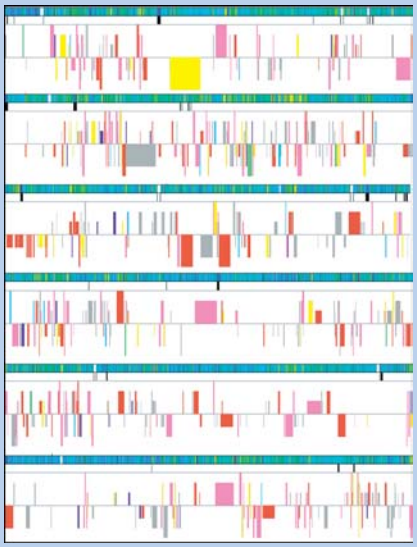
Issues

- Does genetic engineering fundamentally change the biology of an organism?
- Does gene therapy work?
- When should gene therapy be used? When should it not be used?
- Do DNA tests positively identify individuals?
- Why does the U.S. government fund the Human Genome Project?
- What benefits have been derived from the Human Genome Project?
- How could the results of the Human Genome Project be misused? How can we guard against such misuse?

Biological Concepts

- Biotechnology (The Human Genome Project; genetic engineering)
- Molecular biology (genomics; bioinformatics)
- Structure–function relationships (proteomics)

Chapter Outline



4 Genetic Engineering and Genomics

As a result of information published in 2001, humans now know more about themselves, at least at the molecular level, than they ever have before. This watershed date marked the publication of the draft of the nucleotide sequence of all of the DNA in human chromosomes. Along the way, a complete map of the location of these nucleotide sequences on the chromosomes was also produced. All of this information is stored in an enormous database that is publicly available for use by any scientist in the world. A tremendous amount of basic molecular biology has been discovered in the course of the Human Genome Project that produced this database. As a tool for biological research, this database potentially offers new ways of studying everything else in biology. In addition, the project has spawned many practical advances in biotechnology and genetic engineering.

Genetic Engineering Changes the Way That Genes Are Transferred

Genetic engineering is the direct alteration of individual genotypes. It is also called recombinant DNA technology or gene splicing, terms which are used interchangeably. Human genes can be inserted into human cells for therapeutic purposes (gene therapy, p. 100). In addition, because all species carry their genetic information in DNA and use the same genetic code, genes can be moved from one species to another. The uses of genetic engineering in plants are discussed in Chapter 11. Here we see some of the applications of genetic engineering for human medicine.

Methods of genetic engineering

Whether the 'engineered' gene is one from the same species or a different species, the techniques are much the same. All these technologies depend on being able to cut and reassemble the genetic material in predictable ways. This is possible owing to the discovery of special enzymes called restriction enzymes.

Restriction enzymes. **Restriction enzymes** are enzymes used to cut DNA at specific sites. There are several hundred restriction enzymes currently known and each cuts DNA at a different nucleotide sequence; these target sites are generally about four to eight nucleotides long (Figure 4.1). Each of these restriction enzymes is a normal product of a particular bacterial species, and most are named after the bacteria from which they are derived. Thus, in Figure 4.1, *Hae*III is an enzyme from the bacteria *Haemophilus aegypticus* and *Eco*RI is from *Escherichia coli*. They are called restriction enzymes because their normal function within the bacteria is to restrict the uptake of DNA from another bacterial species. Each species' restriction enzyme cuts the DNA from other species, but not its own, because its own DNA does not contain the nucleotide sequence that is the target site for its own enzyme.

Several other enzymes are known that can break apart a DNA molecule, but an enzyme that acts indiscriminately is of little use in genetic engineering. Restriction enzymes act specifically. Each restriction enzyme generally cuts a sample of DNA in several places, wherever the DNA contains a particular sequence of bases that the enzyme recognizes, forming a series of pieces (called restriction fragments). A given restriction enzyme mixed with the same sequence of DNA always produces the same number of fragments. The length of the pieces may vary if there are variable repeat sequences, for example, but the number of pieces and the places cut are always the same. Before the discovery of restriction enzymes, breaking chromosomal DNA into pieces was done mechanically, producing different numbers of pieces every time the procedure was done, making the results of DNA techniques impossible to reproduce from one experiment to the next. Because restriction enzymes always cut at the same sites, they can be used in genetic engineering.

Restriction enzymes in genetic engineering. The first step in inserting a gene for genetic engineering is to isolate the gene in question. This is carried out by using a restriction enzyme to snip out the desired segment of DNA. Each restriction enzyme cuts the DNA at specific places, defined by their DNA sequences. The most useful restriction enzymes are those that cut the two DNA strands at locations that are not directly across from each other, producing short sequences of single-stranded DNA known as sticky ends (see Figure 4.1). For example, the commonly used restriction enzyme *Eco*RI always targets the sequence GAATTC, cutting it between G and AATTC, breaking the two-stranded sequence into fragments that have sticky ends. The ends are called 'sticky' because they can stick together spontaneously with another molecule containing complementary sticky ends. In fragments cut with *Eco*RI, the single-stranded AATT sequences can pair with one another, stick together, and then be joined permanently. (An enzyme such as *Hae*III that cuts at sites directly across from each other forms 'blunt', rather than sticky, ends, as shown in Figure 4.1.)

If a particular restriction enzyme produces sticky ends, all fragments cut with that enzyme will have sticky ends that match one another. Thus, a fragment can be joined to any other fragment cut with the same enzyme. This makes it possible to use restriction enzymes to cut a DNA sequence and insert a functional gene with matching sticky ends.

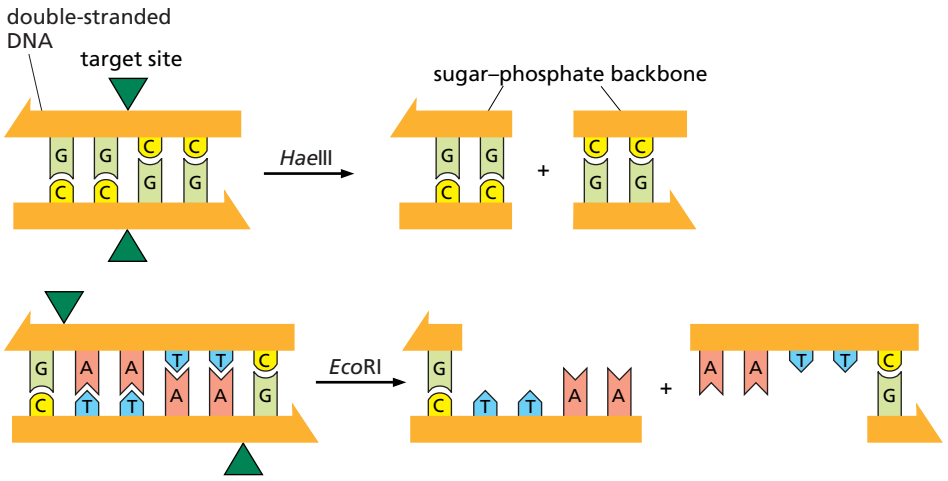


Figure 4.1

Restriction enzymes. The nucleotide sequences recognized and cut by the restriction enzymes *Hae*III and *Eco*RI are shown.

The *Hae*III target site is four bases long and the enzyme cuts the DNA strands at sites directly across from each other, leaving double-stranded ('blunt') ends.

The *Eco*RI target is six bases long and it cuts the DNA between G and AATTC. The sites on the two strands are not directly across from each other, leaving short single-stranded ('sticky') ends.

Restriction enzymes that produce blunt ends are useful in other ways, but are not useful for genetic engineering because the fragments cannot be put back together.

Cutting an entire chromosome with a restriction enzyme produces many fragments, only one of which contains the gene to be isolated. A **DNA probe** specific for the gene will isolate the fragment containing the gene of interest. As we have seen before, such a probe is a complementary DNA strand that carries a radioactive or chemical tag. The probe allows geneticists to isolate the labeled sequences, and then separate the desired genes from the DNA probes that pair with them.

A functional gene isolated in this way can then be inserted into another piece of DNA. The target DNA is cut with the same restriction enzyme, so sticky ends complementary to the fragment are available, and the gene can be incorporated permanently. So far, most genetic engineering of human genes has involved the introduction of these human genes into bacteria. The reasons for this are largely practical: many human gene products are useful in medicine but are more readily produced in large amounts inside genetically engineered bacteria than inside people. For example, the hormone somatostatin, also called growth hormone, is highly valued for the treatment of certain types of dwarfism. The hormone is, however, difficult to obtain from human sources (the traditional way is to extract it from the pituitary glands of dozens of cadavers) and is therefore very expensive. Insulin, the hormone needed by diabetics, is another example of a human gene product. Both of these hormones could be obtained from sheep or pigs or other animals, but the animal hormones are not as active in humans as the human hormones, and some patients are allergic to hormones obtained from other species. Genetic engineering provides a cost-effective way of manufacturing large amounts of these human hormones in bacteria.

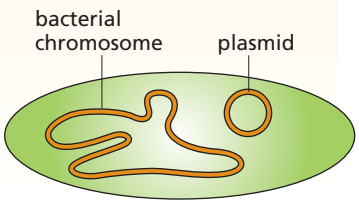
Genetically engineered insulin

Human insulin was the first commercially produced genetically engineered product. The initial step is to grow human cells in tissue culture. Tissue culture is a procedure in which cells that have been removed from an organism are grown in a dish of nutrient-rich medium kept at body temperature in an incubator. After a sufficient number of cells have grown, DNA extracted from the cell nuclei is then exposed to a restriction enzyme that cuts the DNA into desired fragments. One fragment contains the human gene for insulin, which can be isolated using a DNA probe.

The same restriction enzyme is used on nonchromosomal DNA molecules, called **plasmids**. Bacteria have a single chromosome in the form of a closed loop. Many also have a number of plasmids, short circular DNA pieces that are separate from the bacterial chromosome (Figure 4.2). Plasmids are used in genetic engineering because, being short, they have fewer sites at which a given restriction enzyme can cut. Cutting a DNA sequence in the plasmid with the same restriction enzyme that was used on the human DNA creates sticky ends that match the DNA fragment taken from the human cell. This allows incorporation of the human gene for insulin into the bacterial plasmid. The bacteria are then treated so that they take up the engineered plasmid. In most cases, the plasmid also contains another DNA sequence that can be used to

Figure 4.2

A bacterial cell showing its single chromosome and one plasmid.

