Recent years have seen major changes in the way scientists approach cell and molecular biology, with large-scale experimental and computational approaches being applied to understand the complexities of biological systems. Traditionally, cell and molecular biologists studied one or a few genes or proteins at a time. This was changed by genome sequencing projects, which introduced large-scale experimental approaches that generated vast amounts of data to the study of biological systems. The complete genome sequences of a wide variety of organisms, including many individual humans, provide a wealth of information that forms a new framework for studies of cell and molecular biology and opens new possibilities in medical practice. Not only can the sequences of complete genomes be obtained and analyzed, but it is also now possible to undertake large-scale analyses of all of the RNAs and proteins expressed in a cell. These global experimental approaches form the basis of the new field of systems biology, which seeks a quantitative understanding of the integrated behavior of complex biological systems. This chapter considers the development of these new technologies and their impact on understanding the molecular biology of cells.

Genomes and Transcriptomes

Obtaining the complete sequence of the human genome was the first large-scale experimental project undertaken in the life sciences. When it was initiated in the 1980s, it appeared a daunting task to completely sequence all three billion base pairs of the human genome. After all, at that time the largest DNA molecule to have been sequenced was a viral genome of less than 200,000 bases. The Human Genome Project became the largest collaborative undertaking in biology and yielded an initial draft sequence in 2001, with a more refined complete sequence of the human genome published in 2004. Along the way, the complete genome sequences of many other species were obtained, providing important insights into genome evolution. Moreover, tremendous advances in the technology of DNA sequencing have been made, and new sequencing methodologies allow rapid and economical sequencing of individual genomes or transcribed RNAs. These advances have changed the way scientists think about the structure and function of our genomes, as well as allowing new approaches to disease diagnosis and treatment based on personal genome sequencing.

The genomes of bacteria and yeast

The first complete sequence of a cellular genome, reported in 1995 by a team of researchers led by Craig Venter, was that of the bacterium Haemophilus
influenzae, a common inhabitant of the human respiratory tract. The genome of *H. influenzae* is a circular molecule containing approximately $1.8 \times 10^6$ base pairs, more than 1000 times smaller than the human genome. Once the complete DNA sequence was obtained, it was analyzed to identify the genes encoding rRNAs, tRNAs, and proteins. Potential protein-coding regions were identified by computer analysis of the DNA sequence to detect open-reading frames—long stretches of nucleotide sequence that can encode polypeptides because they contain none of the three chain-terminating codons (UAA, UAG, and UGA). Since these chain-terminating codons occur randomly once in every 21 codons (three chain-terminating codons out of 64 total), open-reading frames that extend for more than 100 codons usually represent functional genes. This analysis identified 1743 potential protein-coding regions in the *H. influenzae* genome as well as six copies of rRNA genes and 54 different tRNA genes (Figure 5.1). The predicted coding sequences have an average size of approximately 900 base pairs, so they cover about 1.6 Mb of DNA, corresponding to nearly 90% of the genome of *H. influenzae*. The genome of *E. coli* is approximately twice the size of *H. influenzae*, $4.6 \times 10^6$ base pairs long and containing about 4200 genes, again with nearly 90% of the DNA used as protein-coding sequence. The use of almost all the DNA to encode proteins is typical of bacterial genomes, more than 2000 of which have now been sequenced.

A model for a simple eukaryotic genome, which was sequenced in 1996, is found in the yeast *Saccharomyces cerevisiae*. As discussed in Chapter 1, yeast are simple unicellular organisms, but they have all of the characteristics of eukaryotic cells. The genome of *S. cerevisiae* consists of $12 \times 10^6$ base pairs, which is about 2.5 times the size of the genome of *E. coli*. It contains about 6000
protein-coding genes. Thus, despite the greater complexity of a eukaryotic cell, yeast contain only about 50% more genes than \textit{E. coli}. Protein-coding sequences account for approximately 70% of total yeast DNA, so yeast, like bacteria, have a high density of protein-coding sequence.

\textbf{The genomes of Caenorhabditis elegans, Drosophila melanogaster, and Arabidopsis thaliana}

The next major advance in genomics was the sequencing of the genomes of the relatively simple multicellular organisms, \textit{C. elegans}, \textit{Drosophila}, and \textit{Arabidopsis}. Distinctive features of each of these organisms make them important models for genome analysis: \textit{C. elegans} and \textit{Drosophila} are widely used for studies of animal development, and \textit{Drosophila} has been especially well analyzed genetically. Likewise, \textit{Arabidopsis} is a model for studies of plant molecular biology and development. The genomes of all three of these organisms are approximately 100 × 10^6 base pairs, about ten times larger than the yeast genome but 30 times less than the genome of humans. The determinations of their complete sequences in 1998 (\textit{C. elegans}) and 2000 (\textit{Drosophila} and \textit{Arabidopsis}) were major steps forward, which extended genome sequencing from unicellular bacteria and yeasts to multicellular organisms.

An unanticipated result of sequencing these genomes was that they contained fewer protein-coding genes than expected relative to bacterial or yeast genomes (Table 5.1). The \textit{C. elegans} genome is 97 × 10^6 base pairs and contains about 19,000 predicted protein-coding sequences—approximately eight times the amount of DNA but only three times the number of genes in yeast. When the \textit{Drosophila} genome was sequenced in 2000, it was surprising to find that it contains only about 14,000 protein-coding genes—substantially fewer than the number of genes in \textit{C. elegans}, even though \textit{Drosophila} is a more complex organism. Moreover, it is striking that a complex animal like \textit{Drosophila} has only a little more than twice the number of genes found in yeast. Protein-coding sequences correspond to only about 10% of the \textit{Drosophila} genome and 25% of the \textit{C. elegans} genome, compared with 70% of the

\begin{table}[h]
\centering
\begin{tabular}{|l|c|c|c|}
\hline
Organism & Genome size (Mb)* & Number of protein-coding genes & Protein-coding sequence \\
\hline
\textbf{Bacteria} & & & \\
\textit{H. influenza} & 1.8 & 1700 & 90% \\
\textit{E. coli} & 4.6 & 4200 & 88% \\
\hline
\textbf{Yeast} & & & \\
\textit{S. cerevisiae} & 12 & 6000 & 70% \\
\hline
\textbf{Invertebrates} & & & \\
\textit{C. elegans} & 97 & 19,000 & 25% \\
\textit{Drosophila} & 180 & 14,000 & 10% \\
\hline
\textbf{Plants} & & & \\
\textit{Arabidopsis thaliana} & 125 & 26,000 & 25% \\
\hline
\textbf{Mammals} & & & \\
Human & 3000 & 21,000 & 1.2% \\
\hline
\end{tabular}
\caption{Representative Genomes}
\end{table}

*Mb = millions of base pairs
The organism so far known to have the largest number of genes is the apple, which contains more than twice the number of genes present in the human genome (see Chapter 6).

For many years, scientists generally accepted an estimate of approximately 100,000 genes in the human genome. On publication of the draft genome sequence in 2001, the number was drastically reduced to between 30,000 and 40,000. Current estimates, based on the high-quality sequence published in 2004 and using improved computational tools to identify genes, reduce the number of human protein-coding genes still further, to approximately 21,000. (See Chapter 6.)

yeast genome, so the larger sizes of the Drosophila and C. elegans genomes are substantially due to increased amounts of non-protein-coding sequences rather than protein-coding genes.

