13 GENOMICS AND BIOINFORMATICS

Spencer Muse, PhD

Chapter Contents

At the conclusion of this chapter, the reader will be able to:

- Discuss the basic principles of molecular biology regarding genome science.
- Describe the major types of data involved in genome projects, including technologies for collecting them.

- Describe practical applications and uses of genomic data.
- Understand the major topics in the field of bioinformatics and DNA sequence analysis.
- Use key bioinformatics databases and web resources.

13.1 INTRODUCTION

In April 2003, sequencing of all three billion nucleotides in the human genome was declared complete. This landmark of modern science brought with it high hopes for the understanding and treatment of human genetic disorders. There is plenty of evidence to suggest that the hopes will become reality—1631 human genetic diseases are now associated with known DNA sequences, compared to the less than 100 that were known at the initiation of the Human Genome Project (HGP) in 1990. The success of this project (it came in almost 3 years ahead of time and 10% under budget, while at the same time providing more data than originally planned) depended on innovations in a variety of areas: breakthroughs in basic molecular biology to allow manipulation of DNA and other compounds; improved engineering and manufacturing technology to produce equipment for reading the sequences of DNA; advances in robotics and laboratory automation; development of statistical methods to interpret data from sequencing projects; and the creation of specialized computing hardware and software systems to circumvent massive computational barriers that faced genome scientists. Clearly, the HGP served as an incubator for interdisciplinary research at both the basic and applied levels.

The human genome was not the only organism targeted during the genomic era. As of June 2004, the complete genomes were available for 1557 viruses, 165 microbes, and 26 eukaryotes ranging from the malaria parasite *Plasmodium falciparum* to yeast, rice, and humans. Continued advances in technology are necessary to accelerate the pace and to reduce the expense of data acquisition projects. Improved computational and statistical methods are needed to interpret the mountains of data. The increase in the rate of data accumulation is outpacing the rate of increases in computer processor speed, stressing the importance of both applied and basic theoretical work in the mathematical and computational sciences.

In this chapter, the key technologies that are being used to collect data in the laboratory, as well as some of the important mathematical techniques that are being used to analyze the data, are surveyed. Applications to medicine are used as examples when appropriate.

13.1.1 The Central Dogma: DNA to RNA to Protein

Understanding the applications of genomic technologies requires an understanding of three key sets of concepts: how genetic information is stored, how that information is processed, and how that information is transmitted from parent to offspring.

In most organisms, the genetic information is stored in molecules of DNA, deoxyribonucleic acid (Fig. 13.1). Some viruses maintain their genetic data in RNA, but no

emphasis will be placed on such exceptions. The size of genomes, measured in counts of nucleotides or base pairs, varies tremendously, and a curious observation is that genome size is only loosely associated with organismal complexity (Table 13.1). Most of the known functional units of genomes are called genes. For purposes of this chapter, a gene can be defined as a contiguous block of nucleotides operating for a single purpose. This definition is necessarily vague, for there are a number of types of genes, and even within a given type of gene, experts have difficulty agreeing on precisely where the beginning and ending boundaries of those genes lie. A structural gene is a gene that codes instructions for creating a protein (Fig. 13.2). A second category of genes with many members is the collection of RNA genes. An RNA gene does not contain protein information; instead, its function is determined by its ability to fold into a specific three-dimensional configuration, at which point it is able to interact with other molecules and play a part in a biochemical process. A common RNA gene found in most forms of life is the tRNA gene illustrated in Figure 13.3.

Structural genes are the entities most scientists envision when the word "gene" is mentioned, and from this point on, the term *gene* will be used to mean "structural gene" unless specified otherwise. The number and variety of genes in organisms is a current topic of importance for genome scientists. Gene number in organisms ranges from tiny (470 in mycoplasma) to enormous (60,000 or more in plants).

Figure 13.1 The DNA molecule consists of two strands of nucleotides arranged in a double helix. Complementary pairs of nucleotides (A,T and C,G) form chemical bonds that hold the helix intact.

TABLE 13.1 Genome Size and Genome Content (most from Taft and Mattick, 2003)

Organism	Genome Size (x10^6)	Gene Number
HIV1	0.009	9
SARS	0.03	14
M. genitalium	0.58	470
N. equitans	0.49	552
H. helaticus	1.80	1875
E. coli	4.64	4288
P. falciparum	22.85	5268
S. cerevisiae	12.10	6000
D. melanogaster	122	13,600
C. elegans	97	19,049
A. thaliana	115	25,000
H. sapiens	3200	30,000
M. musculus	2500	37,000

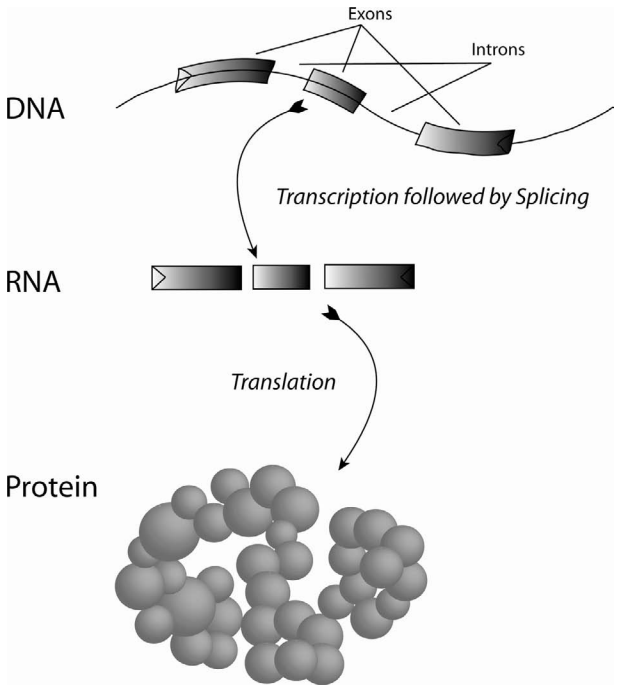

Figure 13.2 The central dogma of molecular biology. Genomic DNA is first transcribed into an RNA copy that contains introns and exons. The introns are spliced from the RNA, and the remaining exons are concatenated. The mature RNA sequence is then translated to produce a protein.

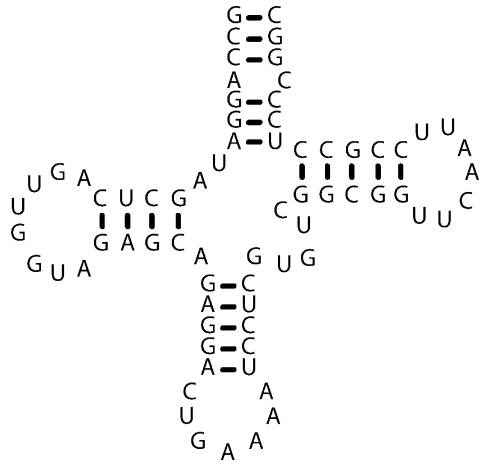

Figure 13.3 The transfer RNA (tRNA) is an example of a non-protein-coding gene. Its function is the result of the specific two- and three-dimensional structures formed by the RNA sequence itself.

Non-free-living organisms have even smaller gene numbers (the HIV virus contains only nine). The number of genes in a typical human genome has been estimated to be about 30,000, perhaps the single most surprising finding from the Human Genome Project. This number was thought to be as large as 120,000 as recently as 1998. The confusion over this number arose in part because there is a not a ''one gene, one protein'' rule in humans, or indeed, in many eukaryotic organisms. Instead, a single gene region can contain the information needed to produce multiple proteins. To understand this fact, the series of steps involved in creating a functional protein from the underlying DNA sequence instructions must be understood. The central dogma of molecular biology states that genetic information is stored in DNA, copied to RNA, and then interpreted from the RNA copy to form a functional protein (Fig. 13.2).

The process of copying the genetic information in DNA into an RNA copy is known as transcription (see Chapter 3). The process is thought by many to be a remnant of an early RNA world, in which the earliest life forms were based on RNA genomes. It is at the level of transcription that gene expression is regulated, determining where and when a particular gene is turned on or off.

The transcription of a gene occurs when an enzyme known as RNA polymerase binds to the beginning of a gene and proceeds to create a molecule of RNA that matches the DNA in the genome. It is this molecule of messenger RNA (mRNA) that will serve as a template for producing a protein. However, it is necessary for organisms to regulate the expression of genes to avoid having all genes being produced in all cells at all times. Transcription factors interact with either the genomic DNA or the polymerase molecule to allow delicate control of the gene expression process. A feedback loop is created whereby an environmental stimulus such as a drug leads to the production of a transcription factor, which triggers the expression of a gene. In addition to this example of a positive control mechanism, negative control is also possible. An emerging theme is that sets of genes are often coregulated by a single or

small group of transcription factors. These sets of genes often share a short upstream DNA sequence that serves as a binding site for the transcription factor.

One of the earliest surprises of the genomic era was the discovery that many eukaryotic gene sequences are not contiguous, but are instead interrupted by DNA sequences known as introns. As shown in Figure 13.2, introns are physically cut, or spliced, from the mRNA sequence before the RNA is converted into a protein. The presence of introns helps to explain the phenomenon that there are more proteins produced in an organism than there are genes present. The process of alternative splicing allows for exons to be assembled in a combinatoric fashion, resulting in a multitude of potential proteins. For example, consider a gene sequence with exons E1, E2, and E3 interrupted by introns I1 and I2. If both introns are spliced, the resulting protein would be encoded as E1-E2-E3. However, it is also possible to splice the gene in a way that produces protein E1-E3, skipping exon E2. Much like transcription factors regulate gene expressions, there are factors that help to regulate alternative splicing. A common theme is to find a single gene that is spliced in different ways to produce isoforms that are expressed in specific tissues.

The process of reading the template in an mRNA molecule and using it to produce a protein is known as translation. Conceptually, this process is much more simple than the transcription and splicing processes. A structure known as a ribosome binds to the mRNA molecule. The ribosome then moves along the RNA in units of 3 nucleotides. Each of these triplets, or codons, encodes one of 20 amino acids. At each codon the ribosome interacts with tRNAs to interpret a codon and add the proper amino acid to the growing chain before moving along to the next codon in the sequence (see Chapter 3).

