8.2 DNA Sequencing

Figure 1 An ancient species, lake sturgeon are Canada's largest freshwater fish. They can grow up to 2 m long and weigh up to 135 kg. Today, however, specimens of this now-endangered species are much smaller.

At one time, lake sturgeon were abundant in many lakes across Canada (**Figure 1**). Currently, however, due to overfishing for their meat and caviar, sturgeon populations are on the decline in many areas. There are serious penalties for fishing and eating this endangered species of fish, and conservation officers now have a reliable tool to help them enforce the ban: DNA.

The lake sturgeon is a protected species under the federal Fisheries Act, and violators of hunting and fishing regulations can be prosecuted. If conservation officers come across campers eating fish, and they suspect that the fish is not in season or is an endangered species, they can take a small sample to a forensic lab. If the DNA matches that of a banned species, the conservation officers have sufficient evidence to give to the police.

Trent University, in Ontario, has a Wildlife DNA Profiling and Forensics Laboratory. One of the major duties of this laboratory is to track and monitor endangered species and support the enforcement of the Convention on International Trade in Endangered Species of Wild Flora and Fauna (CITES). Thanks to a procedure called the polymerase chain reaction (PCR), developed in 1983 by Kary Mullis in California, conservation officers and forensic scientists do not need a large sample of DNA in order to run extensive DNA fingerprinting tests. These tests can be used to identify the species of fish on someone's plate or the source of blood on a hunting knife, or to place a perpetrator at the scene of a crime. CAREER LINK

In this section, you will learn about several molecular biology techniques that can be used to analyze DNA and identify the species that was the source of the DNA.

The Polymerase Chain Reaction

polymerase chain reaction (PCR) a process that is used to make a huge number of copies of a DNA sequence in a laboratory, quickly and without the need for a host organism

The **polymerase chain reaction (PCR)** is a powerful process that can greatly increase the number of copies of DNA from one biological sample in just a few hours. It is important to note that a whole genome of DNA is not replicated using PCR, but a specific region of the chromosome. The PCR method is elegant in its simplicity, making it reliable, fast, and inexpensive; thus, it is in widespread use. PCR has made an enormous difference to molecular biology, and consequently to society. For his innovation, Kary Mullis was awarded the Nobel Prize in Chemistry in 1993.

PCR is DNA replication of a specific region of the genome, outside of the nucleus of a cell, in a laboratory setting. The whole process takes place in a microfuge tube (a small test tube) and consists of three steps: denaturation, annealing, and elongation. These steps are repeated in many cycles, usually 30 to 40 (**Figure 2**, next page). WEB LINK

In the first step, called denaturation, a double-stranded DNA molecule is denatured, or separated, into its two single strands. This occurs when the DNA strand is heated to 94 to 96 °C for 20 to 40 s, causing the hydrogen bonds holding the two strands together to break. (Recall that the purines and pyrimidines in DNA are held together by hydrogen bonds.)

DNA primer a short single-stranded DNA sequence, easily synthesized in a laboratory, that is complementary to a sequence at one end of the target sequence

The second step, called annealing, takes place at a lower temperature, from 50 to 65 °C, for 20 to 40 s. This step facilitates the annealing of synthesized single-stranded **DNA primers** to each of the separated single strands. The primers contain the nucleic acid sequence that is complementary to one end of the targeted fragment of DNA. Two DNA primers are used: one must be complementary to a short portion of DNA on one strand at, or near, a 3′ end of the target sequence, and the other primer must be complementary to a portion of DNA on the other strand at the opposite 3′ end of the target strand. This results in both primers being oriented in the 5′ to 3′ direction toward the target sequence. The specificity of the primers results in the amplification of only the target sequence of the DNA, making the PCR technique highly specific.

In the third step, called elongation or extension, two new DNA strands are synthesized at a temperature of 72 °C. The two original DNA strands, which were separated