The completion of the genome sequence of Arabidopsis thaliana in 2000 extended genome sequencing from animals to plants. The Arabidopsis genome, approximately 125 x 10^6 base pairs of DNA, contains approximately 26,000 protein-coding genes—significantly more genes than were found in either C. elegans or Drosophila. Even more genes have since been found in other plant genomes. For example, the genome of the apple contains about 57,000 protein-coding genes, further emphasizing the lack of relationship between gene number and complexity of an organism.

The human genome

For many scientists, the ultimate goal of the Human Genome Project was determination of the complete nucleotide sequence of the human genome: approximately 3 x 10^9 base pairs of DNA. Because of its large size, determination of the human genome sequence was a phenomenal undertaking, and its publication in draft form in 2001 was heralded as a scientific achievement of historic magnitude.

The draft sequences of the human genome published in 2001 were produced by two independent teams of researchers, each using different approaches. Both of these sequences were initially incomplete drafts in which approximately 90% of the genome had been sequenced and assembled. Continuing efforts then closed the gaps and improved the accuracy of the draft sequences, leading to publication of a high-quality human genome sequence in 2004 by the International Human Genome Sequencing Consortium.

A major surprise from the genome sequence was the unexpectedly low number of human genes (see Table 5.1). The human genome consists of only about 21,000 protein-coding genes, which is not much larger than the number of genes in simpler animals like C. elegans and Drosophila and fewer than in Arabidopsis or other plants. Whereas protein-coding sequences correspond to the majority of the genomes of bacteria and yeast and 10–25% of the genomes of Drosophila and C. elegans, they represent only about 1% of the human genome. The nature of the additional sequences in the human genome and their roles in gene regulation, which may contribute more to biological complexity than simply the number of genes, are discussed in the next chapter.

Over 40% of the predicted human proteins are related to proteins in simpler sequenced eukaryotes, including yeast, Drosophila and C. elegans. Many of these conserved proteins function in basic cellular processes, such as metabolism, DNA replication and repair, transcription, translation, and protein trafficking. Most of the proteins that are unique to humans are made up of protein domains that are also found in other organisms, but these domains are arranged in novel combinations to yield distinct proteins in humans. Compared with Drosophila and C. elegans, the human genome contains expanded numbers of genes involved in functions related to the greater complexity of vertebrates, such as the immune response, the nervous system, and blood clotting, as well as increased numbers of genes involved in development, cell signaling, and the regulation of gene expression.

The genomes of other vertebrates

In addition to the human genome, a large and growing number of vertebrate genomes have been sequenced, including the genomes of fish, frogs,
The Human Genome

Initial Sequencing and Analysis of the Human Genome
International Human Genome Sequencing Consortium

The Sequence of the Human Genome
J. Craig Venter and 273 others

The Context
The idea of sequencing the entire human genome was first conceived in the mid-1980s. It was initially met with broad skepticism among biologists, most of whom felt it was simply not a feasible undertaking. At the time, the largest genome that had been completely sequenced was that of Epstein-Barr virus, which totaled approximately 180,000 base pairs of DNA. From this perspective, sequencing the human genome, which is almost 20,000 times larger, seemed inconceivable to many. However, the idea of such a massive project in biology captivated the imagination of others, including Charles DeLisi, who was then head of the Office of Health and Environmental Research at the U.S. Department of Energy. In 1986 DeLisi succeeded in launching the Human Genome Initiative as a project within the Department of Energy. The project gained broader support in 1988 when it was endorsed by a committee of the U.S. National Research Council. This committee recommended a broader effort, including sequencing the genomes of several model organisms and the parallel development of detailed genetic and physical maps of the human chromosomes. This effort was centered at the U.S. National Institutes of Health, initially under the direction of James Watson (codiscoverer of the structure of DNA), and then under the leadership of Frances Collins.

The first complete genome to be sequenced was that of the bacterium *Haemophilus influenzae*, reported by Craig Venter and colleagues in 1995. Venter had been part of the genome sequencing effort at the National Institutes of Health but had left to head a nonprofit company, The Institute for Genomic Research, in 1991. In the meantime, considerable progress had been made in mapping the human genome, and the initial sequence of *H. influenzae* was followed by the sequences of other bacteria, yeast, and *C. elegans* in 1998.

In 1998 Venter formed a new company, Celera Genomics, and announced plans to use advanced sequencing technologies to obtain the entire human genome sequence in three years. Collins and other leaders of the publicly funded genome project responded by accelerating their efforts, resulting in a race that eventually led to the publication of two draft sequences of the human genome in February 2001.

The Experiments
The two groups of scientists used different approaches to obtain the human genome sequence. The publicly funded team, The International Human Genome Sequencing Consortium, headed by Eric Lander, sequenced DNA fragments derived from bacterial artificial chromosome (BAC) clones that had been previously mapped to human chromosomes, similar to the approach used to determine the sequence of the yeast and *C. elegans* genomes (see figure). In contrast, the Celera Genomics team used a whole-genome shotgun sequencing approach that Venter and colleagues had first used to sequence the genome of *H. influenzae*. In this approach, DNA fragments were sequenced at random, and overlaps between fragments were then used to reassemble a complete genome sequence. Both sequences covered only the euchromatin (Continued on next page)
portion of the human genome—approximately 2900 million base pairs (Mb) of DNA—with the heterochromatin repeat-rich portion of the genome (approximately 300 Mb) remaining unsequenced.

Both of these initially published versions were draft, rather than completed, sequences. Subsequent efforts completed the sequence, leading to publication of a highly accurate sequence of the human genome in 2004.

**The Impact**

Several important conclusions immediately emerged from the human genome sequences. Most strikingly, the number of human genes was surprisingly small and appeared to be between 20,000 and 25,000 in the completed sequence. This unexpected result has led to the recognition of the roles of non-protein coding sequences in our genome, particularly with respect to the multiple mechanisms by which they regulate gene expression.

Beyond the immediate conclusions drawn in 2004, the sequence of the human genome, together with the genome sequences of other organisms, has provided a new basis for biology and medicine. The impact of the genome sequence has been and continues to be felt in discovering new genes and their functions, understanding gene regulation, elucidating the basis of human diseases, and developing new strategies for prevention and treatment based on the genetic makeup of individuals. Knowledge of the human genome may ultimately contribute to meeting what Venter and colleagues refer to as “The real challenge of human biology…to explain how our minds have come to organize thoughts sufficiently well to investigate our own existence.”

chickens, dogs, rodents, and primates (Figure 5.2). The genomes of these other vertebrates are similar in size to the human genome and contain a similar number of genes. Their sequences provide interesting comparisons to that of the human genome and are proving useful in facilitating studies.

**Figure 5.2 Evolution of sequenced vertebrates** The estimated times (millions of years ago) when species diverged are indicated at branch points in the diagram.

© 2015 Sinauer Associates, Inc. This material cannot be copied, reproduced, manufactured or disseminated in any form without express written permission from the publisher.
of different model organisms and in identifying a variety of different types of functional sequences, including regulatory elements that control gene expression. For example, a comparison of the human, mouse, chicken, and zebrafish genomes indicates that about half of the protein-coding genes are common to all vertebrates, whereas approximately 3000 genes are unique to each of these four species (Figure 5.3).