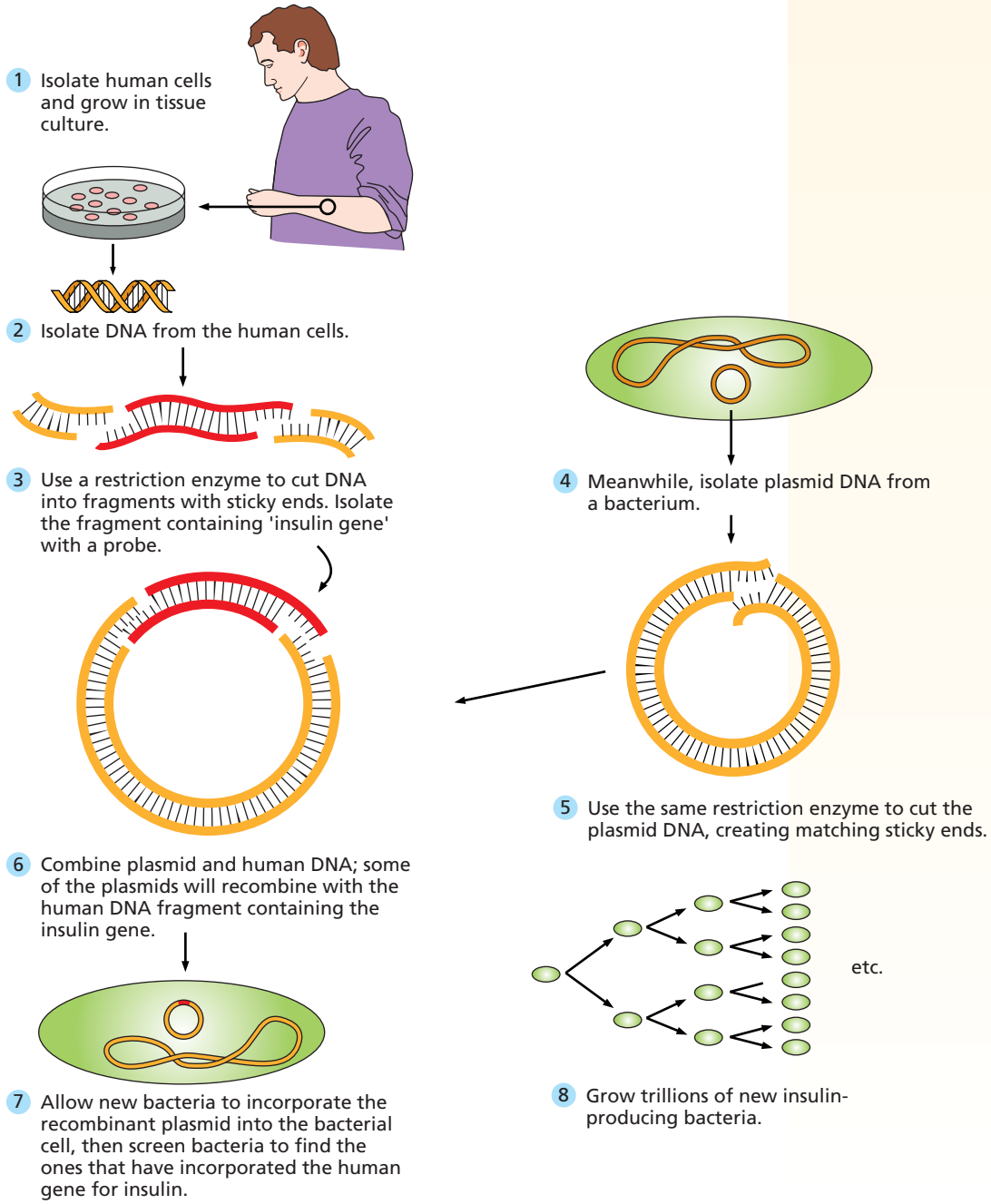


select the bacteria that have incorporated an engineered plasmid. For example, the plasmid might contain the gene for an enzyme that gives the bacteria resistance to a common antibiotic; the antibiotic can then be used to select the bacteria that have incorporated this gene while killing the majority that are still susceptible. The procedures sound easy and straightforward, but each step of the process is technically difficult and only a small proportion of the attempts succeed.

The genetically altered bacterium can now be cloned, that is, allowed to multiply asexually, which produces vast numbers of genetically identical copies of itself and its engineered plasmid. The resultant bacteria then transcribe and translate the human gene to produce human insulin (Figure 4.3). The human insulin extracted from these bacteria, called recombinant human insulin, can be given to diabetic patients.

Figure 4.3

Production of genetically engineered insulin.



Gene therapy

Instead of growing human insulin in bacteria (see Figure 4.3), genetic engineering could theoretically be used to introduce the insulin gene into human cells that do not possess a functional copy. (That would still not cure diabetes unless these cells were also capable of appropriately increasing or decreasing their output of insulin according to conditions.) This type of genetic engineering is called **gene therapy**, the introduction of genetically engineered cells into an individual for therapeutic purposes.

Treatment for hereditary immune deficiency. Human gene therapy has been used successfully to treat severe combined immune deficiency syndrome (SCIDS), a severe and usually fatal disease in which a child is born without a functional immune system. Unable to fight infections, these children will die from the slightest minor childhood disease unless they are raised in total isolation: the 'boy [or girl] in a bubble' treatment. The enzyme that controls one form of SCIDS has been identified; it is called adenosine deaminase (ADA) and its gene is located on chromosome 20. A rare homozygous recessive condition results in a deficiency of this enzyme, which in turn causes the disease.

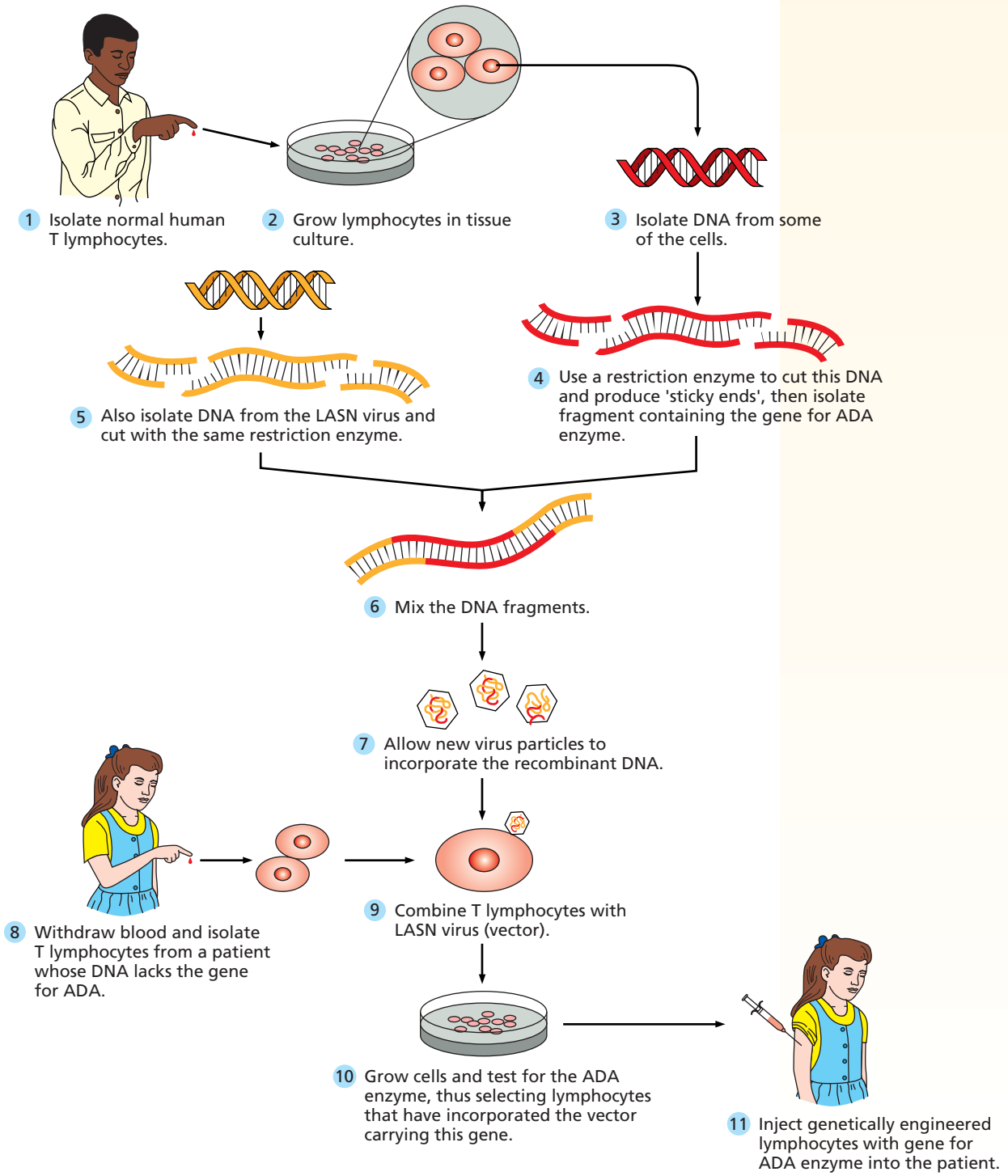
Gene therapy for this condition consists of the following procedural steps, shown in Figure 4.4.

1. Normal human cells are isolated. The cells most often used are T lymphocytes, a type of blood cell that is easy to obtain from blood and easy to grow in tissue culture.
2. The isolated cells are grown in tissue culture.
3. The DNA from these cells is isolated.
4. A restriction enzyme is used to cut the DNA into fragments with sticky ends; one will contain the functional gene for ADA. A probe with a complementary DNA sequence is then used to isolate and identify the fragments bearing the gene.
5. The same restriction enzyme is used to create matching sticky ends in viral DNA isolated from a virus known as LASN. This virus was chosen because it can be used as a **vector**: it can transfer the gene into the desired human cells—the host. (Other vector viruses have also been used; each virus type varies in the size of DNA fragment that can be inserted and the type of cell that it can enter.)
6. The viral DNA is then mixed with the human DNA fragments and allowed to combine with them.
7. The virus is allowed to reassemble itself; it is then ready for further use.
8. Blood is drawn from the patient to be treated and T lymphocytes are isolated from this blood. These lymphocytes, like all of the other cells from this person, are ADA-deficient because they do not possess a functional ADA allele.
9. The virus is now used as a vector to transfer the functional gene. The virus must get the gene not only into the lymphocyte but also into its nucleus. The gene must incorporate into the cell's DNA in a location where it will be transcribed and where it does not break up some other necessary gene sequence.