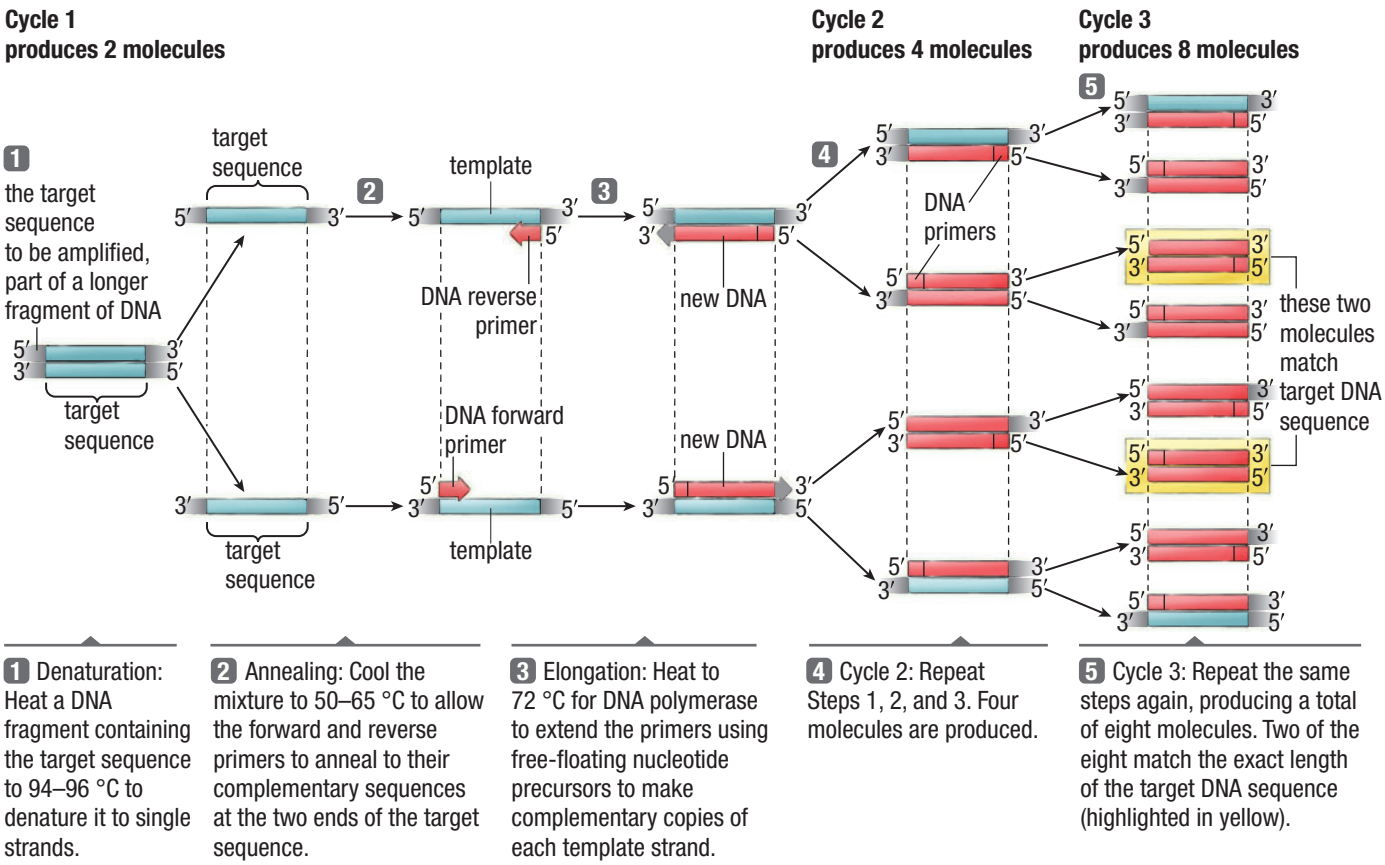

Figure 2 The PCR process has three steps (1, 2, and 3), which are repeated (4 and 5). Beginning with a DNA fragment that contains the target sequence, the process ends up with DNA fragments that exactly match the target sequence, with no extra base pairs.

in the first step, now act as templates. DNA polymerase binds to the primer-template hybrids and moves along each template, adding nucleotides that are complementary to the sequence until it reaches the 5′ end of the template.

A very specific DNA polymerase, called Taq polymerase, is needed for PCR elongation. Taq polymerase is extracted from the bacterium *Thermus aquaticus*. This bacterium is found in hot springs, so its enzymes can withstand high temperatures. The use of Taq polymerase is vital since it is not denatured by the high temperature that is used during the first step of PCR. This would not be the case with DNA polymerase from a typical organism.

After the two double-stranded DNA sequences have been made (at the end of Step 3 in Figure 2), the process can run through another cycle. The temperature is elevated back to 94 °C, and the strands are denatured again. Four strands are used as templates and duplicated to produce new copies. This cycle is repeated 30 to 40 times, resulting in a high number of copies of the specific target sequence.

By the third cycle, two of the eight DNA molecules exactly match the target molecules. As shown in Figure 2 (Step 5), the small piece of DNA at each end of the target sequence is eventually lost, and the target DNA sequence vastly outnumbers any other sequence. Within 20 cycles, the PCR reaction delivers over one million copies of the DNA target sequence. Since each cycle, regulated simply by increasing and decreasing the temperature, takes only about 5 min, millions of copies of a DNA sequence can be synthesized in less than 2 h! PCR is such an efficient process that it is used almost exclusively to amplify a target DNA sequence. As mentioned earlier, it is the process that is used by Fish and Wildlife personnel to identify a species of fish from a minute sample. The equipment that performs the rapid temperature changes and the various enzymes and primers are all available commercially.

WEB LINK

As you have learned, PCR can make millions of copies of a single segment of DNA very quickly. This means that only a very small amount of a biological sample is required. There is enough DNA in a hair follicle, a tiny blood spatter, or a tiny piece of fish tissue to perform PCR. Once the DNA of interest has been amplified using PCR, it is analyzed using different methods. These methods are described on the next few pages.

Gel Electrophoresis

gel electrophoresis a method for separating large molecules, such as DNA, RNA, and proteins

Once the targeted DNA sequence has been amplified by PCR, it needs to be separated and purified from the rest of the contents in the test tube. A chromatography-like process, called **gel electrophoresis**, uses the physical and chemical properties of DNA to separate the fragments.