13.2 CORE LABORATORY TECHNOLOGIES

Genome science is heavily driven by new technological advances that allow for the rapid and inexpensive collection of various types of data. It has been said that the field is data-driven rather than hypothesis-driven, a reflection of the tendency for researchers to collect large amounts of genomic data with the (realistic) expectation that subsequent data analyses, along with the experiments they suggest, will lead to better understanding of genetic processes. Although the list of important biotechnologies changes on an almost daily basis, there are three prominent data types in today's environment: (1) genome sequences provide the starting point that allows scientists to begin understanding the genetic underpinnings of an organism; (2) measurements of gene expression levels facilitate studies of gene regulation, which, among other things, help us to understand how an organism's genome interacts with its environment; and (3) genetic polymorphisms are variations from individual to individual within species, and understanding how these variations correlate with phenotypes such as disease susceptibility is a crucial element of modern biomedical research.

13.2.1 Gene Sequencing

The basic principles for obtaining DNA sequences have remained rather stable over the past few decades, although the specific technologies have evolved dramatically. The most widely used sequencing techniques rely on attaching some sort of ''reporter'' to each nucleotide in a DNA sequence, then measuring how quickly or how far the nucleotide migrates through a medium. The principles of Sanger sequencing, originally developed in 1974, are illustrated in Figure 13.4.

DNA sequences have an orientation. The 5′ end of a sequence can be considered to be the left end, and the 3′ end is on the right. Sanger sequencing begins by creating all possible subsequences of the target sequence that begin at the same 5′ nucleotide. A reporter, originally radioactive but now fluorescent, is attached to the final 3′ nucleotide in each subsequence. By using a unique reporter for each of the four

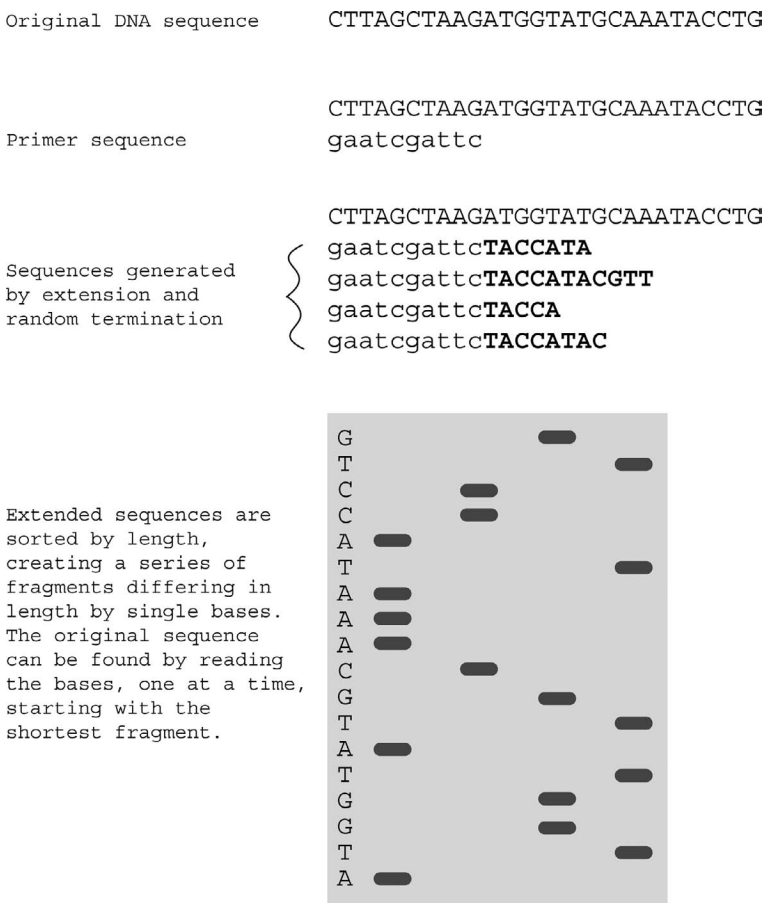

Figure 13.4 The principles of Sanger sequencing of DNA.

nucleotides, it is possible to identify the final 3′ nucleotide in each of the subsequences. Consider the task of sequencing the DNA molecule AGGT. There are four possible subsequences that begin with the 5′ A: A, AG, AGG, and AGGT. The technology of Sanger sequencing produces each of those four sequences and attaches the reporter to the final nucleotide. The subsequences are sorted from shortest to longest based on the rate at which they migrate through a medium. The shortest sequence would correspond to the subsequence A; its reporter tells us that the final nucleotide is an A. The second shortest subsequence is AG, with a final nucleotide of G. By arranging the subsequences in a ''ladder'' from shortest to longest, the sequence of the complete target sequence can be found simply by reading off the final nucleotide of each subsequence.

A series of new advances have allowed Sanger sequencing to be applied in a high-throughput way, paving the way for sequencing of entire genomes, including that of the human. Radioactive reporters have been replaced with safer and cheaper fluorescent dyes, and automatic laser-based systems now read the sequence of fluorescent-labeled nucleotides directly as they migrate. Early versions of Sanger sequencing only allowed for reading a few hundred nucleotides at a time; modern sequencing devices can read sequences of 800 nucleotides or more. Perhaps most important has been the replacement of ''slab gel'' systems with capillary sequencers. The older system required much labor and a steady hand; capillary systems, in conjunction with the development of necessary robotic devices for manipulating samples, have allowed almost completely automated sequencing pipelines to be developed. Not to be ignored in the series of technological advances is the development of automated base-calling algorithms. A laser reads the intensities of each of the four fluorescent reporter dyes as each nucleotide passes it. The resulting graph of those intensities is a chromatogram. Statistical algorithms, including the landmark program phred, are able to accept chromatograms as input and output DNA sequences with very high levels of accuracy, reducing the need for laborious human intervention. By assessing the relative levels of the four curves, the base-calling algorithms not only report the most likely nucleotide at each position, but they also provide an error probability for each site. A single state-of-the-art DNA sequencing machine can currently produce upwards of one million nucleotides per day.

13.2.1 Whole Genome Sequencing

Large regions of DNA are not sequenced in single pieces. Instead, larger contigs of DNA are fragmented into multiple, short, overlapping sequences. The emergence of shotgun sequencing (Fig. 13.5), pioneered by Dr. Craig Venter, has revolutionized approaches for obtaining complete genome sequences. The fundamental approach to shotgun sequencing of a genome is simple: (1) create many identical copies of a genome; (2) randomly cut the genomes into millions of fragments, each short enough to be sequenced individually; (3) align the overlapping fragments by identifying matching nucleotides at the ends of fragments, and finally; (4) read the complete genome sequence by following a gap-free path through the fragments. Until Venter's work, the idea of shotgun sequencing was considered unfeasible for a variety of

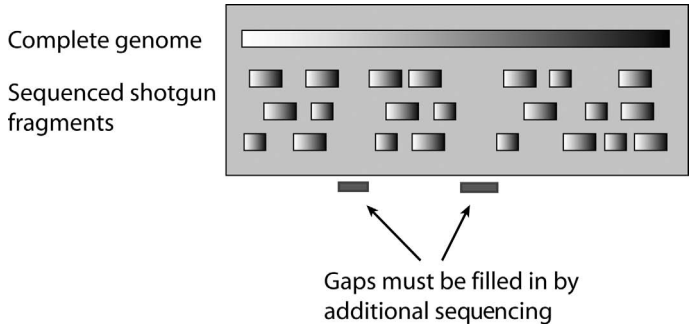

Figure 13.5 Shotgun sequencing of genomes or other large fragments of DNA proceeds by cutting the original DNA into many smaller segments, sequencing the smaller fragments, and assembling the sequenced fragments by identifying overlapping ends.

reasons. Perhaps most daunting was the computational task of aligning the millions of fragments generated in the shotgun process. Specialized hardware systems and associated algorithms were developed to handle these problems.

13.2.3 Gene Expression

Following in the footsteps of high-throughput genome sequencing came technology that allowed scientists to survey the relative abundance of thousands of individual gene products. These technologies are, in essence, a modern high-throughput replacement of the Northern blot procedure. For each member in a collection of several thousand genes, the assays provide a quantitative estimate of the number of mRNA copies of each gene being produced in a particular tissue. Two technologies, cDNA and oligonucleotide microarrays, currently dominate the field, and they have opened the door to many exploratory analyses that were previously impossible.

As a first example, consider taking two samples of cells from an individual cancer patient: one sample from a tumor and one from normal tissue. A microarray experiment makes it possible to identify the set of genes that are produced at different levels in the two tissue types. It is likely that this set of differentially expressed genes contains many genes involved in biological processes related to tumor formation and proliferation (Fig. 13.6).

A second common type of study is a time course experiment. Microarray data is collected from the same tissues at periodic intervals over some length of time. For instance, gene expression levels may be measured in 6-hour increments following the administration of a drug. For each gene, the change in gene expression level can be plotted against time (Fig. 13.7). Groups of coregulated genes will be identified as having points in time where they all experience either an increase or decrease in expression levels. A likely cause for this behavior is that all genes in the coregulated set are governed by a single transcription factor.

A final important medical application of microarray technologies involves diagnosis. Suppose that a physician obtains microarray data from tumor cells of a patient.

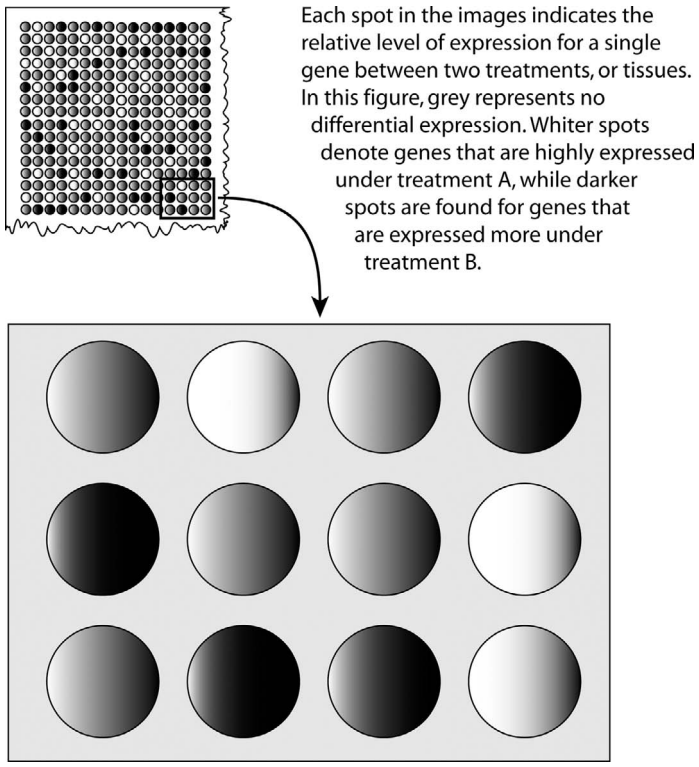

Figure 13.6 Microarray technology allows genome scientists to measure the relative expression levels of thousands of genes simultaneously. A typical microarray slide can hold 5000 or more spots.