10. The lymphocytes are tested to see which ones are able to produce a functional ADA enzyme, showing that they have successfully incorporated the functional ADA allele.
11. The genetically engineered lymphocytes are injected into the patient, where they are expected to outgrow the genetically defective lymphocytes because the ADA-deficient cells do not divide as fast as cells with the ADA enzyme.

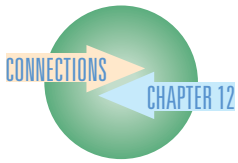
Figure 4.4

An example of gene therapy showing the transfer of the human gene responsible for adenosine deaminase (ADA).



Technical difficulties in gene therapy are numerous. Transferring large pieces of DNA into cells is difficult (most genes are large). Inserting a gene in a location in the DNA where its protein product will be transcribed and translated in a normal way is far more difficult.

The gene therapy described above provides a functional gene that is transcribed and translated by the body cells, producing the missing enzyme in lymphocytes. Because lymphocytes are not the only cells that need the ADA enzyme, the patient must also receive injections of the ADA enzyme coupled to a molecule that permits it to enter cells. (This last step might not be necessary for the treatment of other enzyme defects.) The enzyme controls the symptoms of the disease, but it is not a cure because the underlying disease is still present. Gene therapy for ADA was first successfully used on a 4-year-old girl in 1990. A second patient, a 9-year-old girl, began receiving treatments in 1991. Both patients are being closely monitored, and their immune systems are now working properly. However, because the genetically engineered cells are mature lymphocytes, which have only a limited lifetime, repeated injections of genetically engineered cells are needed.



To get around this problem, in the hope of bringing about a more lasting cure, some Italian researchers have tried using both genetically engineered lymphocytes (as described above) and genetically engineered bone marrow stem cells. Stem cells divide to form all the developed types of blood cells (see Chapter 12) and they maintain this ability throughout life. Therefore, after repaired lymphocytes die off, stem cells with repaired DNA could divide to provide new, ADA-functional lymphocytes, possibly for the lifetime of the individual. This type of therapy was begun on a 5-year-old boy in 1992, and since then several other children have received this treatment.

Questions of safety and ethics. There are legitimate safety concerns with human gene therapy. For example, any virus used as a vector must be capable of entering human cells. Might such a virus cause a disease of its own? To preclude this possibility, the viruses used in human gene therapy have been from viral strains with genetic defects that render them incapable of reproducing and spreading to other cells. Might random insertion into the host DNA destroy some other gene? Methods are being developed for directing the insertion location, but it is still largely a random event. In 1999, gene therapy clinical trials were halted in the United States when an 18-year-old boy died after receiving a viral vector for gene therapy for a metabolic disease. The reasons for his death were not apparent, so clinical studies were halted until issues of safety could be addressed. The boy's father has testified at a U.S. Senate hearing that the boy and his family were not fully informed of the dangers of the experiment. Others have raised ethical objections to the use of the term 'gene therapy' in clinical trials when most of the experiments that have been done so far have not been designed to cure any condition, only to alleviate symptoms (or to test the safety of the procedure itself).

Gene therapy also raises other ethical concerns. New recombinant DNA procedures are very expensive to develop. This raises ethical issues of fairness: will the benefits of genetic engineering be available only to

those who can afford them? Should government programs provide them through Medicare and Medicaid? Should insurance cover their use? How can society's health care resources best be distributed? If medical resources are limited, should an expensive procedure used on one person take up needed resources that could cover inexpensive treatments of other diseases for many people? These particular questions are not unique to genetic engineering; they apply to any expensive form of medical treatment.

Genetic engineering may someday become commonplace in human cells. In theory, gene therapy could be practised either on somatic cells or on gametes. If it were performed on somatic cells, the effects of the gene therapy would last as much as a lifetime, but no longer. For example, insertion of the functional allele for insulin into the pancreatic cells of patients with diabetes might cure them of the disease, but they would still pass on the defective alleles to their children. A general consensus has been reached that using gene therapy on somatic cells has an ethical value if it is used for the purpose of treating a serious disease.

If successful gene therapy is performed on germ cells, then the genetic defect will be cured in the future generations derived from those germ cells. In addition to all the ethical questions raised earlier, gene therapy on germ cells raises many additional ethical questions. Most medical ethicists today advise caution and waiting in the case of germ-cell gene therapy on humans until we have more experience with gene therapy on somatic cells or in other species.

THOUGHT QUESTIONS

1 The use of growth hormone for the treatment of shortness (not dwarfism) in otherwise healthy children is controversial, but its testing for this purpose was approved in 1993 by the Food and Drug Administration. When does a phenotypic condition unwanted by its bearer become a disease to be treated? Who decides? Should the use of human growth factor produced by engineered bacteria to increase someone's height be allowed? Is this simply another form of cosmetic surgery, similar to breast implants or face-lifts?

2 If a person dissatisfied with his or her phenotype suffers from lack of self-esteem on that account, does the lack of self-esteem justify a procedure to correct the phenotype? (This same argument is raised to justify traditional forms of cosmetic surgery.) Do parents have the right to anticipate for a child what the future effects on self-esteem will be with and without corrective procedures? For a phenotype such as height that develops over a period of years, at what age is it appropriate (if ever) to evaluate the phenotype and decide upon corrective measures?

3 A procedure such as gene therapy is expensive. Who should pay for it? Is gene therapy a limited resource? Does giving gene therapy to one patient thereby deprive another of medical care?

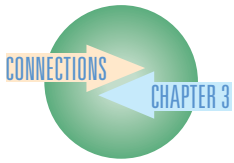
Molecular Techniques Have Led to New Uses for Genetic Information

Molecular biology is an interdisciplinary field that focuses on DNA. Although there are many other kinds of molecules, molecular biologists are concerned mostly with DNA. Molecular biology techniques can tell us a lot about human genetics, and several marker systems have now been discovered for studying human DNA. The first of these marker systems, restriction-fragment length polymorphisms, is described here. More recently other markers, with names such as expressed sequence tags, microsatellites, and single-nucleotide polymorphisms, have been discovered.

Each person has a unique DNA sequence. If it were practical to sequence a person's whole genome, his or her DNA could definitively identify a person. The human genome is far too long for it to be useful for such identification, but the DNA marker techniques that have been so useful in mapping gene regions have also proved useful in distinguishing, with a high probability, any person from another except for identical twins. Two frequent uses of this technique are in the identification of suspects in police investigations and in disputes over paternity.

The first DNA marker: restriction-fragment length polymorphisms

In 1980 a new mapping technique was devised that could readily be used in human studies, as well as in studies on other species. DNA contains, in addition to genes, noncoding regions that vary in length from one individual to another. Short sequences of nucleotides, 3–30 bases long, are repeated over and over anywhere from 20 to 100 times. These are called short tandem repeats. Several thousand different such repeats are now known in humans, each with a unique sequence not found elsewhere in the genome. When DNA containing variable numbers of repeats is cut with a restriction enzyme, fragments of DNA of various lengths are produced (Figure 4.5A). Variations (also called polymorphisms) in the lengths of the fragments produced with restriction enzymes are known as **restriction-fragment length polymorphisms**, or **RFLPs** (pronounced "riflips"). The fragments of different lengths are separated by a technique called electrophoresis (Figure 4.5B). As we saw in Chapter 3 (Figure 3.8, p. 73), because DNA carries an electric charge it moves in an electric field. When a DNA sample that has been cut into fragments is loaded onto a gel and electric current is applied, the fragments move. The gel material retards the movement of the fragments somewhat, and the larger the fragment, the more its movement is retarded by the gel. In the time that the electric current is on, smaller fragments will therefore move farther than large fragments. Because the nucleotide sequence of each short tandem repeat is unique, each can be detected by a specific probe, a piece of DNA with a sequence complementary to the repeat sequence (Figure 4.5C). Probes are specific and cause only those fragments to show up that have sequences complementary to the probe sequence.



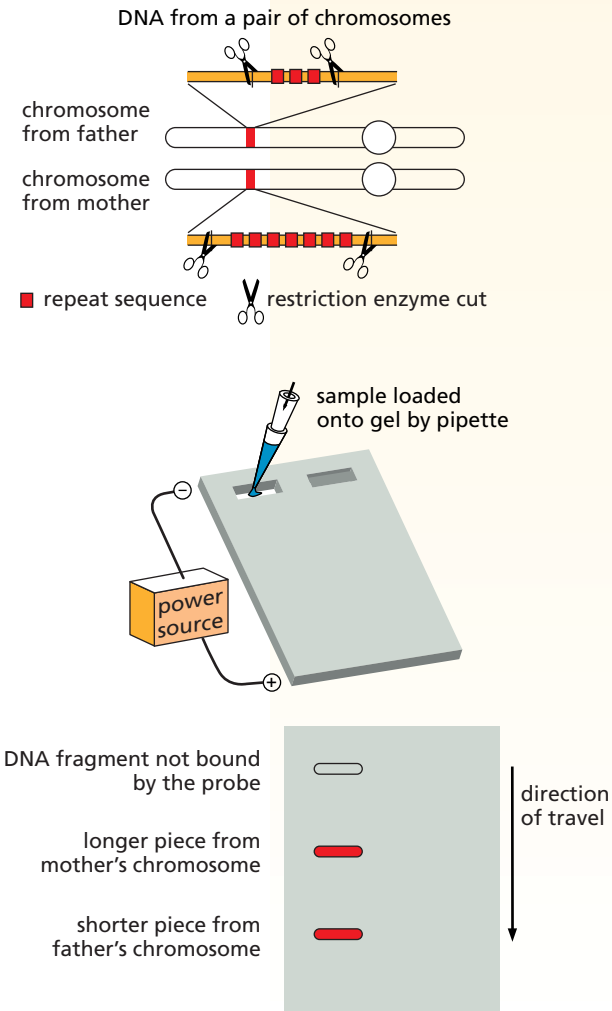
Using DNA markers to identify individuals

Using the same DNA marker techniques that we saw above, geneticists can compare DNA samples from different persons. The samples are cut with restriction enzymes. Pieces are separated according to size by electrophoresis and then transferred to a paper material. Radioactively labeled probes complementary to known DNA sequences are then used to detect the fragments containing particular variable repeats. These fragments appear as bands, with their location indicating the fragment length. Several probes can be used at once so that many bands show up, not just one or two as in the example shown in Figure 4.5, in which just one probe was used.