The mammalian genomes that have been sequenced, in addition to the human genome, include the genomes of the platypus, opossum, mouse, rat, dog, rhesus macaque, and chimpanzee. As discussed in earlier chapters, the mouse is the key model system for experimental studies of mammalian genetics and development, while the rat is an important model for human physiology and medicine. Mice, rats, and humans have 90% of their genes in common, so the availability of the mouse and rat genome sequences provide essential databases for research in these areas. The many distinct breeds of pet dogs make the sequence of the dog genome particularly important in understanding the genetic basis of morphology, behavior, and a variety of complex diseases that afflict both dogs and humans. There are approximately 300 breeds of dogs, which differ in their physical and behavioral characteristics as well as in their susceptibility to a variety of diseases, including several types of cancer, blindness, deafness, and metabolic disorders. These characteristics are highly specific properties of different breeds, greatly facilitating identification of the responsible genes. For example, as discussed in Chapter 6, studies of breed-specific differences in dog legs led to the identification of a specific type of gene rearrangement responsible for this characteristic. Recent analyses of dog genomes have also identified genes responsible for coat color and for the body size of small breeds. Similar types of analysis are under way to understand the genetic basis of multiple diseases, including several types of cancer, that are common in some breeds of dogs. Since many of these diseases afflict both dogs and humans, the results of these studies can be expected to impact human health as well as veterinary medicine. In the future, we can also expect genetic analysis of behavior in dogs. Since many canine behaviors, such as separation anxiety, are also common in humans, psychologists may have much to learn from the species that has been our closest companion for thousands of years.

The sequences of the genomes of other primates, including the chimpanzee, bonobo, orangutan, and rhesus macaque, may help pinpoint the unique features of our genome that distinguish humans from other primates. Interestingly, however, comparison of these sequences does not suggest an easy answer to the question of what makes us human. The genome sequences of humans and chimpanzees are about 99% identical. Perhaps surprisingly, the sequence differences between humans and chimpanzees frequently alter the coding sequences of genes, leading to changes in the amino acid sequences of most proteins. Although many of these amino acid changes may not affect protein function, it appears that there are changes in the structure as well as in the expression of thousands of genes between chimpanzees and humans, so identifying those differences that are key to the origin of humans is not a simple task.

The genome of Neandertals, our closest evolutionary relatives, has also been recently sequenced. It is estimated that Neandertals and modern humans diverged about 300,000 to 400,000 years ago. The genomes of Neandertals and modern humans are more than 99.9% identical, significantly more closely related to each other than either is to chimpanzees. Interestingly, these differences alter the coding sequence of only about 90 genes that are conserved in
modern humans. These include genes that are involved in the skin, skeletal development, metabolism, and cognition. Further studies of these genes may elucidate their potential roles in the evolution of modern humans.

**Next-generation sequencing and personal genomes**

The human genome and most of the genomes of model organisms discussed above were sequenced using the dideoxynucleotide technique discussed in Chapter 4, first described by Fred Sanger in 1977 (see Figure 4.20). Automation of this basic method increased its speed and capacity to allow whole genome sequencing, culminating in the successful completion of a high-quality human genome sequence in 2004. However, even with robust automation, genome sequencing by this approach was slow and expensive, so that the sequencing of a complete genome was a major undertaking. For example, the initial sequencing of the human genome took 15 years at a cost of approximately $3 billion. Starting around 2005, a number of new sequencing methods, collectively called **next-generation sequencing**, were developed that have substantially increased the speed and lowered the cost of genome sequencing (Figure 5.4). Since 2001, the cost of sequencing a human genome has decreased more than 10,000 times—from approximately $100 million to several thousand. The speed of sequencing has increased even more, so it is now possible to sequence a complete human genome in a few days. These dramatic changes in sequencing technology have opened the door to sequencing the complete genomes of large numbers of different individuals, allowing new approaches to understanding the genetic basis for many of the diseases that afflict mankind, including cancer, heart disease, and degenerative diseases of the nervous system such as Parkinson’s and Alzheimer’s disease. In addition, understanding our unique genetic makeup as individuals is expected to lead to the development of new tailor-made strategies for disease prevention and treatment.

Next-generation DNA sequencing (also called massively parallel sequencing) refers to several different methods in which millions of templates are sequenced simultaneously, dramatically increasing the speed of sequencing.

**Figure 5.4 Progress in DNA sequencing**
The cost of sequencing a human genome has dropped from approximately $100 million in 2001 to several thousand dollars in 2015. (Data from the National Human Genome Research Institute.)
sequenced simultaneously in a single reaction. The basic general strategy of these methods is illustrated in Figure 5.5. First, the DNA is fragmented and adapter sequences, which serve as primers for amplification and sequencing reactions, are added to the ends of each fragment. Single DNA molecules are then attached to a solid surface and amplified by PCR to produce clusters of spatially separated templates. Millions of templates can then be sequenced in parallel by using lasers to monitor the incorporation of fluorescent nucleotides. The sequences derived from this large collection of overlapping fragments can then be assembled to yield a continuous genome sequence. Alternatively, if a known genome sequence (for example, the human genome) is already available, it can be used as a reference genome on which fragment sequences from a particular individual can be aligned.

The first individual human genomes to be sequenced included those of Craig Venter and James Watson, reported in 2007 and 2008. Since then, the genome sequences of thousands of individuals have been determined. With the cost of sequencing an individual genome now in the range of several thousand dollars, it can be anticipated that personal genome sequencing will become part of medical practice. This will allow therapies to be specifically tailored to the needs of individual patients, both with respect to disease prevention and treatment. Perhaps the best current example, discussed in Chapter 19, is the development of new drugs for cancer treatment, which are specifically targeted against mutations that can be identified by sequencing the cancer genomes of individual patients. In the future, we may expect genome sequencing of healthy people to play an important role in disease prevention by identifying genes that confer susceptibility to disease, followed by taking appropriate measures to intervene. For example, genome sequencing could identify women carrying mutations in genes that confer a high risk

**Figure 5.5 Next-generation sequencing** Cellular DNA is fragmented and adapters are ligated to the ends of each fragment. Single molecules are then anchored to a solid surface and amplified by PCR, forming millions of clusters of molecules. Four color-labeled reversible chain terminating nucleotides are added together with DNA polymerase and a primer that recognizes the adapter sequence. Incorporation of a labeled nucleotide into each cluster of DNA molecules is detected by a laser. Unincorporated nucleotides are removed, chain termination is reversed, and the cycle is repeated to obtain the sequences of millions of clusters simultaneously.
for development of breast cancer, which might be prevented by mastectomy. There is also little doubt that continuing progress in genomics will not only lead to increasing applications in medicine, but will also help us to elucidate the contribution of our genes to other unique characteristics, such as athletic ability or intelligence, and to better understand the interactions between genes and environment that lead to complex human behaviors.

