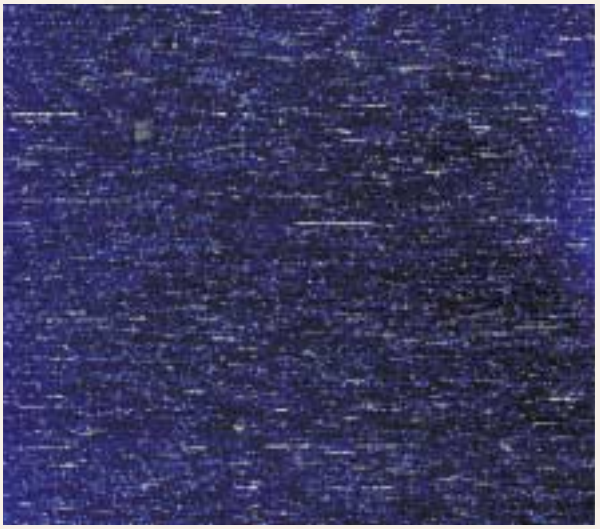
CHAPTER 15

Microbial Genomics



A DNA chip can be used to follow gene expression by a microorganism's complete genome. This chip contains probes for the over 4,200 known open reading frames in the *E. coli* genome.

Outline

Concepts

1. Genomics is the study of the molecular organization of genomes, their information content, and the gene products they encode. It may be divided into structural genomics, functional genomics, and comparative genomics.
2. Individual pieces of DNA can be sequenced using the Sanger method. The easiest way to analyze microbial genomes is by whole-genome shotgun sequencing in which randomly produced fragments are sequenced individually and then aligned by computers to give the complete genome.
3. Because of the mass of data to be analyzed, the use of sophisticated programs on high-speed computers is essential to genomics.
4. Many bacterial genomes have already been sequenced and compared. The results are telling us much about such subjects as genome structure, microbial physiology, microbial phylogeny, and how pathogens cause disease. They will undoubtedly help in preparing new vaccines and drugs for the treatment of infectious disease.
5. Genome function can be analyzed by annotation, the use of DNA chips to study mRNA synthesis, and the study of the organism's protein content (proteome) and its changes.

A prerequisite to understanding the complete biology of an organism is the determination of its entire genome sequence.

—J. Craig Venter, et al.

Chapter 13 provided a brief introduction to microbial recombination and plasmids, including the use of conjugation and other techniques in mapping the chromosome. Chapter 14 described the development and impact of recombinant DNA technology. This chapter will carry these themes further with the discussion of the current revolution in genome sequencing. We will begin with a general overview of the topic, followed by an introduction to the DNA sequencing technique. Next, the whole-genome shotgun sequencing method will be briefly described. This is followed by a comparison of selected microbial genomes and a discussion of what has been learned from them. After we have considered genome structure, we will turn to genome function and the array of transcripts and proteins produced by genomes. The focus will be on annotation, DNA chips, and the use of two-dimensional electrophoresis to study the proteome. The chapter concludes with a brief consideration of future challenges and opportunities in genomics.

15.1 Introduction

Genomics is the study of the molecular organization of genomes, their information content, and the gene products they encode. It is a broad discipline, which may be divided into at least three general areas. **Structural genomics** is the study of the physical nature of genomes. Its primary goal is to determine and analyze the DNA sequence of the genome. **Functional genomics** is concerned with the way in which the genome functions. That is, it examines the transcripts produced by the genome and the array of proteins they encode. The third area of study is **comparative genomics,** in which genomes from different organisms are compared to look for significant differences and similarities. This helps identify important, conserved portions of the genome and discern patterns in function and regulation. The data also provide much information about microbial evolution, particularly with respect to phenomena such as horizontal gene transfer.

It should be emphasized at the beginning that whole-genome sequence information provides an entirely new starting point for biological research. In the future, microbiologists will not have to spend as much time cloning genes because they will be able to generate new questions and hypotheses from computer analyses of genome data. Then they can test their hypotheses in the laboratory.

15.2 Determining DNA Sequences

The most widely used sequencing technique is that developed by Frederick Sanger in 1975. This approach uses dideoxynucleoside triphosphates (ddNTPs) in DNA synthesis. These molecules resemble normal nucleotides except that they lack a 3′-hydroxyl group (**figure 15.1**). They are added to the growing end of the chain, but terminate the synthesis catalyzed by DNA polymerase because more nucleotides cannot be attached to further extend the chain. In the manual sequencing method, a single strand of the DNA to be sequenced is mixed with a primer, DNA polymerase I, four deoxynucleoside triphosphate substrates (one of which is radiolabeled), and a small amount of one of the dideoxynucleotides. DNA synthesis begins with the primer and terminates when a ddNTP is incorporated in place of a regular deoxynucleoside triphosphate. The result is a series of fragments of varying lengths. Four reactions are run, each with a different ddNTP. The mix with ddATP produces fragments with an A terminus; the mix with ddCTP produces fragments with C terminals, and so forth (**figure 15.2**). The radioactive fragments are removed from the DNA template and electrophoresed on a polyacrylamide gel to separate them from one another based on size. Four lanes are electrophoresed, one for each reaction mix, and the gel is autoradiographed (*see p. 322*). A DNA sequence is read directly from the gel, beginning with the smallest fragment or fastest-moving band and moving to the largest fragment or slowest band (figure 15.2*a*). Up to 800 residues can be read from a single gel.

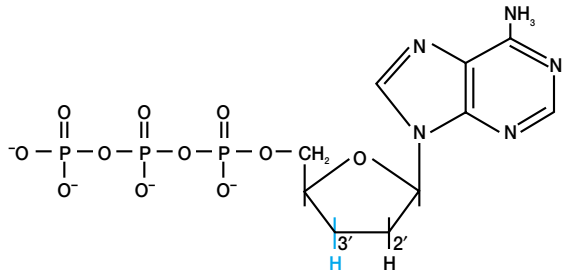


Figure 15.1 Dideoxyadenosine triphosphate (ddATP). Note the lack of a hydroxyl group on the 3′ carbon, which prevents further chain elongation by DNA polymerase.

In automated systems dideoxynucleotides that have been labeled with fluorescent dyes are used (each ddNTP is labeled with a dye of a different color). The products from the four reactions are mixed and electrophoresed together. Because each ddNTP fluoresces with a different color, a detector can scan the gel and rapidly determine the sequence from the order of colors in the bands (figure 15.2*b,c*).

Recently, fully automated capillary electrophoresis sequencers have been developed. These are much faster and allow up to 96 samples to be sequenced simultaneously; it is possible to generate over 350 kilobases of sequences a day. Current systems can sequence strands of DNA around 700 bases long in about 4 hours.

15.3 Whole-Genome Shotgun Sequencing

Although several virus genomes have been sequenced, in the past it has not been possible to sequence the genomes of bacteria. Prior to 1995, whole-genome approaches to sequencing were not

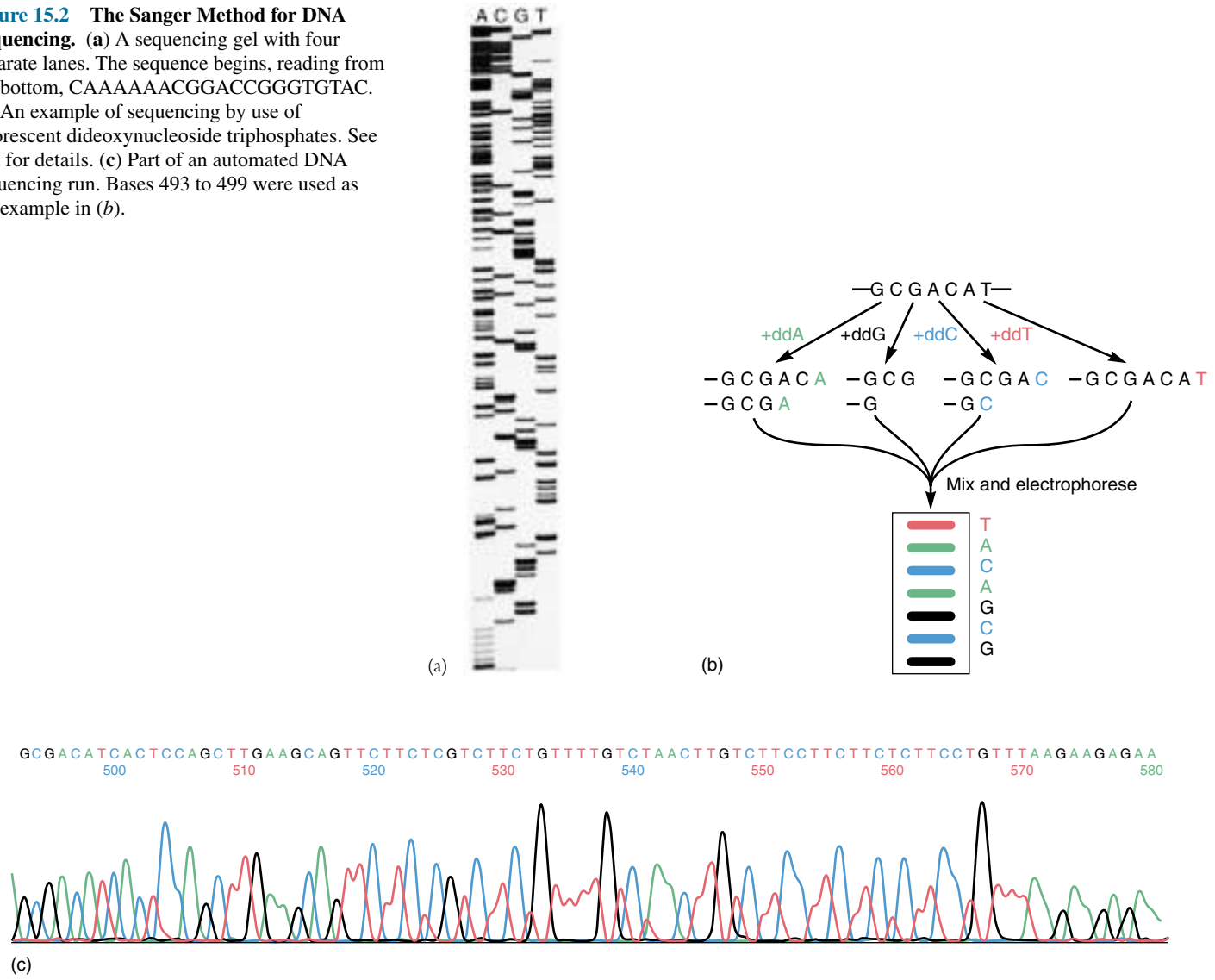


Figure 15.2 The Sanger Method for DNA Sequencing. (**a**) A sequencing gel with four separate lanes. The sequence begins, reading from the bottom, CAAAAAACGGACCGGGTGTAC. (**b**) An example of sequencing by use of fluorescent dideoxynucleoside triphosphates. See text for details. (**c**) Part of an automated DNA sequencing run. Bases 493 to 499 were used as the example in (*b*).