Gel electrophoresis separates nucleic acids and proteins by their rate of migration through a gel. Agarose gel, made from agar (a polysaccharide from seaweed), is normally used for nucleic acids. Negatively charged DNA fragments travel through the pores of the gel, away from a negative electrode at the starting end and toward a positive electrode at the destination end. Recall that DNA is negatively charged, due to the phosphate groups, and is therefore attracted to a positive electrode. Because of the negative charge that is carried by the phosphate groups, each nucleotide is ionized. This means that nucleic acids have about the same charge-to-mass ratio, even if they differ in length. The smaller fragments of nucleotides move more easily, and hence faster, through the pores of the gel than the longer fragments do. Therefore, DNA fragments can be separated by size. The smaller the fragment, the farther it travels in a given period of time.

Figure 3 shows the four steps in gel electrophoresis. The numbered steps here refer to the numbers in Figure 3.

Step 1. Prepare a gel (a thin slab of agarose), and place it in a gel box between two electrodes. The gel has wells for placing the DNA samples to be analyzed. Add a buffer to cover the gel.

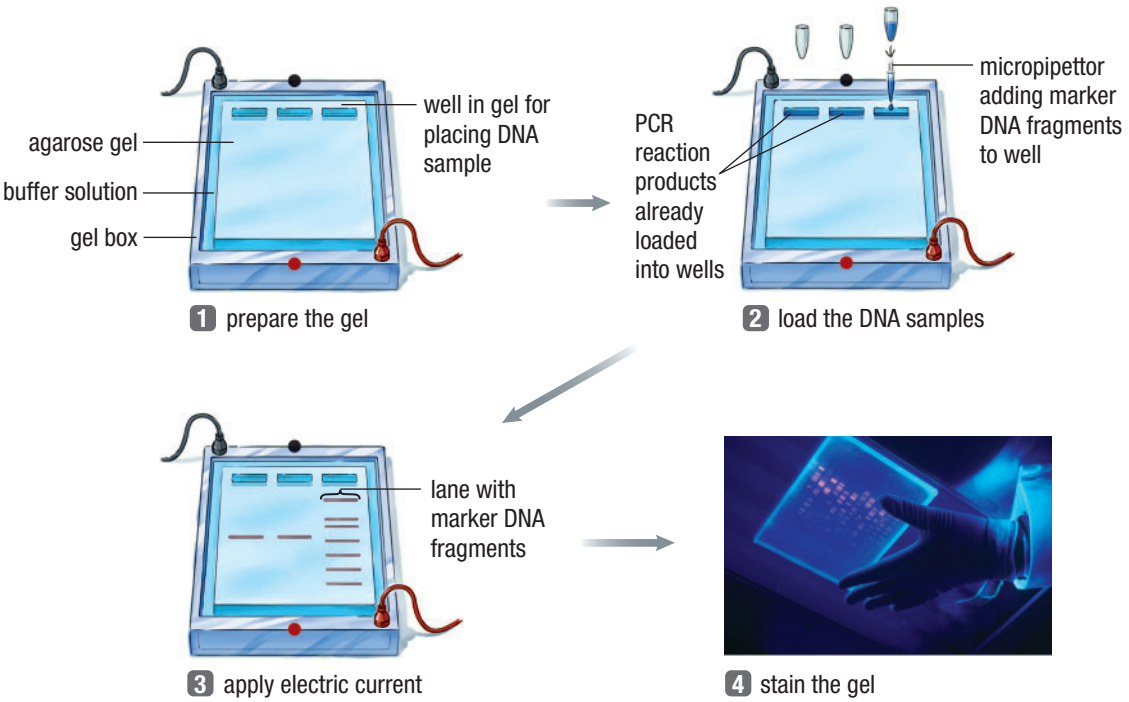

Figure 3 The four steps in the separation of DNA fragments by agarose gel electrophoresis

Step 2. Load DNA sample solutions, such as PCR products, into the wells of the gel, beside a well loaded with DNA "marker" fragments. A **molecular marker** is a DNA fragment of known size that is used to determine the lengths of unknown fragments by comparison. Multiple copies of many different-sized markers are run in a separate lane at the same time for comparison. The sizes of the fragments in the sample are determined by comparing their travelling distances to the travelling distances of the known molecular markers.

molecular marker a fragment of a known size that is run as a comparison standard for gel electrophoresis

A special dye, called a loading dye, is added. The dye has two functions. First, because it is denser than the DNA, the dye keeps the DNA from floating out of the well. It sinks, along with the DNA, into the well. Second, the dye travels slightly faster than the fastest nucleotide fragments. The rate at which the dye is seen to move can be used to judge the pace of the electrophoresis process.

Step 3. Apply an electric current to the gel. DNA fragments are negatively charged, so they migrate toward the positive pole. Shorter DNA fragments migrate faster than longer DNA fragments. When the electric current is stopped, fragments of equal length will have travelled equal distances. The result will be clusters of equal-sized fragments forming invisible bands within the gel.

Step 4. To be able to see the DNA fragments once the process is complete, stain the electrophoresis gel. **Ethidium bromide** is one compound that can be used to stain the gel. It inserts itself inside the DNA double-helical ladder structure (see Section 7.5, page 343, Figure 4). When the gel is viewed under a UV light, the bands of DNA are visible because the ethidium bromide glows, or fluoresces.

ethidium bromide a large molecule that resembles a base pair, which allows it to insert itself into DNA; used for staining electrophoresis gels

Using gel electrophoresis, researchers can compare the size of the DNA fragments from an unknown sample to the size of the DNA fragments from a known sample to reach conclusions about the identity of the unknown sample. These conclusions may include the identity of the tissues of an endangered species, such as the lake sturgeon, or the identity of the perpetrator of a crime or the father in a paternity test. Researchers may also extract a particular DNA fragment from the gel. The extracted fragment can be purified and analyzed to determine its exact base-pair sequence.

