DNA Sequencing

An understanding of the structure, function, and evolutionary history of the human genome will require knowing its primary structure—the linear order of the 3 billion nucleotide base pairs composing the DNA molecules of the genome. Determining that sequence of base pairs is the long-term goal of the 15-year Human Genome Project. Both the merits and the technical feasibility of sequencing the entire human genome are discussed in Parts I and III of "Mapping the Genome." The bottom line is that sequencing technology is not yet up to the job.

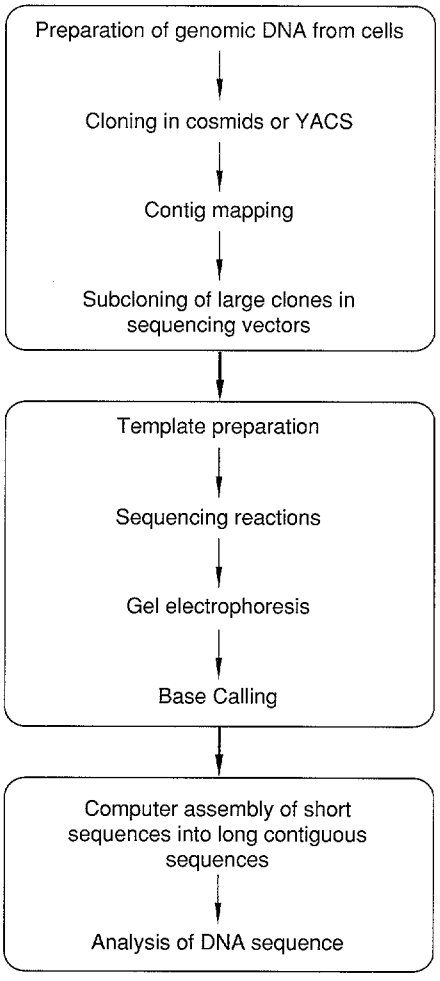
In 1990, when the plans for the Genome Project were being made, the estimated cost of sequencing was $2 to $5 per base. That is, a single person could produce between 20,000 and 50,000 bases of "finished" sequence per year. The term "finished" sequence implies the error rate is very low (the conservatives say an error rate of 1 base in 10^5 is acceptable, and the less conservative say 1 in 10^3 or 10^4). A low rate is achieved, in part, by sequencing a given region many times over. The planners agreed that the costs of sequencing must be substantially reduced and that the rate of producing finished sequence must increase by a factor of 100 to 1000 for sequencing the entire human genome to become an affordable and practical goal.

On the other hand, sequencing technology has been improving steadily for the past two decades. In the early 1970s one person would struggle to complete 100 bases of sequence in one year. Then two very similar techniques were developed—one by Allan Maxam and Walter Gilbert in the United States and the other by Fredrick Sanger and his coworkers in England—that made it possible for one person to sequence thousands of base pairs in a year. Those techniques, for which the inventors were jointly awarded the Nobel Prize, still form the basis of all current sequencing technologies. Both methods are described in greater detail below.

Between 1975 and the present, the number of base pairs of published sequence data grew from roughly 25,000 to almost 100 million. During that time longer and longer contiguous stretches of DNA have been sequenced. In 1991 the longest sequence to be completed was that of the cytomegalovirus genome, which is 229,354 base pairs. By 1992 a cooperative effort in Europe had sequenced an entire chromosome of yeast, chromosome III, which is 315,357 base pairs. And now efforts are underway to sequence million-base stretches of DNA. Accomplishing such large-scale sequencing projects is among the goals for the first five years of the Genome Project.

In order to achieve this goal, each step in the multi-stage DNA sequencing process must be streamlined and smoothly integrated. Figure 1 outlines all the steps involved in the sequencing of long, contiguous stretches of genomic DNA, DNA isolated from the genome. The initial steps include cloning large fragments of genomic DNA in YACs or cosmids and using those clones to construct a contig map for the regions to be sequenced. The contig map arranges the cloned fragments in the order and relative positions in which they appear along the genome. The cloning and mapping steps are described elsewhere in this issue (see "DNA Libraries" and "Physical Mapping").

Figure 1. Steps in Large-Scale Sequencing



To determine the DNA sequence of the mapped region, the large DNA insert in each of the large clones must be broken into smaller pieces of a size suitable for sequencing, and those small pieces must be cloned. This subcloning is often done in the cloning vector M13, a bacteriophage whose genome is a single-stranded DNA molecule. M13 accepts DNA inserts from 500 to 2000 base pairs in length, propagates in the host cell *E. coli*, and is particularly convenient for the Sanger method of sequencing. Each of the small clones is then sequenced.

As mentioned above, all sequencing technologies currently in use are based on the Sanger or the Maxam-Gilbert method, which were developed in 1977. Both methods determine the sequence of only one strand of a DNA molecule at a time, and both methods involve three basic steps. Below we mix and match certain technical details of each method to simplify the description of these three steps. The real methods are described in Figures 4 and 5.