The data, consisting of the relative levels of gene expression for a suite of many genes, can be compared to similar data collected from tumors of known types. If the patient's gene expression profile matches the profile of one of the reference samples, the patient can be diagnosed with that tumor type. The advent of microarray techniques has rapidly improved the accuracy of this type of diagnosis in a variety of cancers.

cDNA microarrays were the first, and are still the most widely used, form of high-throughput gene expression methods. The procedure begins by attaching the DNA sequences of thousands of genes onto a microscope slide in a pattern of spots, with each spot containing only DNA sequences of a single gene. A variety of technologies have emerged for creating such slides, ranging from simple pin spotting devices to technologies using laser jet printing techniques. The RNA of expressed genes is next collected from the target cell population. Through the process of reverse transcription, a cDNA version of each RNA is created. A cDNA molecule is complementary to the genomic DNA sequence in the sense that complementary base pairs will physically bind to one another. For example, a cDNA reading GTTAC could physically bind to the genomic DNA sequence CAATG. During the process of creating the cDNA collection, each cDNA is labeled with a fluorescent dye. The collection of labeled

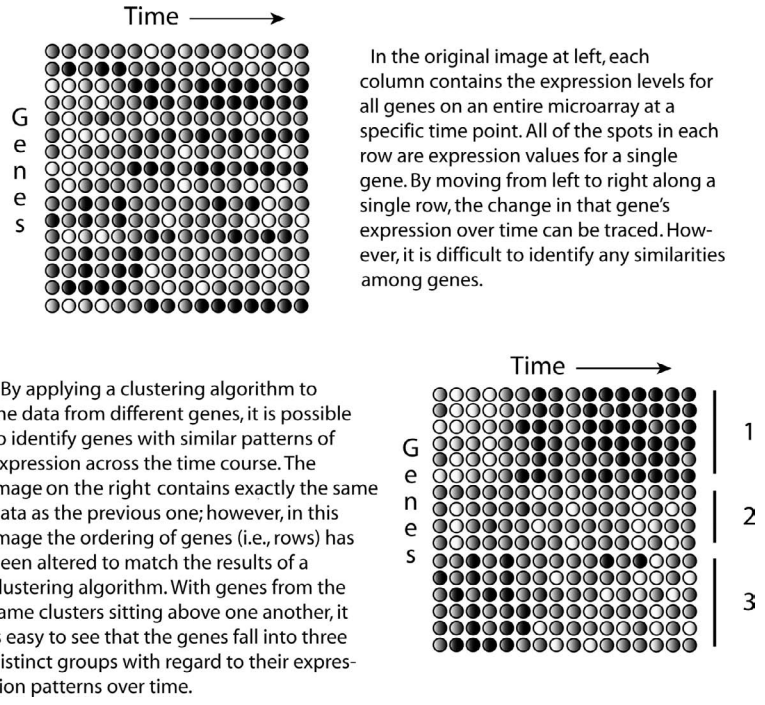

Figure 13.7 Time course expression studies examine patterns of gene expression over time. By finding genes that have similar expression profiles, sets of coregulated genes can be identified.

cDNAs is poured over the microscope slide and its set of attached DNA molecules. The cDNAs that match a DNA on the slide physically bind to their mates, and unbound cDNAs are washed from the slide. Finally, the number of bound molecules at each spot (genes) can be read by measuring the fluorescence level at each spot. Highly expressed genes will create more RNA, which results in more labeled cDNAs binding to those spots.

A more common variant on the basic cDNA approach is illustrated in Figure 13.8. In this experiment, RNA from two different tissues or individuals is collected, labeled with two different dyes, and competitively hybridized on a single slide. The relative abundance of the two dyes allows the scientist to state, for instance, that a particular gene is expressed fivefold times more in one tissue than in the other.

Oligonucleotide arrays take a slightly different approach to assaying the relative abundance of RNA sequences. Instead of attaching full-length DNAs to a slide, oligonucleotide systems make use of short oligonucleotides chosen to be specific to individual genes. For each gene included in the array, approximately 10 to 20 different oligonucleotides of length 20–25 nucleotides are designed and printed onto a chip. The use of multiple oligonucleotides for each gene helps to reduce the effects of a variety of potential errors. Fluorescently labeled RNA (rather than cDNA) is collected from the target tissue and hybridized against the oligonucleotide array. One limitation

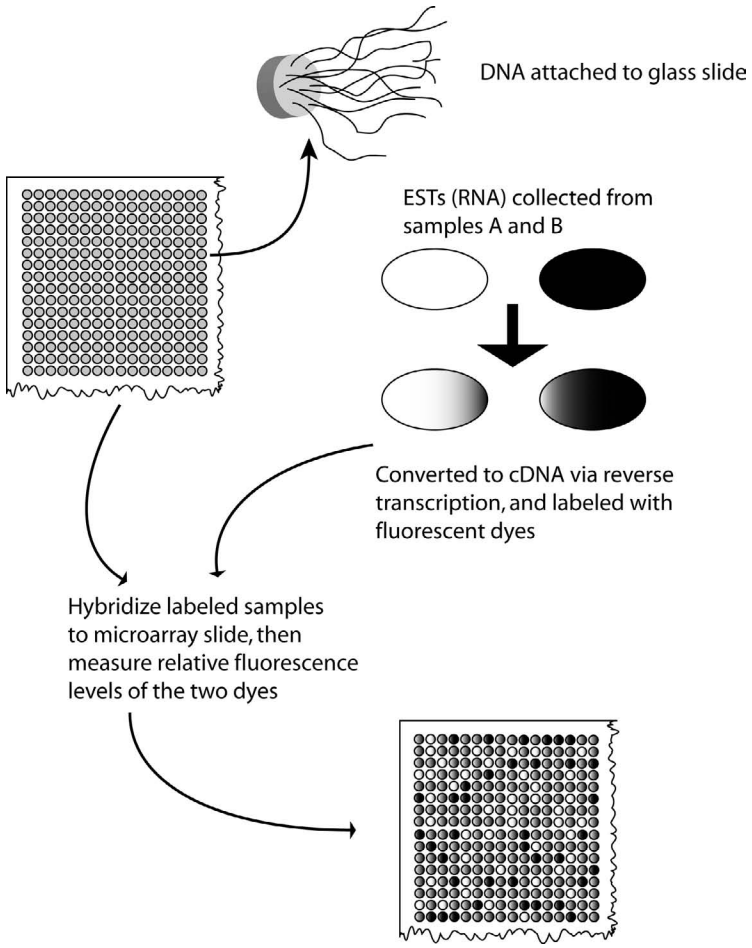

Figure 13.8 A cDNA microarray slide is created by (1) attaching DNA to spots on a glass slide, (2) collecting expressed RNA sequences (expressed sequence tags, ESTs) from tissue samples, (3) converting the RNA to DNA and labeling the molecules with fluorescent dyes, (4) hybridizing the labeled DNA molecules to the DNA bound to the slide, and (5) extracting the quantity of each expressed sequence by measuring the fluorescence levels of the dyes.

of the oligonucleotide approach is that only a single sample can be assayed on a single chip—competitive hybridization is not possible.

Although oligonucleotide and cDNA approaches to assaying gene expression rely on the same basic principles, each has its own advantages and disadvantages. As already noted, competitive hybridization is currently only possible in cDNA systems. The design of oligonucleotide arrays requires that the sequences of genes for the chip are already available. The design phase is very expensive, and oligonucleotide systems are only available for commercially important and model organisms. In contrast, cDNA arrays can be developed fairly quickly even in organisms without sequenced

genomes. In their favor, oligonucleotide arrays allow for more genes to be spotted in a given area (thus allowing more measurements to be made on a single chip) and tend to offer higher repeatability of measurements. Both of these facts reduce the overall level of experimental error rate in oligonucleotide arrays relative to cDNA microarrays, although at a higher per observation cost. Because of the trade-off between obtaining many cheap noisy measurements versus a smaller number of more precise but expensive measurements, it is not clear that either technology has an obvious cost advantage.

Both techniques share the same major disadvantage: only measurements of RNA levels are found. These measurements are used as surrogates for the much more desirable and useful quantities of the amount of protein produced for each gene. It appears that RNA levels are correlated with protein levels, but the extent and strength of this relationship is not understood well. The near future promises a growing role for protein microarray systems, which are currently seeing limited use because of their very high costs.

13.2.4 Polymorphisms

The "final draft" of the human genome was announced in April 2003. It included roughly 2.9 billion nucleotides, with some 30,000 to 40,000 genes spread across 23 pairs of chromosomes. The next phase of major data acquisition on the human genome is to discover how differences, both large and small, from individual to individual, result in variation at the phenotypic level. Toward this end, a major effort has been made to find and document genetic polymorphisms. Polymorphisms have long been important to studies of genetics. Variations of the banding patterns in polytene chromosomes, for instance, have been studied for many decades. Allozyme assays, based on differences in the overall charge of amino acid sequences, were popular in the 1960s. Most modern studies of genetic polymorphisms, though, focus on identifying variation at the individual nucleotide level.

The international SNP Consortium (http://snp.cshl.org) is a collaboration of public and private organizations that discovered and characterized approximately 1.8 million single nucleotide polymorphisms (SNPs) in human populations. In medicine, the expectation is that knowledge of these individual nucleotide variants will accelerate the description of genetic diseases and the drug development process. Pharmaceutical companies are optimistic that surveys of variation will be of use for selecting the proper drug for individual patients and for predicting likely side effects on an individual-to-individual basis.

Most SNPs (pronounced "snips") are the result of a mutation from one nucleotide to another, whereas a minority are insertions and deletions of individual nucleotides. Surveys of SNPs have demonstrated that their frequencies vary from organism to organism and from region to region within organisms. In the human genome, a SNP is found about every 1000 to 1500 nucleotides. However, the frequency of SNPs is much higher in noncoding regions of the genome than in coding regions, the result of natural selection eliminating deleterious alleles from the population. Furthermore, synonymous or silent polymorphisms, which do not result in a change of the encoded

amino acids, are more frequent than nonsynonymous or replacement polymorphisms. The fields of population genetics and molecular evolution provide many empirical surveys of SNP variation, along with mathematical theory, for analyzing and predicting the frequencies of SNPs under a variety of biologically important settings.