Bands at the same position indicate fragments of the same length in samples being compared. If the band patterns are not the same, then it can be stated with certainty that two samples did not come from the same person. In the example from a criminal investigation shown in Figure 4.6, person 1 can be eliminated as a suspect because the band pattern from the evidence is not the same as that from sample 1. The reverse is not true, however; band patterns that are the same are not an absolute guarantee that the samples came from the same individual. What are being visualized are chunks of DNA of variable lengths, not the DNA

Figure 4.5

Restriction-fragment length polymorphisms (RFLPs).



(A) CUTTING DNA WITH RESTRICTION ENZYMES

The pieces differ in length depending on the number of repeats that exist within a piece. In this example, the piece from the father is shorter because it has fewer repeats than the piece from the mother, which is longer because it has more repeats.

(B) SEPARATION BY ELECTROPHORESIS

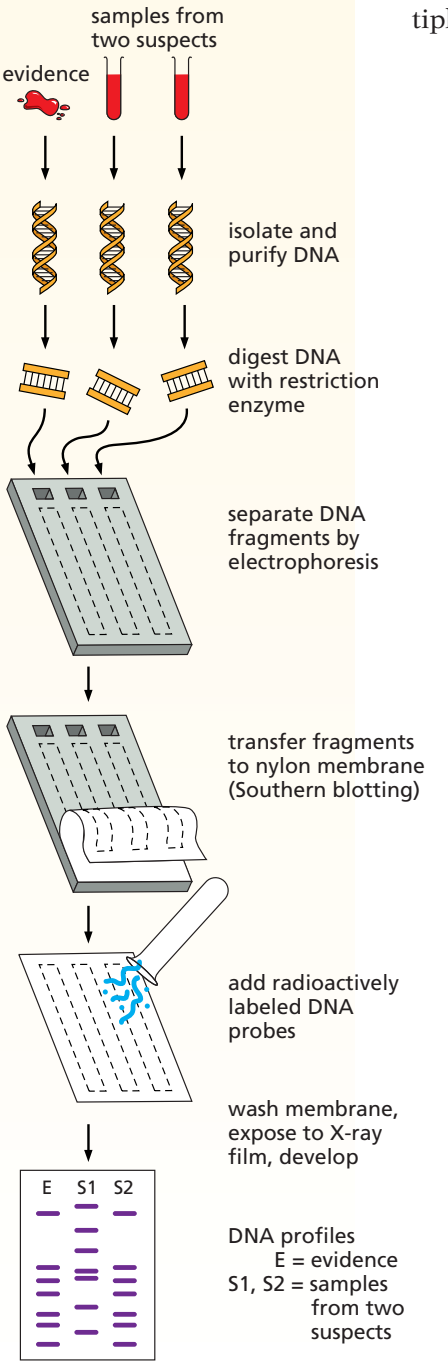
The mixture of pieces is placed on a gel and exposed to an electric field. Because DNA has a negative charge, the pieces move toward the positive electrode. In the time that the current is on, smaller pieces travel farther through the gel than the larger ones do. None of these pieces is visible yet.

(C) DETECTION WITH A PROBE

None of the pieces can be seen; however, they can be detected with a variable-repeat probe tagged radioactively or chemically (bands shown in color). The probe is a small piece of DNA with a sequence complementary to the sequence of that variable repeat, so the probe will bind to those pieces of DNA containing that variable repeat. The probe thus does two things: it identifies pieces with that specific repeat and it indicates whether the sequence is repeated a few times (to give a short DNA piece) or many times (to give a long piece). Other probes will find other sequences that are repeated in other chromosomal locations.

Figure 4.6

Forensic DNA technology. In this example, the evidence sample shows the same pattern of bands as DNA from suspect 2. There is therefore a high probability that the DNA in the evidence is from that suspect. The person from whom sample 1 was taken can be eliminated as a suspect.



sequences of the chunks. A score is calculated that indicates how likely it is that a randomly chosen person, other than the one tested, could have the same band pattern.

The likelihood that another, randomly selected person could have the same banding pattern is made very small in two ways. First, the DNA probes selected are those that pick up specific DNA markers that are rare in a given population. Also, several DNA probes are used, one after another, to produce a composite banding pattern. The probability that the bands produced with just one DNA probe are the same for two people is equal to the frequency of that DNA marker in the population. If more than one DNA probe is used, the probability of both band patterns' matching is equal to the population frequency of the first DNA marker multiplied by the population frequency of the second, and so on for multiple DNA probes and markers.

There are many ways in which the banding pattern can yield flawed or ambiguous results if samples are not properly processed. In samples from crime scenes, there is often DNA from mixed sources, including DNA from several people and from bacteria or fungi. Protein material in the sample may slow the movement of a restriction fragment in the electrophoresis, making the DNA fragment appear as though it were larger than it is. Other chemicals in the samples, such as the dyes in cloth, can interfere with the restriction enzymes cutting the DNA. However, when the tests are done properly and with the proper controls, they can be very reliable. In addition to linking suspects to material taken from crime scenes, the methods can be used to settle questions of disputed parentage. The methods can also be used to identify the dead when an intact corpse is not available, as in the aftermath of the terrorist attacks in the United States on 11 September, 2001.

Using DNA testing in historical controversies

An unusual use of this technique helped shed new light on a historical controversy involving Thomas Jefferson, the third president of the United States. DNA markers were used to investigate whether Thomas Jefferson could have been the father of children borne by one of his slaves, Sally Hemings. Two oral traditions exist: descendants of Hemings's sons, Eston Hemings Jefferson and Thomas Woodson, believe that Jefferson was their ancestor, while descendants of Jefferson's sister believe that one of her children, Jefferson's nephew, fathered Sally Hemings's later children. Researchers compared Y chromosomal DNA from descendants of two of Sally Hemings's sons with DNA from descendants of one of Thomas Jefferson's uncles. No Y chromosomal DNA was available from Thomas Jefferson's direct descendants because he had no sons who survived to have children.

The DNA data show that a set of 19 markers (collectively called the haplotype) is shared by all five of the descendants of Jefferson's uncle who were tested and by the descendants of Eston Hemings Jefferson. The haplotype is not shared by descendants of Hemings's other son, Thomas Woodson, or by the descendants of Jefferson's nephew, nor was it found in almost 1900 unrelated men. Thus, Jefferson may definitively be ruled out as the father of Thomas Woodson.

In the case of the positive match, however, the evidence supports, but does not prove, the idea that Thomas Jefferson could have been Eston Hemings Jefferson's father. As we explained earlier, positive matches indicate probabilities, not definite identity. The researchers state that because "the frequency of the Jefferson haplotype is less than 0.1%," their results are "at least 100 times more likely if the president was the father of Eston Hemings Jefferson than if someone unrelated was the father." They also state that they "cannot completely rule out other explanations of our findings," but that "in the absence of historical evidence to support such possibilities, we consider them to be unlikely." Interestingly, although the authors are very precise in the text of their article, the title, "Jefferson fathered slave's last child," overstates their results (E.A. Foster et al. *Nature* 396: 27, 1998).

THOUGHT QUESTIONS

1 Thomas Jefferson had daughters who survived to have children. Why was the DNA of their descendants not used in the study to determine the paternity of Eston Hemings Jefferson and Thomas Woodson?

2 The authors of the Jefferson study state that they "cannot completely rule out other explanations of our findings." What other explanations are biologically possible?

3 Think about the study done on DNA from descendants of Jefferson's family and Sally Hemings's sons. Why is the title of the study, "Jefferson fathered slave's last child," an overstatement of the results?

4 In the study on Jefferson's descendants, why did the researchers test DNA at 19 DNA marker sites, rather than just at one or two sites?

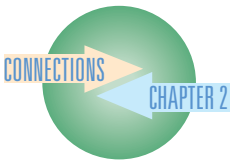
The Human Genome Project Has Changed Biology

The complete genetic material of an entire organism is known as its **genome**. In 1986, scientists proposed a project to make a genetic map, or catalogue, of a prototypical human, including the chromosomal location of all human genes and the complete DNA sequence of the genome. Many scientists and physicians think that many medical and other benefits could flow from knowing the location and sequence of all the genes. Such knowledge would facilitate locating genes that are associated with diseases or disease susceptibility. It will also make possible the development of drugs that are much more specifically tailored to block particular molecules. This effort became known as the Human Genome Project.

The Human Genome Project was funded by the U.S. Congress to begin work in the fall of 1989, and James Watson, co-discoverer of the double-helical structure of DNA, was appointed as the first director. Watson stated his belief that the Human Genome Project would tell us what it means to be human.

It should be noted, however, that although we talk of *the* human genome sequence, the DNA sequence of each person is unique. There is no one DNA sequence that is representative of every human, just as no one person could be said to represent all humans in any other method of describing people. It is estimated that one person differs from another in about 0.1% of the 3 billion base pairs in the human genome. People share the same genes but the nucleotide sequences of those genes vary in different alleles.

Sequencing the human genome



One of the stated goals of the Human Genome Project was to determine the human DNA sequence. When we read in the newspaper or hear on television about a genome being sequenced, what does this mean? The 'sequence' of DNA is the order in which the four nucleotide bases (see Chapter 2, p. 56) appear from one end of the DNA molecule to the other. Because DNA is an unbranched molecule, the sequence of bases can be 'read' from one end to the other.

Determining the order of nucleotides by using fluorescent dyes. Because the amount of DNA in even one chromosome is enormous, it is not practical to work with the whole length of a chromosome in determining sequences. The maximum size of pieces that can be sequenced is currently about 500–700 bases long. The chromosomes are therefore separated and each is cut into overlapping pieces with restriction enzymes. Each piece is inserted into a plasmid which enters a bacterium. The bacteria then divide repeatedly and make large quantities of one piece at a time, as we saw on p. 98 for bacterial production of human insulin.

The nucleotide sequence of each of the pieces can then be determined using an established method (called the di-deoxy method) based on DNA synthesis. The DNA is used as a template for synthesis of new DNA strands in a test tube, as outlined in Figure 4.7. The overall result is the production of a series of smaller pieces, each piece one nucleotide longer than the next. Each of the small pieces is then separated by electrophoresis. The pieces are made visible with a fluorescent dye, a different color used for each of the four nucleotides. Unlike the specific probes used with DNA markers, fluorescent dyes make all of the pieces visible that end in that nucleotide. The sequence of bases in the DNA fragment can thus be read from the gel: the base found at the end of the shortest piece is first (traveled farthest in the gel), followed by the base found at the end of the next longer piece (traveled the second farthest in the gel), and so forth.

