Large-Scale and Automated DNA Sequence Determination

T. HUNKAPILLER, R. J. KAISER, B. F. KOOP, L. HOOD

DNA sequence analysis is a multistage process that includes the preparation of DNA, its fragmentation and base analysis, and the interpretation of the resulting sequence information. New technological advances have led to the automation of certain steps in this process and have raised the possibility of large-scale DNA sequencing efforts in the near future [for example, 1 million base pairs (Mb) per year]. New sequencing methodologies, fully automated instrumentation, and improvements in sequencing-related computational resources may render genome-size sequencing projects (100 Mb or larger) feasible during the next 5 to 10 years.

DNA SEQUENCE ANALYSIS PLAYS AN IMPORTANT ROLE IN three areas of modern biology. First, by revealing the similarities of homologous genes, it provides insights into the possible regulation and functions of these genes as well as into their evolutionary history. Second, the Human Genome Project proposes to sequence the human genome of 3 billion base pairs (bp) and 50,000 to 100,000 genes during the next 15 years (*1*). This endeavor is driving the development and application of powerful new and improved technologies. Third, sequence information leading to an understanding of disease states related to genetic variation should have an enormous impact not only on biological research but also on the practice of medicine. Therefore, DNA sequence analysis has become one of the most important tools in modern biology and has led to the accumulation of more than 65 million bases of DNA and RNA sequence from hundreds of species in the contemporary DNA databases (for example, GenBank Release 68).

DNA sequence analysis involves multiple processes, including DNA cloning, physical mapping, subcloning, sequencing, and information analysis. In large-scale DNA sequencing, all of the steps involved in these processes must be coordinated and where possible automated in order to achieve an optimal throughput with a minimum cost. Thus, DNA sequencing technologies are rapidly evolving. In this article, we emphasize the current technological approaches to sequencing as well as the challenges and benefits that new technologies may hold for generating systems capable of even more rapid and inexpensive DNA sequence analysis in the future.

Two Types of DNA Sequence

Two general forms of DNA are most often the target of sequencing efforts: (i) genomic DNA and (ii) DNA generated as a copy of messenger RNA (complementary or cDNA). Many host vectors are available for cloning various-sized genomic inserts, including yeast artificial chromosomes (YACs, 100,000 to 1,000,000 bp) (*2*), P1 phage (50,000 to 100,000 bp) (*3*), cosmids (30,000 to 45,000 bp) (*4*), and λ phage (100 to 20,000 bp) (*5, 6*). If large genes or extensive chromosomal regions are to be characterized, a physical map (the overlapping order) of a set of inserts must be generated.

For actual DNA sequence analysis, these larger clones are usually converted into smaller, more easily sequenced pieces (subclones) (*6–8*), generally 500- to 2000-bp fragments for single-stranded vector systems such as M13 and 1000- to 5000-bp fragments for double-stranded vectors such as plasmids. Such cDNAs typically range in size from 500 to 8000 bp and, accordingly, can generally be cloned and sequenced more easily than genomic DNA, which often requires significant physical mapping and extensive subcloning. Some scientists believe that the Human Genome Project should concentrate its efforts on characterizing cDNAs as a more efficient approach to identifying the sequences that directly encode genes (*9*). However, many genes that are only expressed at very low levels or in a particularly tissue-restricted manner would be missed by this approach, as would all of the important regulatory sequences, which are not expressed as RNA. Moreover, as sequencing technologies improve, total genomic sequencing should become an extremely efficient strategy for identifying all human genes. However, any approach based on genomic DNA sequencing requires the development of effective computer tools for the discovery of gene-encoding sequences within the predominantly noncoding background (*10*). Certainly, the initial loci to be sequenced by any large-scale effort should be selected for their biologic, genetic, and evolutionary significance.

Large-Scale DNA Sequencing

The Human Genome Project has placed a special focus on large-scale DNA sequencing. "Large-scale" has an operational definition that depends heavily on the current technologies. For the purposes of this discussion, we define large-scale as the capacity of one effort to produce 1 million base pairs or more of finished DNA sequence in 1 year. This is three to four times more than the longest sequences yet obtained (*11*). In order to achieve this goal, it is imperative that each step in the multistage DNA sequencing process be automated and smoothly integrated (Fig. 1) (*12*).