Figure 2. Nested Set of Labeled Fragments for Simplified Example

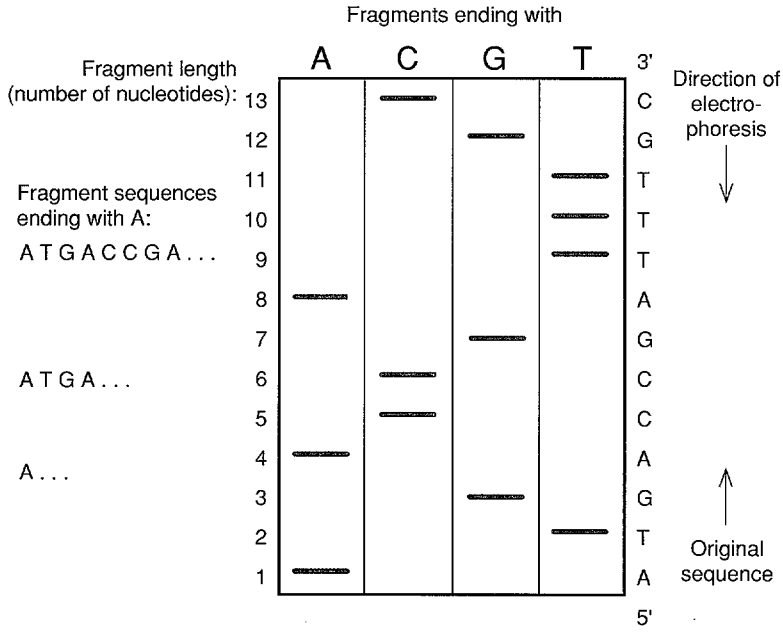
Original Strand	5'-^{32}P-ATGACCGATTTGC-3'
Labeled fragments ending in A	5'-^{32}P-A
	5'-^{32}P-ATGA
	5'-^{32}P-ATGACCGA
Labeled fragments ending in C	5'-^{32}P-ATGAC
	5'-^{32}P-ATGACC
	5'-^{32}P-ATGACCGATTTGC
Labeled fragments ending in G	5'-^{32}P-ATG
	5'-^{32}P-ATGACCG
	5'-^{32}P-ATGACCGATTTG
Labeled fragments ending in T	5'-^{32}P-AT
	5'-^{32}P-ATGACCGAT
	5'-^{32}P-ATGACCGATT
	5'-^{32}P-ATGACCGATTT

- Many copies of the strand to be sequenced are isolated and labeled with, say, the radioisotope ^{32}P, usually at the 5′ end. The strands are chemically manipulated to create a nested set of radio-labeled fragments. By nested, we mean that each fragment in the set has a common starting point, typically at the labeled 5′ end of the original strand, and the lengths of the labeled fragments increase stepwise, or one base at a time. In other words, the shortest fragment contains the radio label and the first base at the 5′ end of the original strand. The next shortest fragment contains the label and the first two bases at the 5′ end, and so on, up to the longest fragment, which is identical to the original strand.

- The fragments that make up the nested set are not prepared in one reaction mixture. Rather, copies of the original labeled strand are divided into four batches. Each batch is subjected to a different reaction, and each reaction produces labeled fragments that end in only one of the four bases A, C, T, or G. For example, if the sequence of the original labeled strand is 5′-^{32}PATGACCGATTTGC-3′, the four reactions produce the four sets of labeled fragments shown in Figure 2. Together those fragments compose the complete set of nested fragments for the original strand. That is, the set includes all fragments that would be obtained by starting at the 5′ end of the original strand and adding one base at a time.

- The fragments from the four reaction mixtures are separated by length using gel electrophoresis. A polyacrylamide gel is prepared with four parallel lanes, one for each reaction mixture. Thus each lane contains labeled fragments that end in only one of the four bases. Since polyacrylmide gels can resolve DNA molecules differing in length by just one nucleotide, the positions of all the labeled fragments can be distinguished. During electrophoresis, shorter fragments travel farther than longer fragments. Thus copies of the shortest fragment form a band farthest from the end at which the fragment batches were loaded into the gel. Successively longer fragments form bands at positions closer and closer to the loading end. Following electrophoresis, the radio-labeled fragments are visualized by exposing the gel to an x-ray filter to make an autoradiogram. Figure 3 shows the pattern of bands that would be created on the autoradiogram by the four sets of labeled fragments in Figure 2. Recall that each band contains many copies of one of those labeled fragments. The end base of those fragments is known by noting the lane in which the band appears, and the length of those fragments is determined from the vertical position of the band; fragment lengths increase from the bottom to the top of the autoradiogram. Therefore, the base sequence of the original long strand can be read directly from the autoradiogram. One starts at the bottom and looks across the four lanes to find the lane containing the band corresponding to the shortest fragments. Those fragments end at the base marked at the top of the lane. Then one continues up and across the autoradiogram, each time identifying the lane containing the band corresponding to the next longer fragments and thus identifying the end base of those fragments. The sequence of the original strand is thus read from its 5′ end, the common starting point, to its 3′ end.