**Global analysis of gene expression**

The availability of complete genome sequences has enabled researchers to study gene expression on a genome-wide global level. It is thus now possible to analyze all of the RNAs that are transcribed in a cell (the *transcriptome*), rather than analyzing the expression of one gene at a time. One commonly used method for global expression analysis is hybridization to *DNA microarrays*, which allows expression of tens of thousands of genes to be analyzed simultaneously. A DNA microarray consists of a glass or silicon chip onto which oligonucleotides are printed by a robotic system in small spots at a high density (Figure 5.6). Each spot on the array consists of a single oligonucleotide. Tens of thousands of unique probes can be printed onto a typical chip, so it is readily possible to produce DNA microarrays containing sequences representing all of the genes in cellular genomes. As illustrated in Figure 5.6, one widespread application of DNA microarrays is in studies of gene expression; for example, a comparison of the genes expressed by two different types of cells. In an experiment of this type, cDNA probes are

---

**Figure 5.6 DNA microarrays**  An example of comparative analysis of gene expression in cancer cells and normal cells. mRNAs extracted from cancer cells and normal cells are used as templates for synthesis of cDNA probes labeled with a fluorescent dye. The cDNA probes are then hybridized to a DNA microarray containing spots of oligonucleotides corresponding to 20,000 or more distinct human genes. The relative level of expression of each gene is indicated by the intensity of fluorescence at each position on the microarray, and the levels of expression in cancer cells and normal cells can be compared. Examples of genes expressed at higher levels in cancer cells are indicated by arrows.
synthesized from the mRNAs expressed in each of the two cell types (e.g., cancer cells and normal cells) by reverse transcription. The cDNAs are labeled with fluorescent dyes and hybridized to DNA microarrays in which 20,000 or more human genes are represented by oligonucleotide spots. The arrays are then analyzed using a high-resolution laser scanner, and the relative extent of transcription of each gene is indicated by the intensity of fluorescence at the appropriate spot on the array. It is then possible to analyze the relative levels of gene expression between the cancer cells and normal cells by comparing the intensity of hybridization of their cDNAs to oligonucleotides representing each of the genes in the cell.

The continuing development of next-generation sequencing has also made it feasible to use DNA sequencing to determine and quantify all of the RNAs expressed in a cell. In this approach, called RNA-seq, cellular mRNAs are isolated, converted to cDNAs by reverse transcription, and subjected to next-generation sequencing (Figure 5.7). In contrast to microarray analysis, RNA-seq reveals the complete extent of transcribed sequences in a cell, rather than just detecting those that hybridize to a probe on a microarray. The frequency with which individual sequences are detected in RNA-seq is proportional to the quantity of RNA in the cell, so this analysis determines the abundance as well as the identity of all transcribed sequences. The sensitivity of RNA-seq is high enough to allow analysis at the single cell level, so the transcriptomes of individual cells can be determined.

One of the surprises revealed by RNA-seq analysis has been that many more RNAs are transcribed than are accounted for by the protein-coding genes of human cells. As discussed in the next chapter, these studies have led to the identification of new classes of RNAs that play critical roles in gene regulation.

**Proteomics**

The analysis of cell genomes and transcriptomes are only the first steps in understanding the workings of a cell. Since proteins are directly responsible for carrying out almost all cell activities, it is necessary to understand not only what proteins can be encoded by a cell’s genome but also what proteins are actually expressed and how they function within the cell. A complete understanding of cell function therefore requires not only the analysis of the sequence and transcription of its genome, but also a systematic analysis of its protein complement. This large-scale analysis of cell proteins (proteomics) has the goal of identifying and quantifying all of the proteins expressed in a given cell (the proteome), as well as establishing the localization of these proteins to different subcellular organelles and elucidating the networks of interactions between proteins that govern cell activities.

**Identification of cell proteins**

The number of distinct species of proteins in eukaryotic cells is typically far greater than the number of genes. This arises because many genes can be expressed to yield several distinct mRNAs, which encode different polypeptides as a result of alternative splicing (discussed in Chapter 6). In addition, proteins can be modified in a variety of different ways, including the addition of phosphate groups, carbohydrates, and lipid molecules. The human genome, for example, contains approximately 20,000 different
protein-coding genes, and the number of these genes expressed in any given cell is thought to be around 10,000. However, because of alternative splicing and protein modifications, it is estimated that these genes can give rise to more than 100,000 different proteins. In addition, these proteins can be expressed at a wide range of levels. Characterization of the complete protein complement of a cell, the goal of proteomics, thus represents a considerable challenge. Although major progress has been made in the last several years, substantial technological hurdles remain to be overcome before a complete characterization of cell proteomes can be achieved.

The first technology developed for the large-scale separation of cell proteins was **two-dimensional gel electrophoresis** *(Figure 5.8)*. A mixture of cell proteins is first subjected to electrophoresis in a tube with a pH gradient running from end to end. Each protein migrates according to charge until it reaches a pH at which the charge of the protein is neutralized, as determined by the protein’s content of acidic and basic amino acids. After denaturation and SDS-binding, the proteins are then subjected to electrophoresis in a second dimension under conditions where they separate according to size, with lower molecular weight proteins moving more rapidly through a gel. This approach is capable of resolving several thousand protein species from a cell extract. However, this is much less than the total number of proteins expressed in mammalian cells, and it is important to note that two-dimensional gel electrophoresis is biased towards the detection of the most abundant cell proteins.

The major tool currently used in proteomics is **mass spectrometry**, which was developed in the 1990s as a powerful method of protein identification.

---

**Figure 5.8 Two-dimensional gel electrophoresis**  A mixture of cell proteins is first subjected to electrophoresis in a pH gradient so that they separate according to charge (horizontal axis). The proteins then undergo another electrophoresis in a second dimension under conditions where they separate by size (vertical axis). The gel is stained to reveal spots corresponding to distinct protein species. The example shown is a gel of proteins from *E. coli*. *(From P. H. O’Farrell. 1975. *J. Biol. Chem.* 250: 4007; courtesy of Patrick H. O’Farrell.)*
Proteins to be analyzed, which could be excised from two-dimensional gels, are digested with a protease to cleave the proteins into small fragments (peptides) in the range of approximately 20 amino acid residues long. A commonly used protease is trypsin, which cleaves proteins at the carboxy-terminal side of lysine and arginine residues. The peptides are then ionized by irradiation with a laser or by passage through a field of high electrical potential and introduced into a mass spectrometer, which measures the mass-to-charge ratio of each peptide. This generates a mass spectrum in which individual tryptic peptides are indicated by a peak corresponding to their mass-to-charge ratio. Computer algorithms can then be used to compare the experimentally determined mass spectrum with a database of theoretical mass spectra representing tryptic peptides of all known proteins, allowing identification of the unknown protein.

More detailed sequence information than just the mass of the peptides can be obtained by tandem mass spectrometry (Figure 5.10). In this technique, individual peptides from the initial mass spectrum are automatically selected to enter a “collision cell” in which they are partially degraded by random breakage of peptide bonds. A second mass spectrum of the partial degradation products of each peptide is then determined. Because each amino acid has a unique molecular weight, the amino acid sequence of the peptide can be deduced from these data. Protein modifications, such as phosphorylation, can also be identified because they alter the mass of the modified amino acid.