Sequencing DNA

The development of DNA technologies for analyzing genes has revolutionized experimental biology. **DNA sequencing** techniques (analyzing and determining the base sequences of DNA) have made it possible for researchers to analyze the base sequences of cloned genes and DNA fragments amplified by PCR. Having the complete sequence of a gene helps researchers understand how the gene functions. Using just the nucleotide sequence, they can identify mutations, locate regulatory sequences and gene sequences, and compare homologous genes of different species.

DNA sequencing a process in which the sequence of base pairs in a DNA strand is determined

Many scientists have worked to develop methods for sequencing DNA (Section 8.5). One efficient method is called the chain termination method. It was developed by Frederick Sanger at Cambridge University in the 1970s. It relies on the addition of a labelled dideoxynucleotide (ddNTP) to a growing DNA strand. This nucleotide, when incorporated into the DNA, prevents the binding of the next nucleotide, thus terminating the elongation of the newly synthesized DNA strand. The labels on the dideoxynucleotides are dyes that fluoresce and can be used to identify the specific base when exposed to laser light.

For the chain termination method in the laboratory, four reaction tubes are set up, each with many copies of normal deoxyribose nucleotides (dATP, dTTP, dGTP, and dCTP), multiple copies of the DNA to be sequenced,

DNA polymerase, and a DNA primer. Only one of the labelled dideoxynucleotides (ddATP, ddTTP, ddGTP, or ddCTP) is added to each tube (**Figure 4**). This dideoxynucleotide is in a much smaller concentration than the normal nucleotide. One tube contains ddATP, another ddTTP, another ddGTP, and the final one ddCTP. Synthesis of a complementary chain produces DNA fragments of different sizes, each with their last nucleotide labelled.

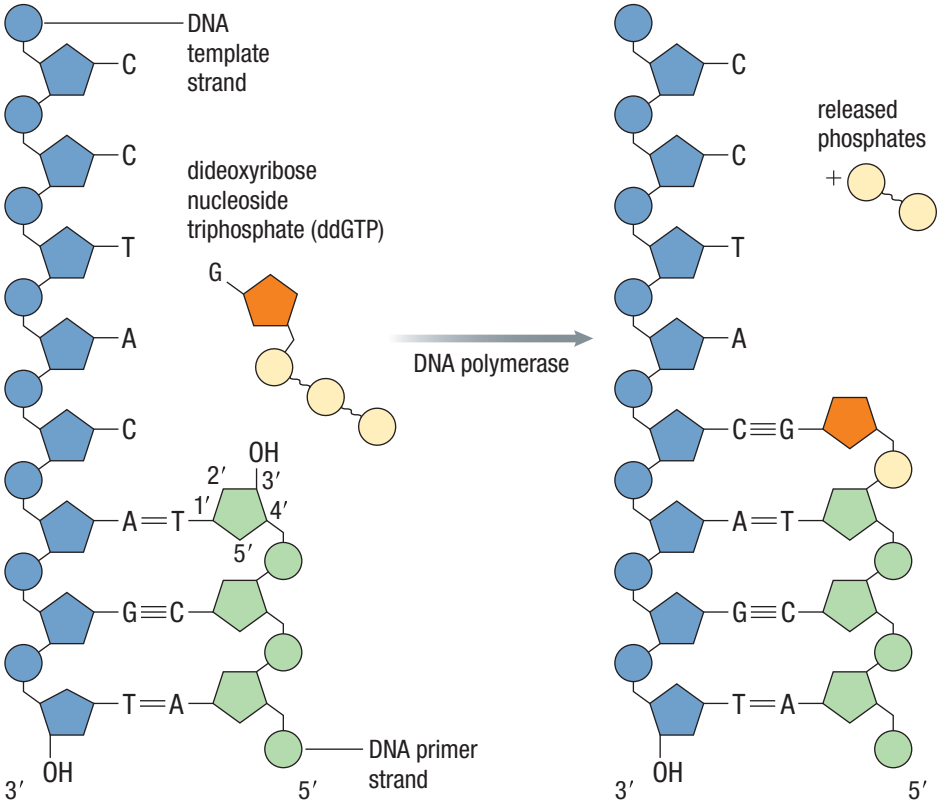

Figure 4 A complementary DNA strand can be terminated during Sanger Dideoxy sequencing. In the diagram on the right, synthesis of the DNA chain has terminated since there is no –OH group on the 3′ carbon for DNA polymerase to add another nucleotide.

The contents of the four reaction tubes are combined and then separated by gel electrophoresis with a resolution of only a single nucleotide. The sequence can then be determined by reading the order of the labelling type that corresponds to the specific dideoxynucleotide. This process is now computer automated (**Figure 5**, next page).