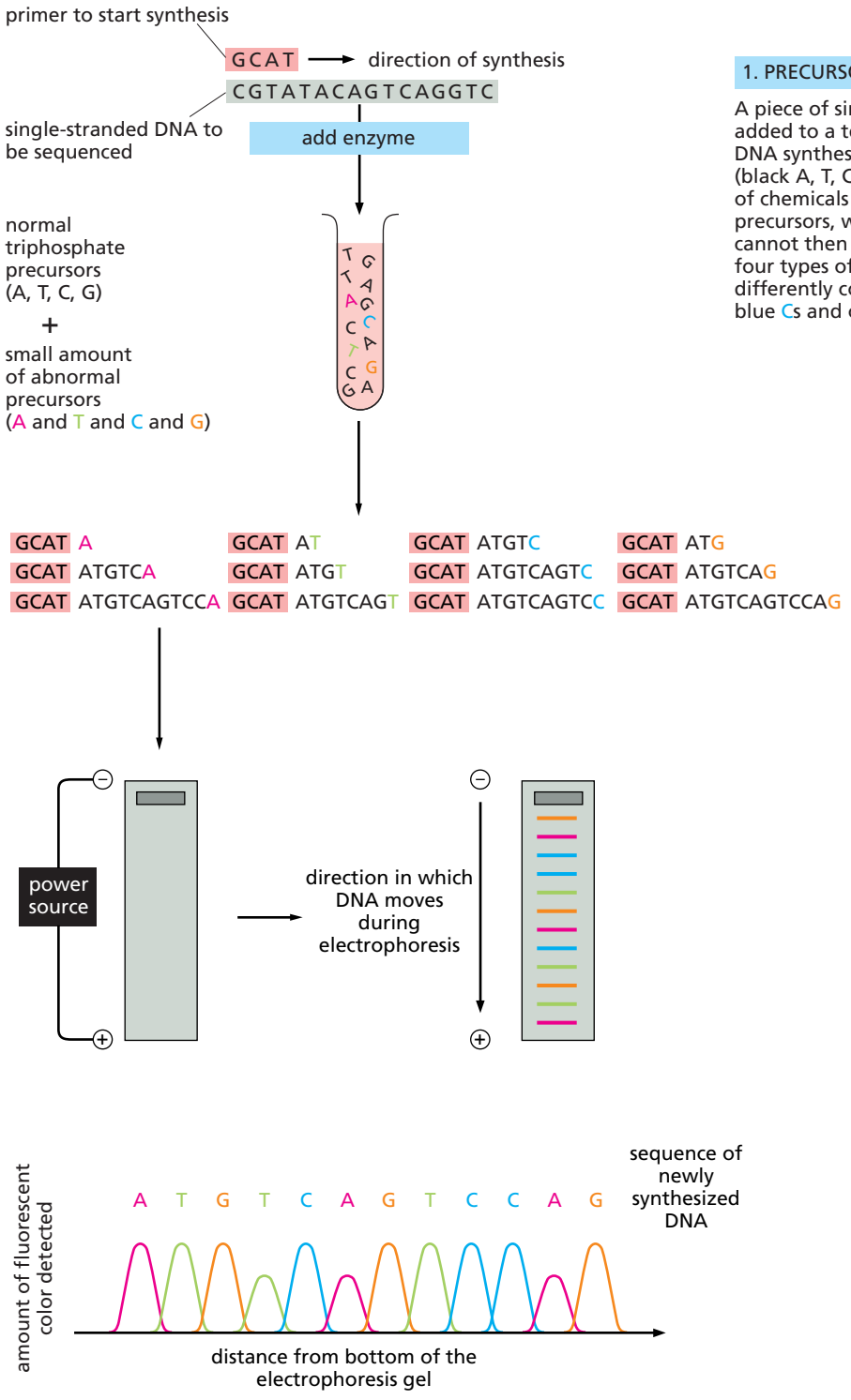
Mistakes can occur in either copying or sequencing, and repeating the process does not always give the same answer, so the technique must be repeated several times by different laboratories until a consensus sequence is established. After the sequence of each piece has been determined, the pieces must be arranged in their original order to get the overall sequence. Remember: this sequence analysis has been carried out on only one fragment of a chromosome at a time. The next challenge is to piece together the sequenced fragments, which is part of the mapping procedure discussed below.

The non-coding DNA. Most of the human chromosomal DNA does not code for genes, however, and the Human Genome Project included the

sequencing of these non-coding regions. The non-gene DNA consists of 'spacer' sequences that are never transcribed, and other kinds of sequences that are transcribed but never translated. The function of most of these non-gene sequences is currently unknown, and the wisdom of spending an estimated $15 billion on their sequencing is a question on which opinion, even among scientists, differs widely. These non-coding regions, however, have turned out to be the locations of many of the DNA markers discussed earlier, which have allowed us to find where specific

Figure 4.7

Discovering the nucleotide sequence of a piece of DNA.



1. PRECURSORS

A piece of single-stranded DNA to be sequenced is added to a test tube with an enzyme to activate DNA synthesis and the four precursor triphosphates (black A, T, C and G). Also added are small amounts of chemicals similar to each of the triphosphate precursors, which can add to the growing chain but cannot then bond to the next precursor. Each of the four types of abnormal precursors is labeled with a differently colored fluorescent dye: red As, green Ts, blue Cs and orange Gs.

2. DNA SYNTHESIS

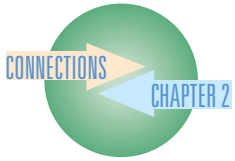
DNA synthesis is then allowed to proceed. When a normal, black precursor is added to the template, the chain keeps growing. When, by random chance, an abnormal precursor gets added instead, synthesis of that chain stops, leaving a strand shorter than the strand being sequenced. Each chain is one nucleotide longer or shorter than the others. Each short sequence ends with a fluorescently tagged molecule.

3. ELECTROPHORESIS

The pieces can then be separated by size using electrophoresis. In the time that the current is on, the fragment that consists of the primer plus a single nucleotide (A in this illustration) will travel the farthest. The fragment that is the primer plus two nucleotides (A + T) will travel not quite as far, and so forth.

4. READING THE SEQUENCE

A fluorescence detector reads each band of the gel, detecting the color of the dye labeling that band.



genes are located. Other scientists suggest that these non-coding regions will also turn out to be important for other reasons. For example, the non-coding regions are the binding sites for proteins, such as the SRY protein (see Chapter 2, p. 48), that regulate DNA folding, and thus regulate when a gene is transcribed.

The human genome draft sequence

In February 2001 two groups simultaneously announced completion of a draft of the sequence of the human genome. One group, the International Human Genome Sequencing Consortium, involving laboratories from the United States, Britain, Japan, France, Germany, and China, published their results in *Nature* (409: 860). The other group, a biotechnology company called Celera Genomics, published their results on the same day in *Science* (291: 1304). The draft covers about 94% of the estimated 3 billion bases in the complete genome. Of those 3 billion bases, 1 billion have been sequenced to completion, including all of those on the smallest paired chromosomes, chromosomes 21 and 22. The other 2 billion bases contain gaps and areas where different efforts at sequencing have resulted in different answers.

Completion of the draft sequence supported some previously established hypotheses, but also produced some surprises. Some key results are:

1. *About 95% of the human genome represents non-coding DNA, a large proportion of which is composed of repetitive sequences.* Less than 5% of the human genome is composed of genes, sequences that code for RNAs or proteins. It has been known for a while that the complexity of an organism does not correlate with the size of its genome. Much of the excess size is due to these non-coding, repeat sequences. Detailed knowledge of these sequences is opening up a new resource for studying evolution. These sequences can be likened to living fossils carried within each of us. They are already used in population genetic studies examining the migrations of human populations.
2. *The actual number of genes is smaller than previously estimated.* In humans it is difficult to predict which sequences represent genes, for reasons we discuss later. Thus, although the draft sequence of the human genome has been published, the number of genes remains unknown. The estimate of the number of genes is currently between 30,500 and 35,500. (Previous estimates had been between 50,000 and 100,000 genes.) The numbers of genes in the fruitfly (*Drosophila melanogaster*) and the roundworm (*Caenorhabditis elegans*) have been ascertained; comparisons reveal that humans are likely to have only twice as many genes as each of them.
3. *The protein products of many human genes remain unknown.* It has been found that many of the known genes can be translated in different ways to produce alternative protein variants from the same gene (see Figure 4.10, p. 117). Thus, although we have only twice as many genes as fruitflies, we may have five times as many different proteins.

4. *A very high percentage of our genes are not unique to humans but are closely similar to comparable genes from other species.* In fact, only 1% of human genes have no sequence similarity to any other organism. Our genes are similar to 46% of the genes in yeast, among the simplest organisms whose cells have a nucleus. Changes within genes over time provide clues to rates and paths of evolution.
5. *More than 200 human genes and their protein products have been found to have significant similarity to those in bacteria.* These genes are not found in intermediate organisms such as fruitflies, and one school of thought suggests that these genes jumped from bacteria to humans or vice versa.
6. *Mutation rates differ in different parts of the genome.* They are also higher in males than in females, although the reason for such a difference is not known.
7. *Within each gene, there is an average of 15 sites at which different individuals carry a different nucleotide, or at which the same individual may have a different nucleotide on each chromosome in a pair.* These variations, called single-nucleotide polymorphisms, are greatly expanding how many alleles we think are possible for different genes. In addition, these small changes may affect the physiology of the organism possessing them. Some of these polymorphisms are associated with disease; most are not, but are instead associated with small changes in protein function or regulation. Knowledge of such small-scale variations continues to challenge our concepts of terms such as 'heterozygous', 'dominant' and 'recessive', and 'allele'. It also makes it clear that there is no such thing as *the* human genome sequence. The genome sequence within each individual is unique.

In April of 2003, only two years after publication of the draft sequence, the sequence of the human genome was completed. Its publication in the journal *Nature* was timed to coincide with the fiftieth anniversary of Watson and Crick's article describing the double helical structure of DNA.

Mapping the human genome

Another goal of the Human Genome Project was to map the human genome. Mapping a species' genome means identifying the chromosomal location of each gene and the order of the genes relative to one another. Just determining the sequence of a piece of DNA does not tell you its location in the genome. The molecular techniques developed as part of the Human Genome Project have accelerated the mapping and identification of genes more generally.