Although powerful, the two-dimensional gel/mass spectrometry approach is limited. As noted above, two-dimensional gels favor detection of the most abundant cell proteins, and membrane proteins are characteristically difficult to resolve by this approach. Because of these limitations, it appears that two-dimensional gels are capable of resolving proteins corresponding to only several hundred genes, representing a small fraction of all cell proteins. An alternative approach is the use of mass spectrometry to analyze mixtures of proteins, thereby eliminating the initial separation of proteins by two-dimensional gel electrophoresis. In this approach, called “shotgun mass spectrometry,” a mixture of cell proteins is digested with a protease (e.g., trypsin), and the complex mixture of peptides is subjected to sequencing by tandem mass spectrometry. The sequences of individual peptides...
are then used for database searching to identify the proteins present in the starting mixture. Additional methods have been developed to compare the amounts of proteins in two different samples, allowing a quantitative analysis of protein levels in different types of cells or in cells that have been subjected to different treatments. Although several problems with the sensitivity and accuracy of these methods remain to be solved, the analysis of complex mixtures of proteins by “shotgun” mass spectrometry provides a powerful approach to the systematic analysis of cell proteins.

**Global analysis of protein localization**

Understanding the function of eukaryotic cells requires not only the identification of the proteins expressed in a given cell type, but also characterization of the locations of those proteins within the cell. As reviewed in Chapter 1, eukaryotic cells contain a nucleus and a variety of subcellular organelles. Systematic analysis of the proteins present in these organelles is an important goal of proteomic approaches to cell biology.

The protein composition of a variety of organelles has been determined by combining classical cell biology methods with mass spectrometry. Organelles of interest are isolated from cells by subcellular fractionation techniques, as discussed in Chapter 1 (see Figures 1.39 and 1.40). The proteins present in isolated organelles can then be determined by mass spectrometry. The proteome of a variety of organelles and large subcellular structures, such as nucleoli, have been characterized by this approach—some examples are illustrated in Table 5.2. For example, more than 700 different proteins have been identified by mass spectrometry of isolated mitochondria and approximately 200 different proteins identified in lysosomes. By performing such studies with organelles isolated from cells of different tissues or grown under different conditions, it is also possible to determine the changes in protein composition that are associated with different cell types or physiological states.

**Table 5.2 Protein composition of cellular structures**

<table>
<thead>
<tr>
<th>Structure</th>
<th>Number of proteins identifieda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plasma membrane</td>
<td>860–2000</td>
</tr>
<tr>
<td>Endoplasmic reticulum</td>
<td>140</td>
</tr>
<tr>
<td>Mitochondria</td>
<td>400–750</td>
</tr>
<tr>
<td>Lysosome</td>
<td>200</td>
</tr>
<tr>
<td>Centrosome</td>
<td>114</td>
</tr>
<tr>
<td>Peroxisome</td>
<td>70</td>
</tr>
<tr>
<td>Golgi apparatus</td>
<td>400</td>
</tr>
</tbody>
</table>

aRanges for numbers of identified proteins represent the results obtained in different cell types.

Protein interactions

Proteins almost never act alone within the cell. Instead, they generally function by interacting with other proteins in protein complexes and networks. Elucidating the interactions between proteins therefore provides important clues as to the function of novel proteins, as well as helps to understand the complex networks of protein interactions that govern cell behavior. Along with global studies of subcellular localization, the systematic analysis of protein complexes and interactions has therefore become an important goal of proteomics.

One approach to the analysis of protein complexes is to isolate a protein from cells under gentle conditions, so that it remains associated with the proteins it normally interacts with inside the cell. Typically, an antibody against a protein of interest would be used to isolate that protein from a cell extract by immunoprecipitation (Figure 5.11). A cell extract is incubated with an antibody, which binds to its antigenic target protein. The resulting antigen-antibody complexes can then be isolated. If the cell extract has been prepared under gentle conditions so that the target protein remains associated with the proteins it normally interacts with inside the cell, then these interacting proteins will also be present in the immunoprecipitates. Such immunoprecipitated protein complexes can then be analyzed, for example by mass spectrometry, to identify not only the protein against which the antibody was directed, but also other proteins with which it was associated in the cell extract (Figure 5.12). This approach to analysis of protein interactions has been used to characterize...

Figure 5.11 Immunoprecipitation A mixture of cell proteins is incubated with an antibody bound to beads. The antibody forms complexes with the protein (green) against which it is directed (the antigen). These antigen-antibody complexes are collected on the beads and the target protein is isolated.

Figure 5.12 Analysis of protein complexes A known protein (blue) is isolated from cells as a complex with other interacting proteins (yellow and red). The entire complex can be analyzed by mass spectrometry to identify the interacting proteins.
a variety of protein-protein interactions in different types of cells and under different physiological conditions, leading to the identification of numerous interactions between proteins involved in processes such as cell signaling or gene expression.

Alternative approaches to systematic analyses of protein complexes include screens for protein interactions in vitro as well as genetic screens that detect interactions between pairs of proteins that are introduced into yeast cells. Expression of cloned genes in yeast is particularly useful because simple methods of yeast genetics can be employed to identify proteins that interact with one another. In this type of analysis, called the yeast two-hybrid system, two different cDNAs (for example, from human cells) are joined to two distinct domains of a protein that stimulates expression of a target gene in yeast (Figure 5.13). Yeast are then transformed with the hybrid cDNA.

**Figure 5.13 The yeast two-hybrid system** cDNAs of two human proteins are cloned as fusions with two domains (designated 1 and 2) of a yeast protein that stimulates transcription of a target gene. The two recombinant cDNAs are introduced into a yeast cell. If the two human proteins interact with each other, they bring the two domains of the yeast protein together. Domain 1 binds DNA sequences at a site upstream of the target gene, and domain 2 stimulates target gene transcription. The interaction between the two human proteins can thus be detected by expression of the target gene in transformed yeast.
clones to test for interactions between the two proteins. If the human proteins do interact with each other, they will bring the two domains of the yeast protein together, resulting in stimulation of target gene expression in the transformed yeast. Expression of the target gene can be easily detected by growth of the yeast in a specific medium or by production of an enzyme that produces a blue yeast colony, so the yeast two-hybrid system provides a straightforward method to test protein-protein interactions.

High throughput yeast two-hybrid screens have been applied to systematic screens of interactions between proteins of higher eukaryotes, including Drosophila, C. elegans, and humans. The screens have identified thousands of protein-protein interactions, which can be presented as maps that depict an extensive network of interacting proteins within the cell (Figure 5.14). Elucidation of these protein networks would be a major step forward in our understanding of the complexities of cell regulation, as well as illuminating the functions of many so-far unidentified proteins.

**Systems Biology**

The genome sequencing projects have led to a fundamental change in the way in which many problems in biology are being approached, with large-scale experimental approaches that generate vast amounts of data now in common use. Handling the enormous amounts of data generated by whole genome sequencing required sophisticated computational analysis and spawned the new field of bioinformatics. This field lies at the interface between biology and computer science and is focused on developing the computational methods needed to analyze and extract useful biological information from the sequence of billions of bases of DNA. Other types of large-scale biological experimentation, including global analysis of gene expression and proteomics, similarly yield vast amounts of data, far beyond the scope of traditional biological experimentation. These large-scale experimental approaches form the basis of the new field of systems biology, which seeks a quantitative understanding of the integrated dynamic behavior of complex biological systems.
biological systems and processes. In contrast to traditional approaches, systems biology is characterized by the use of large scale datasets for quantitative experimental analysis and modeling (Figure 5.15). Some of the research areas that are amenable to large-scale experimentation, bioinformatics, and systems biology are discussed below.