Figure 3. Autoradiogram of Sequencing Gel for Simplified Example



Schematic diagram of autoradiogram showing the positions of labeled fragments generated in four reaction mixtures from the sequence 5'-^{32}P-ATGACCGATTTGC-3'. The sequence in the 5'-to-3' direction is read from the bottom to the top of the autoradiogram.

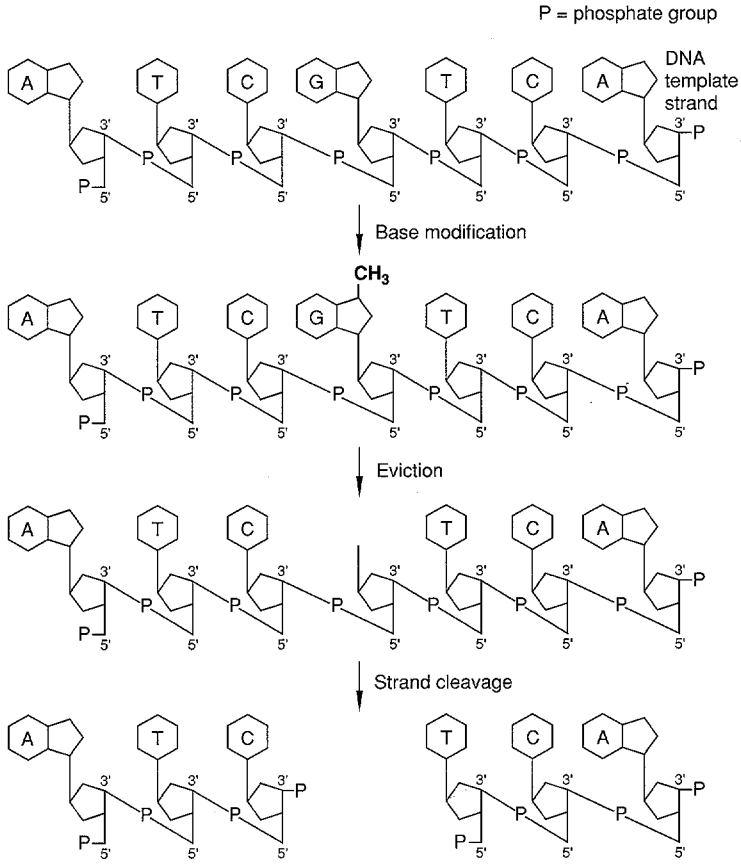
The Sanger and Maxam-Gilbert sequencing protocols differ in the reactions used to generate the four batches of labeled fragments making up the nested set. The Sanger method involves enzymatic synthesis of the radio-labeled fragments from unlabeled DNA strands. The Maxam-Gilbert method involves chemical cleavage of prelabeled DNA strands in four different ways to form the four different collections of labeled fragments. The details of the two procedures are described in Figures 4 and 5.

Figure 4. Maxam-Gilbert Sequencing Method

The Maxam-Gilbert sequencing protocol uses chemical cleavage at specific bases to generate, from pre-labeled copies of the DNA strand to be sequenced, a nested set of labeled fragments. Recall that the fragments in the set increase in length one base at a time from the 5' end of the original labeled strand. Four different cleavage reactions are used, and the reaction products are separated by length on four lanes of a gel to determine the order of the cleaved bases along the original labeled strand.

Two chemical cleavage reactions are employed; one cleaves a DNA strand at guanine (G) and adenine (A), the two purines, and the other cleaves the DNA at cytosine (C) and thymine (T), the two pyrimidines. The first reaction can be slightly modified to cleave at G only, and the second slightly modified to cleave at C only. In each reaction, cleavage of single-stranded DNA is accomplished by chemically modifying a specific base, removing the modified base from its sugar, and then breaking the bonds that hold the exposed sugar in the sugar-phosphate backbone of the DNA molecule.

(a) Cleavage Reaction for Guanine



Dimethylsulfate is used to methylate guanine. After eviction of the modified base, the exposed sugar, deoxyribose, is then removed from the backbone. Thus the strand is cleaved in two.

The reaction that cleaves guanine is shown schematically in (a). A methyl group is added to guanine, the modified base is removed from its sugar by heating, and the exposed sugar is removed from the backbone by heating in alkali. To cleave at both A and G, the procedure is identical except that a dilute acid is added after the methylation step. The reactions that cleave at C, or at C and T, involve hydrazine to remove the bases and piperidine to cleave the backbone. The extent of the reaction shown in (a) can be carefully limited so that, on average, only one G is evicted from each strand, thus each strand is cleaved at only one of its guanine sites.