The entire human genome has now been sequenced, along with many other genomes. Many scientists worked on the project. They used two techniques, both involving the Sanger method for DNA sequencing. However, one technique, called the **whole-genome shotgun method**, proved to be faster than the other. This method was developed by Craig Venter of Celera Genomics. You will read more about Frederick Sanger and Craig Venter in Section 8.5.

whole-genome shotgun method a DNA sequencing method that involves blowing DNA strands into many fragments and then using computer technology to sequence and reassemble the fragments

In the shotgun method, many copies of the DNA are randomly cut into tiny fragments by passing the DNA through a pressurized syringe. The fragments are cloned in plasmids and then sequenced using the Sanger method. Computer software is used to determine the overall sequence by analyzing the many partial sequences (**Figure 6**, next page). The shotgun method is analogous to tearing 10 copies of a book randomly into smaller sets of a few pages each and, by matching overlapping pages, reassembling a complete copy of the book with the pages in the correct order.

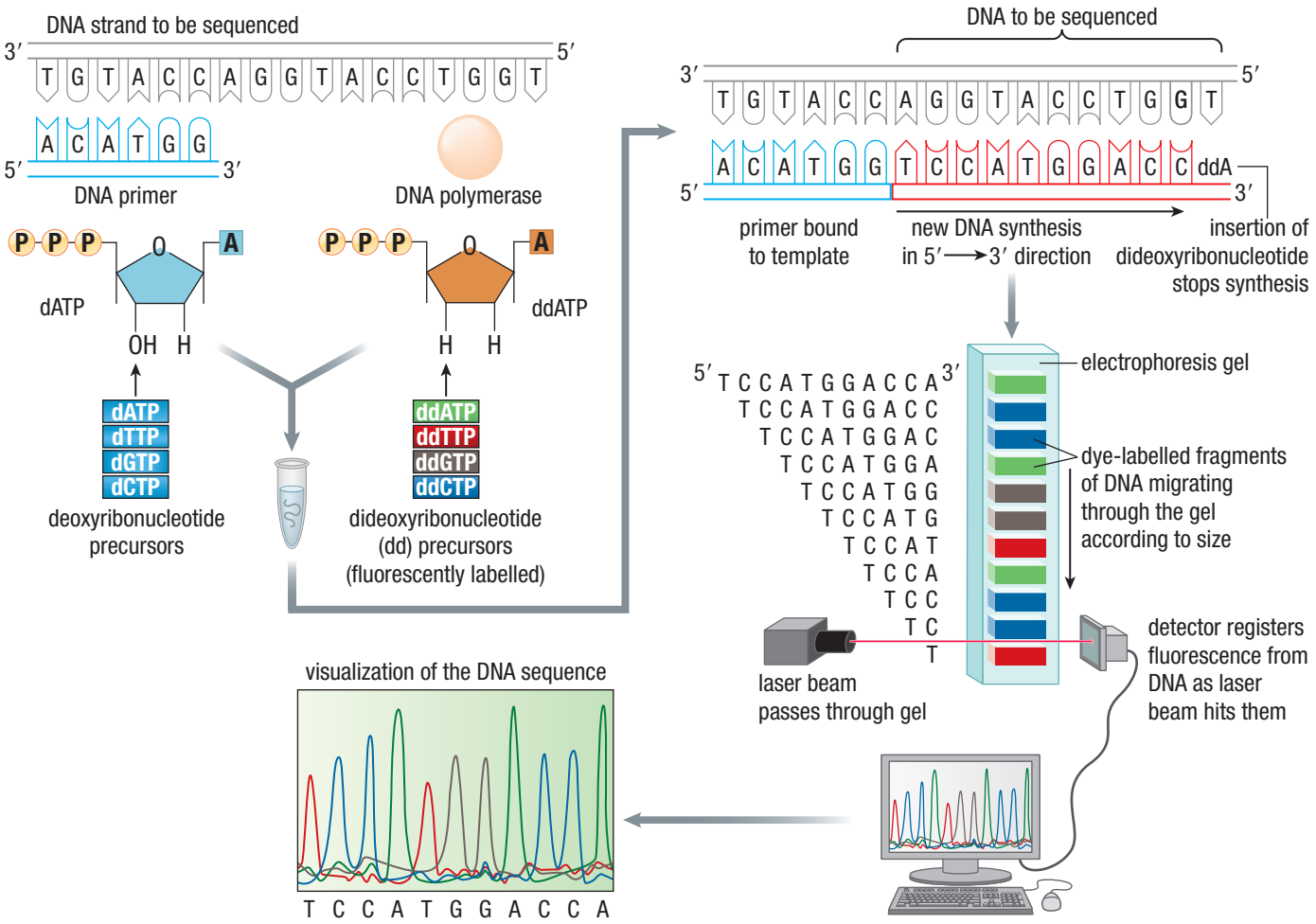

Figure 5 This diagram shows the chain termination (Sanger) method for sequencing DNA.