Simple sequence repeats (SSRs) consist of a moderate (10–50) number of tandemly repeated copies of the same short sequence of 2 to 13 nucleotides. SSRs are an important class of polymorphisms because of their high mutation rates, which lead to SSR loci being highly variable in most populations. This high level of variability makes SSR markers ideal for work in individual identification. SSRs are the markers typically employed for DNA fingerprinting in the forensics setting. In human populations, an SSR locus usually has 10 or more alleles and a per generation mutation rate of 0.001. The FBI uses a set of 13 tetranucleotide repeats for identification purposes, and experts claim that no two unrelated individuals have the exact same collection of alleles at all 13 of those loci.

13.3 CORE BIOINFORMATICS TECHNOLOGIES

As the technology for collecting genomic data has improved, so has the need for new methods for management and analysis of the massive amounts of accumulated data. The term *bioinformatics* has evolved to include the mathematical, statistical, and computational analysis of genomic data. Work in bioinformatics ranges from database design to systems engineering to artificial intelligence to applied mathematics and statistics, all with an underlying focus on genomic science.

A variety of bioinformatics topics may be illustrated using the core technologies described in the preceding section. It is necessary to carry out sequence alignments in order to assemble sequence fragments. All of these sequences, along with the vital information about their sources, functions, and so on, must be stored in databases, which must be readily available to users in a variety of locations. Once a sequence has been obtained, it is necessary to annotate its function. One of the most fundamental annotation tasks is that of computational gene finding, in which a genome or chromosome sequence is input to an algorithm that subsequently outputs the predicted location of genes. A gene sequence, whether predicted or experimentally determined, must have its function predicted, and many bioinformatics tools are available for this task. Once microarray data are available, it is necessary to identify subsets of coregulated genes and to identify genes that are differentially expressed between two or more treatments or tissue types. Polymorphism data from SNPs are used to search for correlations with, for example, the presence or absence of a disease in family pedigrees. These questions are all of fundamental importance and draw on many different fields. By necessity, bioinformatics is a highly multidisciplinary field.

13.3.1 Genomics Databases

Genome projects involve far-reaching collaborations among many researchers in many fields around the globe, and it is critical that the resulting data be easily available

both to project members and to the general scientific community. In light of this requirement, a number of key central data repositories have emerged. In addition to providing storage and retrieval of gene sequences, several of these databases also offer advanced sequence analysis methods and powerful visualization tools.

In the United States, the primary public genomics resource is that of the National Center for Biotechnology Information (NCBI). The NCBI website (http://www.ncbi.nlm.nih.gov) provides a seemingly endless collection of data and data analysis tools. Perhaps the most important element of the NCBI collection is the GenBank database of DNA and RNA sequences. NCBI provides a variety of tools for searching GenBank, and elements in GenBank are linked to other databases, both within and outside of NCBI. Figure 13.9 shows some results from a simple query of the GenBank nucleotide database.

GenBank data files contain a wealth of information. Figure 13.10 shows a simple GenBank file for a prion sequence from duck. The accession number, AF283319, is

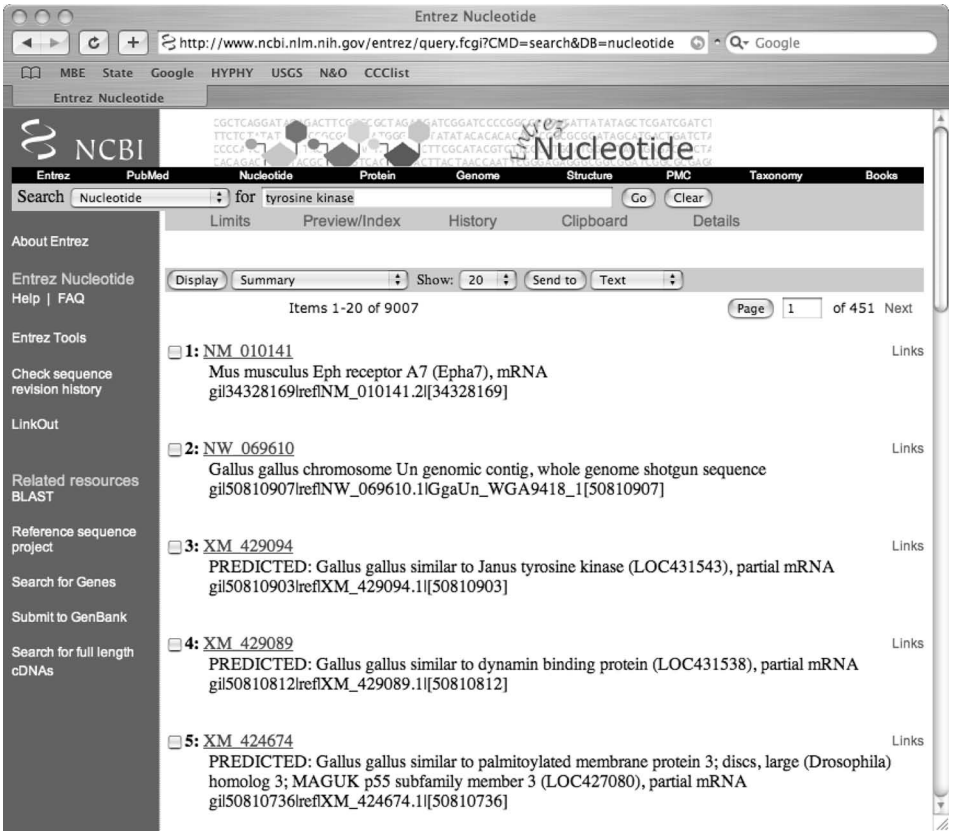

Figure 13.9 The result of a simple query of the GenBank database at NCBI. This query found 9007 entries in the GenBank nucleotide database containing the term ''tyrosine kinase.'' Each entry can be clicked to find additional information.


```
LOCUS       AF283319                819 bp    DNA     linear   VRT 12-JUL-2000
DEFINITION  Anas platyrhynchos prion protein (PrP) gene, complete cds.
ACCESSION   AF283319
VERSION     AF283319.1  GI:9050061
KEYWORDS    .
SOURCE      Anas platyrhynchos
  ORGANISM  Anas platyrhynchos
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Archosauria; Aves; Neognathae; Anseriformes; Anatidae; Anas.
REFERENCE   1  (bases 1 to 819)
  AUTHORS   Wang,Q., Li,N., Zhang,L. and Lian,Z.
  TITLE     Molecular cloning of duck prion gene
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 819)
  AUTHORS   Wang,Q., Li,N., Zhang,L. and Lian,Z.
  TITLE     Direct Submission
  JOURNAL   Submitted (28-JUN-2000) Biochemistry and Molecular Biology, China
            Agriculture University, No. 2 The Western Road of Yanmingyuan,
            Beijing 100094, China
FEATURES             Location/Qualifiers
     source          1..819
                     /organism="Anas platyrhynchos"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:8839"
                     /sex="male"
                     /tissue_type="blood"
                     /dev_stage="adult"
     gene            <1..>819
                     /gene="PrP"
     mRNA            <1..>819
                     /gene="PrP"
                     /product="prion protein"
     CDS             1..819
                     /gene="PrP"
                     /codon_start=1
                     /product="prion protein"
                     /protein_id="AAF82604.1"
                     /db_xref="GI:9050062"
                     /translation="MARLLTTCCLLALLLAACTDVALSKKGKGKPSGGGWGAGSHRQP
                     SYPRQPGYRHNPGYPHNPGYPHNPGYPHNPGYPHNPGYPQNPGYPHNPGYPGWGQGYN
                     PSSGGSYHNQKPWKPPKTNFKHVAGAAAAGAVVGGLGGYAMGRVMSGMSYRFDSPDEY
                     RWWNENSARYPNRVYYRDYSSPVSQDVFVADCFNITVTEYSIGPAAKKNTSEAVPAAN
                     QTDVETENKVVTKVIREVCVQQYREYRLASGIQLHPADTWLAVLLLLTTLFAMH"
ORIGIN
        1 atggctaggc tcctcaccac ctgctgcctg ctggccttgc tgctcgccgc ttgcaccgac
       61 gtcgccctct ccaagaaggg caaaggcaaa cccagtggtg ggggttgggg cgccgggagc
      121 catcgccagc ccagctaccc ccgccagccg ggctaccgtc ataacccagg gtacccccat
      181 aacccagggt acccccataa cccagggtac ccccacaacc ctggctatcc ccataacccc
      241 ggctaccccc agaacctgg ctacccccat aacccaggtt acccaggctg gggacaaggc
      301 tacaacccat ccagcggagg aagttaccac aatcagaagc catggaaacc ccccaaaacc
      361 aacttcaagc acgtggcggg ggcagctgca gctggtgctg tggttggggg cttgggggc
      421 tacgccatgg ggcgtgttat gtcagggatg agctaccgct ttgatagccc cgatgagtat
      481 cggtggtgga acgagaactc ggcgcgttat cccaaccggg tttactacag ggattacagc
      541 agccccgtgt cgcaggatgt cttcgtggcc gactgcttta acatcacagt gactgagtac
      601 agcattggcc ctgctgccaa gaagaacacc tccgaggctg tcccggcagc aaaccaaaca
      661 gacgtggaga cggagaacaa agtggtgacg aaggtgatcc gcgaggtgtg tgtgcagcag
      721 tatcgcgagt accgcctggc ctcgggcatc cagctgcacc ctgctgacac ctggcttgcc
      781 gtcctcctcc tcctcaccac ccttttgcc atgcactga
//
```

Figure 13.10 A simple GenBank file containing the DNA sequence for a prion protein gene.

the unique identifier for this entry. The GenBank file contains a DNA sequence of 819 nucleotides, its predicted amino acid sequence, and a citation to the Chinese laboratory that obtained the data. The "Links" icon in the upper right provides access to related information found in other databases. It is essential for those working in genomics or bioinformatics to become familiar with GenBank and the content of GenBank files.

NCBI is also the home of the BLAST database searching tool. BLAST uses algorithms for sequence alignment (described later in this chapter) to find sequences in GenBank that are similar to a query sequence provided by the user. To illustrate the use of BLAST, consider a study by Professor Eske Willerslev at the University of Copenhagen. Willerslev and his colleagues collected samples from Siberian permafrost that included a variety of preserved plant and animal material estimated to be 300,000–400,000 years old. They were able to extract short DNA sequences from the *rbcL* gene. These short sequences were used as input to the BLAST algorithm, which reported a list of similar sequences. It is likely that the most similar sequences come from close relatives of the organisms that provided the ancient DNA.