(b) Fragments from Single Cleavage at G

5'-^{32}P-ATGACCGATTTGC-3' ——— Labeled template strand

5'-^{32}P-AT-3' 5'-ACCGATTTGC-3'

5'-^{32}P-ATGACC-3' 5'-ATTTGC-3'

5'-^{32}P-ATGACCGATT-3' 5'-C-3'

Six different types of fragments are produced. Only three of those include the labeled 5' end of the original strand.

A radiolabeled strand to be sequenced and the fragments created from that strand by a single cleavage at the site of G are illustrated in (b). Each original strand is broken into a labeled fragment and an unlabeled fragment. All the labeled fragments start at the 5' end of the strand and terminate at the base that precedes the site of a G along the original strand. Only the labeled fragments will be recorded once all the fragments are separated on a gel and visualized by exposing the gel to an x-ray film to create an autoradiogram of the gel.

Given the four chemical cleavage reactions, we can outline the steps involved in Maxam-Gilbert sequencing.

Step 1: Preparation of Labeled Strands. Many copies of the DNA segment to be sequenced are labeled with radioisotope ^{32}P at the 5' end of the strand. If the DNA is cloned in double-stranded form, then the 5' ends of both strands are labeled. The DNA is then denatured, copies of one strand are isolated from copies of the other strand, and each strand is sequenced separately.

Step 2: Generating a Nested Set of Labeled Fragments. Copies of one labeled strand are divided into four batches, and each batch is subjected to one of four chemical cleavage reactions outlined above. The reactions cleave the template strands at G, G and A, C, or C and T, respectively. All labeled fragments in each batch begin at the 5' end of the original strand.

Step 3: Electrophoresis and Gel Reading. The fragments from the four reactions are separated in parallel on four lanes of a gel by electrophoresis. An autoradiogram of the gel shows the positions of the labeled fragments only. A schematic of the autoradiogram is shown in the figure. Each of the four lanes is labeled by the base or bases at which the original strand was cleaved. Fragments cleaved at C show up in two lanes, the one marked C and the one marked C and T. Fragments cleaved at T are identified by noting that they appear in the lane marked C and T, but do not appear in the lane marked C. Fragments ending in A or G can be similarly identified. Note that the fragment cleaved at the first base will not show up on the gel, so the first base at the 5' end of the original strand cannot be determined. As described in the main text, the band corresponding to the shortest fragments is at the bottom of the autoradiogram. The 5'-to-3' sequence of the original strand is read by noting the positions and lanes of the bands from the bottom to the top of the autoradiogram.

(c) Steps in Maxam-Gilbert Sequencing

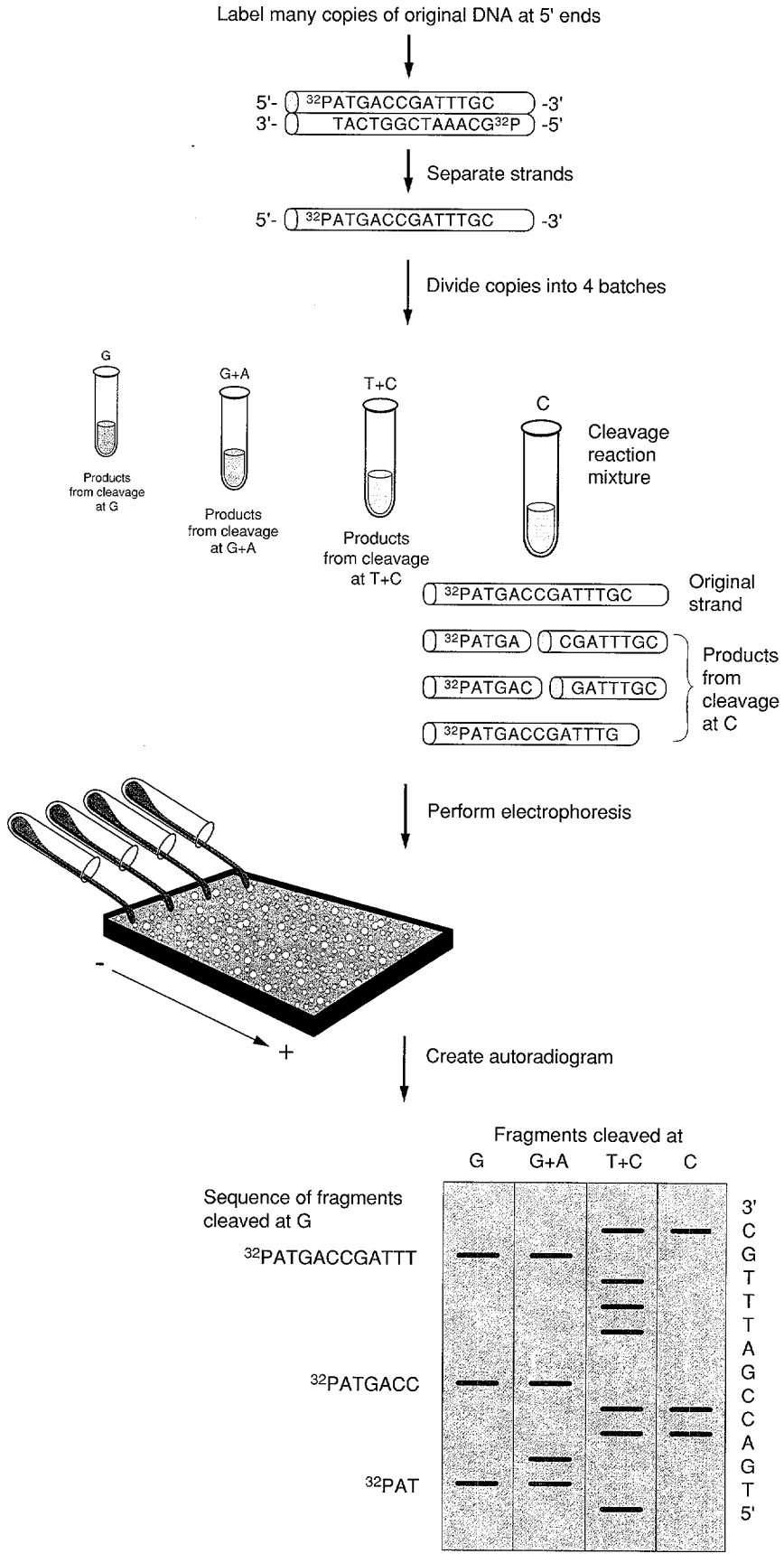
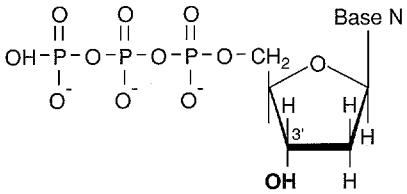