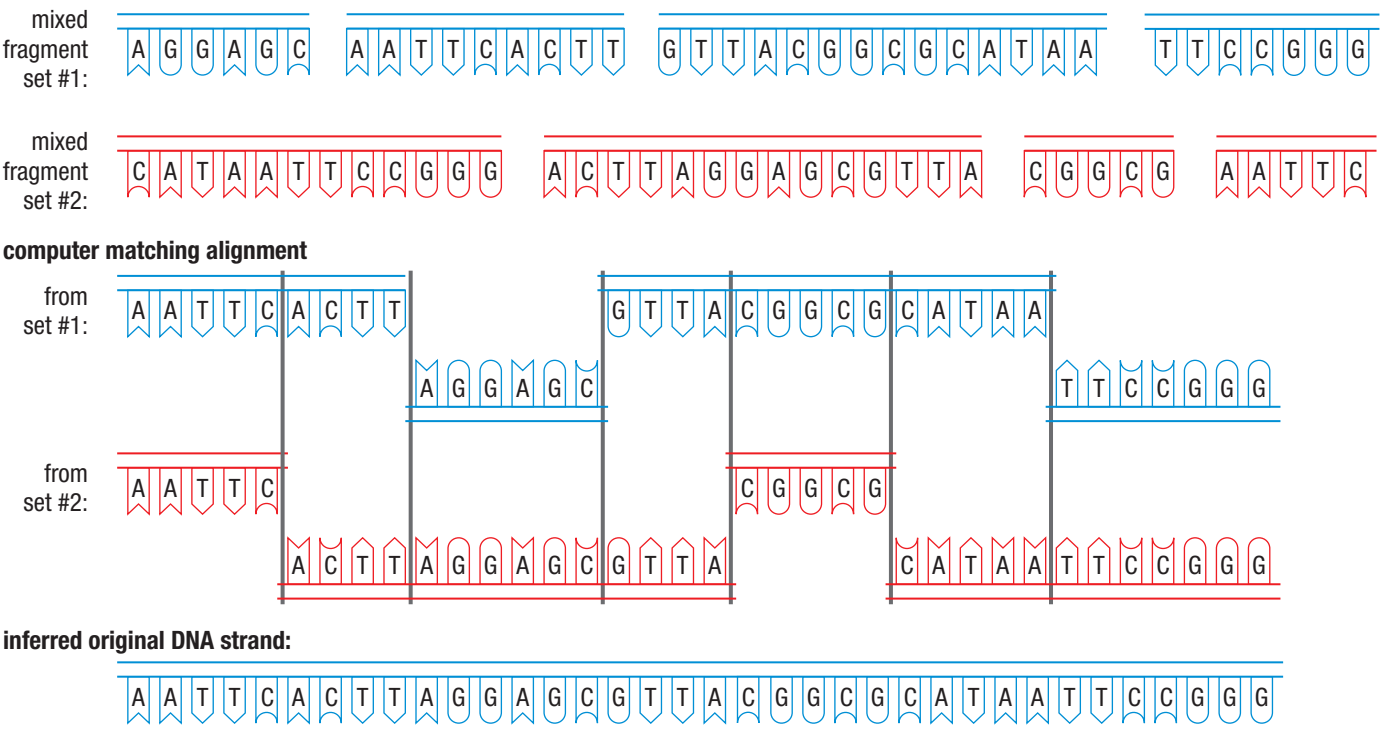

Figure 6 Note how the overlapping sequences of fragments (produced using the shotgun method) from the two sets can be used to infer the original sequence of the DNA molecule.

However, a drawback to the shotgun method was discovered. If the genome has repeating sections, it is difficult (or even impossible) for the computerized analysis to detect where these repeats are. This causes some errors in the results. Since the development of the shotgun method, however, advances in DNA sequencing technologies have allowed scientists to determine the sequences of genomes at a much faster rate.

Analyzing Genomes

structural genomics the study of the structure of genes and their locations in a genome, as well as the analysis of the nucleotide sequences to locate genes within the genome

functional genomics the study of the functions of genes, the proteins they make, and how these proteins function

bioinformatics the use of computer technology to process a large amount of biological data

Once the sequence of a whole genome has been derived, there are two approaches that can be used for its analysis: **structural genomics** (the study of the actual sequencing of genomes and the analysis of the nucleotide sequences to locate genes within the genome) and **functional genomics** (the study of the functions of genes, how they are expressed, what proteins they encode, and what role these proteins play in metabolic processes).

Most research is focused on functional genomics because genes control the functions of cells, and therefore the functions of organisms. Functional genomics relies on **bioinformatics**: the combination of laboratory experiments and sophisticated computer analyses. Bioinformatics can be used, for example, to find a gene within a genomic sequence, align sequences in databases to determine the degree of matching, predict the structure and function of a gene product, and postulate evolutionary relationships for sequences.

Genes that code for proteins are particularly interesting in genome analysis. Once a genome is sequenced, researchers use computer algorithms to search both strands of the sequence for these genes. A genome can be translated by the computer to give the amino acid sequence of the protein it could encode. Researchers may then be able to assign a function to the amino acid sequence by performing a sequence similarity search: a computer-based comparison of a DNA or amino acid sequence with databases of sequences of known genes or proteins. If the sequence resembles that of a previously sequenced gene, the two genes are related in an evolutionary sense and are likely to have similar functions.

Analysis of genome sequencing has led to many new discoveries about genetic organization, and many earlier hypotheses have been verified or disproven. A surprising discovery is that the eukaryotic genomes sequenced to date contain large numbers of previously unknown genes—many more than scientists expected to find. In *Caenorhabditis* (a nematode), for example, 12 000 of the 19 000 genes have no known function. Identifying these genes and their functions is one of the major challenges for molecular geneticists. Another surprising discovery is the degree to which different organisms, some widely separated in their evolution, contain similar genes. For example, even though the yeast *Saccharomyces*, a single-celled fungus, is separated from our species by hundreds of millions of years of evolutionary history, about 2300 of its approximately 6000 genes are related to mammalian genes. Many of these related genes control progress through the cell cycle. The similarities are so close that the yeast and human versions of several genes can be interchanged with little or no effect on cell functions in either organism.

