The Fundamentals of
Chemical Biology

LEARNING OBJECTIVES
• Provide a working definition of chemical biology.
• Apply the central dogma of molecular biology as an organizing principle in chemical biology.
• Describe key features of genes, transcripts, proteins, and the associated genomes, transcriptomes, and proteomes.
• Explain the importance of evolution in chemical biology experiments.
• Introduce the most important model organisms, and provide examples of their applications in chemical biology experiments.
• Provide a brief introduction to viruses and plasmids.

Why organize a book on chemical biology around biooligomers?

Living organisms must obey the rules of chemistry. Chemical biology applies the rules of chemistry—at the level of atoms and bonds—to biological systems. Chemistry is an expansive set of subdisciplines, each with limited dynamic range, that offer insight into the behavior of molecules. Among those subdisciplines, organic chemistry offers a unique compromise between qualitative and quantitative approaches, and is ideal for explaining chemical reactions and nonbonding interactions that generate the molecular diversity needed for evolution.

According to Charles Darwin, selection for beneficial traits is a key step in the evolution of the various species (Figure 1.1). The success of selection as an evolutionary tool depends on a diverse population. If a field of plants, say Cape cowslip lilies, possesses a drought-resistant plant, a flood-resistant plant, a cold-resistant plant, and a heat-resistant plant, the chances are good that at least one member of the population will survive a climate catastrophe. But what happens to the population if there are a series of different climate catastrophes, such as a drought this summer and a flood the next summer? To survive a second type of climate catastrophe, the surviving plants must propagate in a way that re-establishes a diverse population. Thus, evolution is not just driven by survival of the fittest; it is also driven by the propagation and diversification of the fittest.

Generating diverse populations is a key ingredient in the recipe of evolution. Living organisms are engines for the assembly of diverse populations from a limited set of subunits: diverse populations of molecules, diverse populations of cells, diverse populations of organisms, and diverse ecosystems. This modular approach to assembly can be described as combinatorial. We can appreciate the flexibility of this approach by comparing it to writing. Writing is combinatorial at every level: letters can be combined to form words; words can be combined to form sentences; sentences can be combined to form paragraphs; and the meaning of letters, words, and paragraphs is dependent on sequence. In the right combination, letters can express ideas that organize societies, mobilize armies, solidify the bonds of love, and even explain our universe. Words, sentences, and paragraphs are oligomers. They are formed by connecting a limited set of subunits in a nonrepeating fashion. We have rules for each level of writing: rules of spelling for words, rules of grammar for sentences, and rules of
composition for paragraphs. Those rules allow us to encode powerful ideas in a compact form, to store them, to transmit them, to duplicate them, and later to decipher them without ambiguity.

Chemical biologists seek to explain the combinatorial origin of diversity at the molecular level of detail, atom by atom and bond by bond. Most of the molecules in your cells are biooligomers made up of molecular subunits, and each of those molecular subunits is a combinatorial oligomer of atoms. Biooligomers, including proteins, DNA, RNA, glycans, lipids, and terpenes, account for more than 90% of the dry weight of the human cell. To a first-order approximation, such biooligomers are the cell—just add water. Oligomeric architectures are ideally suited to the synthesis of diverse molecules. The goal of this book is to explain the rules of chemical grammar that govern the assembly of diverse biooligomers and, in so doing, to illuminate the workings of the human cell. The biooligomers are organized according to a central dogma of biology, which also provides the thematic organization for this book.

### Problem 1.1
If a thousand chimpanzees typed for a thousand years, is it likely that one of them would turn out a work of Shakespeare? (Courtesy of New York Zoological Society.)

---

### 1.1 THE CENTRAL DOGMA OF MOLECULAR BIOLOGY

The central dogma of molecular biology is an organizing principle for chemical biology

Cells can be simpler to understand than writing because the sequence of every biooligomer is ultimately determined by the sequence of other biooligomers. In the case of DNA, RNA, and proteins, the correspondence between sequences is straightforward. The potential relationships between those biooligomers were first outlined by Francis Crick, a pioneer in determining the structure of DNA. Crick noted a hierarchical flow of information from DNA to other biooligomers and entitled it the central dogma of molecular biology (Figure 1.2). In that hierarchy, DNA can be used as a template to direct the synthesis of RNA and vice versa. RNA can be used as a template to direct the synthesis of proteins, but not the other way around. The structure of proteins makes it difficult to use as a template for directing the biosynthesis of RNA, DNA, or any type of oligomer. The flow of information is not completely unidirectional. Proteins do affect the synthesis of DNA and RNA, but not through a direct encoding mechanism. We will examine those indirect mechanisms in later chapters.

The biopolymer DNA provides a master blueprint for the cell and organism. To follow this blueprint, we will need to master three highly specific biochemical terms: replicate, transcribe, and translate (Figure 1.3). Most of the reactions taking place in biology are catalyzed by proteins, called enzymes. DNA is replicated by the enzyme DNA polymerase, using each strand of DNA as a template for the synthesis of new strands. RNA is transcribed by RNA polymerase, using one of the two strands of DNA as a template. RNA carries out diverse functions in the cell, but the function heralded by the central dogma is that it serves as a template for protein synthesis. Proteins are translated from a messenger RNA (mRNA) template by the ribosome. In organisms, proteins contribute diverse roles. For example, proteins called enzymes can catalyze the formation of other biopolymers, such as oligosaccharides, lipids, and terpenes.
1.2 GENES

A gene is made up of a promoter and a transcribed sequence

In the cell, the transcription of RNA is carefully orchestrated. RNA polymerase does not begin transcription at random places in the genome or even at random genes. At any point in time, RNA polymerase transcribes only specific genes in the genome. Proteins called transcription factors bind to a specific DNA sequence called a promoter. The combination of a promoter and a DNA sequence that encodes an RNA sequence composes a simple gene (Figure 1.4). The promoter and its associated transcription factors control gene expression. Some transcription factors recruit RNA polymerase and the process continues as RNA polymerase translates the transcribed RNA into protein.

Figure 1.3 The central dogma of molecular biology, expanded. The arrows indicate the flow of information as biooligomers of one type serve as templates or catalysts for the synthesis of other types of biooligomers.

Figure 1.4 Gene expression. DNA appears uniform when rendered as a cartoon, but when viewed at the level of atoms and bonds each nanometer of sequence contains rich information. All genes consist of a promoter sequence and a transcribed sequence. Transcription factors bind to the promoter and recruit RNA polymerase to transcribe RNA.
activate transcription; other transcription factors repress transcription (Figure 1.5). As we will see in Chapter 3, the structure of DNA is richly informative, allowing transcription factors to read the sequence of subunits through molecular recognition.

The smallest known gene, mccA, was discovered in *Escherichia coli* and codes for the synthesis of a short peptide called microcin A (Figure 1.6). Microcin A consists of just seven amino acid subunits and, as we will see later in this book, the production of such a short peptide by ribosomal translation is unusual. Microcin A is part of a sleek and lethal antibiotic called microcin C7 that is used by *E. coli* to defend its turf within the human intestinal tract. In addition to the mccA gene, three more genes encode enzymes needed for the final assembly of microcin C7. The genes for the production and export of microcin C7 are organized into an operon or cluster of genes (see Figure 1.6), and the entire McC operon is controlled as a group by transcription factors that are sensitive to intracellular conditions. The grouping of functionally related genes under the control of a common DNA promoter sequence into operons is a strategy used by all organisms.

![Figure 1.5](image1.png)  
**Figure 1.5 A transcription factor binding to DNA.** Proteins called transcription factors (blue) turn genes on and off, by binding to promoters, which are specific DNA sequences (red). In this depiction, the protein is shown as a ribbon tracing the connections between amino acids. DNA is depicted as sticks between the non-hydrogen atoms, and a ribbon traces the phosphodiester backbone of DNA. More information about DNA and protein structure is provided in later chapters of this book.

![Figure 1.6](image2.png)  
**Figure 1.6 The smallest gene makes a chemical weapon.** (A) Microcin A is encoded by the tiny mccA gene grouped together with related genes in the McC operon. (B) The peptide microcin A is assembled into a sleek chemical weapon.