Figure 5. Sanger Sequencing Method

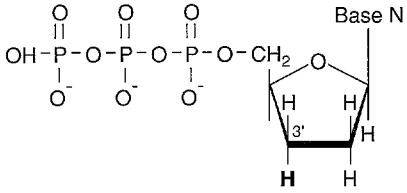
The Sanger method for sequencing, also known as the dideoxy chain termination method, generates the nested set of labeled fragments (see main text) from a template strand by replicating the template strand to be sequenced and interrupting the replication at one of the four bases. Four different replication reactions produce fragments that terminate in A, C, G, or T, respectively.

The replication reaction follows the path described in "DNA Replication" (see box in "Understanding Inheritance"). A DNA primer is attached (by hybridization) to the template strand and deoxynucleoside triphosphates (dNTPs) are sequentially added to the primer strand by a DNA polymerase. However, dideoxynucleoside triphosphates, say, ddATPs, are present in the reaction mixture along with the usual dNTPs. If, during replication, ddATP rather than dATP is incorporated into the growing DNA strand, then replication stops at that nucleotide.

(a) Structure of dNTP and ddNTP



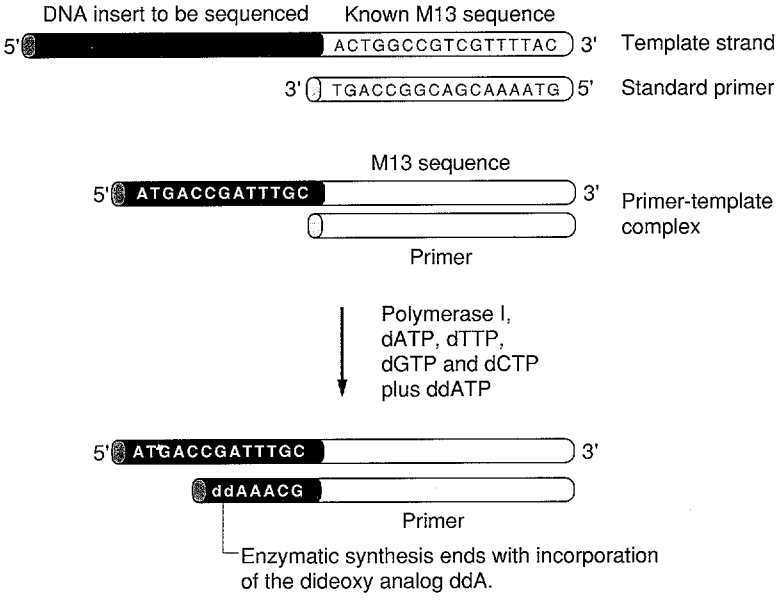
Deoxynucleoside triphosphate (dNTP) (note hydroxyl group on 3' carbon of deoxyribosose).