Gene sequences also show that eukaryotic genomes contain large numbers of noncoding sequences, most of them in the form of repeated sequences of various lengths and numbers. Although these sequences make up about 25 to 50 % of the total genomic DNA in different eukaryotic species, their functions are unknown at this time.

Investigation **8.2.1**

Looking for *Wolbachia* (p. 400)
Now that you have learned about DNA sequencing, you can perform Investigation 8.2.1 and simulate DNA sequencing to research a host–parasite relationship.

Nanopore Sequencing

Nanopore sequencing is a revolutionary method that scientists have been working to perfect since 1995. It involves drawing individual strands of DNA through tiny submicroscopic holes called nanopores and reading the DNA sequence, one base pair at a time, as the strand passes through the nanopore. All of the DNA bases (A, C, T, and G) fit precisely through a 2.5 nm hole. Some transmembrane cellular proteins

work well for nanopore sequencing, although a tiny hole drilled in a superthin sheet of silicon has also been used. Each base is detected as it passes through the pore by an ingenious way of reading the conduction of a current across the silicon sheet. The current that passes is slightly different, depending on which base is passing through the nanopore at the time (**Figure 7**).

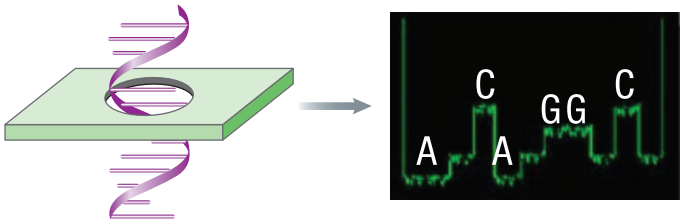

Figure 7 Nanopore sequencing may pave the way to inexpensive gene sequencing in the future.

The development of nanopore sequencing is one of many efforts by genetic technologists around the world to win the "$1000 genome initiative." The race is on to come up with a method that will allow any human to have his or her entire genome sequenced for a fee.

Studying Gene Function Using DNA Microarrays

A **DNA microarray** is a technology that helps scientists pinpoint the functions of specific genes, rather than just their locations, and to compare the expressions of genes in different types of cells or different organisms. Microarrays can be used to understand the processes involved in the development of an organism or to detect genetic mutations by comparing an individual's DNA with so-called normal DNA.

DNA microarray a solid surface that has a microscopic grid of many DNA fragments, called probes, attached; used to determine gene expression

A microarray, or gene chip, is designed to hold many individual DNA samples, ranging from just 10 to 2.1 *million*. These DNA samples, called probes, are known sequences, and the location of each probe on the chip is also known. The technology that is used to create microarrays is similar to the technology that is used to lay out electronic circuits on a computer chip.

Figure 8 (next page) shows how microarray technology works. The microarray protocol starts with the isolation of mRNA from the two types of cells to be compared: a reference cell (a normal cell) and an experimental cell (such as a cancer cell). The extracted mRNA is used to build complementary DNA (cDNA) libraries. The cDNA from each type of cell is labelled with a specific dye, such as green dye for the reference cell and red dye for the experimental cell. The cDNA sequences are then denatured into single strands and placed onto the surface of the DNA microarray.

The DNA microarray contains probes that correspond to the coding genes of the genome being studied. Any labelled cDNA that is complementary to the sequence of a probe will bind (hybridize) to this probe, while cDNA with no matching sequence will be washed off the chip. The fluorescence is then detected, and its location is recorded. Green spots mean that the reference cDNA was hybridized with the probe; hence, the gene for the corresponding protein is expressed in the normal cell. Red spots mean that the experimental cDNA was hybridized with the probe; in other words, the gene was active in the experimental cell (the cancer cell). Yellow spots mean that cDNA for both the reference cell and the experimental cell were hybridized. Both the normal cell and the cancer cell express the same gene. Researchers can then determine which genes have altered expressions in the cancer cell. The advantage of the microarray is that the expressions of thousands of genes can be identified simultaneously in a particular cell.