One way to map a large piece of DNA is to cut the same long piece with two different restriction enzymes, derive the sequence of each of the pieces, then use computers to discover how the two sets of pieces overlap. Figure 4.8 shows how sequence data from overlapping fragments of DNA are used to derive the original order of the fragments. Figure 4.8A shows two sets of fragments of DNA produced by cutting a DNA sample with different restriction enzymes. The first restriction enzyme cut the

DNA into six pieces only; the second resulted in eight pieces. The bases in the sequences of each of the eight pieces can be lined up to match the bases in the six pieces. Can you see how you would use this idea to determine the order that the six pieces had originally been in? Now turn the page and look at Figure 4.8B.

In our example the largest piece contains 40 bases. Actual DNA pieces for sequencing are around 500 bases in length. Because the pieces are so much longer and there are so many of them, computers are needed to line up the overlaps. The accuracy of the method increases with the length of the overlapping region. The longer the sequence of the overlap between two pieces, the higher the probability that the sequence will appear only once in the genome, allowing the unambiguous assignment of the position of the two pieces relative to each other.

Celera used this approach first in 1995 with the complete sequencing of the genome of the bacteria *Haemophilus influenzae*. The same approach was used successfully on the genomes of the 599 viruses, 31 eubacteria, and 7 archaebacteria that were sequenced between 1995 and 2002. They believe that the same approach will work for mapping the human genome.

But there are obstacles to applying this approach to mapping the human genome. One obstacle is size; the human genome is about 25 times larger than any previously sequenced genome, although it is far from being the largest genome known. (One species of single-celled amoeba has a genome 200 times larger than humans!) Another obstacle to accurate reassembly is the fact that much of the non-coding DNA in the human genome is composed of repeated sequences of nucleotides. This enormously complicates the job of putting pieces into unambiguous order. Species whose genomes had previously been sequenced do not contain these repeats, so it was much easier to determine which piece went where in these genomes.

The International Human Genome Sequencing Consortium therefore used DNA markers in addition to sequence overlap to map the locations of the pieces. In the technique used by the Consortium, the total DNA in the genome was split into 29,298 overlapping large fragments with a variety of restriction enzymes. Each large piece was further split into pieces of a size that could be sequenced. Sequencing of the small pieces has been proceeding at the same time as the mapping of the large fragments, and one advantage of this approach is that different laboratories can be simulta-

Figure 4.8(A)

Combining the sequences of small pieces into the sequence of the original whole chromosome. Here are the fragments of a sequence cut with two different enzymes. Can you piece them together to reconstruct the complete sequence? Don't turn the page until you've tried it!

In this example, a DNA sequence of 150 bases is cut with two different restriction enzymes, producing the following fragments, each of which has been sequenced.

Fragments from the first restriction enzyme:

GGTCGGCTATGTAACGAGTTGCC
TCTTGTTCCTAGCTTGTCAACCGGGGATGAATGTTTACTG
CACGCGGACCGTCGGTTCAT
GTCGCAGAGCCTATTGCGAGAAGT
GCCCACCTT
TTATTGAGTTGATGCTCGACGTAGCCAGACTTAA

Fragments from the second restriction enzyme:

ACCGGGGATGAATGTTTACTGGTCGCAGAG
CCTATTGCGAGAAGTGGTCGGCTA
CTTGTCA
TGATGCTCGACGT
CGTCGGTTCAT
AGCCAGACTTAACACGCGGAC
TGTAACGAGTTGCCGCCCACCTTTTATTGAGT
TCTTGTTCCTAG

Try to piece these fragmentary sequences together and determine the entire sequence of 150 bases, before you turn the page.

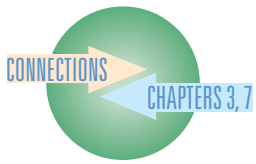
neously working on different pieces of the puzzle. Indeed, the location of each of the large fragments within the genome has now been mapped and the map is publicly available. Mapping of all of the small pieces is still proceeding.

Because Celera started with all small pieces, the Consortium maintains that Celera will not be able to reassemble the sequences of their small pieces without referring to the publicly available data posted by the Consortium. Celera maintains that because the Consortium map and sequence data are publicly available, Celera should use it to help assemble their small pieces more quickly. Why continue to insist on the slow way, when those data can now be used in a more rapid way?

The Consortium requires rapid, public disclosure of all data. Their decision to publish a draft sequence as fast as possible was driven, in their words, by "concerns about commercial plans to generate proprietary databases of human sequences that might be subject to undesirable restrictions on use" (*Nature* 409: 863). These worries have to do with the stated intentions of Celera Genomics to require others to pay for access to their databases.

Some ethical and legal issues

Many of the issues already covered in Chapter 3 regarding genetic testing will become more commonplace as molecular genetics continues to change medicine. How does an individual's right to privacy balance against family members' desire to know the results of genetic tests or an insurance carrier's or employer's desire to cover or to hire only employees who will remain healthy? How does an individual's desire to control their own reproduction balance against possible eugenic aims of society or against further stigmatization of disabled people? How can genetic counseling be value-free while providing education about genetics and not just about the testing procedure itself?



When the Human Genome Project was funded, scientists saw the need for examination of the ethical, legal, and social issues (anticipated and unanticipated) that would be raised by the research. One percent of the funding was set aside for this effort. The issues just mentioned are among those being studied, but there are many others. Social workers, anthropologists, ecologists, ethicists, and others are working together to examine the issues raised by the study of genetic variation in human populations and by the integration of genetic information into health care as well as into non-clinical settings. Others are studying the ways in which socioeconomic factors, race, and ethnicity influence people's understanding, interpretation, and use of genetic information. Simultaneously, new genetic information continues to change our concepts of race and ethnicity (see Chapter 7). Others are examining how genetic knowledge and concepts interact with different philosophical and theological traditions. Many of the working groups have composed reports with their answers to many of these questions and their guidelines for the use of genetic information. These reports are available at the Web site www.genome.gov.

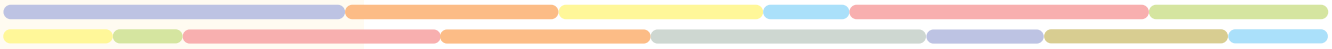
In addition, the data derived from the Human Genome Project raise questions of ownership and patent rights. Who owns the human genome

or the sequence of any particular gene? If a researcher localizes a gene to a particular chromosome, can that researcher patent the information? Can a gene sequence be copyrighted in the manner of a book? Can the genes themselves be patented? Certain biotechnology companies stand to profit greatly from the marketing of gene sequences, tests for gene sequences, or cures for various genetic diseases, but the sharing of information on gene sequences seems at first glance to threaten their competitive position. Several corporations intend to determine as many gene sequences as possible and then copyright them and sell the information at a profit. Other scientists think that the human genome should be public information, and that scientists should share this information cooperatively, particularly if public money in the form of research grants has been used in production of the knowledge. A middle ground is developing, wherein most sequences are posted in data banks with public access, but fees are sometimes charged for that access.

Figure 4.8(B)

Here is the complete sequence of 150 bases. Geneticists usually work with hundreds of fragments at once, each of them longer than this entire sequence, so the task of piecing them together is much more difficult.

When the two sets of fragments are lined up in this way, the order of the bases in the first row is the same as the order of the bases in the second row.



and the complete sequence is therefore as follows:

```
TCTTGTTCCTAGCTTGTCAACCGGGGATGAATGTTTACTGGTCGCAGAGC-
TCTTGTTCCTAGCTTGTCAACCGGGGATGAATGTTTACTGGTCGCAGAGC-
TCTTGTTCCTAGCTTGTCAACCGGGGATGAATGTTTACTGGTCGCAGAGC-

CTATTGCGAGAAGTGGTCGGCTATGTAACGAGTTGCCGCCCACCTTTTAT-
CTATTGCGAGAAGTGGTCGGCTATGTAACGAGTTGCCGCCCACCTTTTAT-
CTATTGCGAGAAGTGGTCGGCTATGTAACGAGTTGCCGCCCACCTTTTAT-

TGAGTTGATGCTCGACGTAGCCAGACTTAACACGCGGACCGTCGGTTCAT   deduced sequence
TGAGTTGATGCTCGACGTAGCCAGACTTAACACGCGGACCGTCGGTTCAT   fragments from first enzyme
TGAGTTGATGCTCGACGTAGCCAGACTTAACACGCGGACCGTCGGTTCAT   fragments from second enzyme
```

THOUGHT QUESTIONS

1. To what extent do you agree with Watson's statement that sequencing the human genome will tell us what it means to be human? Suppose you knew the exact gene sequence of part or all of your genome; what would you really know about yourself?
2. If only stretches of DNA 500–700 bases long can be sequenced at a time, how many of these small sections of DNA must be sequenced to determine the sequence of the entire human genome? (Think also about the overlaps required to piece the sequences together; assume an average of 10% overlap.)
3. Will the DNA sequence of the human genome tell us what traits are controlled by each part of the sequence? Will it tell us which sequences represent genes and which sequences represent spacers?
4. If you have a certain rare genetic condition, and scientists use cell samples from your body to determine the gene's DNA sequence, what rights (if any) does this give you to the information? Do the scientists have the right to publish your gene sequence, or any part of it? Is it an invasion of your privacy? Can the scientists sell the information? If they do, are you entitled to a share of the profits?

Genomics Is a New Field of Biology Developed as a Result of the Human Genome Project

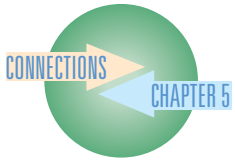
The Human Genome Project also funded the sequencing of the genomes of many other species. This may seem odd at first because the name of the project specifies the *human* genome, but there were several reasons for including these other species. The study of the genomes of species has become an entire new area of biology called **genomics**. This field has arisen to help unfold the mysteries of human genes now that the sequences and mapping are nearing completion. One focus of genomics is the identification of individual human genes. The combination of molecular biology and computer science that has been necessary to navigate through the tremendous amounts of data produced by the various genome projects is called **bioinformatics**.

Bioinformatics

Just as the NASA space program led to many unexpected 'spin-off' technologies in the 1960s and 1970s, the Human Genome Project is doing so as well, with new computer technologies and genetic engineering having wide applications outside genetics. DNA sequencing and mapping would not have been practical before the advent of large computers. Although the techniques for determining sequences of short pieces of DNA are rather simple (see Figure 4.7), finding the overlaps that indicate how the small sequenced pieces were originally arranged (see Figure 4.8) requires massive computer power. Then, when the longer sequences have been determined, storing the data has necessitated the development of larger and larger computer databases and new methods for searching them. Genomics requires the development of new types of computer software. The need for people who are trained in both molecular biology and computer science who can work with these data has made bioinformatics a fast-growing new field of employment.