The *McC* operon has one more gene, which codes for a protein transporter that aggressively expels microcin C7 from the cell. Microcin C7 resembles the binary chemical weapons invented by humans but is much more sophisticated (Figure 1.7). Microcin C7 is a booby trap. Hapless strains of *E. coli* see the tasty peptide and actively import it. When those bacteria try to strip it for usable parts, one of the fragments turns out to be toxic, targeting an enzyme that is essential for protein synthesis, called aspartyl tRNA synthetase. The bacterium that produces microcin C7 is well prepared if the peptide reaches dangerous levels within its own cell. The mccC gene adjacent to the *McC* operon, but not part of it, encodes an enzyme that cleaves a C–N amide bond of microcin C7, thus preventing formation of the toxic fragment. If any of the toxic fragment is accidentally produced, the enzyme coded by the mccE gene has a second active site that acetylates the toxic fragment, rendering it inactive.

![Figure 1.7](image3.png)  
**Figure 1.7 Chemical weapon.** Binary chemical munitions generate phosphonate nerve agents from two nonlethal precursors. This pattern of assembly mimics the biosynthesis of microcin C7. Top: diagram of a 155-mm projectile that generates the nerve agent sarin (bottom) from two reactants. (Adapted from Department of the Army, Chemical Systems Laboratory, 1981.)
1.3 GENOMES

We have sequenced the human genome and many others. Now what?

The turn of the millennium saw one of the greatest scientific achievements in the history of humankind—the structural elucidation of the human genome. A genome is a collection of all of the DNA in an organism. The human genome carries instructions for all human biosynthesis: DNA, RNA, proteins—everything. Because DNA is a linear oligomer of four chemical subunits, the process of structure elucidation involves determining the sequence of the subunits. We will refer to DNA subunits as base pairs, or bases for short, for reasons that will become apparent in Chapter 3. Technology for genomic sequencing was first applied to a diminutive bacterium, *Haemophilus influenzae*, and within a few years a rough draft of the human genome was available. The ability to sequence entire genomes has given us unprecedented access to the genetic differences that distinguish bacteria from yeast, yeast from worms, worms from flies, flies from mice, mice from humans, and healthy humans from diseased humans.

Scientists routinely study simpler organisms to understand important principles about human biology. We still endeavor to understand the minimum requirements for life. The smallest known genome, with about 200 genes, belongs to the microorganism *Carsonella rudii*, but *C. rudii* is a symbiont that cannot live outside its insect host. The smallest known genome for a free-living organism is that of the bacterium *Mycoplasma genitalium*, which sometimes infects the human respiratory or genital tract (Figure 1.8). The *M. genitalium* genome has only 521 genes efficiently coded within 582,970 DNA bases; 482 of those genes code for proteins. From the gene sequences, we can predict the chemical structures of the resulting RNA transcripts and, ultimately, the chemical structures of the proteins. Only 382 of those proteins are essential for life. However, the bacterium probably would not be viable if all of the 100 nonessential genes were removed at once. J. Craig Venter, one of the leaders of the human genome project, recently generated a viable cell from a chemically synthesized genome of *M. genitalium*—the first synthetic organism! It would seem that we are well on our way to understanding a minimal organism—how it survives, evolves, and responds to widely varying conditions—but we are not. Scientists have been staring at the chemical structures of the *M. genitalium* proteins for almost two decades, and we still cannot figure out what one-quarter of the essential proteins do.

Why is there a yawning gap between protein sequence and protein function? There are two main reasons. First, the function of many of those proteins relates to other types of biooligomers (such as glycans, lipids, terpenes, and metabolites) that are not easy to study. Second, the protein may only serve a relevant function under a certain set of conditions. This is why an alien making a brief visit to your home might not understand the function of the ceiling fire alarm. Furthermore, the function of many proteins can be understood only when observed within a dynamic system. For example, it is virtually impossible to predict the function of a synchronizer gear once it is removed from a 1967 four-speed Toploader transmission. As long as the top is bolted on the transmission, you cannot see what the synchronizer is doing as you shift gears (Figure 1.9). Similarly, no one can see what is going on in *M. genitalium*. Chemical biologists are uniquely equipped for such problems. They are trained to think at the level of atoms and bonds, and can design molecular tools to probe and peer into complex systems such as cells.

We are far from understanding cells that we understand the best—*Escherichia coli*

*Escherichia coli* is the best-understood species of organism, and it is larger and more complex than a mycoplasmic bacterium. Some strains of *E. coli* live harmoniously within the human gastrointestinal tract, and even thrive there; one-third of the dry weight of human feces is *E. coli* cells. The task of understanding *E. coli* is sometimes complicated by the fact that there are numerous strains. The common laboratory strain, *E. coli* K-12, is harmless, but other strains of *E. coli* are associated with disease.

![Figure 1.8 The smallest genome. *Mycoplasma genitalium* has the smallest genome of any free-living organism. This image was taken by scanning electron microscopy and then colored. Optical microscopy does not have sufficient resolution to resolve the features of *M. genitalium*. (From C. McGowin et al., BMC Microbiol. 9:139, 2009. With permission from BioMed Central Ltd.)](image)

![Figure 1.9 What does it do? Sometimes the function of a component—such as a synchronizer from an automobile transmission—is difficult to discern if you cannot see inside. Even with an accurate static diagram, the roles of the components are cryptic if you cannot see the system respond dynamically. (Courtesy of Roland Dudley and Mark Olson, eds., Classic Tiger. http://www.classictiger.com)](image)
Uropathogenic strains such as CFT073 sometimes lead to human bladder infections. The enterohemorrhagic strain O157:H7 sometimes contaminates the beef sold in grocery stores.

The bacterium *E. coli* is a small rod-shaped organism (Figure 1.10), typically a little more than 1 μm in diameter. It has a complex coating composed of two fluid lipid membranes. A tough, web-like network, called the cell wall, is sandwiched in the periplasmic space between the two membranes. Most of the bacterial contents, such as the DNA, are housed within the inner membrane as opposed to the cramped periplasmic space. *E. coli* K-12 is capable of a solitary itinerant existence. It is equipped with rotating whip-like flagella that propel it toward nutrients and away from toxins. In spite of its itinerant tendencies, *E. coli* is also quite sociable. Under nutrient-rich conditions, *E. coli* undergoes binary fission to generative cooperative colonies akin to superorganisms. Chemical signaling within the colony maximizes the use of environmental resources (related to invasiveness) and inhibits overpopulation. Each member of the colony is a genetic clone, equally fit yet equally unfit to survive. *E. coli* cells rely on promiscuity to improve genetic diversity and increase the chances of survival. The bacterial capsule is covered with thin hair-like projections—much smaller than the flagella—that allow the bacteria to exchange genetic material with other strains. Many genotypic variants of K-12 have been engineered for laboratory experiments such as viral transfection, selecting clones, and producing proteins. The genome of the MG1655 variant of *E. coli* K-12 is composed of 4377 genes, encoded within 4,639,221 bases of DNA. Both the cell and the genome of *E. coli* are about 10 times larger than that of *M. genitalium*. Like *M. genitalium*, more than 95% of the *E. coli* K-12 genes encode proteins. Advances in DNA sequencing technology have reduced the sequencing of bacterial genomes to an expensive exercise, and the genomes of more than 50 strains of *E. coli* have now been sequenced. You would be hard pressed to distinguish the various strains of *E. coli* by looking at micrographs (Figure 1.11).

We are far from understanding any one strain of *E. coli* and much further from understanding all the different strains. As with *M. genitalium*, we do not know the function of about 20% of the genes in *E. coli* K-12. Surprisingly, only about 20% of the *E. coli* genes are conserved among all of the strains. The genetic diversity of *E. coli* is further increased by the presence of smaller circles of DNA called plasmids. The plasmids contained within bacteria vary, even within the same strain. Plasmids typically contain one or several genes, usually beneficial. Bacteria trade plasmids during their conjugal visits. There is only one copy of the bacterial genome per cell, but many copies of each of the plasmids.