**Systematic screens of gene function**

The identification of all of the genes in an organism opens the possibility for a large-scale systematic analysis of gene function. One approach is to systematically inactivate (or knockout) each gene in the genome by homologous recombination with an inactive mutant allele (see Figure 4.35). Complete collections of strains with mutations in all known genes are available for several model organisms, including *E. coli*, yeast, *Drosophila*, *C. elegans*, and *Arabidopsis thaliana*. These collections of mutant strains can be analyzed to determine which genes are involved in any biological property of interest. A large-scale international project to systematically knockout all genes in the mouse is also under way, and targeted mutagenesis has now indicated functions of more than 7000 mouse genes.

Alternatively, large-scale screens based on RNA interference (RNAi) are being used to systematically dissect gene function in a variety of organisms, including *Drosophila*, *C. elegans*, and mammalian cells in culture. In RNAi screens, double-stranded RNAs are used to induce degradation of the homologous mRNAs in cells (see Figure 4.38). With the availability of complete genome sequences, libraries of double-stranded RNAs can be designed and used in genome-wide screens to identify all of the genes involved in any biological process that can be assayed in a high-throughput manner. For example, genome-wide RNAi analysis can be used to identify genes required for the growth and viability of cells in culture (Figure 5.16). Individual double-stranded RNAs from the genome-wide library are tested in microwells in a high-throughput format to identify those that interfere with the growth of cultured cells, thereby characterizing the entire set of genes that are required for cell growth or survival.

**Figure 5.15 Systems biology**  Traditional biological experiments study individual molecules and pathways. Systems biology uses global experimental data for quantitative modeling of integrated systems and processes.

**Figure 5.16 Genome-wide RNAi screen for cell growth and viability**  Each microwell contains siRNA corresponding to an individual gene. Tissue culture cells are added to each well and incubated to allow cell growth. Those wells in which cells fail to grow identify genes required for cell growth or viability.
under particular sets of conditions. Similar RNAi screens have been used to identify genes involved in a variety of biological processes, including cell signaling pathways, protein degradation, and transmission at synapses in the nervous system. More recently, genome-wide screens using the CRISPR/Cas system (see Figure 4.36) have similarly been applied to systematically identify sets of genes in human cells that are responsible for properties such as survival or resistance to anticancer drugs.

**Regulation of gene expression**

Genome sequences can, in principle, reveal not only the protein-coding sequences of genes, but also the regulatory elements that control gene expression. As discussed in subsequent chapters, regulation of gene expression is critical to many aspects of cell function, including the development of complex multicellular organisms. Understanding the mechanisms that control gene expression, including transcription and alternative splicing, is therefore a central undertaking in contemporary cell and molecular biology, and the availability of genome sequences contributes substantially to this task. Unfortunately, it is far more difficult to identify gene regulatory sequences than it is to identify protein-coding sequences. Most regulatory elements are short sequences of DNA, typically spanning only about 10 base pairs. Consequently, sequences resembling regulatory elements occur frequently by chance in genomic DNA, so physiologically significant elements cannot be identified from DNA sequence alone. The identification of functional regulatory elements and elucidation of the signaling networks that control gene expression therefore represent major challenges in bioinformatics and systems biology.

The availability of genomic sequences has enabled scientists to undertake global studies of gene expression, using microarrays or RNA-seq (see Figures 5.6 and 5.7), in which the expression levels of all genes in a cell can be assayed simultaneously. Such studies of global patterns of transcription have become extremely valuable in revealing the overall changes in gene regulation associated with discrete cell behaviors, such as cell differentiation or the response of cells to a particular hormone or growth factor. Since genes that are coordinately regulated within a cell may be controlled by similar mechanisms, analyzing changes in the expression of multiple genes can help pinpoint shared regulatory elements. Moreover, changes in gene expression that occur over time can reveal networks of gene expression.

A variety of computational approaches are also being used to characterize functional regulatory elements. One approach is comparative analysis of the genome sequences of related organisms. This is based on the assumption that functionally important sequences are conserved in evolution, whereas nonfunctional segments of DNA diverge more rapidly. Approximately 5% of the genome sequence is conserved among mammals: since ~1.2% of the genome corresponds to protein-coding sequence, the remaining 4% may represent functionally important regulatory sequences. For example, computational analysis to identify noncoding sequences that are conserved between the human, mouse, rat, and dog genomes has proven useful in delineating sequences that control gene transcription (Figure 5.17). In addition, functional regulatory elements often occur in clusters, 

Figure 5.17 Conservation of functional gene regulatory elements Human, mouse, rat, and dog sequences near the transcription start site of a gene contain a functional regulatory element that binds the transcriptional regulatory protein Err-α. These sequences (highlighted in yellow) are conserved in all four genomes, whereas the surrounding sequences are not. (After X. Xie et al., 2005. Nature 434: 338.)
reflecting the fact that genes are generally regulated by the interactions of multiple transcription factors (see Chapter 8). Computer algorithms designed to detect clusters of transcription factor binding sites in genomic DNA have also proven useful in identifying sequences that regulate gene expression.

As discussed further in Chapter 8, large-scale experimental approaches for genome-wide analysis of the binding sites of regulatory proteins have also been developed. In addition, specific modifications of histones that are associated with transcriptionally-active genes or gene regulatory regions have been characterized, and genome-wide analysis of the sites of these histone modifications can be used to provide a global experimental identification of gene regulatory sequences. The application of these global approaches provides a new window to understanding the control of gene expression in complex eukaryotic cells. A prominent example is a large scale project called ENCODE (Encyclopedia of DNA Elements), which utilized RNA-seq to characterize all transcribed RNAs as well as global methods for determining gene regulatory sequences in 147 different types of human cells. The results of this global analysis have changed our views of the structure and function of mammalian genomes, as will be discussed in subsequent chapters. Most notably, even though protein coding-sequences account for only about 1% of the human genome, more than 50% of our genome is transcribed, with many of these transcribed noncoding sequences playing important roles in gene regulation.

**Networks**

Classical experimental biology is focused on single genes or proteins, which often act sequentially to catalyze a series of reactions that constitute a pathway. Several metabolic pathways consisting of such series of reactions were discussed in Chapter 3. For example, glycolysis (the breakdown of glucose to pyruvate) is mediated by the sequential action of nine different enzymes (see Figure 3.2). Signaling pathways act similarly to transmit information from the environment, such as the presence of a hormone, to targets within the cell. For example, the hormone epinephrine signals the breakdown of glycogen to glucose in muscle cells (see Figure 2.30). This response to epinephrine is mediated by a signaling pathway consisting of six different proteins (Figure 5.18).

The activities of individual pathways have been elucidated by classical methods of analysis of small numbers of genes or proteins. However, signaling within the cell is far more complicated than the activities of individual pathways. Metabolic or signaling pathways do not operate in isolation; rather, there is extensive crosstalk between different pathways, so that multiple pathways interact with one another to form networks within the cell. Computational modeling of such networks is currently a major challenge in systems biology, which will be necessary to understand the dynamic response of cells to their environment.