The European Bioinformatics Institute (EBI, http://www.ebi.ac.uk) is the European "equivalent" of NCBI. Users who explore the EBI website will find much of the same type of functionality as provided by NCBI. Of particular note is the Ensembl project (http://www.ensembl.org), a joint venture between EBI and the Sanger Institute. Ensembl has particularly nice tools for exploring genome project data through its Genome Browser. Figure 13.11 shows a portion of the display for a region of human chromosome 7. Ensembl provides comparisons to other completed genome sequences (rat, mouse, and chimpanzee), along with annotations of the locations of genes and other interesting features. Most of the items in the display are clickable and provide links to more detailed information on each display component.

Many other databases and Web resources play important roles in the day-to-day working of genome scientists. Table 13.2 includes a selection of these resources, along with short descriptions of their unique features.

13.3.2 Sequence Alignment

The most fundamental computational algorithm in bioinformatics is that of pairwise sequence alignment. Not only is it of immediate practical value, but the underlying dynamic programming algorithm also serves as a conceptual framework for many other important bioinformatics techniques. The goal of sequence alignment is to accept as input two or more DNA, RNA, or amino acid sequences; identify the regions of the sequences that are similar to one another according to some measure; and output the sequences with the similar positions aligned in columns. An alignment of six sequences from HIV strains is shown in Figure 13.12.

Sequence alignments have numerous uses. Alignments of pairs of sequences help us to determine whether or not they have the same or similar functions. Regions of alignments with little sequence variation likely correspond to important structural or functional regions of protein coding genes. By studying patterns of similarity in an alignment of genes from several species, it is possible to infer the evolutionary history

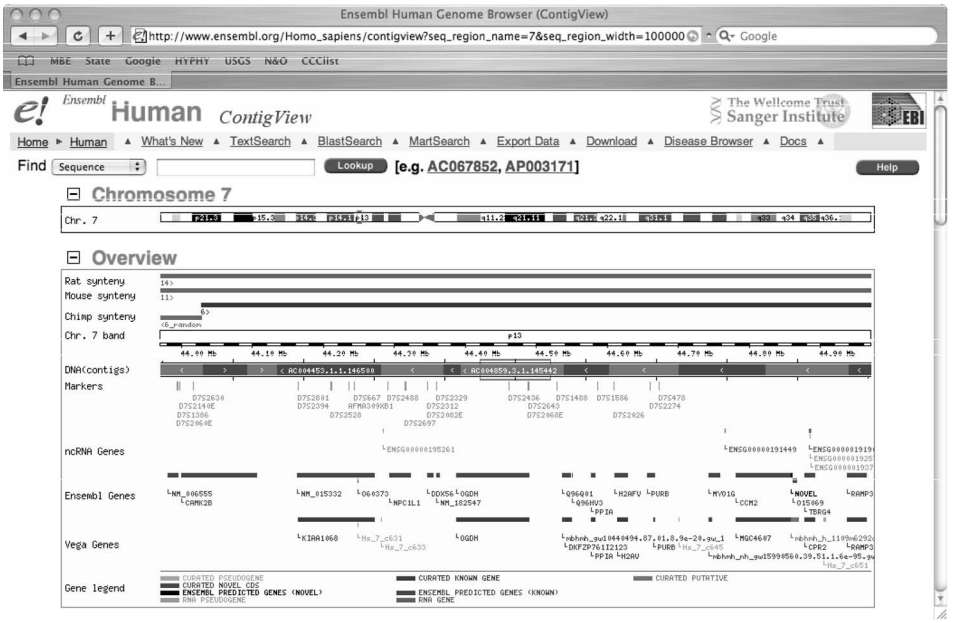

Figure 13.11 A view of a portion of human chromosome 7 in the Ensembl genome browser.

TABLE 13.2 Examples of Online Genome Resources

Comprehensive Genome Resources	
www.ncbi.nlm.nih.gov	NCBI: GenBank and many others
www.tigr.org	TIGR: Microbial genomes: Rich software collection
www.ebi.ac.uk	European Bioinformatics Institute: Databases and analysis tools
Microbial Genomes	
www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl	Comprehensive Microbial Resource (CMR) Home Page
www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html	NCBI Microbial Genomes Databases
Model Organism Genomic Databases and Resources	
www.arabidopsis.org	TAIR: The Arabidopsis Information Resource
www.wormbase.org	*C. elegans* database
flybase.bio.indiana.edu	FlyBase: A database of the *Drosophila* genome
www.yeastgenome.org	*Saccharomyces* Genome Database
genome.ucsc.edu	Genome browsers for a variety of completed genomes
Annotation and Analysis Resources	
www.sanger.ac.uk/Software/Pfam	PFAM: Protein family database and HMM software
genes.mit.edu/GENSCAN.html	GENSCAN HMM-based gene finder
www.ebi.ac.uk/arrayexpress	Microarray databases and software
www.ncbi.nlm.nih.gov/BLAST	NCBI BLAST database search server
www-hto.usc.edu/software/seqaln	Sequence alignment software and server
www.cellbiol.com	Resources and software indices

```
Strain1  ATAGTAATTA GATCTGAAAA CTTCTCGAAC AATGCTAAAA CCATAATAGT ACAGCTAAAT
Strain2  GTAGTAATTA GATCTGAAAA CTTCTCGAAC AATGCTAAAA CCATAATAGT ACAGCTAAAT
Strain3  ATAGTA---- --TCTGAAAA CTTCACGAAC AATGCTAAAA CCATAATAGT ACATCTAAAT
Strain4  GTAGTAATTA GATCT---AA CTTCACGAAC AATGCTAAGA CCATAATAGT ACAGCTAAAT
Strain5  ATAGTAATTA GATCT---AA CTTCACGAAC AATGCTAAAA CCATAATAGT ACAGCTAAAG
Strain6  ATAGTAATCA GATCT---AA CTTCTCGGAC AATGCTAAAA CCATAATAGT ACAGCTAAAC
```

Figure 13.12 A multiple sequence alignment of six HIV sequences. Indels, which are insertions or deletions of nucleotides into or from existing sequences, are indicated with the symbol ''-''.

of the species, and even to reconstruct DNA or amino acid sequences that were present in the ancestral organisms. Many methods for annotation, including assigning protein function and identifying transcription factor binding sites, rely on multiple sequence alignments as input.

To illustrate the principles underlying sequence alignment, consider the special case of aligning two DNA sequences. If the two sequences are similar, it is most likely because they have evolved from a common ancestral sequence at some time in the past. As illustrated in Figure 13.13, the sequences differ from the ancestral sequence and from each other because of past mutations. Most mutations fall into one of two classes: nucleotide substitutions, which result in these two sequences being different at the location of the mutation (Fig. 13.13a), and insertions or deletions of short sequences (Fig. 13.13b). The term *indel* is often used to denote an insertion or deletion mutation. Figure 13.13b shows that indels lead to one sequence having nucleotides present at certain positions, whereas the second sequence has no nucleotides at those positions. To align two sequences without error, it would be necessary to have knowledge of the entire collection of mutations in the history of the two sequences. Since this information is not available, it is necessary to rely on computational algorithms for reconstructing the likely locations of the various mutation events. A score function is chosen to evaluate alignment quality, and the algorithms attempt to find the pairwise alignment that has the highest numerical score among all possible alignments.

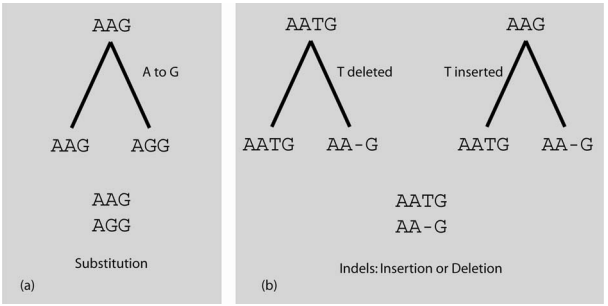

Figure 13.13 Types of mutations. A substitution (a) involves the replacement of one nucleotide with another. Note that a single alignment might be consistent with either an insertion or deletion mutation (b).

Consider aligning the two short sequences CAGG and CGA. It can be shown that there are 129 possible ways to align these two sequences, several of which are shown in Figure 13.14. How does one determine which of the 129 possibilities is best? Alignments (a) and (b) each have two positions with matching nucleotides; however, alignment (b) includes three columns with indels, whereas (a) has only one. On the other hand, alignment (a) has one mismatch to (b)'s zero. There is no definitive answer to the question of which alignment is best; however, it makes sense that "good" alignments will tend to have more matches and fewer mismatches and indels. It is possible to quantify that intuition by invoking a scoring scheme in which each column receives a score, s_i, according to the formula

$$s_i = \begin{cases} m, & \text{the bases at column } i \text{ match} \\ d, & \text{the bases at column } i \text{ do not match} \\ i, & \text{there is an indel at column } i \end{cases}$$

Using this scheme with match score $m = 5$, mismatch score $d = -1$, and indel score $i = -3$, the alignment in Figure 13.14a would receive a score of $5 - 3 + 5 - 1 = 6$. Similarly, the alignment in Figure 13.14b has a score of $5 - 3 - 3 + 5 - 3 = 1$. The remaining alignments in Figure 13.14 have scores of 0, 0, 1, 0, -5, and 1, respectively. Alignment (a) is considered best under the standards of this scoring scheme, and, in fact, it has the best score of all 129 possible alignments. This example suggests an algorithm for finding the best scoring alignment of any two sequences: enumerate all possible alignments, calculate the score for each, and select the alignment with the highest score. Unfortunately, it turns out that this approach is not practical for real data. It can be shown that the number of possible alignments of 2 sequences of length n is approximately $2^{2n}/\sqrt{2\pi n}$ when n is large. Even for a pair of short sequences of length 100, the number of alignments is 6×10^{58}, 35 orders of magnitude larger than Avogadro's number! Techniques such as the Needleman–Wunsch and Smith–Waterman algorithms, which allow for computationally efficient identifications of the optimal alignments, are important practical and theoretical components of bioinformatics.

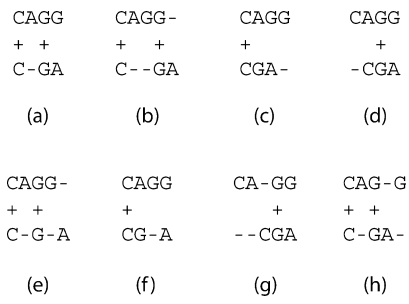

Figure 13.14 Eight of the 129 possible alignments of the sequences CAGG and CGA.