We are even farther from understanding human cells

Human cells are about a thousand times larger than *E. coli* cells and have about a thousand times more DNA. Like *E. coli*, human cells are topologically separated by two lipid membranes (Figure 1.12). The outer membrane, called the plasma membrane,
is exposed to the extracellular milieu. The nuclear envelope is actually a double mem-
brane, with a thin space between the layers; it houses the DNA. Human cells have
nothing like a bacterial cell wall. Instead, the shape of human cells is maintained by
internal protein scaffolding. There is plenty of room in the cytoplasm, between the
nuclear membrane and the plasma membrane, for the things that distinguish one
human cell type from another. In addition, the cytoplasm also contains other mem-
brane-enclosed organelles. Among the most distinctive organelles are the mitochon-
dria, which produce chemical energy for the cell. Mitochondria are like symbiotic
organisms, replicating independently of the cell. They have two membranes, and
within the inner membrane they have their own DNA, their own enzymes, and their
own ribosomes. In addition to mitochondria, microscopy also reveals a network of
bubble-like vesicles (the endoplasmic reticulum and Golgi complex) that carry pro-
tein and glycan products to the exterior of the cell.

Human DNA exists in two cellular locations: the nucleus and the mitochondria.
The genomic DNA within the nucleus is divided between 23 pairs of homologous
chromosomes. The mitochondrial DNA is a smaller circle of DNA, much like the
\textit{E. coli} genome. Whereas most of the DNA in \textit{E. coli} encodes proteins, only a small fraction
(1.5\%) of the DNA in human cells encodes proteins. The human genome encodes only
about 20,000–25,000 transcribed proteins—only five times as many proteins as one
finds in \textit{E. coli}. Thus, humans, accustomed to thinking of themselves as tremendously
more complex than unicellular life, have a surprisingly modest parts list. How can a
mere 25,000 genes give rise to a multicellular organism as complex as a human being
(Figure 1.13)? To understand this complexity, we will need to further explore the ori-
gins of molecular diversity at the level of RNA, proteins, and beyond.
You cannot judge a cell by its genome

Strains of *E. coli* have similar structures but very different genomes, whereas the opposite is true for mammals such as mice and humans, which share 99% of the same genes. If *E. coli* were geneticists, they would easily mistake you for a mouse. All humans of the same sex, regardless of age, race, or ethnic origin, are 99.9% identical (Figure 1.14). The wide variation in human appearance and disease susceptibilities results from differences in a mere 0.1% of the human genome. Ultimately, analysis of genome sequences from many different humans may reveal genetic differences that underlie conditions such as Alzheimer’s disease, osteoporosis, and diabetes. Many of these differences will lead to targets for the next generation of pharmaceuticals.

Each human individual is made up of roughly 220 different cell types, which have a dazzling array of sizes, shapes, and functions. With a few special exceptions, all of the somatic cells in your body share 100% genetic identity: nerve cells, muscle cells, white blood cells, photoreceptor cells, excretory cells, fat cells, bone-making cells, hair-producing cells, and skin cells (Figure 1.15). A major focus of developmental biology is to understand how the various human cell types arise during embryonic cell division. Ultimately, the differences between these cells are attributable not to differences in the genomic DNA sequences but to differences in gene expression—that is, which of these genes are “turned on” and which are “turned off.”
The observable phenotype belies the hidden genotype

Each gene or combination of genes, referred to as a genotype, results in an observable set of characteristics, called the phenotype (such as red hair or antibiotic resistance). Resistance to the antibiotic erythromycin is a phenotype, but there are many genetic variations that can confer resistance to erythromycin. Bacteria with the mef gene (abbreviated mef+ and spoken m-e-f-positive) have a protein pump that exports erythromycin. Bacteria with the ermA+ and ermC+ genotypes produce an enzyme that methylates a single atom in the ribosome, preventing the binding of erythromycin (Figure 1.16). A mutation to the DNA sequence within an existing gene can result in erythromycin resistance. Mutations to the gene that encodes the ribosomal protein L4 (for example a mutation termed K63E) can make bacteria resistant to erythromycin. Mutation of a single DNA subunit in the gene that codes for the ribosome (for example a mutation termed A2058G) also makes bacteria resistant to erythromycin.

Cancer is a phenotype of human cells that arises from a combination of genetic changes. The phenotype of cancer is invasive, uncontrolled cell proliferation. Each cancer cell line usually exhibits a unique genotype. In the previous millennium, most of the drugs prescribed for leukemia targeted the phenotype of uncontrolled cell growth in the fervent hope that they would kill the cancer faster than the patient. That strategy often failed. Newer anticancer drugs are intended to distinguish cancer cells from normal cells at the level of genotypic variation, resulting in fewer side effects.

The field of chemical biology has entered this millennium armed with an abundance of genomic information, but we cannot yet make useful predictions about the workings of cells based solely on DNA structure. Genomic structure is essential knowledge for understanding cellular properties, but it does not enable us to predict the molecular workings of even the simplest cell. To use this information to the fullest extent, we will need to master the chemical principles that underlie all of the processes embodied in the central dogma and obeyed by each and every molecule in the universe.

1.4 SOURCES OF DIVERSITY BEYOND GENOMES

The transcriptome is the collection of all of the RNA transcripts in a cell

In Chapter 4 we will immerse ourselves in the world of RNA, where we will meet an ancient and secretive class of molecules whose chemical similarity to DNA belies a pervasive set of functions. The central dogma (see Figure 1.3) emphasizes the role of RNA as a messenger, but the structural diversity available to RNA molecules allows it to exert control over gene expression at every level.

The complete collection of all RNA found in a cell is called a transcriptome (Figure 1.17). The transcriptional profile of a human cell is a direct readout of the active genes and thus depends on the tissue type and many other variables, such as the nutrients, extracellular signaling molecules, and temperature. The transcriptional profiles vary between cells, even from the same tissue. Thus, the exact details of the sample source and preparation become vitally important. By comparing the transcriptional profiles of different cells, we gain insight into which genes are being expressed and how the resulting RNA transcripts are being processed. For example, growing human cells in the laboratory at a slightly elevated temperature markedly alters the transcriptome, as the cells scramble to adjust for the shock of the higher temperature. The condition, termed heat shock, equips the cell with new specialized
mRNA transcripts; in addition, the levels of the standard mRNA transcripts are altered. Some genes will be transcribed at lower levels, but others critical for cell stability at the higher temperature will be produced more often. For example, some proteins fall apart at slightly elevated temperatures (increases of just 5 °C in some cases), and other proteins are deployed to stabilize and help remove such breakdowns. Thus, the identity and concentrations of individual mRNA transcripts within the transcriptome control the cell’s future fate and current activities.

**RNA splicing amplifies the diversity of the transcriptome**

In bacterial cells, mRNA is immediately seized by ribosomes for translation into protein, but in human cells, RNA transcripts undergo substantial editorial processing, called splicing ([Figure 1.18](#)), that removes various stretches of RNA before it is ready to serve as mRNA. We cannot yet predict which pieces of human RNA will be spliced out by looking at the original DNA sequence. Because we cannot predict which parts of an RNA transcript will be spliced, it is impossible to predict protein sequences based solely on DNA gene sequences. Our predictive power is further diminished by the fact that splicing depends on environmental conditions. For example, the transcription factor ATF3 is spliced in two different ways depending on whether growth factors are present in the culture medium. Differential splicing allows a single human gene to code for more than one protein and amplifies the diversity of the human genome.

**Post-translational modification of proteins amplifies the diversity of the proteome**

Proteins are assembled from 20 universal building blocks called amino acids. Amino acids have a much broader range of chemical functionality than the nucleotide subunits that make up DNA and RNA. Folded proteins accomplish the bulk of activity in the cell, including the catalysis of specific reactions and forming structural supports. Analogous to the transcriptome, the proteome is the complete collection of all proteins in a cell, organism, or tissue sample. Simple methods are available for isolating and sequencing mRNA and the proteins they encode, but it is still difficult to predict the full range of proteins within a human cell.

Analogously to the regulation of gene transcription by transcription factors, translation of mRNA into proteins is regulated by multiple translational control proteins. First, protein translation begins at a specific sequence near the beginning of the mRNA transcript. In humans, the eIF-2 protein assists in recognition of this start sequence. Illustrating the complexity inherent in the regulation of protein production, levels of protein production in eukaryotic cells can be decreased by covalent modification of eIF-2. This attenuation can take place in response to viral attack or heat shock. This chemical modification to alter the functional role of a protein is an example of a post-translational process; these are described in the next section and in Chapter 6.