possible because available computational power was insufficient for assembling a genome from thousands of DNA fragments. J. Craig Venter, Hamilton Smith, and their collaborators initially sequenced the genomes of two free-living bacteria, *Haemophilus influenzae* and *Mycoplasma genitalium.* The genome of *H. influenzae,* the first to be sequenced, contains about 1,743 genes in 1,830,137 base pairs and is much larger than a virus genome.

Venter and Smith developed an approach called **whole-genome shotgun sequencing.** The process is fairly complex when considered in detail, and there are many procedures to ensure the accuracy of the results, but the following summary gives a general idea of the approach originally employed by The Institute of Genomic Research (TIGR). For simplicity, this approach may be broken into four stages: library construction, random sequencing, fragment alignment and gap closure, and editing.

1. *Library construction.* The large bacterial chromosomes were randomly broken into fairly small fragments, about the size of a gene or less, using ultrasonic waves; the fragments were then purified (**figure 15.3**). These fragments were attached to plasmid vectors (*see pp. 334–35*), and plasmids with a single insert were isolated. Special *E. coli* strains lacking restriction enzymes were transformed with the plasmids to produce a library of the plasmid clones.
2. *Random sequencing.* After the clones were prepared and the DNA purified, thousands of bacterial DNA fragments were sequenced with automated sequencers, employing special dye-labeled primers. Thousands of templates were used, normally with universal primers that recognized the plasmid DNA sequences just next to the bacterial DNA insert. The nature of the process is such that almost all stretches of genome are sequenced several times, and this increases the accuracy of the final results.
3. *Fragment alignment and gap closure.* Using special computer programs, the sequenced DNA fragments were clustered and assembled into longer stretches of sequence by comparing nucleotide sequence overlaps between fragments. Two

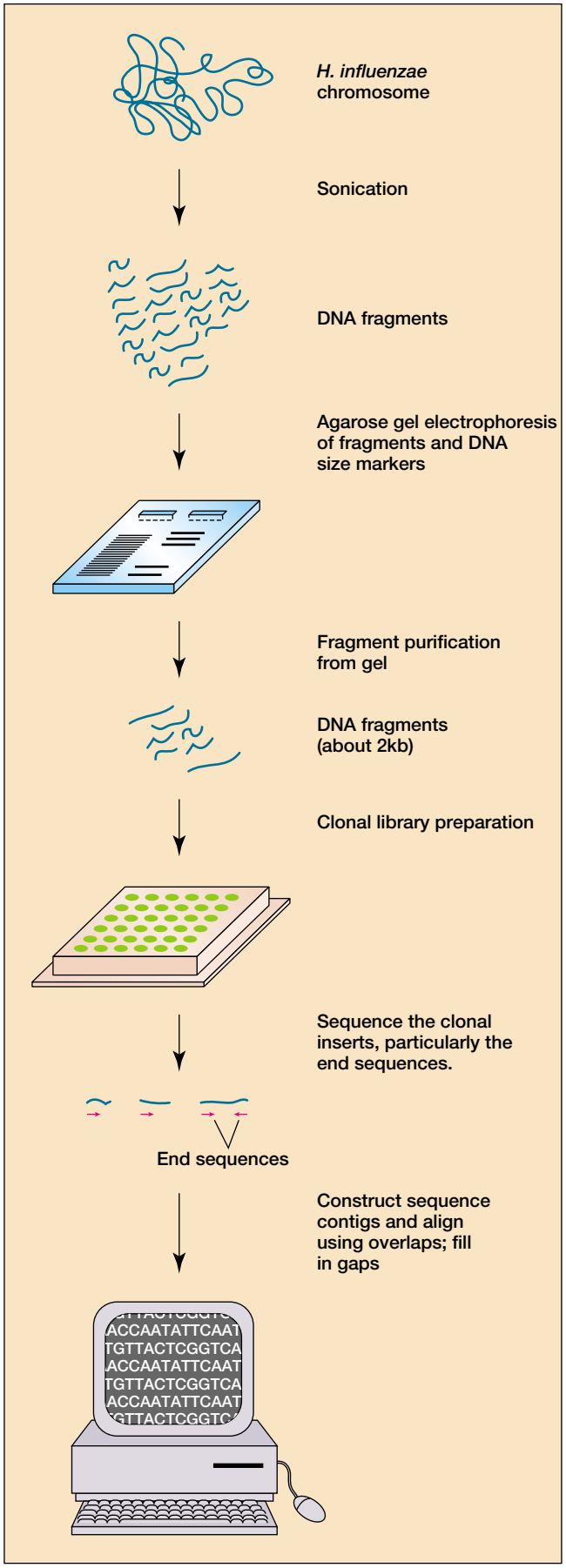


Figure 15.3 Whole-Genome Shotgun Sequencing. This general overview shows how the *Haemophilus influenzae* genome was sequenced. See text for details.

fragments were joined together to form a larger stretch of DNA if the sequences at their ends overlapped and matched (i.e., were the same). This overlap comparison process resulted in a set of larger contiguous nucleotide sequences or contigs.

Finally, the contigs were aligned in the proper order to form the completed genome sequence. If gaps existed between two contigs, sometimes fragment samples with their ends in the two adjacent contigs were available. These fragments could be analyzed and the gaps filled in with their sequences. When this approach was not possible, a variety of other techniques were used to align contigs and fill in gaps. For example, λ phage libraries containing large bacterial DNA fragments were constructed (*see pp. 330–332, 335*). The large fragments in these libraries overlapped the previously sequenced contigs. These fragments were then combined with oligonucleotide probes that matched the ends of the contigs to be aligned. If the probes bound to a λ library fragment, it could be used to prepare a stretch of DNA that represented the gap region. Overlaps in the sequence of this new fragment with two contigs would allow them to be placed side-by-side and fill in the gap between them.

4. *Editing.* The sequence was then carefully proofread in order to resolve any ambiguities in the sequence. Also the sequence was checked for unwanted frameshift mutations and corrected if necessary.

The approach worked so well that it took less than 4 months to sequence the *M. genitalium* genome (about 500,000 base pairs in size). The shotgun technique also has been used successfully by Celera Genomics in the Human Genome Project and to sequence the *Drosophila* genome.

Once the genome sequence has been established, the process of **annotation** begins. The goal of annotation is to determine the location of specific genes in the genome map. Every **open reading frame (ORF)**—a reading frame sequence (*see p. 241*) not interrupted by a stop codon—larger than 100 codons is considered to be a potential protein coding sequence. Computer programs are used to compare the sequence of the predicted ORF against large databases containing nucleotide and amino acid sequences of known enzymes and other proteins. If a bacterial sequence matches one in the database, it is assumed to code for the same protein. Although this comparison process is not without errors, it can provide tentative function assignments for about 40 to 50% of the presumed coding regions. It also gives some information about transposable elements, operons, repeat sequences, the presence of various metabolic pathways, and other genome features. The results of genome sequencing and annotation for *Mycoplasma genitalium* and *Haemophilus influenzae* are shown in figures 15.5 and 15.6. Often the results of annotation are expressed in a diagram that summarizes the known metabolism and physiology of the organism. An example of this is given in figure 15.7.

1. Define genomics. What are the three general areas into which it can be divided.
2. Describe the Sanger method for DNA sequencing.
3. Outline the whole-genome shotgun sequencing method. What is annotation and how is it carried out?

15.4 Bioinformatics

DNA sequencing techniques have developed so rapidly that an enormous amount of data has already accumulated and genomes are being sequenced at an ever-increasing pace. The only way to organize and analyze all these data is through the use of computers, and this has led to the development of a new interdisciplinary field that combines biology, mathematics, and computer science. **Bioinformatics** is the field concerned with the management and analysis of biological data using computers. In the context of genomics, it focuses on DNA and protein sequences. The annotation process just described is one aspect of bioinformatics. DNA sequence data is stored in large databases. One of the largest genome databases is the International Nucleic Acid Sequence Data Library, often referred to as GenBank. Databases can be searched with special computer programs to find homologous sequences, DNA sequences that are similar to the one being studied. Protein coding regions also can be translated into amino acid sequences and then compared. These sequence comparisons can suggest functions of the newly discovered genes and proteins. The gene under study often will have a function similar to that of genes with homologous DNA or amino acid sequences.