Conceptually, the task of aligning three or more sequences is essentially the same as that of aligning pairs of sequences. The computational task, however, becomes enormously more complex, growing exponentially with the number of sequences to be aligned. No practically useful solutions have been found, and the problem has been shown to belong to a class of fundamentally hard computational problems said to be NP-complete. In addition to the increased computation, there is one important new concept that arises when shifting from pairwise alignment to multiple alignment. Scoring columns in the pairwise case was simple; that is not the case for multiple sequences. Complications arise because the evolutionary tree relating the sequences to be aligned is typically unknown, which makes assigning biologically plausible scores difficult. This problem is often ignored, and columns are scored using a sum of pairs scoring scheme in which the score for a column is the sum of all possible pairwise scores. For example, the score for a column containing the three nucleotides CGG, again using the scores $m = 5$, $d = -1$, and $i = -3$, is $-1 - 1 + 5 = 3$. Other algorithms, such as the popular CLUSTALW program, use an approach known as progressive alignment to circumvent this issue.

Almost all widely used methods for finding sequence alignments rely on a scoring scheme similar to the one used in the preceding paragraphs. Clearly, this formula has very little biological basis. Furthermore, how does one select the scores for matches, mismatches, and indels? Considerable work has addressed these issues with varying degrees of success. The most important improvement is the replacement of the simple match and mismatch scores with scoring matrices obtained from empirical collections of amino acid sequences. Rather than assigning, for example, all mismatches a value of -1, the BLOSUM and PAM matrices provide a different penalty for each possible pair of amino acids. Since these penalties are derived from actual data, mismatches between chemically similar amino acids such as leucine and isoleucine receive smaller penalties than mismatches between chemically different ones.

A second area of improvement is in the assignment of indel penalties. The alignments in Figure 13.14b and 13.14e each have a total of three sites with indels. However, the indels at sites 2 and 3 of Figure 13.14b could have been the result of a single insertion or deletion event. Recognizing this fact, it is common to use separate open and extension penalties for indels. If the open penalty is $o = -5$ and the extension penalty is $e = -1$, then a series of three consecutive indels would receive a score of $-5 - 1 - 1 = -7$.

13.3.3 Database Searching

The most common bioinformatics task is searching a molecular database such as GenBank for sequences that are similar to a query sequence of interest. For example, the query sequence may be a gene sequence from a newly isolated viral outbreak, and the search task may be to find out if any known viral sequences are similar to this new one. It turns out that this type of database searching is a special case of pairwise sequence alignment. Essentially, all sequences in the database are concatenated end to end, and this new "supersequence" is aligned to our query sequence. Since the supersequence is many, many times longer than the query sequence, the resulting

pairwise alignment would consist mostly of gaps and provide relatively little useful information. A more useful procedure is to ask if the supersequence contains a short subsequence that aligns well with the query sequence. This problem is known as local sequence alignment, and it can be solved with algorithms very similar to those for the basic alignment problem. The Smith–Waterman algorithm is guaranteed to find the best such local alignment.

Even though the Smith–Waterman algorithm provides a solution to the database search problem for many applications, it is still too slow for high-volume installations such as NCBI, where multiple query requests are handled every second. For these settings, a variety of heuristic searches have been developed. These tools, including BLAST and FASTA, are not guaranteed to find the best local alignment, but they usually do and are, therefore, valuable research tools.

It is no exaggeration to claim that BLAST (http://www.ncbi.nlm.nih.gov/BLAST) is one of the most influential research tools of any field in the history of science. The algorithm has been cited in upwards of 30,000 studies to date. In addition to providing a fast and effective method for database searching, the use of BLAST spread rapidly because of the statistical theory developed to accompany it. When searching a very large database with a short sequence, it is very likely that one or more instances of the query sequence will be found in the database simply by chance alone. When BLAST reports a list of database matches, it sorts them according to an E-value, which is the number of matches of that quality expected to be found by chance. An E-value of 0.001 indicates a match that would only be found once every 1000 searches, and it suggests that the match is biologically interesting. On the other hand, an E-value of 2.0 implies that two matches of the observed quality would be found every search simply by chance, and therefore, the match is probably not of interest.

Consider an effort to identify the virus responsible for SARS. The sequence of the protease gene, a ubiquitous viral protein, was isolated and stored under GenBank accession number AY609081. If that sequence is submitted to the tblastx variant of BLAST at NCBI, the best matching non-SARS entries in GenBank (remember that the SARS entries would not have been in the database at the time) all belong to coronaviruses, providing strong evidence that SARS is caused by a coronavirus. This type of comparative genomic approach has become invaluable in the field of epidemiology.

13.3.4 Hidden Markov Models

Much work in genomic science and bioinformatics focuses on problems of identifying biologically important regions of very long DNA sequences, such as chromosomes or genomes. Many important regions such as genes or binding sites come in the form of relatively short contiguous blocks of DNA. Hidden Markov models (HMMs) are a class of mathematical tools that excel at identifying this type of feature. Historically, HMMs have been used in problems as diverse as finding sources of pollution in rivers, formal mathematical descriptions of written languages, and speech recognition, so there is a rich body of existing theory. Predictably, many successful applications of HMMs to new problems in genomic science have been seen in recent years. HMMs have proven to be excellent tools for identifying genes in newly sequenced genomes,

predicting the functional class of proteins, finding boundaries between introns and exons, and predicting the higher-order structure of protein and RNA sequences.

To introduce the concept of an HMM in the context of a DNA sequence, consider the phenomenon of isochores, regions of DNA with base frequencies unique from neighboring regions. Data from the human genome demonstrate that regions of a million or more bases have G+C content varying from 20% to 70%, a much higher range than one would expect to see if base composition were homogeneous across the entire genome. A simple model of the genome assigns each nucleotide to one of three possible classes (Fig. 13.15a): a high G+C class (H), a low G+C class (L), or a normal G+C class (N). In the normal class, each of the four bases A, C, G, and T is used with equal frequency (25%). In the high G+C class, the frequencies of the four bases are 15% A, 35% C, 35% G, and 15% T, and in the low G+C class, the frequencies are 30% A, 20% C, 20% G, and 30% T. In the parlance of HMMs, these three classes are called hidden states, since they are not observed directly. Instead, the emitted characters A, C, G, and T are the observations. Thus, this simple model of a genome consists of successive blocks of nucleotides from each of the three classes (Fig. 13.15b).

The formal mathematical details of HMMs will not be discussed, but it is useful to understand the basic components of the models (Fig. 13.16). Each hidden state in an HMM is able to emit characters, but the emission probabilities vary among hidden states. The model must also describe the pattern of hidden states, and the transition probabilities determine both the expected lengths of blocks of a single hidden state and the likelihood of one hidden state following another (e.g., is it likely for a block of high G+C to follow a block of low G+C?). The transition probabilities play important roles in applications such as gene finding.

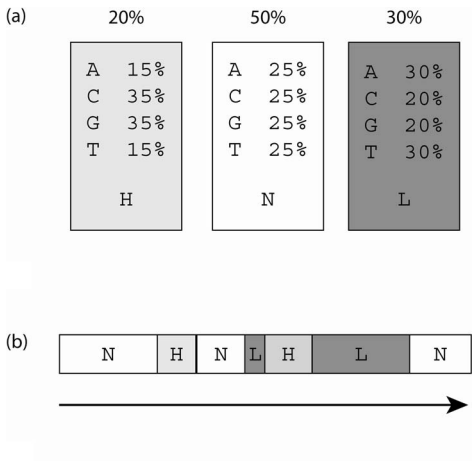

Figure 13.15 (a) Three hidden states: H, N, and L, along with their emission probabilities. In normal (N) regions of the genome, each base appears 25% of the time. Fifty percent of the genome is in a normal region. (b) The genome consists of patches of the three types of hidden states: H, N, and L.

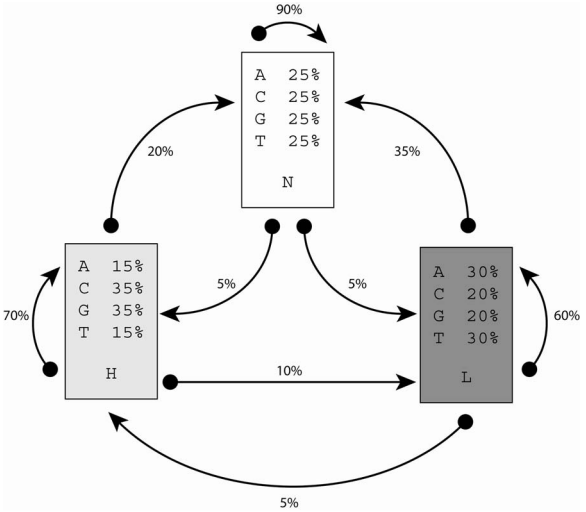

Figure 13.16 An HMM for the states in Fig. 13.15. Transition probabilities govern the chance that one hidden state follows another. For example, an N state is followed by another N state 90% of the time, by an L state 5% of the time, and by an H state 5% of the time. Emission probabilities control the frequency of the four nucleotides found at each type of hidden state. In the hidden state L, there is 30% A, 20% C, 20% G, and 30% T.

With the components of the model in place, it is possible to take a string of nucleotides and apply existing statistical methods to answer questions such as

- Where are the boundaries between the three hidden states?
- Are nucleotides 1200–1500 in a high G+C region?
- What is the most likely hidden state for nucleotide 3872?
- What is the chance of seeing a block of high G+C nucleotides shorter than 5000?

These types of questions will be addressed in the examples discussed in the next section.

13.3.5 Gene Prediction

The task of gene prediction is conceptually simple to describe: given a very long sequence of DNA, identify the locations of genes. Unfortunately, the solution of the problem is not quite as simple. As a first pass, one might simply find all pairs of start (ATG) and stop (TAG, TGA, TAA) codons. Blocks of sequence longer than, say, 300 nucleotides that are flanked by start and stop codons and that have lengths in multiples of three are likely to be protein coding genes. Although this simple method will be likely to find many genes, it will probably have a high false positive rate

(incorrectly predict that a sequence is a gene), and it will certainly have a high false negative rate (fail to predict real genes). For instance, the method fails to consider the possibility of introns, and it is unable to predict short genes. Gene finding algorithms rely on a variety of additional information to make predictions, including the known structure of genes, the distribution of lengths of introns and exons in known genes, and the consensus sequences of known regulatory sequences.