Traditional Methods of Obtaining DNA Sequence

All of the current strategies for obtaining DNA sequence share a fundamental, three-step approach. First, a complete nested set of DNA fragments is generated from a DNA clone, each fragment having a common starting point and each successive fragment being one base longer than the preceding one. Second, these fragments are

The authors are members of the NSF Science and Technology Center for Molecular Biotechnology in the Division of Biology of the California Institute of Technology, Pasadena, CA 91125.

separated by size in a system that can resolve fragments differing in length by a single base. Third, the base at the unique end of each fragment is characterized. When ordered by the size of the fragments that they terminate, these end bases provide the DNA sequence that spans the length of the longest fragment in the nested set. The immediate methodological questions concern the methods of fragment generation, separation, and labeling to identify the end base of each fragment.

Generation of the nested fragment set. Nested DNA fragments are produced either by limited chemical cleavage or by enzymatic synthesis of progressively longer copies of a larger DNA segment. In the chemical, or Maxam and Gilbert method (*13*), the DNA sample is divided into four aliquots and then four different chemical reactions are used that cleave the DNA only at a particular base (A, C, G, or T) or base type [pyrimidine (C,T) or purine (A,G)] (Fig. 2A). Proper titration of the reactants as well as careful timing of the reactions means that each molecule of the starting sample is cleaved on average only once. Thus, all possible fragment lengths are generated by the four reactions.

In the enzymatic, or Sanger, method (*14*), instead of chemically cleaving existing molecules, DNA fragments are synthesized by DNA polymerase, which incorporates deoxynucleotide monomers into a polymeric complementary copy of a template DNA strand (Fig. 2B). A small, generally synthetic piece of DNA (an oligonucleotide primer) is used to initiate synthesis of the new DNA strand from the template DNA at a single, specific location. Four separate reactions are performed, each containing all four deoxynucleotides and one of four dideoxynucleotide analogs. These analogs lack a chemical functionality that is critical for further chain elongation but that does not grossly interfere with its incorporation by the enzyme. The synthesis of any new strand is terminated by the incorporation of an analog into the growing strand instead of the normal nucleotide. Since each synthesis is begun with a common primer, use of the proper ratios of normal and analog nucleotides generates a nested set of fragments, each ending in a particular dideoxynucleotide. Although the enzymatic approach is most effectively used with single-stranded templates, it is also compatible with double-stranded templates (*15*). The Sanger method is used for most sequencing done today.

Separation of the nested DNA fragments. Once generated, the nested set of DNA fragments must be separated by size to determine the order of the end bases. Traditionally electrophoresis through a sieving matrix of cross-linked polyacrylamide is used. Polyacrylamide slab gels of the appropriate concentrations can resolve single-base length differences over a range of a very few bases to several hundred bases. Under constant current conditions, fragment mobility is an inverse logarithmic function of fragment length, thereby reducing the resolution of fragments as their size increases and limiting the range of sizes that can be resolved in a single gel to provide accurate DNA sequence information. Various voltage ramping and pulsed-field protocols can narrow or extend this range (*16*).

Likewise, the thickness, dimensions, and geometry of the gel itself can affect the mobility of fragments, as can the composition of the fluid portion of the gel matrix. For example, gels with a wedge-shaped thickness dimension and gels with a continuous gradient of polyacrylamide concentration in the direction of separation have both been used to obtain a more linear separation function, thus increasing the amount of accurate sequence data that can be obtained from a single separation (*6, 7*). Separation in gels of normal thickness (0.2 to 0.4 mm) is a relatively slow process, minimally requiring several hours to resolve a few hundred bases.

In an attempt to address both time and resolution issues, the use of gel-filled capillaries or ultrathin slab gels (less than 0.1-mm thick) has recently been described (*17*). In both cases, a significant decrease in electrophoresis time was obtained, with several hundred bases being resolved in 15 to 60 min. Separations also exhibit a positive effect on overall resolution, resulting in the potential to read perhaps 1000 bases from a single separation.