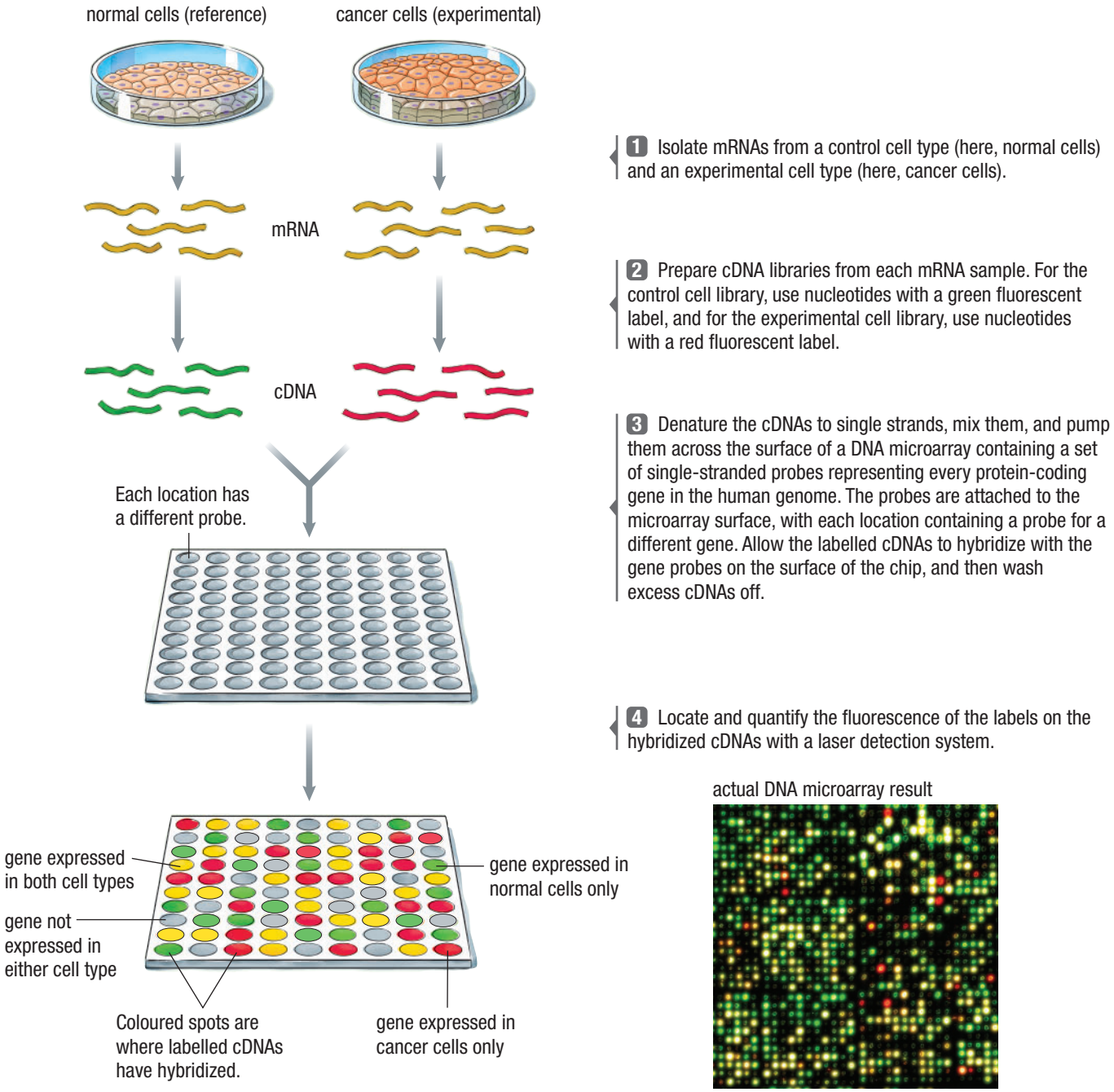

Figure 8 DNA microanalysis of gene expression levels

8.2 Review

Summary

- The polymerase chain reaction (PCR) is used to make many copies of short DNA sequences rapidly. This is useful for applications in which only a tiny sample is available, such as forensic applications.
- DNA fragments are separated by size using gel electrophoresis.
- Genomes are analyzed structurally to determine the DNA sequence. DNA sequences can be compared to detect mutations and also to study sequences that do not contain a gene but can have an important role.
- DNA sequencing can be done using the Sanger method (chain termination method) and the shotgun method.
- Genomes are analyzed to understand the function of each gene, the protein it codes for, and the function of the protein in the cell.
- Nanopore sequencing is a promising method that could be inexpensive, quick, and easy. It involves running a DNA strand through a tiny hole and measuring the current.
- DNA microarray technology can be used to detect mutations, study gene functions, and compare the expressions of genes in different tissues, individuals, and organisms.

Questions

1. Explain the role of each of the following in the PCR. K/U T/I
   (a) the DNA primer
   (b) the choice of Taq polymerase
   (c) the cycling through three different temperatures
2. Explain, using at least two or three details, how PCR technology has revolutionized genetic testing. What are the benefits for society? What are some possible disadvantages for society? K/U T/I
3. Why does the double-stranded DNA have to be raised to a temperature of 94 to 96 °C for denaturation to occur? T/I
4. PCR requires the use of two DNA primers: one complementary to the 3′ region of the target area on one strand, and the other complementary to the 3′ region of the target area on the opposing strand. Why must two different primers be used? T/I
5. How can materials as simple as agarose gel, an electric field, and a DNA stain be used to separate and compare different fragments of DNA? K/U
6. How and when is gel electrophoresis used in general society? K/U
7. A plasmid was digested using restriction enzymes. The following DNA fragments were produced: 1200 bp, 1800 bp, 200 bp, and 750 bp. T/I C
   (a) Which fragment will travel the fastest in an agarose gel? Why?
   (b) Which fragment will travel the slowest in an agarose gel? Why?
   (c) Sketch the gel pattern that would result after running this plasmid digestion in an agarose gel.
8. Compare and contrast the Sanger method and the shotgun method of DNA sequencing. How is a computer used in each method? K/U
9. Compare and contrast structural and functional genomics. K/U T/I
10. What benefits do microarrays add to the variety of genetic technologies at our disposal? K/U
11. The genomes of several species, such as *Caenorhabditis*, *Saccharomyces*, and *Drosophila*, were sequenced about the same time as the human genome was sequenced. Why do you think the choices of these organisms did, or did not, make sense? Share your reasons with a partner. T/I C