Proteins are modified in various ways after ribosomal translation. Post-translational modifications include trimming, splicing, phosphorylation, glycosylation, oxidation, addition of membrane anchors, fusion with other proteins, alkylation, acetylation—in fact, too many modifications to list here ([Figure 1.19](#)). Thus, the number of different proteins in a human cell will always exceed the number of genes.
being expressed. The potential diversity of the human proteome greatly exceeds the approximately 23,000 genes in the genome, as a result of splicing and the potential for post-translational modifications. Like transcriptomes, proteomes vary considerably between different tissue samples from the same organism. Chemical biology seeks to understand such differences; for example, characterizing differences between seemingly identical cells from young and old humans could solve many fundamental questions in biology.

**Beyond template-directed synthesis of biooligomers**

Enzymes catalyze the synthesis of the three other types of biooligomers—namely, polyketides, oligosaccharides, and terpenes. For these three types of biooligomers, each subunit is added by a unique enzyme that does not use another biooligomer as a template, in contrast with the way in which RNA polymerase can transcribe DNA or the ribosome can translate mRNA. In general, the order of subunits in polyketides, oligosaccharides, and terpenes does not correlate with the order of the genes that encode the biosynthetic enzymes. Some polyketide synthase genes are important exceptions, and we discuss them in Chapter 8.

Many oligosaccharides are nonlinear and/or branched; thus, the sequence of subunits cannot correlate in a simple way with a linear DNA sequence or protein sequence. For example, the synthesis of a branched oligosaccharide with 11 types of bonds could require 11 different enzymes, encoded by 11 different genes (Figure 1.20). The genes can be in any order within the genome, even if the enzymes...
assemble the oligosaccharide subunits in a defined sequence. We discuss the structures of oligosaccharides in much more detail in Chapter 7.

Polyketides, oligosaccharides, and terpenes are further modified by additional enzymes after the subunits have been linked together. The additional enzymes catalyze many reactions that mask the structure of the initial biooligomer, such as cyclization, oxidation, reduction, and cleavage.

### 1.5 COMBINATORIAL ASSEMBLY GENERATES DIVERSITY

**Combinatorial assembly of linear biooligomers can generate massive diversity**

There are two key steps in evolution: first generate a diverse population, and then select for fitness. We can readily observe these evolutionary steps in *E. coli* populations. When *E. coli* cells are exposed to the antibiotic ciprofloxacin, the bacteria initiate an SOS response characterized by a 10,000-fold increase in the rate of genetic mutations. Almost all of the resulting mutants are less fit for survival, even under ideal growth conditions. A few mutants, however—and it only takes one—are resistant to the antibiotic. The mutation rates subside as these healthy, resistant bacteria begin to thrive and regenerate healthy colonies.

All organic molecules have combinatorial architectures based on atomic subunits. For example, 31 stable compounds can be assembled from four carbon atoms, eight hydrogen atoms, and one oxygen atom. However, these varying molecules cannot be accessed through a single efficient synthetic route. The origin of diverse populations can be found at the molecular level of detail in the one-dimensional oligomeric structures of DNA, RNA, proteins, oligosaccharides, polyketides, and terpenes. One-dimensional oligomeric architectures make it possible to generate massive, diverse collections of molecules from a few simple joining reactions and a small number of subunits (*Figure 1.21*). Like words in the English language, the meaning (that is, the function) of a biooligomer depends on both the constitution and the order of the letters. The total number of possible “words” of a particular length depends on the number of letters in the alphabet. In a similar way, one can calculate the number of possible oligomers based on the number of types of subunits and the number of subunits in the oligomer (Equation 1.1). With the 26 letters of the English alphabet, one could construct $26^4$, or 456,976, four-letter words. The order of the letters is essential, too; the letters “N-O-I-T-U-L-O-V-E” have no particular meaning in the English language, but spelled backward they produce a word with a well-defined meaning: evolution. As with spoken words, the order of subunits in biooligomers determines their functions.

$$\text{number of possible oligomers} = (\text{number of types of subunits})^{\text{length of oligomer}} \text{ (1.1)}$$

We can customize Equation 1.1 for DNA, RNA, and proteins. That is, the number of types of subunits is four for DNA and RNA, whereas it is 20 for proteins synthesized by the human ribosome.

**Problem 1.2**

Calculate the potential diversity of the following one-dimensional constructs and give the answer as a base 10 number (for example $3.7 \times 10^4$).

- A English words with 10 letters.
- B An unusually small gene with 1000 DNA subunits.
- C A microRNA transcript with 100 RNA subunits.
- D A protein with 100 amino acids.
Combinatorial synthesis can be used to synthesize DNA libraries

The principles of combinatorial assembly can be used to synthesize libraries of biooligomers. In Chapter 3 we describe the chemical synthesis of DNA. The chemistry is so highly optimized that DNA oligomers with more than 50 subunits can be reliably synthesized by machines on a 50 nmol scale. That may not sound like much, but 50 nmol is $3 \times 10^{16}$ molecules. If you carry out each coupling step with an equal mixture of all four DNA subunits, you will generate a combinatorial library of DNA molecules in which each strand has the same length but a different sequence. The potential diversity of a DNA strand with 50 subunits is $4^{50}$, or about $10^{30}$ molecules. If you somehow made all of the possible DNA 50-mers, collectively they would weigh more than 25,000 kilotons. Thus, automated DNA synthesis makes it easy to assemble combinatorial libraries of DNA, but practical considerations usually keep us from preparing DNA samples that contain all of the possible variants. The tools of molecular biology dramatically amplify the power of DNA libraries. If you have a DNA library, you can use RNA polymerase to generate a complementary RNA library. Similarly, that RNA library can be used as a template for ribosomal translation to generate a protein library.

**Problem 1.3**

What length of DNA (based on the number of subunits) has a potential diversity close to $3 \times 10^{16}$?

Modular architecture lends itself to the synthesis of non-natural chemical libraries

For example, a non-natural class of biooligomers known as peptoids can be assembled by organic synthesis. Up to 8000 different trimeric peptoids can be created from 20 different subunits through the iterative application of just one efficient bond-forming sequence. Collections of molecules are often referred to as libraries. When a library of 5120 peptoid trimers was screened for activity against apoptosis, a form of cell suicide, potent inhibitors were identified (Figure 1.22). We discuss apoptosis in more detail in Chapters 6 and 9. The apoptosis inhibitor was found to bind selectively to the protein Apaf-1 in vitro, but the inhibitor was unable to rescue living cells because it could not cross cell membranes. Fortunately, a cyclic, membrane-permeable version of the inhibitor, with fewer hydrogen-bonding groups, was shown to act on living cells.

![Figure 1.22 Synthetic peptoid oligomers.](image)

**Figure 1.22 Synthetic peptoid oligomers.** An inhibitor of cell suicide was identified from a combinatorial library of 5120 synthetic peptoids; the individual peptoid subunits are highlighted. The compound binds selectively to the protein Apaf-1. A membrane-permeable version of the inhibitor was active against U937 lymphoma cells.

The concepts of combinatorial architecture can be applied to any type of molecule that is assembled from modular subunits, even when the subunits are connected using completely different types of chemical reactions (Figure 1.23). For example, a
A synthetic route involving an acylation reaction followed by an imine condensation followed by an $S_n$-2 reaction could be used to make large numbers of synthetic molecules from a small set of starting materials. As long as the reactions are high yielding, with no side reactions, solid-phase synthesis (discussed further in Chapters 3 and 5) and robotics can be used to carry out the synthesis. The important advantage of chemical synthesis is that it offers direct access to small molecules with drug-like properties, such as oral bioavailability, low immunogenicity, and slow metabolism. Since the introduction of parallel synthesis in the 1990s, the synthesis and screening of synthetic libraries has become a standard tool in the development of pharmaceuticals.

The human immune system uses combinatorial biosynthesis

Your immune system synthesizes combinatorial libraries of antibody proteins (sometimes called immunoglobulins) through the combinatorial assembly of genetic modules. This allows your body to generate libraries of different B lymphocytes, ready to fend off a wide range of infections (Figure 1.24). Each antibody can potentially bind to a unique nonhuman molecule, such as those found on viral and bacterial pathogens. When an influenza virus binds to a surface-bound antibody on a B lymphocyte, the binding event turns on the genes for cell proliferation (Figure 1.25). The transduction of extracellular signals into changes in transcription are fascinating, and we discuss...
many more examples in Chapter 9. As the B lymphocytes proliferate, they generate additional variations of the antibody gene, some that bind viruses more tightly and some that bind viruses less tightly. The tighter binders continue to proliferate and ultimately release antibodies into the extracellular medium. The human B-lymphocyte response is very much a model for evolution, involving diversity-generating steps and selection steps. A similar system of combinatorial gene assembly is used to generate libraries of T lymphocytes, each displaying a different T-cell receptor, except that T-cell receptors are not released into the surrounding medium.