End base determination. In order to observe the results of the separation process, each fragment must be labeled or marked in some manner. This labeling can be done prior to, during, or after the generation of fragments. Traditionally, labeling has been accomplished with radioisotopes such as ^{32}P or ^{35}S prior to or during

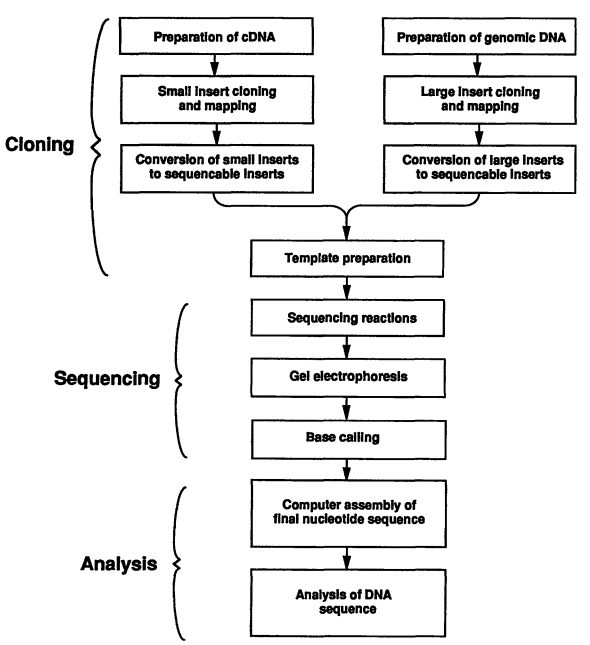

Fig. 1 (left). A schematic diagram of the steps involved in large-scale DNA sequencing (see text).

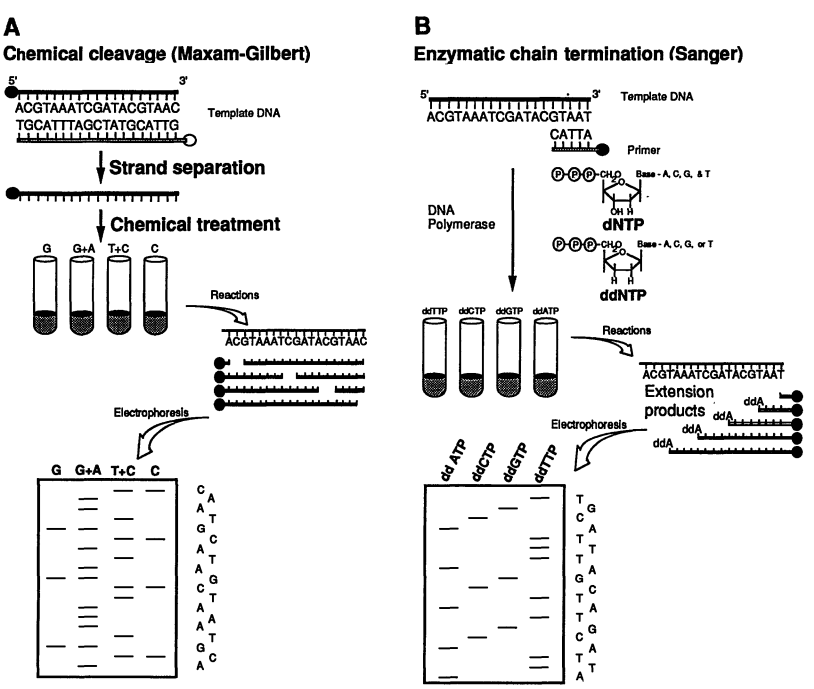

Fig. 2 (right). Chemical (**A**) and enzymatic (**B**) sequencing procedures (see text).

fragment production in the sequencing reactions (*6*). After separation, a photographic film is exposed to the gel containing the labeled fragments in order to visualize the separation pattern. The radioactive decay of the label produces a band on the film at the position of each fragment in the gel. Since only a single label is used, each base-specific reaction set is separated in a different, adjacent lane of the gel; thus, a single sequence determination requires four lanes. The sequence is "read" by determining the order of fragment sizes among the four lanes (Fig. 2). Recently, high-sensitivity nonisotopic labeling strategies based on light detection (chemiluminescence) have been used to visualize the band patterns in sequencing gels (*18*).