Dideoxynucleoside triphosphate (ddNTP). Dideoxy analog lacks the hydroxyl group on the 3' carbon.

In (a) we show the difference between dNTP and ddNTP. The dideoxy analog lacks the hydroxyl group that is present on the 3' carbon of the sugar in dNTP and is needed to form an O-P-O bridge to the next nucleotide. Thus, the addition of a ddNTP to the growing strand prevents the polymerase from adding additional nucleotides, and the new synthesized strand terminates with the base N. Thus all the strands synthesized in the presence of ddATP have sequences that terminate at A. These strands are complementary to the template strand, and terminate opposite the site of a T on the template strand. Complementary strands terminating in either A, G, C, or T are produced by the inclusion in the reaction mixture of ddATP, ddGTp, ddCTP, or ddTTP, respectively.

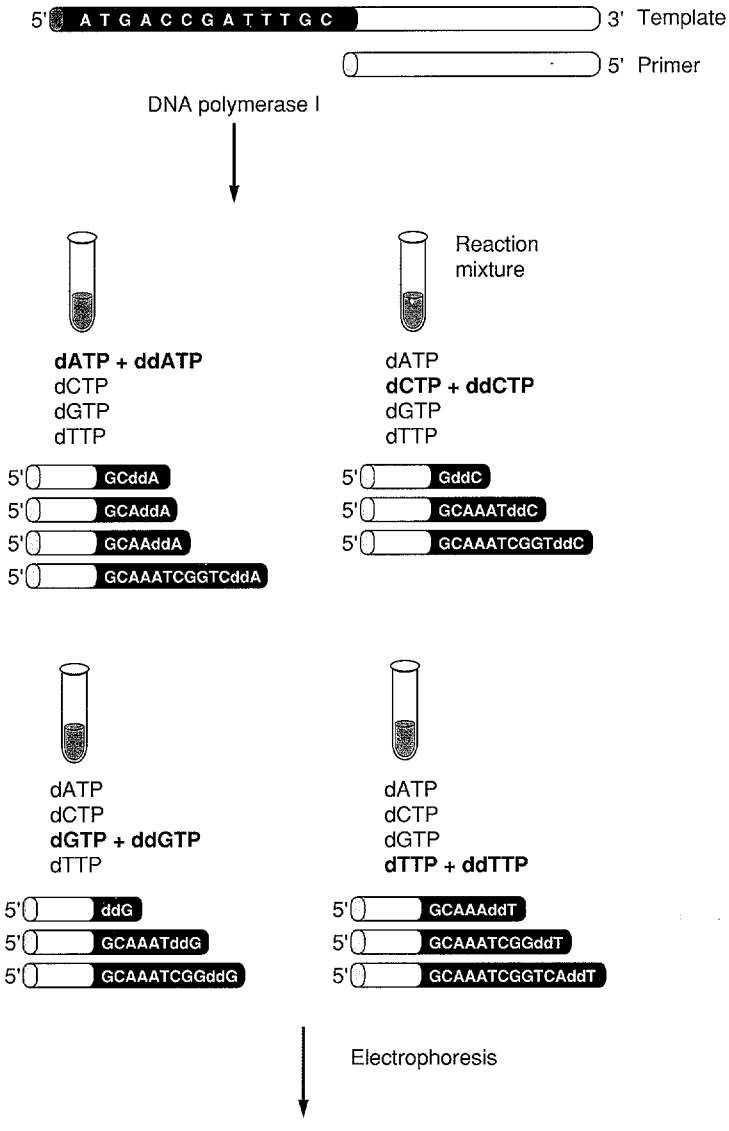
(b) Dideoxy Chain Termination Reaction with ddATP



Incorporation of ddATP rather than dATP is random so all possible strands ending at ddATP are synthesized in the reaction.

As illustrated in (b), copies of the template strand to be sequenced must be prepared with a short known sequence at the 3' end of the strand. That short sequence will then hybridize to a DNA primer whose sequence is exactly complementary to that of the known sequence. The primer is essential to initiate replication of the templates by DNA polymerase. The most convenient method for adding a known sequence to the 3' end of the template strand is to clone the strand in the single-stranded cloning vector M13 so that a known M13 sequence will always flank the unknown DNA insert and can serve as the site for binding a standard primer. Also, the M13 cloning protocol automatically creates two types of clones, each type containing a DNA insert whose sequence is complementary to that of the other DNA insert. Thus, the two complementary strands may be sequenced and the two sequences cross-checked to ensure sequence accuracy.