15.5 General Characteristics of Microbial Genomes

The development of shotgun sequencing and other genome sequencing techniques has led to the characterization of many procaryotic genomes in a very short time. Many genome sequences of procaryotes have been completed and published, and some of these are given in **table 15.1.** These procaryotes represent great phylogenetic diversity (**figure 15.4**). At least 100 more procaryotes, many of them major human pathogens, are being sequenced at present. Comparison of the genomes from different procaryotes will contribute significantly to the understanding of procaryotic evolution and help deduce which genes are responsible for various cellular processes. Genome sequences will aid in our understanding of genetic regulation and genome organization. In some cases, such information will also aid in the search for human genes by the Human Genome Project because of the similarities between procaryotic and human biochemistry.

Currently published genome sequences already have provided new and important insights into genome organization and function. *Mycoplasma genitalium* grows in human genital and respiratory tracts and has a genome of only 580 kilobases in size, one of the smallest genomes of any free-living organism (**figure 15.5**). Thus the sequence data are of great interest because they help establish the minimal set of genes needed for a free-living existence. There appear to be approximately 517 genes (480 protein-encoding genes and 37 genes for RNA species). About 90 proteins are involved in translation, and only around 29 proteins for DNA replication. Interestingly, 140 genes, or 29% of those in the genome, code for membrane proteins, and up to 4.5% of the genes seem to be involved in evasion of host immune responses. Only 5 genes have regulatory

Table 15.1 Examples of Complete Published Microbial Genomes

Genome	Domain[a]	Size (Mb)	% G + C
Aquifex aeolicus	B	1.50	43
Archaeoglobus fulgidus	A	2.18	48
Bacillus subtilis	B	4.20	43
Borrelia burgdorferi	B	1.44	28
Campylobacter jejuni	B	1.64	31
Chlamydia pneumoniae	B	1.23	40
Chlamydia trachomatis	B	1.05	41
Deinococcus radiodurans	B	3.28	67
Escherichia coli	B	4.60	50
Haemophilus influenzae Rd	B	1.83	39
Helicobacter pylori	B	1.66	39
Methanobacterium thermoautotrophicum	A	1.75	49
Methanococcus jannaschii	A	1.66	31
Mycobacterium tuberculosis	B	4.40	65
Mycoplasma genitalium	B	0.58	31
Mycoplasma pneumoniae	B	0.81	40
Neisseria meningitidis	B	2.27	51
Pseudomonas aeruginosa	B	6.3	67
Pyrococcus horiksohii	A	1.80	42
Rickettsia prowazekii	B	1.10	29
Saccharomyces cerevisiae	E	13	38
Synechocystis sp.	B	3.57	47
Thermotoga maritima	B	1.80	46
Treponema pallidum	B	1.14	52
Vibrio cholerae	B	4.0	48

[a]The following abbreviations are used: A, Archaea; B, Bacteria; E, Eucarya.

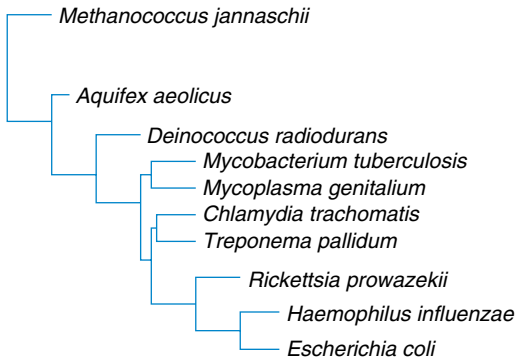


Figure 15.4 Phylogenetic Relationships of Some Procaryotes with Sequenced Genomes. These procaryotes are discussed in the text. *Methanococcus jannaschii* is in the domain *Archaea,* the rest are members of the domain *Bacteria.* Genomes from a broad diversity of procaryotes have been sequenced and compared. *Source: The Ribosomal Database Project.*

Figure 15.5 **Map of the *Mycoplasma genitalium* genome.** The predicted coding regions are shown with the direction of transcription indicated by arrows. The genes are color coded by their functional role. The rRNA operon, tRNA genes, and adhesin protein operons (MgPa) are indicated. *Reprinted with permission from Fraser, C. M., et al. Copyright 1995. The minimal gene complement of* Mycoplasma genitalium. Science *270:397–403. Figure 1, page 398 and The Institute for Genomic Research.*

functions. Even in this smallest genome, 22% of the genes do not match any known protein sequence. Comparison with the *M. pneumoniae* genome and studies of gene inactivation by transposon insertion suggests that about 108 to 121 *M. genitalium* genes may not be essential for survival. Thus the minimum gene set required for laboratory growth conditions seems to be approximately 265 to 350 genes; about 100 of these have unknown functions. *Haemophilus influenzae* has a much larger genome, 1.8 megabases and 1,743 genes (**figure 15.6**). More than 40% of the genes have unknown functions. It has already been found that the bacterium lacks three Krebs cycle genes and thus a functional cycle. It does devote many more genes (64 genes) to regulatory functions than does *M. genitalium. Haemophilus influenzae* is a species capable of transformation (*see pp. 305–7*). The process must be very important to this bacterium because it contains 1,465 copies of the recognition sequence used in DNA uptake during transformation. *Methanococcus jannaschii,* a member of the Archaea, also has been sequenced. Only 44% of its 1,738 genes match those of other organisms, an indication of how different this archaeon is from bacteria and eucaryotes. Despite this profound difference, many of its genes for DNA replication, transcription, and translation are similar to eucaryotic genes and quite different from bacterial genes. However, the metabolism of *M. jannaschii* is more similar to that of bacteria than to eucaryotic metabolism. More recently the sequence of the 4.6 megabase *Escherichia coli* K12 genome has been published. About 5 to 6% of the genes code for proteins involved in cell and membrane structure; 12 to 14% for transport proteins; 10% for the enzymes of energy and central intermediary metabolic pathways; 4% for regulatory genes; and 8% for replication, transcription, and translation proteins. The genome contains about 4,288 predicted genes, almost 2,500 of which do not resemble known genes. The large number of unknown genes in *Escherichia coli, Haemophilus influenzae,* and other procaryotes has great significance. It shows how little we know about microbial biology. Clearly there is much more to learn about the genetics, physiology, and metabolism of procaryotes, even of those that have been intensively studied.

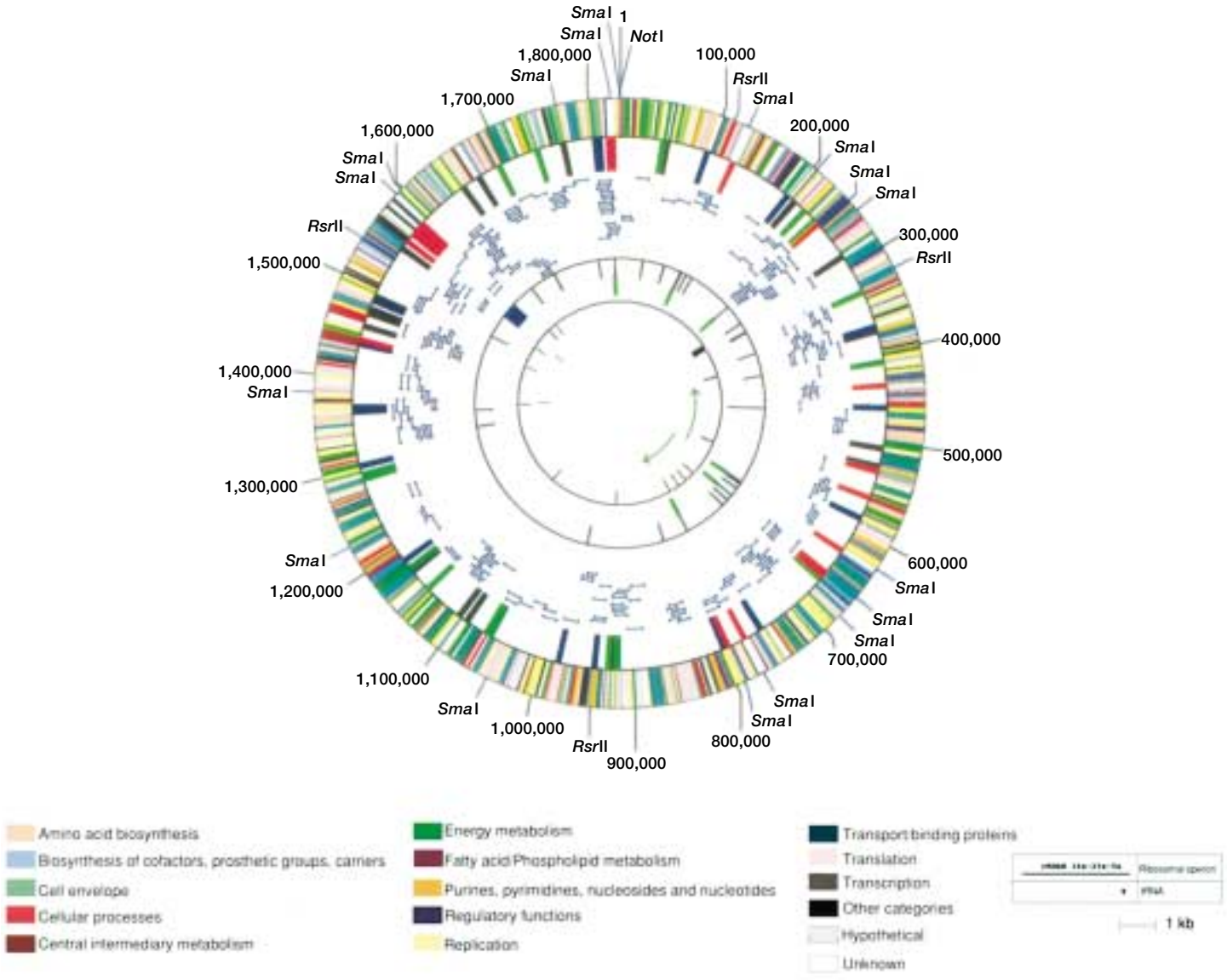


Figure 15.6 **Map of the *Haemophilus influenzae* genome.** The predicted coding regions in the outer concentric circle are indicated with colors representing their functional roles. The outer perimeter shows the *NotI, RsrII,* and *SmaI* restriction sites. The inner concentric circle shows regions of high G + C content (red and blue) and high A + T content (black and green). The third circle shows the coverage by λ clones (blue). The fourth circle shows the locations of rRNA operons (green), tRNAs (black), and the mu-like prophage (blue). The fifth circle shows simple tandem repeats and the probable origin of replication (outward pointing green arrows). The red lines are potential termination sequences. *Reprinted with permission from Fleischman, R. D., et al. 1995. Whole-genome random sequencing and assembly of* Haemophilus influenzae *Rd.* Science *269:496–512. Figure 1, page 507 and The Institute of Genomic Research.*

Comparison of these and other genome sequences shows large differences. Not surprisingly, *E. coli* is most similar to *H. influenzae,* which is also a member of the γ-proteobacteria (1,130 similar genes). It differs more from the cyanobacterium *Synechocystis* sp. PCC6803 (675 similarities) and *Mycoplasma genitalium* (468 similarities). These four bacteria have only 111 proteins in common. *Escherichia coli* is even more unlike the archaeon *M. jannaschii* (231 similar genes) and the eucaryotic yeast *Saccharomyces cerevisiae* (254 similar genes). Only 16 proteins, mostly translation proteins such as ribosomal proteins and aminoacyl synthetases, are essentially the same in all six organisms. There have been many gene losses and changes during the course of evolution.