An antibody is composed of four peptide chains: two identical heavy chains and two identical light chains. The gene encoding the heavy chain is constructed from three genetic modules, each selected from a pool. There are 40 variable (V) modules, 25 diversity (D) modules, and six joining (J) modules, offering a theoretical diversity of $40 \times 25 \times 6 = 6000$ heavy chains (Figure 1.26). There are two types of light chains, each constructed from two genetic modules, each selected from a pool. Kappa ($\kappa$) light chains arise from 40 variable (V) modules and five joining (J) modules; lambda ($\lambda$) light chains arise from 30 variable (V) modules and four joining (J) modules. When summed, there are $(40 \times 5) + (30 \times 4) = 320$ possible types of light chains. Overall, there are almost 2 million ways to combine the heavy chains with the light chains. The human immune system has harnessed an additional source of diversity that scares away most chemists: it is sloppy. Instead of joining genetic modules in precise ways, it does so in an imprecise way that further amplifies the potential diversity by a factor of more than 10 million. B lymphocytes employ one additional diversity-generating trick. When they proliferate in response to antibody binding, the V, D, and J modules are much more subject to point mutations than the rest of the genome, about one mutation per module per cell division. This hypermutation accesses further genetic variation that was not present in the original genome.

![Figure 1.26 Combinatorial antibody gene assembly](image)

**Figure 1.26 Combinatorial antibody gene assembly.** Combinatorial assembly of V, D, and J gene modules with a constant gene module generates a final antibody heavy-chain gene.

### 1.6 SOME COMMON TOOLS OF CHEMICAL BIOLOGY

**Chromophores reveal invisible molecules**

Most biological molecules are colorless, making them difficult to quantify or locate by light microscopy. Molecules that absorb or emit light at visible wavelengths are particularly valuable for assays and microscopy. Cross-conjugation is a common feature of molecules with high extinction coefficients, particularly when charge separation competes with aromaticity. For example, the $p$-nitrophenolate anion is stabilized by resonance delocalization of the negative charge, but delocalization of the negative charge into the nitro group generates a cross-conjugated nonaromatic pi system (Figure 1.27). Solutions of $p$-nitrophenolate strongly absorb violet light (405 nm) and, when you subtract violet light from white light, the material appears yellow.

When visible chromophores absorb a photon of light, they shake off the excess energy through bond vibrations. When a fluorescent chromophore absorbs a photon, only a small amount of the energy is lost to vibrations; the remaining energy is emitted as a photon of lower energy (and higher wavelength) than the photon that was absorbed. Detecting emission at a different wavelength from excitation minimizes background caused by reflections and scattering. Many fluorescent chromophores exhibit frustration between resonance and aromaticity (Figure 1.28). Fluorescent

![Figure 1.27 A simple color test](image)

(A) A colorimetric substrate releases yellow $p$-nitrophenolate anion, revealing the presence of the enzyme $\alpha$-galactosidase. (B) The intensity of the $p$-nitrophenolate anion is attributable to frustration between charge delocalization and aromaticity. (A, from S.-F. Chien et al., *J. Nanomaterials* Article ID 391497:1–9, 2008.)
Chromophores such as aminocoumarins absorb ultraviolet photons (which you cannot see) and emit violet photons (which you can see). Fluorescein absorbs cyan light and emits green light. Derivatives of tetramethylrhodamine emit red light. BODIPY derivatives are available with a wide range of emission colors. Most of the fluorescent micrographs of cells that show dazzling colors were generated with chemical derivatives of these types of fluorophores. The process involves highly specific antibodies that have been chemically linked to synthetic fluorophores.

Assays connect molecular entities to readily visible phenomena

Chemists are obsessed with the identities and quantities of the molecules that they study. Usually, they use spectroscopic techniques (such as NMR, IR, and UV-vis spectroscopy) or mass spectrometry to establish the structural identity of pure molecules and then employ assays to determine the concentration within mixed samples—some assays are even based on spectroscopy. Selective detection of molecules in biological samples is difficult because the biological molecules rarely have unique chromophores; they are also present in small amounts and exist as mixtures of chemical homologs. By coupling chemical entities such as genes, mRNA, enzymes, and carbohydrates to observable phenomena, we can infer their presence or absence, even in a system as complex as a cell. Typical examples of observable phenomena are precipitation, colony growth, and absorption of light. Recall from Figure 1.16 that bacterial growth on agar plates is a reliable indicator of the presence or absence of antibiotic resistance genes.

Antibodies are a useful class of reagents for the detection of specific biological molecules. Hemaggutination assays, for example, are based on the ability of specific antibodies to crosslink erythrocytes or to prevent viruses from crosslinking erythrocytes. Cellular crosslinking is easily detected by the precipitation of a cellular mass. In the early 1950s the Hungarian physician Gyula Takátsy initiated the modern form of the high-throughput assay by developing a method for accurately dispensing small amounts of serum. To make the best use of his dispensing technique, he machined plates with 8 rows and 12 columns of miniature wells so that up to 96 hemaggutination assays could be carried out in parallel. The 96-well microplate is now a standard tool in chemical biology.

The essential advance for high-throughput assays came from the ability to manufacture uniform plates with flat, optically transparent bottoms, allowing the absorbance of the solution in each well to be measured with a vertical beam of light (Figure 1.29). Each well of a 96-well plate can serve as a reaction vessel, a cuvette, or...
enzyme-catalyzed reactions or living cells. To screen tens of thousands of compounds in 96-well or 384-well format for their effects for drug development leads, pharmaceutical companies use automated systems to test their candidates. Commercial robotic systems can fill all the wells of a microplate simultaneously. In the quest for drug development leads, pharmaceutical companies use automated systems to screen tens of thousands of compounds in 96-well or 384-well format for their effects on enzyme-catalyzed reactions or living cells.

**Powerful microbiological screens reveal interesting chemical phenomena**

Robotics and miniaturization have greatly increased the capacity of high-throughput assays, but they do not represent a truly parallel approach. Each compound to be tested must be aliquoted into an individual well, and most plate readers measure absorbance one well at a time, albeit very fast. If you want to screen truly massive collections of molecules—say, a billion—you cannot do anything one at a time or the logistics would be unmanageable. For example, testing a billion compounds in 384-well microplates would require enough plates to cover four soccer fields. Screening high-diversity collections requires parallel assays on a miniature scale. Cells are nearly ideal for small-scale work because a single tiny cell can proliferate to produce large numbers of identical clones, enough to determine the structures of the molecules inside. Auxotrophic bacterial selections are based on strains of bacteria that are lacking a gene necessary for growth. For example, when a clonal population of Δfes E. coli, lacking an enzyme needed for the release of iron, was spread onto nutrient agar, the bacteria failed to generate colonies (Figure 1.30). However, when a diverse library of 1.5 million different variants of Δfes E. coli, each expressing a different test protein, were plated on the same medium, some of the bacteria benefited from the test protein and thrived. These types of auxotrophic selections, where one screens for cell growth/ division, are extremely space-efficient because the nonviable bacteria take up virtually no room on the plate. A single milliliter of bacterial culture readily holds more than a billion bacteria and it is straightforward to fit up to a billion different bacteria on a single Petri dish.

The classical Ames test for chemical mutagens is a screen that tests the ability of chemicals to induce genetic mutations into *Salmonella typhimurium* bacteria. The auxotrophic T100 *Salmonella* strain used in the Ames test has a single mutation in the gene that encodes an enzyme essential for histidine biosynthesis. Transcription and translation of the mutant gene leads to a defective enzyme. When plated on histidine-deficient nutrient agar, T100 cells cannot proliferate unless they undergo a DNA reversion mutation back to the correct coding sequence. Chemical mutagens increase the rate of random mutations, including the mutation that rescues histidine biosynthesis. The Ames test has often been used as evidence for the potential carcinogenicity of synthetic chemicals. However, in large-scale random tests of natural and synthetic compounds that repel insects from crops, Ames found that natural compounds were just as likely as synthetic compounds to have potential carcinogenic effects.