One research project within bioinformatics has been the development of computer programs to locate genes within a genome. In the past, as we have seen, scientists worked from a trait, back to finding a gene. Now that the genome sequence is complete for many species and nearly complete for humans, the method of gene discovery has changed. Now people are examining the sequence data itself and trying to determine which parts may be genes, without any prior knowledge of a trait or a function for those genes. Many such genes have already been found in bacteria and yeast, and they are referred to as "orphan genes" because, at the time of their discovery, no function was known. (The later identification of their function is part of the research program of functional genomics, described later.)

Within bioinformatics, people are programming computers to scan the sequence data to locate genes, meaning areas that code for RNAs and proteins. To do so, programmers must discern 'rules' of the genetic code: what characteristics of a sequence distinguish a coding region from a non-coding region? The computer search for genes within sequences is called gene scanning.



Gene scanning in different organisms. Interestingly, most genes start with the codon ATG and end with one of three 'stop codons': TAA, TAG, or TGA. If the nucleotides A, T, G and C were distributed randomly, each of the stop codon triplets would be expected to occur on average every 4^3 or 64 bases. But nucleotides are not distributed randomly within genes; they are retained in a non-random pattern as a result of evolution because they code for a product conferring advantage to the organism. In bacteria, genes are typically 300–500 codons long, are contiguous, and do not overlap. In addition, bacteria have very little non-coding DNA. These factors make gene scanning in bacteria relatively easy. A computer can scan the sequences that follow any ATG and find those areas where the next stop codon occurs a few hundred bases further along.

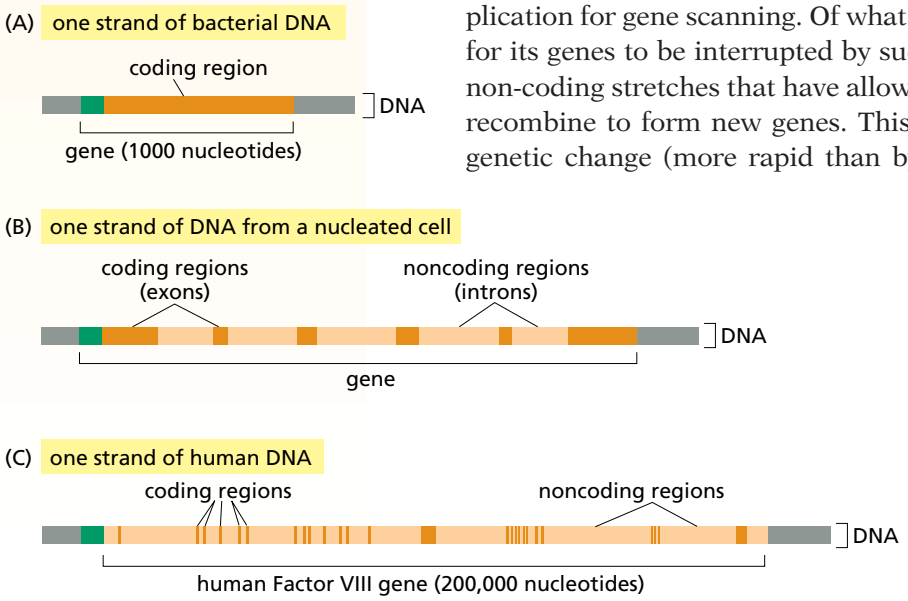
Gene scanning is much more difficult in other organisms, namely the nucleated organisms (eucaryotic organisms; see Chapter 5). In contrast to bacterial species, they have long, non-coding stretches of nucleotides (called **introns**) dispersed among much shorter regions that correspond to codons. While the coding regions (called **exons**) are roughly the same lengths in different species, the size of the non-coding introns is much greater in humans than in other species (Figure 4.9).

Within the human genome, less than 5% is located within genes; furthermore, within these human genes, only about 5% of the nucleotides comprise coding sequences. This makes it difficult to use raw sequence data to predict which nucleotide regions represent genes. Thus, gene-scanning programs are continuing to be refined to include up-to-date knowledge about the characteristics of the 'departures from randomness' within genes in species other than bacteria. One such departure from randomness is called 'codon bias:' not all codons are used equally by a given species. For example, the amino acid alanine can be coded by four different codons in humans; furthermore, three of those are used much more frequently than the fourth. In a non-coding region of the genome, all of the codons have an equal probability of being represented, but in a coding region the one codon is present less frequently.

The presence of non-coding regions within genes is clearly a complication for gene scanning. Of what benefit could it be to an organism for its genes to be interrupted by such non-coding regions? It is these non-coding stretches that have allowed the shorter coding segments to recombine to form new genes. This provides a mechanism for rapid genetic change (more rapid than by mutation). New genes are produced by the novel assembly of parts. There is another way in which the division of genes into many coding regions is adaptive, and that is in providing a mechanism by which slightly different versions of a protein can be made in different tissues, adapted to the cellular environment and function of that tissue. An example is shown in Figure 4.10. The

Figure 4.9

Single strands of DNA showing the differences in gene structures in bacteria compared with eucaryotic cells. (A) Bacterial genes contain only coding regions; that is, the DNA is all transcribed to mRNA. (B) In eucaryotic cells non-coding regions that are not transcribed are located within the coding regions of genes. (C) In humans (a eucaryotic species) the amount of non-coding DNA is much greater than the amount of DNA that codes for a protein product.



human gene for a protein called α-tropomyosin contains many coding regions scattered among non-coding regions. This gene can be transcribed to mRNA in different ways, so that in the cells of one tissue one set of coding regions is used, and in the cells of another tissue, a different set of coding regions is transcribed. This results in different mRNAs in the different cells, and therefore in slightly different proteins after translation. Each protein is still α-tropomyosin, but with a slightly different amino acid sequence and therefore a slightly different functional capability.

Although scientists think these large non-coding regions within genes are adaptive for the organism, they do present a significant obstacle to identifying genes by gene scanning. In fact, it does not appear that gene-scanning programs alone will be able to identify all of the genes in a eucaryotic genome. Hence, the Human Genome Project also funded work on the genomes of other species, so that human genes could be located by comparison with the genes of other species, a field now known as comparative genomics.

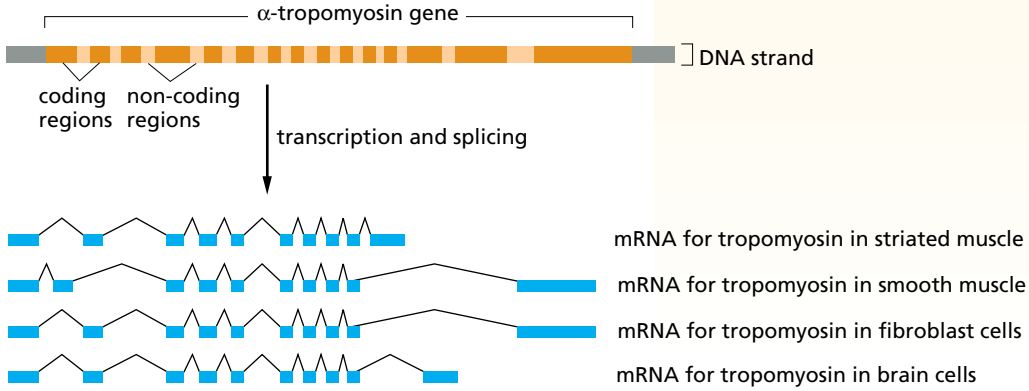


Figure 4.10
Within human genes all of the nucleotides are transcribed into RNA but only some of the RNA nucleotides are translated into protein.

Comparative genomics

When scientists compare sequences of genes from one species with those from another, they are working in the field of comparative genomics. The size of the genome of many species has been determined. As we saw earlier, the overall size does not always correlate with the complexity of the organism. This is due to the very great differences in the amount of non-coding DNA in various genomes, so that overall size does not correlate with the numbers of genes present.

As we have just discussed, genes are much easier to identify in some species than in others. Once a gene has been identified and its sequence determined in one species, there is often enough sequence similarity for its counterpart gene (or genes) to be located in other species. This is the major reason why other species' genomes were also examined as part of the Human Genome Project. Another reason was that sequencing the genomes of other species allowed scientists to develop the technology that was later used to analyze human genome sequences.

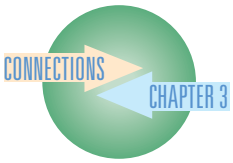
Many human genes have been located by their similarity to yeast genes. A yeast cell, like a human cell, has a nucleus and many of its genes have remained very similar to the counterpart genes in humans. Animals are even more similar and one animal that is proving to be quite useful in comparative genomics is the pufferfish, *Fugu rubripes*. Its genome is only one-seventh the size of the human genome, yet it is estimated to have the same number of genes. Because of its small genome size, gene location is much easier in pufferfish, and may subsequently allow mapping of the human counterpart genes. The mouse genome is almost complete, and

many more human genes will be found by comparison with those in the mouse. Many known genes in mice are located in the same order on their chromosomes as they are on human chromosomes, and this correspondence is extremely helpful in mapping genes. The mouse genome, however, is even larger than the human genome, so the problems of working with a large genome still pertain. See our Web site for information on the sizes of the genomes of various species (under Resources: Genome sizes).

Aside from its usefulness in locating human genes, comparative genomics has produced new data for evolutionary biologists. Species that have a common ancestor have more genes and more nucleotide sequences in common than species that do not. Unfortunately, the scientists working on a particular species have often independently devised the database for each species' genome. Consequently, another goal of bioinformatics is to devise ways of making the different databases compatible and interactive, thereby facilitating comparative genomics.

In addition to finding similarities between species, comparative genomics has led to the realization that within a species there are groups of genes that share large portions of their sequences. These 'gene families' are presumed to have evolved from a common ancestral gene. Finding one gene in the family enables the others to be located, and most often the different protein products of the family members to be identified. For example, the human hormones oxytocin and vasopressin (both proteins) belong to the same gene family, and they have very similar amino acid sequences and genes that code for them. The same is true of the oxygen-carrying proteins hemoglobin and myoglobin.

Functional genomics



In Chapter 3 we described how Archibald Garrod and other scientists studied "inborn errors of metabolism," disease conditions caused by changes in biochemical pathways. The study of similar changes in bacteria or yeast have often led to the discovery of entire chains of biochemical reactions. In the past, scientists looking for the molecules involved in such a biochemical pathway, would start with a trait and work backwards to a protein. Pedigrees such as we saw in Chapter 3 would be linked with different forms of a protein. After purifying the protein and discovering its amino acid sequence, its gene sequence could be inferred. Gene sequencing turns this whole process around. Genes are found by linkage to DNA markers, and only later is the protein product found. However, finding a gene, mapping its location, sequencing it, and even deriving the amino acid sequence of its protein product, will not tell us its function.