---

**Figure 5.18 Example of a signaling pathway**  The binding of epinephrine (adrenaline) to its cell surface receptor triggers a signaling pathway that leads to the breakdown of glycogen to glucose-1-phosphate.
Some of the ways in which pathways can connect within a network are illustrated in Figure 5.19. The activity of many pathways is controlled by feedback loops. An example of a negative feedback loop is provided by feedback inhibition of metabolic pathways (see Figure 2.28). In addition to negative feedback loops, signaling networks can contain positive feedback loops and feedforward relays in which the activity of one component of a pathway stimulates a distant downstream component. Individual pathways communicate to form networks by crosstalk, which refers to the interaction of one pathway with another. The crosstalk between different pathways can be either positive (where one pathway stimulates the other) or negative (where one pathway inhibits the other).

A full understanding of cell signaling will require the development of network models that predict the dynamic behavior of the interconnected pathways that ultimately result in a biological response. In this view of the cell as an integrated system, mathematical models and computer simulations will clearly be needed to deal with the complexity of the problem. For example, many signaling pathways (discussed in Chapter 16) ultimately affect gene expression. The human genome encodes more than 4000 proteins that function in these pathways, so the potential for cross-regulation between pathways formed from combinations of these elements is enormous. A recently developed model of a gene regulatory network responsible for development of an embryonic cell lineage in sea urchins provides a graphical representation of this complexity (Figure 5.20). Although clearly a daunting task, understanding cell regulatory networks in quantitative terms using mathematical and computational approaches that view the cell as an

**Figure 5.20 A gene regulatory network** The network includes all regulatory genes required for development of the embryonic cells that differentiate into skeletal cells of the sea urchin. (From P. Oliveri, Q. Tu and E. H. Davidson, 2008. Proc. Natl. Acad. Sci. USA 105: 5955.)
An integrated biological system is a critically important problem at the cutting edge of research in cell biology.

**Synthetic biology**

Synthetic biology is an engineering approach to understanding and manipulating biological systems. Rather than studying existing (natural) biological systems, the goal of synthetic biology is to design and create new (unnatural or synthetic) systems. By designing such new systems, synthetic biologists hope to not only create useful products but also to better understand how the behavior of existing cells is controlled.

One approach taken by synthetic biologists is to synthesize new molecules with biological properties. An example is provided by the synthesis of RNA molecules that are capable of self-replication, which provided a critical demonstration of the ability of RNA to serve as the first self-replicating molecule in prebiotic evolution (see Chapter 1). Alternatively, many synthetic biologists engineer new systems using components of existing cells, which can be assembled in different ways that result in distinct functional properties. The engineering approach of synthetic biology is not only important in designing biological systems with useful functions but is also complementary to the analytical approach of systems biology in understanding naturally occurring cells. The ability to engineer a novel biological system tests and expands our understanding of the principles that govern the function of natural systems.

The first systems designed by synthetic biologists were genetic circuits engineered in *E. coli* by the laboratories of James Collins and Stanislas Leibler in 2000. One of these systems, called a genetic toggle switch, was designed to confer stability and memory on a network regulating gene expression. The system consists of three genes: a reporter gene encoding an easily detectable indicator of gene expression and two regulatory genes encoding repressors that can be inactivated by the addition of small molecule inducers (Figure 5.21). The key feature responsible for the stability of this engineered circuit is that the two repressors control expression of each other as well as the

---

**Figure 5.21** A genetic toggle switch. The circuit includes genes encoding two repressors (A and B) that regulate each other and a reporter controlled by repressor B. Inactivation of repressor B leads to a stable state in which the reporter is expressed, whereas inactivation of repressor A leads to a stable state in which the reporter is repressed. (After T. S. Gardner, C. R. Cantor and J. J. Collins, 2000. *Nature* 403: 339.)
reporter. Repressor A blocks expression of repressor B, and repressor B blocks expression of both repressor A and the reporter gene. As an example of the system’s stability, consider the effects of adding a small molecule inducer that inactivates repressor B. In the absence of repressor B, both the reporter gene and repressor A are expressed. Importantly, the system will remain in this state even if the inducer is withdrawn, because repressor A continues to prevent expression of repressor B. Conversely, if repressor A is inactivated, repressor B will be expressed and reporter gene expression as well as further expression of repressor A will be stably shut down. Thus, the system can switch between two stable states. Similar genetic circuits have since been engineered in eukaryotic models, including both yeast and mammalian cells. Engineering these systems has substantially advanced our understanding of how a regulatory circuit can alternate between two stable states—a common feature of the networks involved in many aspects of cell signaling and the regulation of cell proliferation, as will be discussed in later chapters.

The practical applications of synthetic biology include the engineering of metabolic pathways to efficiently produce therapeutic drugs. A good example is provided by the production of the antimalarial drug artemisinin (Figure 5.22). Malaria is a major global health problem, affecting more than 200 million people and responsible for more than 600,000 deaths per year. It is caused by infection with parasites belonging to the genus Plasmodium, most commonly P. falciparum. The most effective treatment for malaria caused by P. falciparum is artemisinin in combination with other anti-malarial drugs (called artemisinin-combination therapy or ACT). Unfortunately, artemisin is produced by a plant (sweet wormwood) that takes about eight months to grow to full size, and its supply has been unstable, leading to shortages and substantial price fluctuations. To address this problem and develop a non-botanical source of artemisin, Jay Keasling and his collaborators engineered strains of yeast that produced high-yields of artemisinic acid, which could then be efficiently converted to artemisinin by a chemical process. In 2013, the pharmaceutical company Sanofi announced the launch of a new facility to produce artemisinin by this method, and more than 16 million treatments have been produced since 2014.

The ultimate goal of synthetic biology might be viewed as the creation of a fully synthetic cell. A milestone was thus reached with the creation of the first cell with a completely synthetic genome in 2010. Starting from the known nucleotide sequence of the 1.08-Mb Mycoplasma mycoides genome, Craig Venter and his colleagues synthesized overlapping oligonucleotides corresponding to the complete genome sequence. These synthetic oligonucleotides were then assembled in several steps to yield a complete synthetic genome of 1,077,947 bases that also contained sequences required for propagation as a plasmid in yeast. The synthetic genome was propagated as a plasmid in yeast and then introduced into a different mycoplasma subspecies, M. capricolum, by gene transfer techniques. Cells containing the synthetic genome were selected by tetracycline resistance and propagated in culture. Sequencing their genomes indicated that they were entirely derived from the synthetic M. mycoides DNA, and cells with the synthetic genome were found to grow normally and show the morphology of normal M. mycoides (Figure 5.23). These mycoplasma thus represent the first cells with a purely synthetic genome. Since all of the
Malaria and Synthetic Biology

The Disease
Malaria is caused by a parasite transmitted by infected mosquitoes and is the most significant human parasitic disease. Symptoms are severe chills, vomiting and fever, resembling flu. If the disease is left untreated, it may lead to severe complications and death. The World Health Organization estimates that there were 198 million cases of malaria and 584,000 deaths in 2013, with 90% of the deaths occurring in Africa.