Many of the genomes already sequenced belong to procaryotes that are either major human pathogens or of particular biological interest. Some recent sequences have yielded interesting discoveries and will be briefly discussed as examples of the kind of important information that can be obtained from genomics. Because of their practical importance, our focus will be primarily on human pathogens. As we will see, the results often pose more new questions than they answer old ones and open up many new areas of research.

The deinococci are soil bacteria of great interest because of their ability to survive a dose of radiation thousands of times greater than the amount needed to kill humans. They survive by stitching together their splintered chromosomes after radiation exposure. The genome consists of two circular chromosomes of different size (2.6 Mb and 0.4 Mb), a megaplasmid (177,466 bp), and a regular plasmid (45,704 bp). One would think that this genome should have some quite different DNA repair genes. However, despite its remarkable resistance to radiation, *Deinococcus radiodurans* has the same array of DNA repair mechanisms as other bacteria. It differs in simply having more of them. For example, most organisms have one MutT gene, which is involved in disposing of oxidized nucleotides; *Deinococcus* has 20 MutT-like genes. The genome also possesses many repeat sequences, which may be important in the repair process. It should be emphasized that many of the bacterium's genes have unknown functions, and some of these genes may aid in its unusually great resistance to radiation. The deinococci (p. 468)

Rickettsia prowazekii is a member of the α-proteobacteria that is an obligate intracellular parasite of lice and humans. It is the causative agent of typhus fever and killed millions during and following the First and Second World Wars. Many microbiologists think that mitochondria may have arisen when a member of the α-proteobacteria established an endosymbiotic relationship with the ancestral eucaryotic cell (*see pp. 424–25*). The sequenced genome of *R. prowazekii* is consistent with this hypothesis. Its protein-encoding genes show similarities to mitochondrial genes. Glycolysis is absent, but genes for the TCA cycle and electron transport are present, and ATP synthesis is similar to that in mitochondria. Both *Rickettsia* and the mitochondrion lack many genes for the biosynthesis of amino acids and nucleosides, in contrast with the situation in free-living α-proteobacteria. Thus aerobic respiration in eucaryotes may have arisen from an ancestor of *Rickettsia.* *Rickettsia* biology and clinical aspects (pp. 488–90, 909–10)

Chlamydiae are nonmotile, coccoid, gram-negative bacteria that reproduce only within cytoplasmic vesicles of eucaryotic cells by a unique life cycle. *Chlamydia trachomatis* infects humans and causes the sexually transmitted disease, nongonococcal urethritis, probably the most commonly transmitted sexual disease in the United States. It also is the leading cause of preventable blindness in the world. The sequencing of its genome has revealed several surprises. The bacteria's life cycle is so unusual (*see pp. 477–78*) that one would expect its genome to be somewhat atypical. This has turned out not to be the case; the genome is similar to that of many other bacteria. Microbiologists have called *Chlamydia* an "energy parasite" and believed that it obtained all its ATP from the host cell. The genome results show that *Chlamydia* has the genes to make at least some ATP on its own, although it also has genes for ATP transporters. Another surprise is the presence of enzymes for the synthesis of peptidoglycan. Chlamydial cell walls lack peptidoglycan and microbiologists have been unable to explain why the antibiotic penicillin, which disrupts peptidoglycan synthesis, is able to inhibit chlamydial growth. The presence of peptidoglycan biosynthetic enzymes helps account for the penicillin effect, but no one knows the purpose of peptidoglycan synthesis in this bacterium. Another major surprise is the absence of the *FtsZ* gene, which is thought to be required by all bacteria and archaea for septum formation during cell division (*see p. 286*). The absence of this supposedly essential gene makes one wonder how *Chlamydia* divides. It may be that some of the genes with unknown functions play a major role in cell division. Perhaps *Chlamydia* employs a mechanism of cell division different from that of other procaryotes. Finally, the genome contains at least 20 genes that have been obtained from eucaryotic host cells (most bacteria have no more than 3 or 4 such genes). Some of these genes are plantlike; originally *Chlamydia* may have infected a plantlike host and then moved to animals. *Chlamydia* (pp. 477–78; section 39.3)

One of the most difficult human pathogens to study has been the causative agent of syphilis, *Treponema pallidum.* This is because it has not been possible to grow *Treponema* outside the human body. We know little about its metabolism or the way it avoids host defenses, and no vaccine for *Treponema* has yet been developed. Naturally the sequencing of the *Treponema pallidum* genome has generated considerable excitement and hope. It turns out that *Treponema* is metabolically crippled. It can use carbohydrates as an energy source, but lacks the TCA cycle and oxidative phosphorylation (**figure 15.7**). *Treponema* also lacks many biosynthetic pathways (e.g., for enzyme cofactors, fatty acids, nucleotides, and some electron transport proteins) and must rely on molecules supplied by its host. In fact, about 5% of its genes code for transport proteins. Given the lack of several critical pathways, it is not surprising that the pathogen has not been cultured successfully. The genes for surface proteins are of particular interest. *Treponema* has a family of surface protein genes characterized by many repetitive sequences. Some have speculated that these genes might undergo recombination in order to generate new surface proteins and allow the organism to avoid attack by the immune system, but this is not certain. However, it may be possible to develop a vaccine for syphilis using some of the newly discovered surface proteins. We also may be able to identify strains of *Treponema* using these surface proteins, which would be of great importance in syphilis epidemiology. The genome results have not provided much of a clue about how *Treponema* causes syphilis. About 40% of the genes have unknown functions. Possibly some of them are responsible for avoiding host defenses and for the production of toxins and other virulence factors. *Treponema* and syphilis (pp. 479–81, 923–24)

For centuries, tuberculosis has been one of the major scourges of humankind and still kills about 3 million people annually. Furthermore, because of the spread of AIDS and noncompliance in drug treatment, *Mycobacterium tuberculosis* is increasing in frequency once again and is becoming ever more drug resistant. Anything that can be learned from genome studies could be of great importance in the fight to control the renewed spread of tuberculosis. The *Mycobacterium tuberculosis* genome is one of the largest yet found (4.40 Mb), exceeded only by *E. coli* (4.60 Mb genome) and *Pseudomonas aeruginosa* (6.26 Mb), and contains around 4,000 genes. Only about 40% of the genes have been given precise functions and 16% of its genes resemble no known proteins; presumably they are responsible for specific mycobacterial functions. More than 250 genes are devoted to lipid metabolism (*E. coli* has only about 50 such genes), and *M. tuberculosis* may obtain much of its energy by degrading host lipids. There are a surprisingly large number of regulatory elements in the genome. This may mean that the infection process is much more complex and