The major weakness of bacterial selections is that cell populations are resilient, and natural mutations offer many unexpected mechanisms for surviving stringent selections. Imagine creating a library of enzymes (in a large population of *E. coli*) in the hope that one will catalyze the hydrolysis of phosphorus–nitrogen bonds. If you plate the library of bacteria on nutrient agar containing microcin C7, a small number of the bacteria would probably generate colonies. However, the bacteria that grow into colonies would probably do so for reasons unrelated to P–N bond cleavage, and it would require years of effort to figure out the strategy that a colony of bacteria used to survive. Some of the survivors will have mutant enzymes that acetylate the microcin C7. Other survivors will produce more pumping proteins that export microcin C7 faster than it enters the cell. Other survivors will produce vast amounts of natural enzymes that have weak phosphoramide-cleaving activity.

One can readily screen bacterial libraries for phenotypic traits other than survival. For example, when the biochemical machinery for the production of lycopene (a red pigment that confers the red color of tomatoes) is disrupted, it is a phenotypic trait that a bacterial colony can be selected for. The pigments are produced in two steps: first, the pigments are synthesized (in bacteria containing a gene for the synthesis of the pigment) in a cell-free system, and then the pigments are transformed into the desired colors (in bacteria containing a gene for the production of the pigment). When a diverse library of enzymes (in a large population of *E. coli*), each expressing a different test protein, are plated on the same medium, some of the bacteria will be able to synthesize the pigments (Δfes E. coli). When each member of the Δfes bacterial population expresses a different test protein (out of 1.5 million variants), some of the bacteria thrive and generate colonies. (From M.A. Fisher et al., PLoS ONE 6:e15364, 2011.)
nutrient discussed in Chapter 8) was moved from plants to *E. coli*, it became possible to screen for enzyme mutations that improve lycopene production (Figure 1.31). The advantage of this assay over a survival assay is that the color of the colonies correlates unambiguously with lycopene content. The disadvantage is that the uninteresting colonies take up just as much space on the Petri dish as the interesting deep red colonies, and that limits the diversity of the library to about 10,000 members per plate.

Viruses deliver genes efficiently

Viruses are small packages of genetic material that hijack the biosynthetic machinery of a cell to make additional copies of the virus. Some contain DNA; others contain RNA. They are usually extremely specific to one cell type. For example, the human immunodeficiency virus targets only one type of T lymphocyte in the human body. It does not target other cells and it does not target other organisms. Viruses are well known as the pathogenic agents that cause all manner of human diseases, such as Ebola, smallpox, AIDS, severe acute respiratory syndrome, papillomas, herpes, influenza, and the common cold. They also attack crops, livestock, and pets—even the gastrointestinal bacteria that assist human digestion. As far as we know, viruses infect every type of free-living organism. They also infect computers. Computer viruses are aptly named because they mimic biological viruses in almost every respect except for a general lack of evolutionary capability.

In spite of their notoriety, viruses have several important properties that make them useful for molecular biology. First, they deliver genetic material efficiently to cells (Figure 1.32). Viral gene delivery has been ruthlessly honed by evolution. So far, all of the nonviral methods (such as carrier molecules, heating, electrocution, and ballistics) that have been developed for delivering genes into cells are both disruptive and inefficient. A second important property of viruses is that they encode enzymes that are better translated, fold more efficiently, and catalyze better than the enzymes of the host cell. Viral enzymes have to be superior for the virus to take over the cell. Some of the most efficient enzymatic tools used by molecular biologists were obtained from viruses. Third, viruses are lean and simple. Many human viruses gain added complexity by coating themselves in membrane material from the host cell (a complex mixture of lipids, proteins, and glycans), but ultimately all viruses encapsulate relatively small genetic payloads that encode a small repertoire of proteins. Some viruses are small enough for us to model the positions of all of the atoms in the virus. No living cell has that level of simplicity.

Viruses that infect bacteria, called bacteriophage, or phage for short, are particularly useful, because the bacterial hosts are cheap and easy to grow. Phages are valuable for packaging DNA and making combinatorial libraries. Packaging a synthetic DNA library into phage is inefficient, but once accomplished, each member of the phage library is capable of efficiently transfecting a bacterium to produce a limitless supply of copies (Figure 1.33). Importantly, you can package libraries of phage genes...
in phage that display the corresponding proteins on the viral surface. Libraries of proteins displayed on bacteriophages are a powerful tool for selection in chemical biology.

**Vast libraries of proteins can be screened in vitro using bacteriophages**

*In vitro* selections involving viruses accommodate even greater diversity than selections involving living cells. For example, one can conveniently generate and work with solutions containing more than 10 billion infective bacteriophages per milliliter. Like a bacterium, a single virus has the potential to proliferate to give a vastly larger population of identical clones. Wild-type filamentous phages have proteins designed to bind tightly and specifically to bacteria, but one can alter the viral genome so that novel proteins are displayed on the virus. Using this phage display technique, one can conveniently screen vast libraries of viruses, each displaying a novel protein, for the ability to cling to a surface. Thus, if you wanted to screen a library of proteins displayed on phages for the ability to bind tightly to estrogen, you would chemically attach estrogen molecules to the surface of, for example, a microplate well (*Figure 1.34*). Bacteriophages with proteins that bind to estrogen would stick to the derivatized well, whereas other bacteriophage would not bind at all. The unbound phages are readily rinsed away. Once the selection is complete, the phages that bound to the well could be detached and used to infect bacteria, re-establishing a vast population of clones. Many companies sell microplates with chemically reactive surfaces for use in these kinds of experiments.

**In vitro screens of DNA and RNA push the limits of library diversity**

The largest libraries that have been created and screened are based on DNA. Combinatorial chemical synthesis can conveniently be used to create libraries containing more than $10^{13}$ DNA molecules. Some DNA molecules fold into unique three-dimensional shapes that are capable of binding tightly and specifically to other biomolecules or even catalyzing chemical reactions, much like protein-based enzymes. DNA libraries are usually screened for the ability to bind to chemically immobilized molecules, much like phage libraries. Those that bind tightly and specifically can be amplified into larger quantities with DNA polymerase. DNA and RNA libraries are usually much larger than phage-based protein libraries; however, DNA possesses a limited repertoire of chemical functionality. DNA libraries can be transcribed with RNA polymerase to generate RNA libraries, which can be screened and amplified in a manner similar to DNA (*Figure 1.35*).

**Small molecules take control**

Chemical biologists prefer to use small molecules to control cell function and to report on cellular processes. The main advantage of small-molecule effectors is that they can
be membrane permeable. It is difficult to overstate the advantages of a molecule that can slip undetected through the defensive system of membranes that protect all living cells, gaining access to high-value targets that are deep inside the cell. In contrast, few biological macromolecules can cross cell membranes without some kind of clumsy assistance.

The main disadvantage of small-molecule effectors is that it is hard to find ones with high selectivity. High potency is also desirable, but it is not as important as selectivity. Potency is less important than selectivity because, in theory, you can always increase the concentration of a weak inhibitor, provided it is selective. In practice, solubility limits the concentrations that can be achieved with small molecules. Selectivity is paramount if you want to use a small molecule as a tool to understand cell biology. For example, if you want to know whether the phosphorylation of the transcription factor STAT1 is essential for macrophage activation, you could add staurosporine, a potent inhibitor of protein phosphorylation, and see whether macrophages respond to bacteria. Unfortunately, macrophages contain hundreds of phosphorylating enzymes that target hundreds of proteins, and staurosporine inhibits most of those phosphorylating enzymes. If you use a nonselective inhibitor such as staurosporine, you cannot draw any specific conclusions, even if the compound inhibits macrophage activation.