HMMs turn out to be exceptionally well-suited for gene finding, and the basic structure of a simple gene finding HMM is shown in Figure 13.17. Note that the HMM includes hidden states for promoter regions, the start and stop codons, exons, introns, and the noncoding DNA falling between different genes. Also note that not all hidden states are connected to one another. This fact reflects an understanding of gene and genome structure. The sequence of states Start – Intron – Stop – Exon – Promoter is not biologically possible, whereas the series Noncoding – Promoter – Start – Exon – Intron – Exon – Intron – Exon – Stop – Noncoding is. Good HMMs incorporate this type of knowledge extensively.

To put the HMM of Figure 13.17 to use, the model must first be trained. The training step involves taking existing sequences of known genes and estimating all of the transition and emission probabilities for each of the model's hidden states. For example, if a training data set included 150 introns, the observed frequency of C in those intron sequences could be used to come up with the emission probability for C in the hidden state intron. The average lengths of introns and exons would be used to estimate the transition probabilities to and from the Exon and Intron hidden states. Once the training step is complete, the HMM machinery can be used to predict the locations of genes in a long sequence of DNA, along with their intron/exon boundaries, promoter sites, and so forth.

Gene finding algorithms in actual use are much more complex than the one shown in Figure 13.17, but they retain the same basic structure. The performance of gene

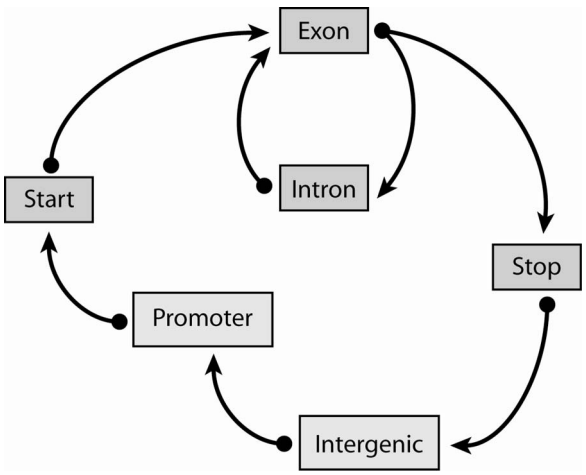

Figure 13.17 A simple HMM for gene finding.

finders continues to get better and better as more genomes are studied and the quality of the underlying HMMs is improved. In bacteria, modern gene finding algorithms are rarely incorrect. Upwards of 95% of the predicted genes are subsequently found to be actual genes, and only 1–2% of true genes are missed by the algorithms. The situation is not as rosy for eukaryotic gene prediction, however. Eukaryotic genomes are much larger, and the gene structure is more complex (most notably, eukaryotic genes have introns). The effectiveness of gene finding algorithms is usually measured in terms of sensitivity and specificity. If these quantities are measured on a per nucleotide basis, an algorithm's sensitivity, S_n, is defined to be the percentage of nucleotides in real genes that are actually predicted to be in genes. The specificity, S_p, is the percentage of nucleotides predicted to be in genes that truly are in genes. Good gene predictors have high sensitivity and high specificity. The best gene eukaryotic gene finders today have sensitivities and specificities around 90% at the individual nucleotide. If the quantities are measured at the level of entire exons (e.g., did the algorithm correctly predict the location of the entire exon or not?), the values drop to around 70%.

An emerging and powerful approach for predicting the location of genes uses a comparative genomics approach. The entire human genome sequence is now available, and the locations of tens of thousands of genes are known. Suppose that a laboratory now sequences the genome of the cheetah. Since humans and cheetahs are both mammals, they should have reasonably similar genomes. In particular, most of the gene sequences should be quite similar. Gene prediction can proceed by doing a pairwise sequence alignment of the two genomes and then predicting that positions in the cheetah genome corresponding to locations of known human genes are also genes in the cheetah. This approach is remarkably effective, although it will obviously miss genes that are unique to one species or the other. The degree of relatedness of the two organisms also has a major impact on the utility of this approach. The human genome could be used to predict genes in the gorilla genome much better than it could be used to predict genes in the sunflower or paramecium genomes.

In addition to HMMs and comparative genomics approaches, a variety of other techniques are being used for gene prediction. Neural networks and other artificial intelligence methods have been used effectively. Perhaps most intriguing, as more and more genomes become available, are hybrid methods that integrate, for example, HMMs with comparative genomic data from two or more genomes.

13.3.6 Functional Annotation

Once a genome is sequenced and its genes are found or predicted, the next step in the bioinformatics pipeline is to determine the biological function of the genes. Ideally, molecular biological work would be carried out in the laboratory to study each gene's function, but clearly that approach is not feasible. Two basic computational approaches will be described, one using comparative genomics and the other using HMMs.

Comparative genomics approaches to assigning function to genes rely on a simple logical assumption: if a gene in species A is very similar to a gene in species B, then the

two genes most likely have the same or related functions. This logic has long been applied at higher biological levels (e.g., the kidneys of different species have the same basic biological function even though the exact details may differ in the two species). At the level of genes, the inference is less accurate, especially if the species involved in the comparison are not closely related, but the approach is nonetheless useful and usually effective.

Simple database searches are the most straightforward comparative genomic approach to functional annotation. A newly discovered gene sequence that returns matches to cytochrome oxidase genes when input to BLAST is likely to be a cytochrome oxidase gene itself. Complications arise when matches are to distantly related species, when the matching regions are very short, or when the sequence matches members of a multigene family. In the first case, the functions of the genes may have changed during the tens or hundreds of millions of years since the two organisms shared a common ancestor. However, if two or more such distantly related organisms have gene sequences that are nearly identical, a strong argument can be made that the gene is critical in both organisms and that the same function has been maintained throughout evolutionary history. Short matches may arise simply as a result of elementary protein structure. For example, two sequences may have regions that match simply because they both encode alpha helical regions. Such matches provide useful structural information, but the stronger inference of shared function is not justified. Multigene families are the result of gene duplications followed by functional divergence. Examples include the globin and amylase families of genes. At some point in the past, a single gene in one organism was completely duplicated in the genome. At that point, the duplicated copy was free to evolve a new, but often related, function. Subsequent duplications allow for the growth and diversification of such families. Because of their shared ancestry, all members of a gene family tend to have similar DNA sequences. This fact makes it difficult to assign function with high accuracy when matches appear in database searches, but it often provides a general class of functions for the query sequence.

Efforts have been made to classify all known proteins into functional groups using comparative genomics. Suppose that the GenBank protein database is queried with protein sequence A and the result is that its closest match is protein sequence B. If the database is next queried using sequence B and the closest match for B is found to be sequence A, then these two proteins are said to be reciprocal best matches, and they are likely to have the same function. Likewise, if the best match to sequence A is B, the best match to B is C, and the best match to C is A, then A, B, and C are likely to have the same function. This general principle has been used to create clusters of genes that are predicted to have similar or identical functions. The COGs (Clusters of Orthologous Groups of proteins) database at NCBI (http://www.ncbi.nlm.nih.gov/COG) represents a comprehensive clustering of the entire GenBank protein database using this type of scheme.

There are many known examples of proteins or individual protein domains that have the same function or structure. The PFAM (Protein Family) database (http://www.sanger.ac.uk/Software/Pfam) includes multiple sequence alignments of almost 7500 such protein families. Using the sequence data for each alignment, the PFAM

project members created a special type of HMM called a profile HMM. This database makes it possible to take a query sequence and, for each of the 7500 families and their associated profile HMMs, ask the question, "Is the query sequence a member of this gene family?" A query to the PFAM results in a probability assigned to each of the included protein families, providing not only the best matches but also indications of the strength of the matches. Currently, about 74% of the proteins in GenBank have a match in PFAM, indicating a fairly high likelihood of any newly discovered protein having a PFAM match. PFAM is of interest not only because of its effectiveness, but also because of its theoretical approach of combining comparative genomic and HMM components.

13.3.7 Identifying Differentially Expressed Genes

A common experiment is to use microarray or oligonucleotide array technology to measure the expression level for several thousand genes under two different "treatments." It is often the case that one treatment is a control while the other is an environmental stimulus such as a drug, chemical, or change in a physical variable such as temperature or pH. Other possibilities include comparisons between two tissue types (e.g., brain vs. heart), between diseased and undiseased tissues (e.g., tumor vs. normal), or between samples at two developmental phases (e.g., embryo vs. adult). One of the primary reasons to carry out such an experiment is to identify the genes that are differentially expressed between the two treatments.

The basic format of the data from a simple two-treatment microarray experiment is the following:

Spot	Trt 1	Trt 2	Ratio
1	350	250	1.4
2	1200	1250	0.96
3	900	300	3.00
4	14,000	52,200	0.27

Spot: Location on microarray
Trt 1, Trt 2: The raw expression levels measured for each of the two treatments
Ratio: Ratio of the treatment 1 measurement to the treatment 2 measurement

Each spot on a microarray corresponds to a single gene, and in competitive hybridization experiments, a single spot usually provides measurements of gene expression under two different treatments. Note that the first column has been intentionally labeled "Spot" instead of "Gene." It is important that the same gene be used and measured multiple times; therefore, a number of different spots will typically correspond to the same gene. The final column of data is the most important for interpreting this experiment. The most extreme difference in *relative* expression levels is found at spot 4, where the gene is expressed almost fourfold higher under treatment 2. The question now becomes, "How large (or small) must the ratio be to say that the expression levels are really different?" This question is one of variability

and of statistical significance. Phrased differently, would a ratio near 0.27 for spot 4 be likely if the experiment were repeated? The data in the table do not provide the necessary information to answer this question, and this fact points out the importance of replication in experimental design. Whenever quantitative measurements are to be compared, replication is needed in order to estimate the variance of the measurements. This fundamental tenet of experimental design was largely ignored during the early history of microarray studies. Fortunately, recent work has included careful attention to experimental design and proper analysis using the analysis of variance (ANOVA). Typical experiments now include five or more replicate measurements of each gene. In order to detect very small treatment effects on levels of expression, even larger amounts of replication are needed.

13.3.8 Clustering Genes with Shared Expression Patterns

A second type of microarry experiment is designed not to find differentially expressed genes, but to identify sets of genes that respond to two or more treatments in the same manner. This type of study is best illustrated with a time course study in which expression levels are measured at a series of time intervals. Examples of such studies might involve measuring expression levels in laboratory mice each hour following exposure to a toxic chemical, expression levels in a mother or fetus at each trimester of a pregnancy, or expression levels in patients each year following infection with HIV. If plots of expression levels (y axis) against time (x axis) for each gene are overlaid as shown in Figure 13.18, it is possible to visually compare the expression profiles of genes. The desired pattern is a group of genes that tend to increase or decrease their expression levels in unison. In Figure 13.18 it appears that genes 2 and 5 have very similar expression profiles, as do genes 1 and 4.