New sequences can be compared with those whose function is already known. This is the field of functional genomics. Species that can be easily manipulated experimentally have been most useful in discovering gene functions. The zebrafish is a vertebrate that reproduces rapidly, and many of its internal structures are visible in the living fish because overlying structures are transparent. For these reasons, zebrafish have become an experimental species of great interest to scientists working on the genetics of development. An even simpler species, the yeast *Saccharomyces cerevisiae*, has been found to share many genes with humans. Gene functions that were discovered in yeast have proved to

have parallels with disease-associated gene mutations in humans. For some examples see our Web site, under Resources: Yeast genes. The functions of the yeast genes are discovered by several methods. One is to examine which genes are transcribed to mRNA when the yeast undergoes a particular response or function; another is to inactivate (mutate) a gene and see what effect this has. Once the sequence of a gene is known, it is relatively easy to mutate it by manipulating the DNA causing a change in the protein product, which is now not functional. The opposite approach can also be used: extra copies of the gene can be inserted and observations made of the change in function under different environmental conditions. These approaches are not confined to yeast, but are also done to discover gene functions in mice and other species.

Earlier in this chapter we saw how a gene could be added to a genome, using a vector. In Figure 4.4 we saw how the functional gene for ADA was added to the genome in human cells. This method adds a gene, but in an unpredictable location within the genome. The non-functional gene is still present, and indeed one of the possible dangers of the technique is that the new gene may get added in a place that disrupts some other gene. More recently, methods have been developed for changing the sequence of a specific gene. In theory, this technique could be used to repair a non-functional gene, to mutate a gene in a specific way, or to disrupt a functional gene. A vector is used to carry into the cell a piece of DNA partly complementary to the gene to be altered. The introduced double-stranded DNA becomes substituted for the gene region as a result of crossing-over at two sites where the insert and the gene have the same sequences (Figure 4.11). If the inserted DNA is non-functional, as shown in this example, the normal gene is disrupted. The effect of deletion of that gene's protein can then be studied in the offspring cells (yeast or tissue culture of human cells). If the gene disruption is carried out on a cell from a very early stage of development, an entire organism can develop that is lacking the gene and its protein product. (This topic, part of stem cell research, is covered in more detail in Chapter 12.) Mutated mice with particular genes nonfunctional or 'knocked out', or mice with overexpressed genes produced by the insertion of additional copies of a functional gene, have led to important clues to the functions of human genes.

Families of genes have been found within species that have structurally related protein products but very often quite different functions. This has led scientists to realize that gene duplication and mutation can occur first, and that new functions can follow. Of course, a gene that is present in just a single copy cannot change to a new form (possibly with a new function) without giving up its original form and function. A duplicated gene, however, can undergo changes in one copy (possibly evolving new functions) while the other copy remains unchanged.

Proteomics

Just as the complete DNA sequence of an organism is its genome, the complete protein content of that organism is its proteome. **Proteomics** is the study of how the protein content changes over time in a cell and in an organism, how it differs in different tissues,

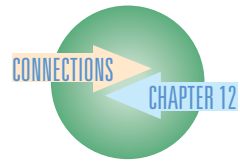
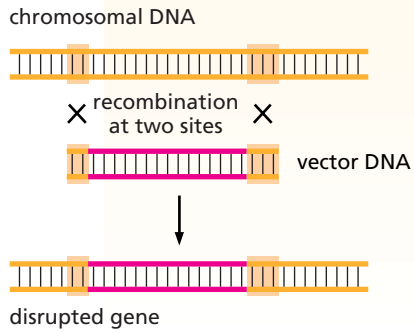


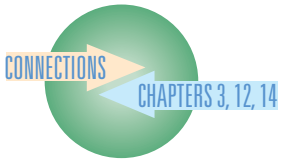
Figure 4.11

In human genes, different combinations of coding regions can be transcribed to produce different mRNAs in different tissues.



and how it relates to the health and function of the organism. Proteins are synthesized as a result of transcription of genes and translation of mRNA, as we saw in Chapter 3. However, there are further modifications to a protein after it has been translated that affect both its activity and its concentration. No protein stays in a cell forever; all are degraded and removed. We will see more about these aspects of protein function in cells in Chapter 12.

Knowing the DNA sequence of genes has hastened the discovery of the amino acid sequence of proteins. Computer programs use the known energies and bond angles of chemical bonds to turn amino acid sequence data into molecular models of the three-dimensional shape of a protein or portion of a protein. Having the ability to visualize these shapes by computer graphics has led to new strategies for the design of medicines. In the past, natural products and synthetic compounds were randomly tested in functional assays to see which would work for a particular need. Now small molecules can be designed to exactly fit a critical enzymatic site of a protein. Once the molecule has been designed by computer simulation, medicinal chemists then synthesize it and biologists test to see whether it has the desired outcome of blocking the protein's function. The action of such drugs is far more specific, and the drug will therefore have fewer side-effects than traditionally developed drugs, for reasons we will study in Chapter 14.



To synthesize a protein with even a slightly different structure can be very difficult and costly. However, once the sequence for the gene coding for that protein is known, it becomes relatively easy to modify the protein by changing the sequence of its gene. Roughly the same technique as that in Figure 4.11 is used, but the inserted piece of DNA differs from the normal piece by only a few nucleotides. Such modifications can, for example, lead to the development of proteins that are stable under a wider variety of conditions. These proteins find a variety of industrial applications. Stain removers in laundry detergents, altered enzymes for food processing, and cleanup of pollution, are just a few examples.

Rather than studying one protein change at a time, proteomics also has another goal: to study all of a cell's proteins in the aggregate. Such a goal has been unattainable in the past, and is a big factor explaining why reductionism (reducing a problem to its simplest form) has been a widespread experimental approach in biology. Proteomics is in its infancy, but promises to be a much more integrative approach.

THOUGHT QUESTIONS

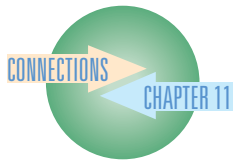
1 In what ways are humans poor subjects for genetic research? In what ways are humans good subjects? Which of your reasons are purely biological, and which have ethical components?

2 Why are certain traits studied in some species and not others?

3 Will genomics allow the findings in one species to be applied in other species? Why or why not?

Concluding Remarks

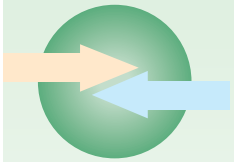
As genomics has discovered genes with useful properties within one species, genetic engineering has given us the tools to transfer those genes into another species. The Human Genome Project has discovered that transfer of genes from one species to another does also occur in nature. Viral and bacterial genes are found in the human genome, for example. Because we have almost always studied the effect of one gene or one protein at a time, transferring a gene into a new genomic environment may lead to different results from those that we expect. As we develop the tools to alter genomes, proteomics may give us the ways to study the effects of such changes throughout the cell. We also need to be mindful of effects at the level of the whole organism and effects of genetic engineering on ecosystems as well, which we will explore further in Chapter 11.



Chapter Summary

- **Restriction enzymes** cut DNA into fragments with known sequences at their ends. Restriction enzymes that produce fragments with single-stranded sticky ends are used in genetic engineering to splice new genes into genomes.
- Variations in the lengths of these fragments are called **restriction-fragment length polymorphisms** or **RFLPs**. RFLPs have helped in finding the location of many genes, as well as in the identification of individuals and in genetic engineering.
- **Genetic engineering** consists of inserting functional genes into cells, thereby altering the cell's genotype. The recipient cells may be bacterial cells that may then acquire the ability to make certain human proteins, or they may be human cells that acquire a functional allele and are injected into a patient as **gene therapy**.
- Bacterial **plasmids** are used to carry genes into a new species.
- A **genome** is the total genetic information carried by a particular organism. The Human Genome Project has now produced a draft sequence and map of the human genome.
- DNA markers of various kinds have allowed the mapping of genes to locations within the genome. Markers also allow the identification of individuals.
- **Genomics** is the study of the genome, either the comparison of genomes of different species or as a method of discovering gene functions.
- **Bioinformatics** combines computer science and molecular biology in the analysis of genomes and the identification of genes within a genome.
- **Proteomics** is the study of all of the proteins present within a cell.

CONNECTIONS TO OTHER CHAPTERS



Chapter 1	Genetic engineering and gene therapies raise many ethical issues.
Chapter 2	The genome is the blueprint for the proteome.
Chapter 3	We have learned a lot about human genetics by studying comparative genomics.
Chapter 5	Comparative genomics is a new tool for discovering the evolutionary relationships among organisms.
Chapter 6	Comparative genomics may give us important data for use in studying classifications.
Chapter 7	The amount of possible variation within each gene is much greater than was previously thought.
Chapter 11	Genetic engineering of crop species is increasing agricultural productivity.
Chapter 17	The same DNA testing techniques as those that are used to identify individual humans can also identify the bacterial species involved in new infectious outbreaks and can sometimes also identify its source.
Chapter 18	Comparative genomics is increasing our knowledge of biodiversity.

PRACTICE QUESTIONS

1. If one individual human differs from another in 0.1% of the genome, how many bases are different?

2. In the following stretch of DNA, how many fragments will result from digestion with the *Hae*III restriction enzyme shown in Figure 4.1? How many will result from digestion with *Eco*RI?

strand 1

A T C C G T A G G C C T A A C C A T C C T A G T G C

T A G G C A T C C G G A T T G G T A G G A T C A C G

strand 2

3. Why are restriction enzymes that produce fragments with 'sticky ends' more useful in genetic engineering than restriction enzymes that produce fragments with 'blunt ends?'

4. Could the following sequence be used as an insert into genomic DNA? Why or why not?

strand 1

A A G C T T A A C G G A T T A G C A A G C

C G A A T T G C C T A A T C G T T C G A A

strand 2

5. Could the following sequence be used as an insert into genomic DNA? Why or why not?

strand 1

A A G C U U A A C G G A U U A G C A A G C

C G A A U U G C C U A A U C G U U C G A A

strand 2

6. When a plasmid is being cut with a restriction enzyme in preparation for inserting a DNA fragment, the plasmid needs to be cut with the same restriction enzyme as was used to make the DNA fragment. Why?

7. Can DNA marker band patterns be used to identify maternity, as well as paternity?

8. Can DNA marker testing be used to identify individual organisms in other species besides humans?