Small molecules have already been identified that inhibit the main steps in biooligomer assembly (Figure 1.36). Most were isolated from organisms that use them to fend off predators, but synthetic chemistry is beginning to overtake nature as a source of selective inhibitors. The fungal natural product aphidicolin, for example, is a selective inhibitor of DNA polymerase $\alpha$ and $\varepsilon$. The mushroom toxin $\alpha$-amanitin is a selective inhibitor of RNA polymerase II. The bacterial natural product cycloheximide inhibits ribosomal protein synthesis (in humans, but not in bacteria). These inhibitors may seem large and complex compared with the molecules one typically sees in an introductory organic chemistry class; however, as we will see shortly, they are minute compared with typical biooligomers such as DNA, RNA, and proteins. There are many enzymes involved in the construction of terpenes, polyketides, glycosides, and other molecules that make up the rest of the cell. A few can be inhibited selectively with small molecules. For example, the osteoporosis drug alendronate is a specific inhibitor of terpene assembly. The bacterial natural product platensimycin inhibits polyketide assembly. A soy natural product, soyasaponin I, selectively inhibits the glycosylation of proteins.

Arguably, the most interesting steps in the control of biooligomer synthesis are those that control the transcription of genes—specifically, the process by which extracellular signals are transduced into changes in gene expression. These signal
transduction pathways determine the differences between various cell types and between healthy cells and diseased cells. Decades of convoluted experiments have outlined the rough details of the seven main signal transduction pathways, and we discuss them in more detail in Chapter 9. Many of the steps are still poorly understood. Selective inhibitors of signal transduction pathways would allow biologists to test hypotheses better and offer medicinal chemists ideal leads for drug design.

**Short RNA molecules silence gene expression**

Short RNA molecules, about 20 subunits in length, elicit a destructive response in human cells that leads to the cleavage of related RNA sequences. This phenomenon prevents the translation of mRNA transcripts and is called RNA interference. We will talk much more about RNA interference in Chapter 4, but RNA interference merits attention here because it has the potential to displace small molecules as a tool for chemical biology. Short synthetic RNA molecules can enter human cells grown in laboratory culture, and that ability was once thought to be limited to small molecules. Moreover, it is very easy to design and synthesize short RNA molecules without developing new synthetic strategies, as is required for small drug-like molecules.

Will RNA interference replace small molecules as drugs and biological tools? Universal application of RNA interference has been hindered by three major obstacles. First, the process of RNA cell entry is poorly understood and not completely general. When injected directly into the eye, small RNA molecules are effective against viruses or over-vascularization, but the same injection approach is unlikely to work on other human organs. Second, because RNA interference directs the cleavage of both identical and similar sequences, one has to check the RNA transcriptome carefully to anticipate and avoid collateral damage. Finally, RNA interference does not inhibit proteins that have already been translated; it only prevents new proteins from being translated. Until these problems are addressed, small molecules are likely to remain the central tool in modern chemical biology.

**Monoclonal antibodies bind specifically**

Recall that the human immune system can rapidly evolve highly specific antibodies that are capable of binding foreign molecules. The same is true for other mammals, including laboratory strains of mice. When mice are challenged with foreign proteins, or proteins that have been chemically derivatized with foreign molecules, the mice generate B cells that produce highly specific antibodies against the foreign molecules. If you remove B cells from a mouse spleen and try to culture them in the laboratory, they will eventually die. This is because most differentiated mammalia cells have a limited capacity for cell division; however, by fusing antibody-producing B cells with immortal mouse myeloma cells, one can obtain hybrid cell lines called hybridomas that live forever in culture yet produce antibodies (Figure 1.37). Monoclonal antibodies against many important proteins are widely available for laboratory use and can be chemically derivatized to create highly selective reagents for chemical biology experiments. The human immune system would react violently to antibodies from a mouse, but clever techniques have enabled human-like monoclonal antibodies against disease proteins to be obtained and used as human drugs. Most clinically used protein drugs are based on antibodies, including Herceptin™ (breast cancer), Humira™ (autoimmune diseases), Erbitux™ (colorectal cancer), Rituxan™ (non-Hodgkin’s lymphoma), and Avastin™ (macular degeneration). Growing bacteria is easy. In contrast, culturing mammalian cells in the laboratory is a costly endeavor that requires special nutrients and highly controlled, sterile conditions. Thus, protein drugs obtained from mammalian cells command a premium.

**Immortal cancer cell lines serve as mimics of human organs**

Mammalian cells require controlled conditions for laboratory culture—namely sterile conditions, special carbon dioxide incubators, and complex mixtures of growth factors. A wide range of immortal mammalian cell lines are available for laboratory use in the laboratory under controlled conditions (top), typically in an atmosphere of carbon dioxide that creates a carbon dioxide/bicarbonate buffer system in the growth medium. The media is usually supplemented with phenol red as a harmless pH indicator. Hybridoma cell lines combine the immortality of cancer cells and the antibody-producing ability of an evolved B cell (bottom). This false color image of a hybridoma reveals the nucleus in blue and the extensive vesicular network that is important for trafficking antibodies out of the cell. (From A. Karpas, A. Dremucheva, and H. Czepulkowski, *Proc. Natl. Acad. Sci. USA* 98:1799–1804, 2001. With permission from the National Academy of Sciences.)
studies (Table 1.1). Thus, much of our understanding of mammalian cell function has come from laboratory studies of cancer cell lines with cryptic designations such as HeLa cells, MCF-7 cells, and Jurkat cells. The inner workings of these cells are not necessarily the same as the workings of cells in healthy tissue, but they are often the best systems for study and testing, short of living human or animal subjects. A few other common cell lines have been derived from noncancerous tissue. The immortality of NIH 3T3 cells arose spontaneously while culturing embryonic fibroblasts from a Swiss mouse. The Chinese hamster ovary (CHO) cell line is widely used for the expression of protein pharmaceuticals that require the unique machinery of mammalian cells for biosynthesis.

**Human stem cells are highly valuable tools for research and medicine**

Many of the cells in the developing human embryo retain their capacity to regenerate most or all of the various tissue types. These stem cells may have the capacity to replace injured or diseased organs such as spinal cords, hearts, kidneys, eyes, and skin. It is the dream of chemical biologists to use small molecules to control the differentiation of stem cells (Figure 1.38). The most powerful stem cells, capable of regenerating virtually any type of tissue, must be harvested early in embryonic development and therefore require one to destroy the embryo. The use of such embryonic stem cells is mired in moral controversy, so tremendous effort has been invested in the development of stem-cell lines that are obtained without sacrificing an embryo or fetus.

The various cells of the bone marrow differentiate throughout human life through the process of hematopoiesis. We discuss the signal transduction pathways that control this differentiation later, in Chapter 9. Pluripotent hematopoietic stem cells divide and differentiate to form red blood cells, T lymphocytes, B lymphocytes, macrophages, platelets, neutrophils, and eosinophils as needed. Differentiation and proliferation along these different pathways occurs in response to various protein hormones. In most cases of human cell differentiation, the expression of genes changes, but the genotype does not. In response to viral infection, antibody-producing B lymphocytes undergo a dramatic process of differentiation, at both the genetic level and the cellular level. Thus, the genotype of mature B lymphocytes is permanently altered through this special type of differentiation.

**Model organisms teach us about humans**

The study of human physiology, human cell biology, and human development is driven not merely by curiosity but also by the desire to improve the human condition. However, just as human cells are much more difficult to grow in the laboratory than bacterial cells, human beings create significant logistical and ethical challenges as organisms of study. Thus, most of the scientific literature is focused on simpler organisms. These range from simple, single-celled organisms, such as bacteria and yeast, to

### Table 1.1 Examples of immortal mammalian cell lines.

<table>
<thead>
<tr>
<th>Cell line</th>
<th>Tissue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jurkat cells</td>
<td>Human T-cell lymphoma</td>
</tr>
<tr>
<td>HeLa</td>
<td>Human cervical cancer</td>
</tr>
<tr>
<td>MCF-7</td>
<td>Human breast adenocarcinoma</td>
</tr>
<tr>
<td>HepG2</td>
<td>Human hepatocarcinoma</td>
</tr>
<tr>
<td>NIH 3T3</td>
<td>Embryonic mouse fibroblast</td>
</tr>
<tr>
<td>MDCK</td>
<td>Madin-Darby canine kidney</td>
</tr>
<tr>
<td>CHO</td>
<td>Chinese hamster ovary</td>
</tr>
</tbody>
</table>

**Figure 1.38 Controlling differentiation.**

complex animals, such as mice (Table 1.2 and Figure 1.39). The DNA, RNA, and proteins generated by bacteria are very similar to their counterparts found in human cells. Bacterial cells are useful for the production of DNA and proteins in the laboratory, but are a bit too simple to be good models for human cells. At a cellular and genetic level, yeast cells work in a way that is much more relevant than *E. coli*. Moreover, humans have been working with yeast for thousands of years—for as long as we have been baking bread and brewing beer. Furthermore, some genera of yeast cause harmful infections, particularly in immunosuppressed patients.