Molecular and Cellular Basis
The parasites that cause malaria are protozoans belonging to the genus Plasmodium. Five species, *P. falciparum*, *P. malariae*, *P. ovale*, *P. vivax* and *P. knowlesi*, cause malaria in humans. Approximately 75% of cases and almost all deaths are caused by *P. falciparum* (see figure). The disease is transmitted by the bite of an Anopheles mosquito. The parasite then travels to the liver and replicates in liver cells, producing thousands of progeny. These progeny then infect and replicate in red blood cells, which lyse to release more parasites that infect additional red blood cells and undergo additional rounds of replication. The blood stage parasites cause the symptoms of the disease, and can also be picked up by a mosquito for further transmission.

Prevention and Treatment
Development of a vaccine against malaria is a major public health research priority, but no anti-malaria vaccine is currently available. Instead, efforts at disease prevention focus on measures to prevent mosquito bites, such as using insect repellants, sleeping under mosquito nets, or spraying insecticides. People traveling to areas where malaria is prevalent can take a variety of anti-malarial drugs, most of which are also used for treatment. However, prevention with these drugs is not practical for residents of areas where malaria exists because of their cost and the side effects of long term use.

The first effective treatment for malaria was quinine, which was replaced by chloroquine in the 1940s. Unfortunately, strains of *P. falciparum* that were resistant to chloroquine developed in the 1950s and have now become widespread. The most effective current therapy for *P. falciparum* malaria is artemisinin, which was discovered by the Chinese scientist Tu Youyou in the 1970s. It is administered as combination therapy with other antimalarial drugs, in order to prevent the development of strains of *P. falciparum* that are resistant to artemisinin.

Artemisinin was discovered in the leaves of *Artemisia annua* (sweet wormwood), which unfortunately provides a limited supply of the drug. The plants take about eight months to grow to full size and, because of instability in its availability, the market price for artemisinin has fluctuated over a tenfold range in recent years. The work of synthetic biologists has made an important contribution to this problem by engineering a strain of yeast that produces high levels of artemisinic acid, which can be efficiently converted chemically to artemisinin. Based on the efficient production of artemisinin by this engineered yeast, the pharmaceutical company Sanofi established a new production facility for artemisinin in 2013 and has produced more than 16 million treatments since 2014.

Molecular Medicine

cellular proteins are specified by that genome, with the proteins present in the original recipient cell diluted out during replication, these mycoplasma also represent the first synthetic cells. Current efforts are under way to engineer a synthetic yeast genome as a model for design of a eukaryotic cell. Given the steadily decreasing costs of DNA synthesis, the possibility of engineering plant and animal genomes may also become feasible undertakings in the foreseeable future.

P. falciparum in a blood smear
SUMMARY

Genomes and Transcriptomes

- **The genomes of bacteria and yeast**: Bacterial genomes are extremely compact, with protein-coding sequences accounting for nearly 90% of the DNA. The *E. coli* genome consists of $4.6 \times 10^6$ base pairs of DNA and contains about 4200 genes. The first eukaryotic genome to be sequenced was that of the yeast *S. cerevisiae*. The *S. cerevisiae* genome contains about 6000 genes, and protein-coding sequences account for approximately 70% of the genome.

- **The genomes of Caenorhabditis elegans, Drosophila melanogaster, and Arabidopsis thaliana**: The genome of *C. elegans* contains about 19,000 protein-coding genes, which account for about 25% of the genome. The genome of *Drosophila* contains approximately 14,000 genes, with protein-coding sequences accounting for only about 10% of the genome. The numbers of genes in these species indicate that gene number is not simply related to the biological complexity of an organism. Likewise, the genome of the small flowering plant *Arabidopsis thaliana* contains approximately 26,000 genes—surprisingly more genes than in either *Drosophila* or *C. elegans*. Other plant genomes also contain large numbers of genes, including approximately 57,000 genes in the apple. The large number of genes in these plants further emphasizes the lack of relationship between gene number and complexity of an organism. See Video 5.1.

- **The human genome**: The human genome contains approximately 21,000 protein-coding genes—not much more than the number of genes found in simpler animals like *Drosophila* and *C. elegans*, and fewer than in *Arabidopsis* and other plants.

- **The genomes of other vertebrates**: The genomes of fish, chickens, mice, rats, dogs, rhesus macaques, chimpanzees, and Neandertals provide important comparisons to the human genome. All of these vertebrates contain similar numbers of genes.

- **Next-generation sequencing and personal genomes**: Enormous progress in the technology of DNA sequencing has now made it feasible to determine the complete sequence of individual genomes. See Animation 5.1.

- **Global analysis of gene expression**: The RNAs expressed in a cell can be analyzed globally by hybridization to DNA microarrays or by next-generation sequencing. See Animation 5.2.

Proteomics

- **Identification of cell proteins**: Characterization of the complete protein complement of cells is a major goal of proteomics. Mass spectrometry provides a powerful tool for protein identification, which can be used to identify either isolated proteins or proteins present in mixtures.

- **Global analysis of protein localization**: Isolated subcellular organelles can be analyzed by mass spectrometry to determine their protein constituents.

- **Protein interactions**: The purification of protein complexes from cells and analysis of interactions of proteins introduced into yeast can identify interacting proteins and may lead to elucidation of the complex networks of protein interactions that regulate cell behavior.
**Systems Biology**

- **Systematic screens of gene function:** The genome sequencing projects have introduced large-scale experimental and computational approaches to research in cell and molecular biology. Genome-wide screens using RNA interference or the CRISPR/Cas system can systematically identify all of the genes in an organism that are involved in any biological process that can be assayed in a high-throughput format.

- **Regulation of gene expression:** The identification of gene regulatory sequences and understanding the mechanisms that control gene expression are major challenges in bioinformatics and systems biology. These problems are being approached by genome-wide studies of gene expression, histone modification, and regulatory protein binding, combined with the development of computational approaches to identify functional regulatory elements.

- **Networks:** The activity of signaling pathways within the cell is regulated by feedback loops that control the extent and duration of signaling. Different signaling pathways also interact to regulate each other’s activity. The extensive crosstalk between individual pathways leads to the formation of complex signaling networks that ultimately control cell behavior.

- **Synthetic biology:** Synthetic biology is an engineering approach to understanding biology, with the goal of designing new biological systems. The systems designed by synthetic biologists include genetic circuits, metabolic pathways, and synthetic genomes.

**Questions**

1. How can protein-coding sequences be identified in the DNA sequence of a genome?
2. How does the fraction of protein-coding sequence in the human genome compare to that in the yeast genome?
3. How do the number of genes in the human genome compare to the number of genes in apples?
4. Explain how mass spectrometry can be used to identify a protein.
5. How could you compare the protein composition of mitochondria from cancer cells and normal cells?
6. How can immunoprecipitation be used to study protein-protein interactions?
7. How could you use RNA interference to identify genes that might be good targets for the development of drugs for breast cancer treatment?
8. Explain how next-generation sequencing is used to study gene expression.
9. Why is it more difficult to identify regulatory sequences than protein-coding sequences?
10. How can synthetic biology contribute to understanding natural biological systems?
References and Further Reading

(Key review articles for each major section are highlighted in bold.)

Genomes and Transcriptomes


Prüfer, K. and 44 others. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505: 43–49. [P]


Rhesus Macaque Genome Sequencing and Analysis Consortium. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science 316: 222–234. [P]


Proteomics


Systems Biology