(c) Steps in Sanger Sequencing

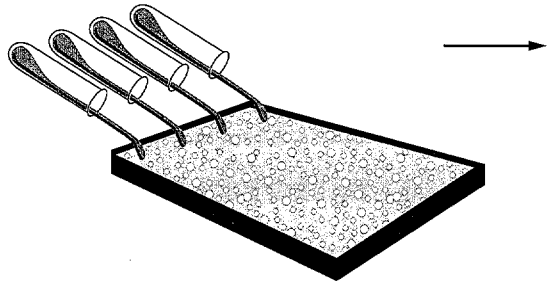


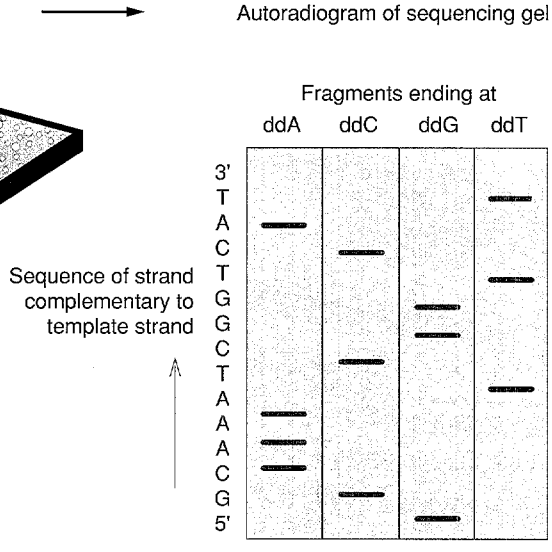
In (c) we outline the three steps involved in the Sanger dideoxy sequencing method.

Step 1: Template Preparation. Copies of the template strand are cloned in M13. They are thus flanked at their 3' ends by a known sequence that will bind to a standard primer.

Step 2: Generating a Nested Set of Labeled Fragments. Copies of each template strand are divided into four batches, and each batch is used for a different replication reaction. Copies of the same standard primer and DNA polymerase I is used in all four reactions. To synthesize fragments, all of which terminate at A, the dideoxy analog ddATP is added to the reaction mixture along with dATP, dGTP, dCTP, dTTP the standard primer and DNA polymerase I. The ddATPs and one of the dNTPs are labeled with a radioactive isotope to produce radiolabeled strands. The figure shows a short template strand, the primer, the four reaction mixtures, and the labeled strands produced by each reaction. Note that the synthesized fragments from the four reaction mixtures compose the set of nested fragments needed to determine the order of the bases in the strand complementary to the template strand.

Step 3: Electrophoresis and Gel Reading. The fragments from the four reaction mixtures are loaded into four parallel lanes of a polyacrylamide gel and separated by length using electrophoresis.





An autoradiogram of the gel is read as described in the main text to determine the order of the bases in the strand complementary to that of the template strand. Again, since the bands corresponding to the shortest fragments are at the bottom of the autoradiogram, the 5'-to-3' sequence of the strand complementary to the template strand is read from the bottom to the top of the autoradiogram.

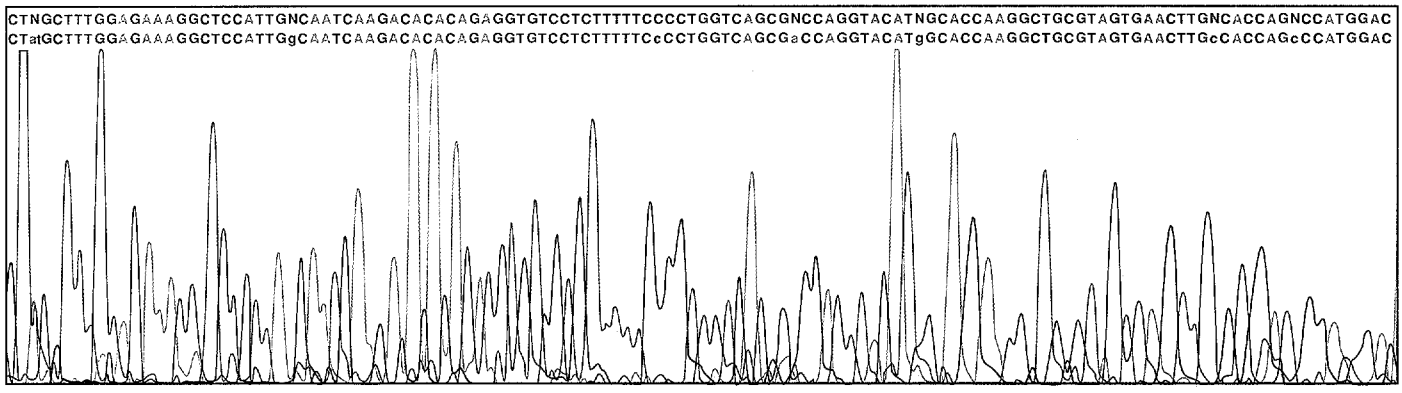
The final step in both procedures is to separate the labeled fragments by length using gel electrophoresis (see "Gel Electrophoresis" in "Understanding Inheritance"). Since the fragment mobility in the gel varies as the reciprocal of the logarithm of the fragment length, shorter fragments are more widely separated from one another than longer fragments. That is, the resolution of fragment lengths decreases as the fragment length increases. Therefore, the range of fragment lengths that can be resolved in a single gel is limited to several hundred bases. Moreover, the separation of fragments in a standard gel (0.2 to 0.4 millimeters thick) is a relatively slow process. At least several hours are required to resolve fragment lengths from one to several hundred bases long. [More recently, very narrow gel-filled capillary tubes have been used to decrease the time needed for fragment separation. Several hundred bases can be resolved in tens of minutes and the resolution is high enough to read 1000 bases from a single gel.] The average error rate in a single sequencing run is about 1 base in 100. The errors are often due to inhomogeneities in the gel and various sequence-dependent conformational changes in the single-stranded fragments that affect their mobility in the gel.