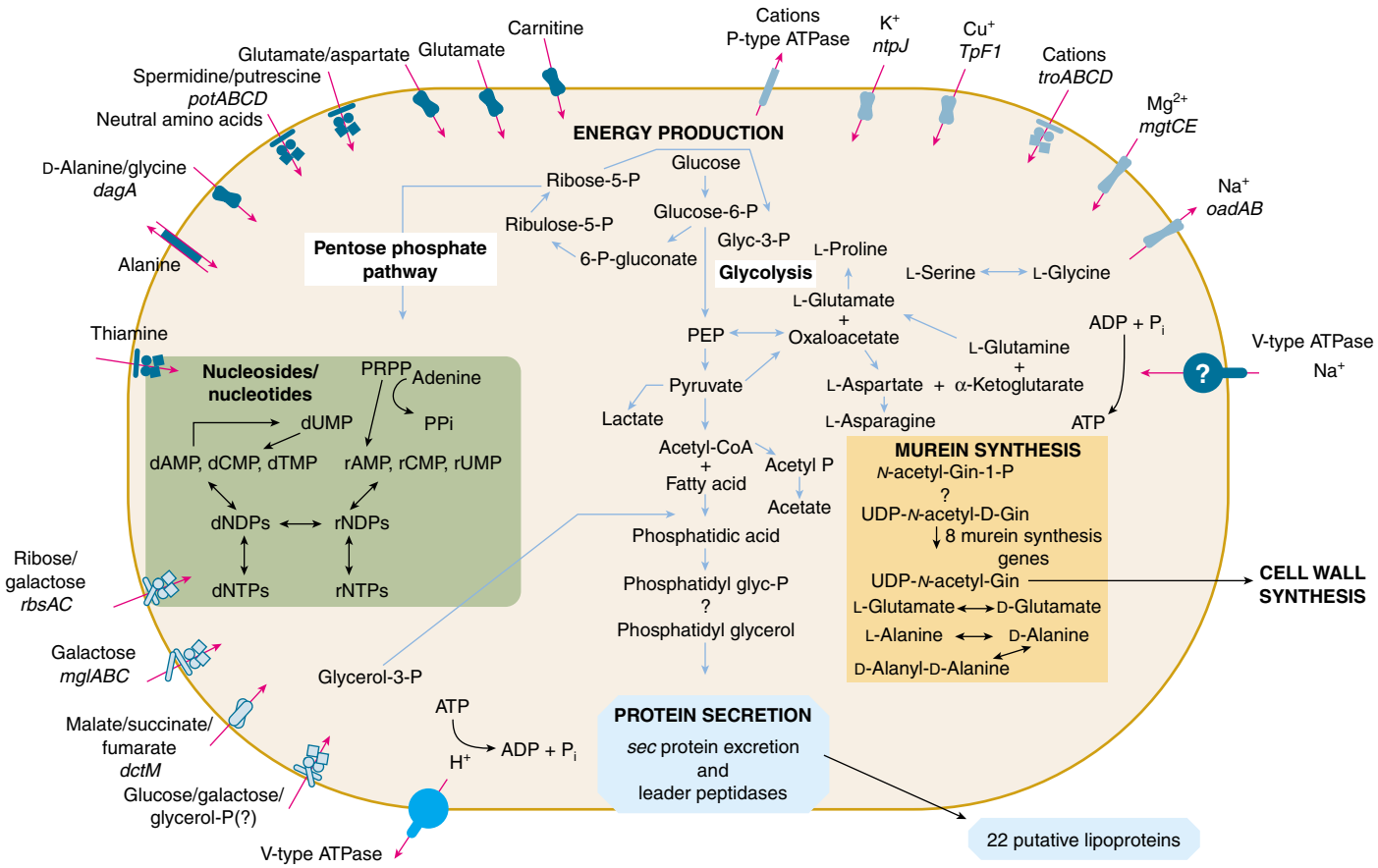


Figure 15.7 Metabolic Pathways and Transport Systems of *Treponema pallidum*. This depicts *T. pallidum* metabolism as deduced from genome annotation. Note the limited biosynthetic capabilities and extensive array of transporters. Although glycolysis is present, the TCA cycle and respiratory electron transport are lacking. Question marks indicate where uncertainties exist or expected activities have not been found.

sophisticated than previously thought. Two families of novel glycine-rich proteins with unknown functions are present and represent about 10% of the genome. They may be a source of antigenic variation and involved in defense against the host immune system. One major medical problem has been the lack of a good vaccine. A large number of proteins that are either secreted by the bacterium or on the bacterial surface have been identified from the genome sequence. It is hoped that some of these proteins can be used to develop new, effective vaccines. This is particularly important in view of the spread of multiply drug resistant *M. tuberculosis.* *Mycobacterium tuberculosis* (pp. 543–44, 906–8)

It is tempting to think that closely related and superficially similar bacteria must have similar genomes. Although the genome of the leprosy bacillus, *Mycobacterium leprae,* has only been about 90% sequenced, it is already clear that this assumption can be mistaken. The whole *M. leprae* genome is a third smaller than that of *M. tuberculosis.* About half the genome seems to be devoid of functional genes; it consists of junk DNA that represents over 1,000 degraded, nonfunctional genes. In total, *M. leprae* seems to have lost as many as 2,000 genes during its career as an intracellular parasite. It even lacks some of the enzymes required for energy production and DNA replication. This might explain why the bacterium has such a long doubling time, about two weeks in mice. One hope from the genomics study is that critical surface proteins can be discovered and used to develop a sensitive test for early detection of leprosy. This would allow immediate treatment of the disease before nerve damage occurs. Leprosy (pp. 916–17)

Analysis and comparison of the genomes already sequenced have disclosed some general patterns in genome organization. Although protein sequences are usually conserved (i.e., about 70% of proteins contain ancient conserved regions), genome organization is quite variable in the Bacteria and Archaea. Sometimes two genes can fuse to form a new gene that has a combination of the functions possessed by the two separate genes. Less often, a gene can split or undergo fission; this seems to be more prevalent in thermophilic procaryotes. There also appears to be considerable horizontal gene transfer, particularly of housekeeping or operational genes. Informational genes, primarily those essential to transcription and translation, are not transferred as often. Perhaps, as James Lake has proposed, genes whose protein products are parts of large, complex systems and interact with many other molecules are not often transferred successfully. About 18% of the genes in *E. coli* seem to have been acquired by horizontal transfer after its divergence from *Salmonella.* There also is gene transfer

Table 15.2 Estimated Number of Genes Involved in Various Cell Functions[a]

Gene Function	*Escherichia coli* K12	*Bacillus subtilis*	*Mycoplasma genitalium*	*Treponema pallidum*	*Rickettsia prowazekii*	*Chlamydia trachomatis*	*Mycobacterium tuberculosis*	*Methanococcus jannaschii*	*Pyrococcus abyssi*
Approximate total number of genes[b]	2,933	2,232	477	757	523	847	2,095	1,271	1,345
Cellular processes[c]	179	123	6	77	27	43	65	26	44
Cell envelope components	146	86	29	53	36	42	50	25	25
Transport and binding proteins	304	223	33	59	18	57	87	56	67
DNA metabolism	97	80	29	51	39	53	57	53	33
Transcription	38	45	13	25	23	23	26	21	19
Protein synthesis	121	105	90	97	87	100	90	117	99
Regulatory functions	159	163	5	22	6	15	77	18	19
Energy metabolism[d]	351	230	33	54	48	61	211	158	116
Central intermediary metabolism[e]	64	61	7	6	6	12	57	18	25
Amino acid biosynthesis	89	97	0	7	9	13	72	64	51
Fatty acid and phospholipid metabolism	67	53	8	11	11	25	78	9	8
Purines, pyrimidines, nucleosides, and nucleotides	75	68	19	21	12	15	48	37	40
Biosynthesis of cofactors and prosthetic groups	97	79	4	15	17	31	84	49	31

[a]Data adapted from TIGR (The Institute for Genomic Research) databases.

[b]The number of genes with known or hypothetical functions.

[c]Genes involved in cell division, chemotaxis and motility, detoxification, transformation, toxin production and resistance, pathogenesis, adaptations to atypical conditions, etc.

[d]Genes involved in amino acid and sugar catabolism, polysaccharide degradation and biosynthesis, electron transport and oxidative phosphorylation, fermentation, glycolysis/gluconeogenesis, pentose phosphate pathway, Entner-Doudoroff, pyruvate dehydrogenase, TCA cycle, photosynthesis, chemoautotrophy, etc.

[e]Amino sugars, phosphorus compounds, polyamine biosynthesis, sulfur metabolism, nitrogen fixation, nitrogen metabolism, etc.

between domains. The bacterium *Aquifex aeolicus* probably received about 16% of its genes from the Archaea, and 24% of the genes in *Thermotoga maritima* are similar to archaeal sequences. Some microbiologists have proposed that new species are created by the acquisition of genes that allow exploitation of a new ecological niche. For example, *E. coli* may have acquired the lactose operon and thus become able to metabolize the milk sugar lactose. This capacity would aid in colonization of the mammalian colon. Existence of extensive gene transfer between species and domains may require reevaluation of the bacterial taxonomic schemes that are based only on rRNA sequences (*see sections 19.6 and 19.7*). The comparison of many more genome sequences may clarify these phylogenetic relationships. Horizontal gene transfer (section 13.1)

1. What sorts of general insights have been provided by the analysis of the genomes of *M. genitalium, H. influenzae, M. jannaschii,* and *E. coli?*
2. The genomes of *D. radiodurans, R. prowazekii, C. trachomatis, T. pallidum, M. tuberculosis,* and *M. leprae* have been briefly discussed. Give one or two surprises or interesting insights that have arisen from each genome sequence.
3. Discuss what has been learned about horizontal gene transfer from genome comparisons.

15.6 Functional Genomics

Clearly, determination of genome sequences is only the start of genome research. It will take years to learn how the genome actually functions in a cell or organism (if that is completely possible) and to apply this knowledge in practical ways such as the conquest of disease and increased crop production. Sometimes the study of genome function and the practical application of this knowledge is referred to as postgenomics because it builds upon genome sequencing data. Functional genomics is a major postgenomics discipline. As mentioned earlier, functional genomics is concerned with learning how the genome operates. We will consider a few of the many approaches used to study genome function. First we will discuss annotation, which has already been introduced in the context of genome sequencing. Then techniques for the study of RNA- and protein-level expression will be described.

Genome Annotation

After sequencing, annotation can be used to tentatively identify many genes and this allows analysis of the kinds of genes and functions present in the microorganism (figure 15.7). **Table 15.2** summarizes some of the data for several important procaryotic genomes, seven bacterial and two archaeal (*Methanococcus* and

Pyrococcus). Even with these few examples, patterns can be seen. Genes responsible for essential informational functions (DNA metabolism, transcription, and protein synthesis) do not vary in number as much as other genes. There seems to be a minimum number of these essential genes necessary for life. Second, complex free-living bacteria such as *E. coli* and *B. subtilis* have many more operational or housekeeping genes than do most of the parasitic forms, which depend on the host for a variety of nutrients. Generally, larger genomes show more metabolic diversity. Parasitic bacteria derive many nutrients from their hosts and can shed genes for unnecessary pathways; thus they have smaller genomes.

