	B	670 kb	1 circular	