Since only short stretches of DNA, several hundred to a thousand base pairs in length, can be obtained from a single sequencing gel, many short sequences must be generated separately and then combined to determine the sequence of a much longer DNA fragment. Various strategies have been developed to generate these short sequences from the larger fragment.

The "shotgun" approach is the most widely used in the larger sequencing projects. Copies of a long fragment to be sequenced are broken into much shorter fragments that overlap one another, and the short fragments are cloned. Those clones are then picked at random and sequenced. The sequence of the long fragment is determined by finding overlaps among the short sequences and assembling those sequences into the most likely order. Numerous computer algorithms have been developed to facilitate the assembly of long sequences.

Inevitably, gaps remain in the sequence of the long fragment, and they are filled by switching to a directed sequencing strategy. That is, the short clones are no longer sequenced at random, but rather, short sequences at the end of a continuous stretch of known sequence provide the information necessary to construct a probe to pick out a clone, or region of a clone, whose sequence will extend the known sequence. Most of the large sequencing projects to date have used a mixture of random and directed sequencing strategies to complete the sequence of long, contiguous stretches of DNA. The advantage of the random, or "shotgun," strategy is that in the course of picking clones at random and sequencing them, any given region is usually sequenced many times, thereby reducing the errors in the final sequence.

Almost all steps involved in sequencing are amenable to automation, and through automation many groups hope to increase both the throughput and the consistency of large-scale sequencing efforts. Several automatic sequencing machines have been



on the market for a number of years. Those machines automate the steps of gel electrophoresis, gel reading, and the "calling" of the end bases of the successively longer fragments. The machines designed for high throughput require that the fragments produced by the four sequencing reactions be labeled with fluorescent dyes rather than radioisotopes, and they employ laser-induced fluorescence to detect the order of the labeled fragments as they migrate through the gel. Some machines use four parallel lanes for the fragments of the four reaction mixtures; others use a single gel lane for all the fragments. The output of a high-throughput sequencing machine includes a plot of the fluorescence signals versus time produced as the fragments migrate past the laser as well as the sequence of bases corresponding to the time sequence of the variously colored fluorescence peaks. Ambiguities in the data are also noted automatically (see Figure 6).

Under optimal conditions, the automatic sequencers are capable of producing 12,000 base pairs of raw data per day. However, much work remains to improve reliability and to organize the efficient use of those machines in large-scale sequencing projects. For example, problems associated with the preparation of clones for sequencing, the checking of the short sequences and assembling them into longer contiguous sequences, and the tracking of all procedures involved in sequencing need increased attention. So far, despite the availability of automatic sequencing machines, production of finished sequence remains a slow and expensive process. Those working on improving existing technologies and streamlining their use expect to achieve a tenfold increase in sequencing throughput within the next few years, and perhaps a hundredfold increase in ten years. Others are involved in developing radically new sequencing technologies that, if successful, might achieve the hundredfold to thousandfold increase needed to sequence the entire human genome. (See the discussion of new technologies in Part III of "Mapping the Genome" as well as "Rapid DNA Sequencing Based on Single Molecule Detection.") ■

Figure 6. Output of Automatic Sequencing Machine

Each of four dideoxy sequencing reactions produces fragments labeled with a dye that fluoresces at a different wavelength. As the fragments from the four reactions migrate down a single lane of a polyacrylamide gel, they pass through a laser beam and produce a fluorescence signal. The machine automatically records the signal and calls the end base of the fragments based on the color (wavelength) of the fluorescence signal. The sequence of the strand complementary to the template strand is read from right to left corresponding to the 5′-to-3′ direction. The machine automatically generates the top sequence, recording any ambiguity in the base call as an N. A technician can resolve most such ambiguities by direct examination of the fluorescence signals. If the technician concludes with high certainty that a particular N is, for example, the base G, he or she replaces that N with a g in the bottom sequence.

Further Reading

T. Hunkapiller, R.J. Kaiser, B.F. Koop, L. Hood "Large-Scale and Automated DNA Sequence Determination." *Science*, October 4, 1991.