Evaluation of RNA-Level Gene Expression

One of the best ways to evaluate gene expression is through the use of **DNA microarrays (DNA chips).** These are solid supports, typically of glass or silicon and about the size of a microscope slide, that have DNA attached in highly organized arrays. The chips can be constructed in several ways. In one approach, a programmable robotic machine delivers hundreds to thousands of microscopic droplets of DNA samples to specific positions on a chip using tiny pins to apply the solution (*see figure 42.26, p. 1020.*) The spots are then dried and treated in order to bind the DNA tightly to the surface. Any DNA fragment can be attached in this way; often cDNA (*see p. 321*) about 500 to 5,000 bases long is used. A second procedure involves the synthesis of oligonucleotides directly on the chip in the following way (**figure 15.8**):

1. Coat the glass support with light-sensitive protecting groups that prevent random nucleoside attachment.
2. Cover the surface with a mask that has holes corresponding to the sites for attachment of the desired nucleosides.
3. Shine laser light through the mask holes to remove the exposed protecting groups.
4. Bathe the chip in a solution containing the first nucleoside to be attached. The nucleoside will chemically couple to the light-activated sites. Each nucleoside has a light-removable protecting group to prevent addition of another nucleoside until the appropriate time.
5. Repeat steps 2 through 4 with a new mask each time to add nucleosides until all sequences on the chip have been completed.

This procedure can be used to construct any sequence. The commercial chip contains oligonucleotide probes that are 25 bases long. It is about 1.3 cm on a side and can have over 200,000 addressable positions (**figure 15.9**). The probes are often expressed sequence tags. An ***e*xpressed *s*equence *t*ag (EST)** is a partial gene sequence unique to the gene in question that can be used to identify and position the gene during genomic analysis. It is derived from cDNA molecules. There are now chips that have probes for every expressed gene or open reading frame in the genome of

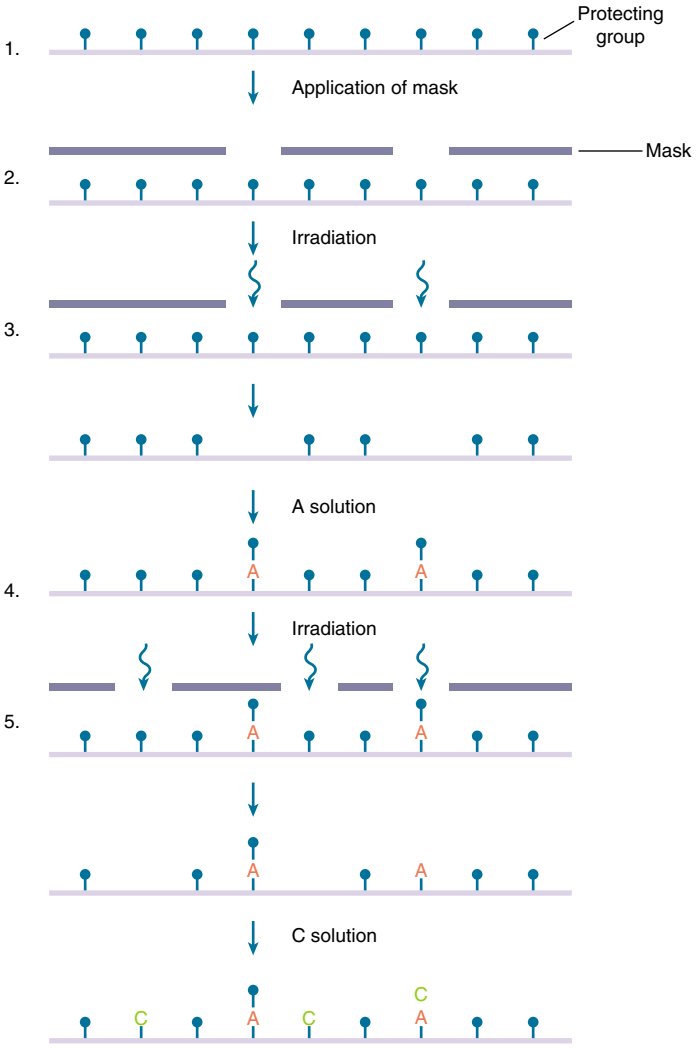


Figure 15.8 Construction of a DNA Chip with Attached Oligonucleotide Sequences. Only two cycles of synthesis are shown. See text for description of the steps.

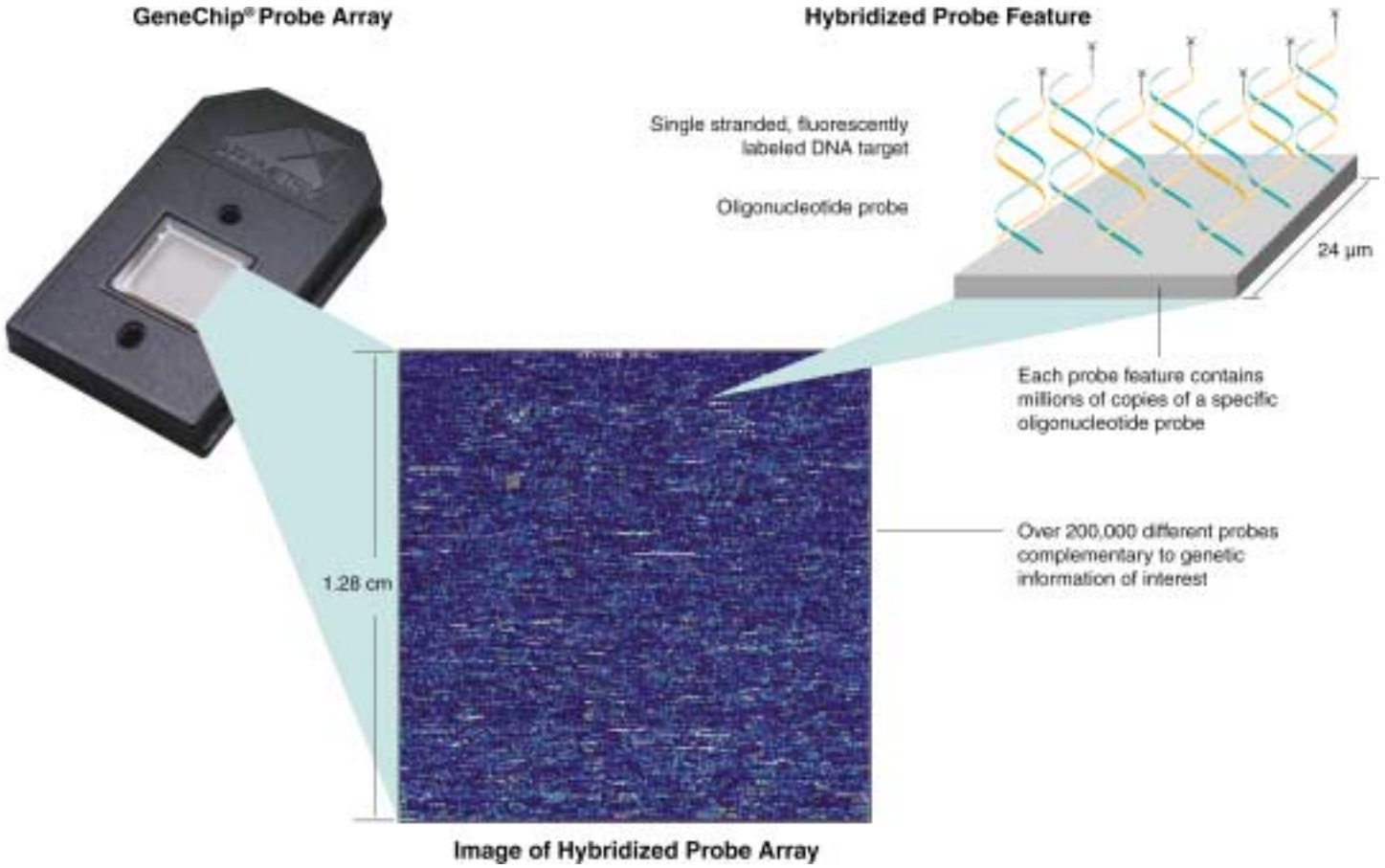


Figure 15.9 **The GeneChip Expression Probe Array.** The DNA chip manufactured by Affymetrix, Inc. contains probes designed to represent thousands or tens of thousands of genes.

E. coli (about 4,200 open reading frames) and the yeast *Saccharomyces cerevisiae* (approximately 6,100 open reading frames).

The nucleic acids to be analyzed, often called the targets, are isolated and labeled with fluorescent reporter groups. Nucleic acid targets may be mRNA or cDNA produced from mRNA by reverse transcription (*see figure 14.3*). The chip is incubated with the target mixture long enough to ensure proper binding to probes with complementary sequences. The unbound target is washed off and the chip is then scanned with laser beams. Fluorescence at an address indicates that the probe is bound to that particular sequence. Analysis of the hybridization pattern shows which genes are being expressed. Target samples from two experiments can be labeled with different fluorescent groups and compared using the same chip. Figure 15.9 and the chapter opening figure provide examples of fluorescently labeled DNA microarrays.

DNA chip results allow one to observe the characteristic expression of whole sets of genes during differentiation or in response to environmental changes. In some cases, many genes change expression in response to a single change in conditions. Patterns of gene expression can be detected and functions can be tentatively assigned based on expression. If an unknown gene is

expressed under the same conditions as genes of known function, it is coregulated and quite likely shares the same general function. DNA chips also can be used to study regulatory genes directly by perturbing a regulatory gene and observing the effect on genome activity. Of course, only mRNAs that are currently expressed can be detected. If a gene is transiently expressed, its activity may be missed by a DNA chip analysis.

Evaluation of Protein-Level Gene Expression

Genome function can be studied at the translation level as well as the transcription level. The entire collection of proteins that an organism produces is called its **proteome.** Thus **proteomics** is the study of the proteome or the array of proteins an organism can produce. It is an essential discipline because proteomics provides information about genome function that mRNA studies cannot. There is not always a direct correlation between mRNA and protein levels because of the posttranslational modification of proteins and protein turnover. Measurement of mRNA levels can show the dynamics of gene expression and tell what might occur in the cell, whereas proteomics discovers what is actually happening.