The similarity between the expression profiles of two genes can be described using the correlation coefficient,

$$r_{X,Y} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right),$$

where X_i and Y_i are the expression levels of genes X and Y at time point i. Values near 1 or -1 indicate that the two genes have very similar profiles. When faced with thousands of profiles, the task becomes a bit more problematic. A common theme is to cluster genes on the basis of the similarity in their profiles, and many algorithms for carrying out the clustering have been published. All of these algorithms share the objective of assigning genes to clusters so that there is little variation among profiles within clusters, but considerable variation between clusters. Top down clustering begins with all genes in a single cluster, then recursively partitions the genes into smaller and smaller clusters. Bottom up methods start with each gene in its own cluster and progressively merge smaller clusters into larger ones. Clustering algorithms may also be supervised, meaning that the user specifies ahead of time the final number of clusters, or unsupervised, in which case the algorithm determines the final number of clusters.

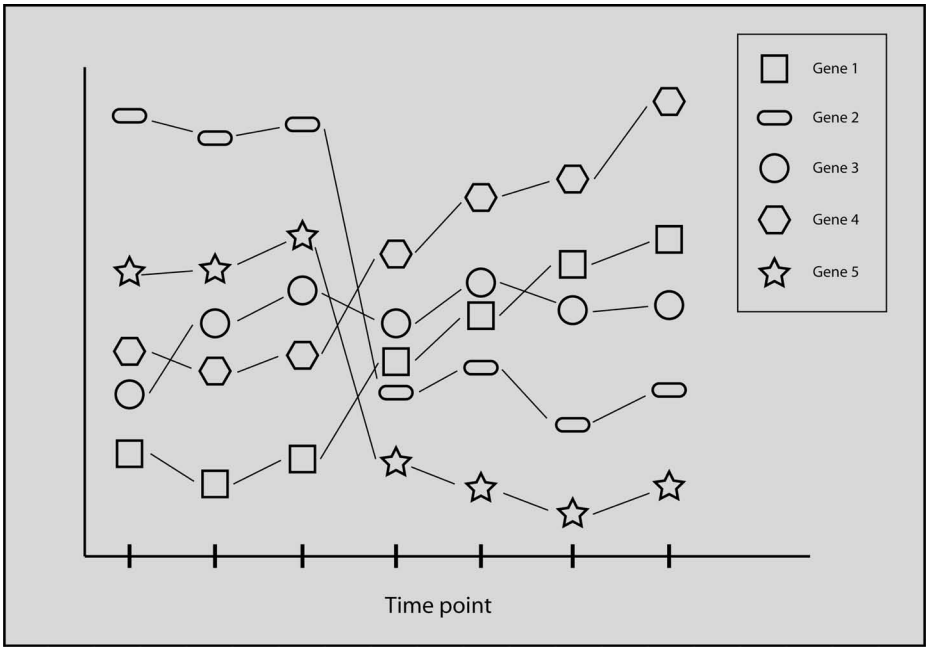

Figure 13.18 Overlaid expression profiles for 5 genes. Note that genes 2 and 5, as well as genes 10 and 4, have similar profiles, suggesting that each pair may be regulated by the same transcription factors.

13.4 CONCLUSION

The emergence of genomic science has not simply provided a rich set of tools and data for studying molecular biology. It has been the catalyst for an astounding burst of interdisciplinary research, and it has challenged long-established hierarchies found in most institutions of higher learning. The next generation of biologists will need to be as comfortable at a computer workstation as they are at the lab bench. Recognizing this fact, many universities have already reorganized their departments and their curricula to accommodate the demands of genomic science.

From a more practical point of view, the results of genomic research will begin to trickle into medicine. Already, diagnostic procedures are changing rapidly as a result of genomics. The next phase of genomics will focus on relating genotypes to complex phenotypes, and as those connections are uncovered, new therapies and drugs will follow. Consider, for example, a drug that is of significant benefit to 99% of users, but causes serious side effects in the remaining 1%. Such drugs currently have difficulty remaining in the marketplace. However, the use of genetic screens to identify the patients likely to suffer side effects should make it possible for these drugs to be used safely and effectively. Less imminent, but certainly in the foreseeable future, are gene therapies that will allow for repair of genetic defects. The continued interplay of

biology, engineering, and the mathematical sciences will be responsible for exploration of these frontiers.

Exercises

1. How many possible proteins could be formed by a gene region containing four exons?
2. In general, eukaryotes have introns, whereas prokaryotes do not. What are possible advantages and disadvantages of introns?
3. Most amino acids are encoded by more than a single codon. If one of these synonymous codons is energetically more efficient for the organism to use, what effect would that have on the organism's genome content? How might this fact be used in gene finding algorithms?
4. What is the chance that a 20-nucleotide oligonucleotide matches a sequence other than the one it was designed to match? Assume for simplicity that all nucleotides have frequency 25%. How many matches to that oligonucleotide would one expect to find in the human genome?
5. If each of the 13 SSR markers used by the FBI for identification purposes has 20 equally frequent alleles, what is the chance that two randomly chosen individuals have the same collection of alleles at those 13 markers?
6. How many mammalian genomes have been completely sequenced? What are they?
7. What is the size of the *Anopheles gambiae* genome? How many chromosomes does it have? How many genes does it have?
8. What is the length of the *Drosphila melanogaster* alcohol dehydrogenase gene?
9. Consider the following two alignments for the sequences CGGTCA and CAGCA:

```
C-GGTCA        C-GGTCA
CA-G-CA        CA-GCA.
```

   a. Find the score of each alignment using a match score of 5, mismatch penalty of −2, and gap penalty of −4.
   b. Find the score of each if the gap penalty is −5 for opening and −1 for extending.

10. Suppose a computer can calculate the scores for one million alignments per second. How long would it take to find the best alignment of two 1000 bp sequences by exhaustive search?
11. Find an example of a zinc finger gene sequence using GenBank. Use BLAST to discover how many GenBank sequences are similar to the sequence you found. What does the result tell you about zinc finger genes?
12. What are some additional features that might be added to the simple gene finding HMM of Fig. 13.17?

13. Draw a diagram of a simple gene finding HMM that might be useful for prokaryotes. The HMM should contain hidden states for exons and intergenic regions, and it should guarantee that exons have lengths that are multiples of three.

14. Use the PFAM website to give a brief description of the structure and function of members of the hamartin gene family.

15. In Fig. 13.18, gene 2 seems to be expressed at higher levels than gene 5. Justify the claim that the two genes have similar profiles and might be coregulated.

16. The expression levels for three genes measured at four times are:

Gene 1	5	7	20	22
Gene 2	15	12	18	17
Gene 3	10	15	50	45

Compute the correlation coefficient for each pair of genes. Do any of them have similar profiles?

17. The expression levels for two genes measured at four times are:

Gene 1	12	16	24	28
Gene 2	64	16	16	64

Compute the correlation coefficient for this pair of genes. Make a plot of the gene 1 levels against the gene 2 levels. Is the information in the plot consistent with the implication of the correlation coefficient? Comment on potential hazards of using the correlation coefficient based on your findings.

18. Often, the gene sequences placed on microarray slides are of unknown function. Suppose that an experiment identifies such a gene as being important for formation of a particular type of tumor. Explain how you might use resources at NCBI to identify the gene's function.

19. In Fig. 13.7, genes in a time course study are clustered according to their expression profiles. Develop an algorithm for assigning genes to clusters.

20. When carrying out a database search using BLAST with a protein coding gene as the query sequence, there are two possible approaches. First, it is possible to query using the original DNA sequence. Second, one could translate the coding DNA and query using the amino acid sequence of the encoded protein. Which approach is likely to be better? What are the advantages and disadvantages of each?

Suggested Reading

Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.

Bernardi, G. (2000). Isochores and the evolutionary genomics of vertebrates. *Gene* **241,** 3–17.

Brown, P.O. and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nat. Genetics* **21 (suppl),** 33–37.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94.

Burris, J., Cook-Deegan, R. and Alberts, B. (1998). The Human Genome Project after a decade: Policy issues. *Nat. Genetics* **20,** 333–335.

Cantor, C.R. and Smith, C.L. (1999). *Genomics: The Science and Technology Behind the Human Genome Project*. John Wiley and Sons, New York, NY.

Collins, F.S. et al. (1998). New goals for the US Human Genome Project: 1998–2003. *Science* **262,** 682–689.

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

Gibson, G. and Muse, S.V. (2002). *A Primer of Genome Science*. Sinauer Associates, Sunderland, MA.

Hartl, D. and Clark, A. (1998). *Principles of Population Genetics,* 3rd Ed. Sinauer Associates, Sunderland, MA.

Henikoff, S. and Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89,** 10915–10919.

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921.

International SNP Map Working Group. (2001). A map of human genome sequence variation containing 1.42 million single-nucleotide polymorphisms. *Nature* **409,** 928–933.

Kerr, M.K., Martin, M. and Churchill, G. (2000). Analysis of variance for gene expression in microarray data. *J. Comput. Biol.* **7,** 819–837.

Lee, C., Weindruch, R. and Prolla, T. (2000). Gene-expression profile of the aging brain in mice. *Nat. Genetics* **25,** 294–297.

Mount, D.W. (2001). *Bioinformatics: Sequence and genome analysis*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Needleman, S. and Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.* **48,** 443–453.

Scherf, U. et al. (2000). A gene expression database for the molecular pharmacology of cancer. *Nat. Genetics* **24,** 236–244.

Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197.

Sonnhammer, E., Eddy, S., Birney, E., Bateman, A. and Durbin, R. (1998). Pfam: Multiple sequence alignments and HMM profiles of protein domains. *Nucl. Acids Res.* **26,** 320–322.

Taft, R.J. and Mattick, J.S. (2003). Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. *Genome Biology* **5,** P1.

Venter, J.C., Adams, M., Sutton, G., Kerlavage, A., Smith, H. and Hunkapiller, M. (1998). Shotgun sequencing of the human genome. *Science* **280,** 1540–1542.

Venter, J.C. et al. (2001). The sequence of the human genome. *Science* **291,** 1304–1351.

Waterman, M.S. (1995). *Introduction to Computational Biology.* Chapman and Hall, London.

Wheeler, D.L. et al. (2000). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **28,** 10–14.

Willerslev, E. et al. (2003). Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science* **300,** 791–795.

Zhang, L. et al. (1997). Gene expression profiles in normal and cancer cells. *Science* **276,** 1268–1272.