Microscopic soil worms and fruit flies seem to be extravagant research targets. They are not pathogenic, so why study them? Worms and flies serve as economical models for the study of human biology, physiology, and behavior. They are faster, easier, and more ethical to breed than humans. In general, the reproduction times for model organisms tend to scale inversely with the simplicity of the organism. “Simple”

### Table 1.2 Commonly studied model organisms.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Common name</th>
<th>Generation time</th>
<th>Primary importance</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Escherichia coli</em></td>
<td>Bacterium</td>
<td>20 minutes</td>
<td>Production of DNA and proteins</td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>Yeast</td>
<td>2 hours</td>
<td>Genetics, production of complex human-like proteins</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td>Roundworm</td>
<td>36 hours</td>
<td>Genetic model for cell development/differentiation</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>Fruit fly</td>
<td>10 days</td>
<td>Genetic model for organismal development/differentiation</td>
</tr>
<tr>
<td><em>Mus musculus</em></td>
<td>Mouse</td>
<td>3 months</td>
<td>Model for human physiology</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>Human</td>
<td>14 years</td>
<td>Target organism for the development of new medicines</td>
</tr>
</tbody>
</table>

![Figure 1.39 Model organisms.](image) Model organisms help us understand humans. Clockwise from the upper left: *E. coli* bacteria, *Saccharomyces* yeast, a mouse, humans, *Drosophila*, and *C. elegans*. (*E. coli*, courtesy of Janice Hamel Carr; yeast, courtesy of Masur, Wikimedia Commons; mouse, courtesy of Florence Fortescue, Wikimedia Commons; humans, from National Heart, Lung, and Blood Institute, 1987; *Drosophila*, courtesy of André Karwath, Wikipedia; C. elegans, courtesy of Bob Goldstein, UNC Chapel Hill.)
organisms are not that simple: *Drosophila* and *Caenorhabditis elegans* pack an impressive number of genes, about 14,000 and about 20,000, respectively. As a consequence of evolution, worms, flies, and humans share similar genes, similar proteins, and similar biochemical mechanisms.

The soil worm *C. elegans* is transparent, allowing one to see the cells and the organs. Developmental biologists have traced the development of every one of the 959 cell types in *C. elegans* starting from a single cell. Fruit flies are not transparent, but the stages of embryological development more closely mimic that of humans than do microscopic soil worms. In particular, the use of irradiation to induce distinct phenotypic mutations motivated our early understanding of the relationship between genes and physiological development. Had it not been for *Drosophila*, we would not have a molecular-level understanding of how humans develop the basic body plan of two arms and two legs. Model organisms can also allow us to study genes that relate to human behavior. On ingestion of alcohol, fruit flies experience hyperactivity, decreased coordination, and ultimately sleepiness, much like their human counterparts. An inebriometer was devised to screen a diverse population of fruit flies, to identify mutations that would make them more susceptible to alcohol (Figure 1.40). Using the inebriometer, researchers demonstrated that increased sensitivity to alcohol can result from mutation to a gene they named *cheapdate*. Mutations to the gene *cheapdate* affect levels of a molecule called cyclic AMP (cAMP) in neuronal cells.

Finally, the resilient house mouse has become indispensable for studies of mammalian physiology, metabolism, immunity, cognition, and behavior. In the drug development process, as one progresses from cellular models to human subjects, mice are usually the first animal model in which drugs are tested. No other mammal has been subjected to as much study. Thus we have cured cancer, obesity, heart disease, bacterial infections—all manner of diseases—in mice. New, useful strains of mice have been bred in the laboratory. Many knockout strains, lacking a particular gene, have been created to provide a better understanding of the importance of defective proteins in humans. Mice with defective immune systems have been particularly important because they can be used as hosts for the growth of either diseased or normal tissue from other mammals (Figure 1.41).

### 1.7 SUMMARY

Combinatorial assembly readily accesses unlimited diversity from limited sets of building blocks, and the creation of diversity is central to evolution. It is possible to see the power of combinatorial assembly at all levels of life. The biomolecules that make us human are constructed from a limited set of atom types—mostly hydrogen, oxygen, carbon, nitrogen, phosphorus, and sulfur. The biooligomers that make up cells—DNA, RNA, proteins, oligosaccharides, polyketides, and terpenes—are each assembled from a limited set of subunits, about 40 in total. Further assembly of cells to organisms, to populations, and to ecosystems is accompanied by a marked increase in complexity.

As chemical biologists, we are chiefly concerned with atoms and molecules, and in this book we focus primarily on organic molecules. The central dogma serves as a blueprint for understanding the combinatorial assembly of the biooligomers that make up human cells. The structure of every protein is encoded in a simple way by an RNA molecule and, ultimately, by a fragment of the DNA that makes up the genome. The combinatorial assembly of other biooligomers is more powerfully encrypted. Glycans, lipids, fats, and steroids each require unique sets of catalytic protein modules for their assembly.
LEARNING OUTCOMES

- Contrast the architectural efficiency of one-dimensional oligomers such as DNA, RNA, and proteins with multidimensional oligomers such as organic molecules.
- Identify each type of biooligomer in the central dogma and the flow of chemical information.
- Diagram the elements of a gene and how the expression of genes is controlled.
- Contrast the human genome with that of other organisms.
- Contrast the size and morphology of human cells with those of bacterial cells.
- Understand the nongenomic origins of the diversity of transcriptomes and proteomes.
- Contrast the template-directed biosynthesis of DNA, RNA, and proteins with the nontemplated biosynthesis of oligosaccharides, polyketides, and terpenes.
- Calculate the potential diversity of biooligomers of a given length.
- Understand the combinatorial genetic architecture of antibodies.
- Recognize the chemical structures of common fluorophores.
- Diagram the process of selection and amplification using a phage library.
- Contrast the advantages of small organic molecules over large biooligomers, such as antibodies, as medicines and tools in chemical biology.
- Contrast the advantages and disadvantages of immortal mammalian cell lines and bacteria for understanding normal human cells.
- Identify major applications of common model organisms, such as E. coli, Saccharomyces, C. elegans, Drosophila, and mouse.

ADDITIONAL PROBLEMS

1.5 Throughout this book we will refer to macromolecular structures whose atomic coordinates can be accessed through the Protein Data Bank (PDB) Web site. The Web site has numerous built-in applications that allow you to view and render molecules of DNA, RNA, and proteins. Use the PDB Web site to answer the following questions:

A How many phosphorus atoms are present in the DNA dodecamer (PDB 1BNA)?
B How many monomeric RNA subunits are present in the crystal structure of the hepatitis B virus encapsidation signal (PDB 2IXY)?
C How many iron atoms are present in the enzyme aconitase (PDB 1ACO)?

*1.6

A How many membranes separate the DNA of E. coli from the extracellular environment?
B How many membranes separate the nuclear DNA of human cells from the extracellular environment?
C How many membranes separate the mitochondrial DNA of human cells from the extracellular environment?

1.7 Use the Internet to determine the major organelle that is missing from mature human erythrocytes (red blood cells).

*1.8 Show the starting materials and steps (in the correct order) that would be needed to synthesize the ester shown below. Using that synthetic route, how many different esters could be synthesized using the starting materials shown below?

1.9 If one synthesized every possible DNA molecule, each with 20 subunits, how much would your collection of molecules weigh? The average molecular mass of a DNA subunit is 165 g mol⁻¹.

1.10 Use the Internet to identify the organism and tissue of origin of the following laboratory tumor-cell lines:

A COS7
B LNCaP
C LA-4
D SF21

1.11 Biologists often use Latin animal adjectives (for example human or feline) to describe the origin of tissues, cells, and molecules. For each of the following viruses, identify the common name of the animal it infects.

Avian influenza
Bovine respiratory syncytial virus
Canine parvovirus
Caprine arthritis–encephalitis virus
Equine infectious anemia
Feline panleukopenia
Abelson murine leukemia virus
Ovine pulmonary carcinoma
Piscine mycobacteriosis
Porcine circovirus disease
Simian immunodeficiency virus