Although new techniques in proteomics are currently being developed, we will focus briefly only on the traditional approach. A mixture of proteins is separated using two-dimensional electrophoresis. The first dimension makes use of isoelectric focusing, in which proteins move electrophoretically through a pH gradient (e.g., pH 3 to 10 or 4 to 7). The protein mixture is applied to a strip with an immobilized pH gradient and electrophoresed. Each protein moves along the strip until the pH on the strip equals its isoelectric point. At this point, the protein's net charge is zero and the protein stops moving. Thus the technique separates the proteins based on their content of ionizable amino acids. The second dimension is SDS *p*oly*a*crylamide *g*el *e*lectrophoresis (SDS-PAGE). SDS (sodium dodecyl sulfate) is an anionic detergent that denatures proteins and coats the polypeptides with a negative charge. After the first electrophoretic run has been completed, the pH gradient strip is soaked in SDS buffer and then placed at the edge of an SDS-PAGE gel sheet. Then a voltage is applied at right angles to the strip at the edge of the sheet. Under these circumstances, polypeptides migrate through the polyacrylamide gel at a rate inversely proportional to their masses. That is, the smallest polypeptide will move the farthest in a particular length of time. This two-dimensional technique is very effective at separating proteins and can resolve thousands of proteins in a mixture (**figure 15.10**). If radiolabeled substrates are used, newly synthesized proteins can be distinguished and their rates of synthesis determined. Gel electrophoresis (pp. 327–28)

Two-dimensional electrophoresis is even more powerful when coupled with mass spectrometry. The unknown protein spot is cut from the gel and cleaved into fragments by trypsin digestion. Then fragments are analyzed by a mass spectrometer and the mass of the fragments is plotted. This mass fingerprint can be used to estimate the probable amino acid composition of each fragment and tentatively identify the protein. When the two techniques are employed together, the proteome and its changes can be studied very effectively.

Proteomics has been used to study the physiology of *E. coli.* Some areas of research have been the effect of phosphate limitation, proteome changes under anaerobic conditions, heat-shock protein production, and the response to the toxicant 2,4-dinitrophenol. One particularly useful approach in studying genome function is to inactivate a specific gene and then look for changes in protein expression. Because changes in the whole proteome are followed, gene inactivation can tell much about gene function and the large-scale effects of gene activity. A gene-protein database for *E. coli* has been established and provides information about the conditions under which each protein is expressed and where it is located in the cell.

The preceding discussion of functional genomics has emphasized areas of investigation with a record of success and a bright future. However it should be noted that many problems remain to be solved, and there may be limits to how much genomics can tell us about the living cell for a variety of reasons. For example, sequence information does not specify the nature and timing of gene regulation. Regulation of protein activity in living cells is extraordinarily complex and involves regulatory networks, which we do not yet understand completely. Functional assignments from annotation and other approaches sometimes may be inadequate because the function of a gene product often depends on its cellular context. Cells are extremely complex structural entities permeated by various physical compartments in which many processes are restricted to surfaces of membranes and macromolecular complexes (*see p. 165*). Thus localization of proteins also affects function, and genomics cannot account for this. These and other problems should be kept in mind when thinking about future progress in genomics.

1. What general lessons about genome function have been learned from the annotation results?
2. How are DNA microarrays or chips constructed and used to analyze gene expression? What sorts of things can be learned by this approach?
3. Describe two-dimensional electrophoresis and how it is used in the study of proteomics. What kinds of studies can be carried out with this technique?

15.7 The Future of Genomics

Although much has been accomplished in the past few years, the field of genomics is just beginning to mature. There are challenges ahead and many ways in which genomics can advance our knowledge of microorganisms and their practical uses. A few of these challenges and opportunities are outlined here.

1. We need to develop new methods for the large-scale analysis of genes and proteins so that more organisms can be studied.
2. All the new information about DNA and protein sequences, variations in mRNA and protein levels, and protein interactions must be integrated in order to understand

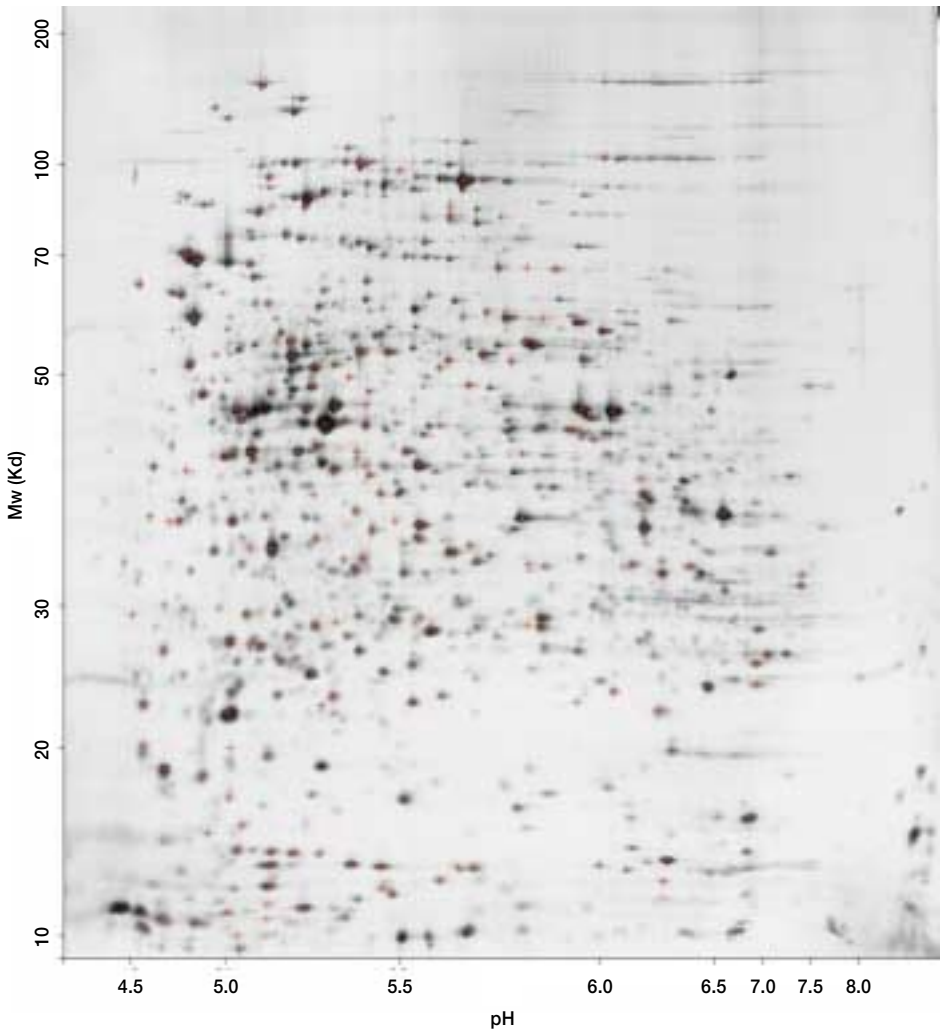


Figure 15.10 **Two-Dimensional Electrophoresis of Proteins.** The SWISS-2D PAGE map of *E. coli* K12 proteins. The first dimension pH gradient ran from pH 3 to 10. The second dimension comprised an 8 to 18% acrylamide gel for separation based on molecular weight. Identified proteins are indicated by red crosses.

genome organization and the workings of a living cell. One goal would be to have sufficient knowledge to model a cell on a computer and make predictions about how it would respond to environmental changes.

3. Genomics can be used to provide insights into pathogenicity and suggest treatments for infectious disease. Possible virulence genes can be identified, and the expression of genes during infection can be studied. Host responses to pathogens can be examined. More sensitive diagnostic tests, new antibiotics, and different vaccines may come from genomic studies of pathogens.
4. The field of pharmacogenomics should produce many new drugs to treat disease. Databases of human gene sequences can be searched both for proteins that might have therapeutic value and for new drug targets. Genomics also can be used to study variations in drug-metabolizing enzymes and individual responses to medication.
5. The nature of horizontal gene transfer and the process of microbial evolution can be studied by comparing a wide variety of genomes. Comparative genomics will aid in the study of microbial biodiversity.
6. The industrial applications are numerous. For example, genomics can be used to identify novel enzymes with industrial potential, enhance the bioremediation of hazardous wastes, and improve techniques for the microbial production of methane and other fuels.
7. Genomics will profoundly impact agriculture. It can be used to find new biopesticides and to improve sustainable agricultural practices through enhancements in processes such as nitrogen fixation.

It is clear that the possibilities are great and genomics will profoundly impact many areas of microbiology. Advances in our understanding of microorganisms also will aid in the genomic study of more complex eucaryotic organisms.

Summary

1. Genomics is the study of the molecular organization of genomes, their information content, and the gene products they encode. It may be divided into three broad areas: structural genomics, functional genomics, and comparative genomics.
2. DNA fragments are normally sequenced using dideoxynucleotides and the Sanger technique (**figure 15.2**).
3. Most often microbial genomes are sequenced using the whole-genome shotgun technique of Venter, Smith, and collaborators. Four stages are involved: library construction, sequencing of randomly produced fragments, fragment alignment and gap closure, and editing the final sequence (**figure 15.3**).
4. After the sequence has been determined, it is annotated. That is, computer analysis is used to identify genes and their functions by comparing them with gene sequences in databases.
5. Analysis of vast amounts of genome data requires sophisticated computers and programs; these analytical procedures are a part of the discipline of bioinformatics.
6. Many microbial genomes have already been sequenced (**table 15.1**) and about 100 more procaryotic genomes currently are being sequenced.
7. The genome of *Mycoplasma genitalium* is one of the smallest of any free-living organism. Analysis of this genome and others indicates that only about 265 to 350 genes are required for growth in the laboratory.
8. *Haemophilus influenzae* lacks a complete set of Krebs cycle genes and has 1,465 copies of the recognition sequence used in DNA uptake during transformation.
9. The archaeon *Methanococcus jannaschii* is quite different genetically from bacteria and eucaryotes; only around 44% of its genes match those of the bacteria and eucaryotes it was compared against. Its informational genes (replication, transcription, and translation) are more similar to eucaryotic genes, whereas its metabolic genes resemble those of bacteria more closely.
10. Even in the case of *E. coli,* perhaps the best-studied bacterium, about 58% of the predicted genes do not resemble known genes.
11. The genome sequence of *Rickettsia prowazekii* is very similar to that of mitochondria. Aerobic respiration in eucaryotes may have arisen from an ancestor of *Rickettsia.*
12. The genome of *Chlamydia trachomatis* has provided many surprises. For example, it appears able to make at least some ATP and peptidoglycan, despite the fact that it seems to obtain most ATP from the host and does not have a cell wall with peptidoglycan. The presence of plantlike genes indicates that it might have infected plantlike hosts before moving to animals.
13. *Treponema pallidum,* the causative agent of syphilis, has lost many of its metabolic genes, which may explain why it hasn't been cultivated outside a host.
14. *Mycobacterium tuberculosis* contains more than 250 genes for lipid metabolism and may obtain much of its energy from host lipids. Surface and secretory proteins have been identified and may help vaccine development.
15. There has been a great deal of horizontal gene transfer between genomes in both Bacteria and Archaea. This is particularly the case for housekeeping or operational genes.
16. Annotation of genomes can be used to identify many genes and their functions. There seem to be patterns in gene distribution. For example, parasitic forms tend to lose genes and obtain nutrients from their hosts.
17. DNA microarrays (DNA chips) can be used to follow gene expression and mRNA production (**figures 15.9** and **15.10**).
18. The entire collection of proteins that an organism can produce is the proteome, and its study is called proteomics. The proteome often is analyzed by two-dimensional electrophoresis followed in some cases by mass spectrometry. Proteomic experiments sometimes provide more evidence about gene expression than the use of DNA chips.
19. Despite great success in sequencing genomes, many problems still need to be resolved before the data can be interpreted adequately and applied to the understanding of organisms.
20. In the future genomics will positively impact many areas of microbiology.

Key Terms

annotation *347*
bioinformatics *348*
comparative genomics *345*
DNA microarrays (DNA chips) *354*
expressed sequence tag (EST) *354*
functional genomics *345*
genomics *345*
open reading frame (ORF) *347*
proteome *356*
proteomics *356*
structural genomics *345*
whole-genome shotgun sequencing *346*

Questions for Thought and Review

1. What impact might genome comparisons have on the current phylogenetic schemes for Bacteria and Archaea that are discussed in chapter 19?
2. How would you use genomics data to develop new vaccines and antimicrobial drugs?
3. Why are proteomic studies necessary when one can use DNA chips to follow mRNA synthesis?
4. Discuss the importance of bioinformatics for genomics and the information it can supply.
5. Contrast informational and housekeeping or operational genes with respect to function and variation in quantity between genomes. How do free-living and parasitic microorganisms differ with respect to these genes?
6. Discuss some of the more important problems for postgenomic studies of microorganisms. What areas of microbiology do you think will be most positively impacted by genomics?

Critical Thinking Questions

1. Propose an experiment that can be done easily with a microchip that would have required years before this new technology.
2. What are the pitfalls of searches for homologous genes and proteins?

Additional Reading

General

Brown, T. A. 1999. *Genomes.* New York: John Wiley.

Brown, K. 2000. The human genome business today. *Sci. Am.* 283(1):50–55.

Charlebois, R. L., editor 1999. *Organization of the prokaryotic genome.* Washington, D.C.: ASM Press.

Dougherty, B. A. 2000. DNA sequencing and genomics. In *Encyclopedia of microbiology,* 2d ed., vol. 2, J. Lederberg, editor-in-chief, 106–16. San Diego: Academic Press.

Downs, D. M., and Escalante-Semerena, J. C. 2000. Impact of genomics and genetics on the elucidation of bacterial metabolism. *Methods* 20(1):47–54.

Ezzell, C. 2000. Beyond the human genome. *Sci. Am.* 283(1):64–69.

Field, D.; Hood, D.; and Moxon, R. 1999. Contribution of genomics to bacterial pathogenesis. *Curr. Opin. Genet. Dev.* 9(6):700–703.

Haseltine, W. A. 1997. Discovering genes for new medicines. *Sci. Am.* 276(3):92–7.

Lander, E. S., and Weinberg, R. A. 2000. Genomics: Journey to the center of biology. *Science* 287:1777–82.

Strauss, E. J., and Falkow, S. 1997. Microbial pathogenesis: Genomics and beyond. *Science* 276:707–12.

15.4 Bioinformatics

Ashburner, M., and Goodman, N. 1997. Informatics—genome and genetic databases. *Curr. Opin. Genet. Dev.* 7:750–56.

Howard, K. 2000. The bioinformatics gold rush. *Sci. Am.* 283(1):58–63.

Patterson, M., and Handel, M., editors. 1998. Trends guide to bioinformatics. New York: Elsevier Science, Ltd.

Rashidi, H. H., and Buehler, L. K. 2000. *Bioinformatics basics: Applications in the biological sciences and medicine.* Boca Raton, Fla.: CRC Press.

15.5 General Characteristics of Microbial Genomes

Andersson, S. G. E., et al. 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* 396:133–40.

Blattner, F. R., et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* 277:1453–62.

Bult, C. J., et al. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science* 273:1058–1107.

Cole, S. T., et al. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393:537–44.

Fleischmann, R. D., et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269:496–512.

Fraser, C. M., et al. 1995. The minimal gene complement of *Mycoplasma genitalium. Science* 270:397–403.

Fraser, C. M., et al. 1998. Complete genome sequence of *Treponema pallidum,* the syphilis spirochete. *Science* 281:375–88.

Galperin, M. Y., and Koonin, E. V. 1998. Sources of systematic error in functional annotation of genomes: Domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biology* 1:55–67.

Gaasterland, T. 1999. Archaeal genomics. *Curr. Opin. Microbiol.* 2(5):542–47.

Gogarten, J. P., and Olendzenski, L. 1999. Orthologs, paralogs and genome comparisons. *Curr. Opin. Genet. Dev.* 9:630–36.

Heidelberg, J. F., et al. 2000. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae. Nature* 406:477–83.

Jain, R.; Rivera, M. C.; and Lake, J. A. 1999. Horizontal gene transfer among genomes: The complexity hypothesis. *Proc. Natl. Acad. Sci.* 96:3801–6.

Koonin, E. V., and Galperin, M. Y. 1997. Prokaryotic genomes: The emerging paradigm of genome-based microbiology. *Curr. Opin. Genet. Dev.* 7:757–63.

Lawrence, J. G., and Ochman, H. 1998. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci.* 95:9413–17.

Makarova, K. S.; Aravind, L.; Wolf, Y. I.; Tatusov, R. L.; Minton, K. W.; Koonin, E. V.; and Daly, M. J. 2001. Genome of the extremely radiation-resistant bacterium *Deinococcus radiodurans* viewed from the perspective of comparative genomics. *Microbiol. Mol. Biol. Rev.* 65(1):44–79.

Ochman, H.; Lawrence, J. G.; and Groisman, E. A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 408:299–304.

Pollack, J. D. 1997. *Mycoplasma* genes: A case for reflective annotation. *Trends Microbiol.* 5(10):413–19.

Riley, M., and Serres, M. H. 2000. Interim report on genomics of *Escherichia coli. Annu. Rev. Microbiol.* 54:341–411.

Snel, B.; Bork, P.; and Huynen, M. 2000. Genome evolution: Gene fusion versus gene fission. *Trends Genet.* 16(1):9–11.

Stephens, R. S., et al. 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis. Science* 282:754–59.

Travis, J. 2000. Pass the genes, please. *Science News* 158:60–61.

White, O., et al. 1999. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* 286:1571–77.

15.6 Functional Genomics

Blackstock, W., and Mann, M., editors. 2000. Proteomics: A trends guide. New York: Elsevier Science, Ltd.

Brenner, S. 2000. The end of the beginning. *Science* 287:2173–74.

Eisenberg, D.; Marcotte, E. M.; Xenarios, I.; and Yeates, T. O. 2000. Protein function in the post-genomic era. *Nature* 405:823–826.

Ferea, T. L., and Brown, P. O. 1999. Observing the living genome. *Curr. Opin. Genet. Dev.* 9:715–22.

Galperin, M. Y., and Koonin, E. V. 1999. Functional genomics and enzyme evolution. *Genetica* 106(1–2):159–70.

Gingeras, T. R., and Rosenow, C. 2000. Studying microbial genomes with high-density oligonucleotide arrays. *ASM News* 66(8):463–69.

Hamadeh, H., and Afshari, C. A. 2000. Gene chips and functional genomics. *American Scientist* 88:508–15.

Huang, S. 2000. The practical problems of post-genomic biology. *Nature Biotechnol.* 18:471–72.

Lockhart, D. J., and Winzeler, E. A. 2000. Genomics, gene expression and DNA arrays. *Nature* 405:827–36.

Phimister, B., editor. 1999. The chipping forecast. *Nature Genet.* 21(1) supplement.

Pandey, A., and Mann, M. 2000. Proteomics to study genes and genomes. *Nature* 405:837–46.

Rastan, S., and Beeley, L. J. 1997. Functional genomics: Going forwards from the databases. *Curr. Opin. Genet. Dev.* 7:777–83.

Schena, M., editor. 1999. *DNA microarrays: A practical approach.* New York: Oxford University Press.