Sequencing Strategies

Since the length of sequence data obtainable from a single reaction set as described above is limited to several hundred to perhaps 1000 bases, the characterization of long regions of DNA, such as genes, must be accomplished through the successive generation of many smaller sequences. The manner in which these smaller sequences are obtained from the larger whole is the fundamental issue of choice of overall sequencing strategy. Current strategies fall into two general categories, directed and random.

Directed strategies. These strategies permit the direct and sequential sequence analysis of a large DNA fragment from one end to the other. Traditionally a fine-resolution restriction enzyme map identifies overlapping smaller restriction fragments that are then subcloned into appropriate cloning vectors and subsequently sequenced (*6*). More recently, three additional approaches have been developed (Fig. 3A).

1) In primer-directed sequencing (*6, 7*) an initial round of DNA sequencing by the enzymatic method with the use of a vector-specific "universal" primer is followed by repetitive cycles of synthesis of a new sequencing primer generated from the just-acquired sequence and subsequent new sequence determination with this primer. In order to facilitate large-scale sequencing by primer-directed walking, various proposals have been made to first synthesize a general library of primers instead of synthesizing specific primers as the walk proceeds (*19*). Calculations and tests indicate that a panel of as few as 4000 small primers might suffice for the bulk of the sequencing effort.

2) Exonucleases can be used to create a nested set of deletional clones with one common end (*20*). In this process, a cloned DNA fragment of interest is subjected to end-specific digestion for successively longer periods of time. The resulting clones are thus characterized by a common end and sequentially larger insert sizes. These clones can then be sequenced at their deletional ends with a universal primer.

3) Various strategies have been developed that first create a set of DNA clones from a single source clone that each contain a single, randomly inserted transposable element (transposon) (*21*). The set is characterized to obtain a subset of clones whose transposons occur at relatively evenly spaced positions in the source DNA and that are separated by distances such that sequencing is possible. The transposon sequence provides a common primer site sequence for subsequent enzymatic sequencing.

Random methods. Random or "shotgun" strategies generate a library of subclones through random cleavage or shearing of a large piece of DNA (Fig. 3B) (*8, 22*). Because subclones are selected for sequencing from this library at random, the process is unordered. Thus, the order of the resulting sequence fragments relative to that of the original source clone must be determined by identifying the probable overlaps of subclone sequences and assembling them into the most likely order. Numerous algorithms have been developed to facilitate this reconstruction or sequence assembly (*10, 23*). With random methods, a significant number of extra (redundant) subclones must be sequenced to ensure that the majority of the larger sequence has been included and characterized in one or more subclones. For large sequences (such as cosmids), the minimum number of subclones required for complete coverage is prohibitive. Therefore, in practice the process is usually continued until 75 to 90% of the original fragment is covered by nested clusters of clones called contigs, which are separated by relatively short gaps. Directed methods can then be used to obtain the sequence of these gaps. Most of the large sequencing projects undertaken to date have used a mixture of random and directed sequencing (*12*). An example of the number of clones and final redundancy in a large shotgun sequencing project, 94.6 kb of DNA in the mouse T cell receptor and locus, is given in Table 1. Although precise error levels are difficult to assess, we believe the error in this DNA segment is less than 1 in 1000 bases.

Multiplex sequencing. An interesting variation on shotgun sequencing has been described that maximizes the amount of information obtainable from a single gel (*24*). In this approach, multiple random libraries are generated from the same initial source of DNA (Fig. 4). The subcloning vector used for each distinct library has a unique tag sequence downstream from the common priming site. Single clones from each of the different libraries are combined into a pool. Each pool is sequenced chemically or enzymatically, and the reaction products are separated as a four-lane set. The separated DNA fragments are then transferred to a blotting membrane. Hybridization of the blot with a labeled oligonucleotide probe complementary to a particular tag sequence illuminates only the sequence reactions from the subclones with that particular vector. Successive steps of washing and rehybridization with different

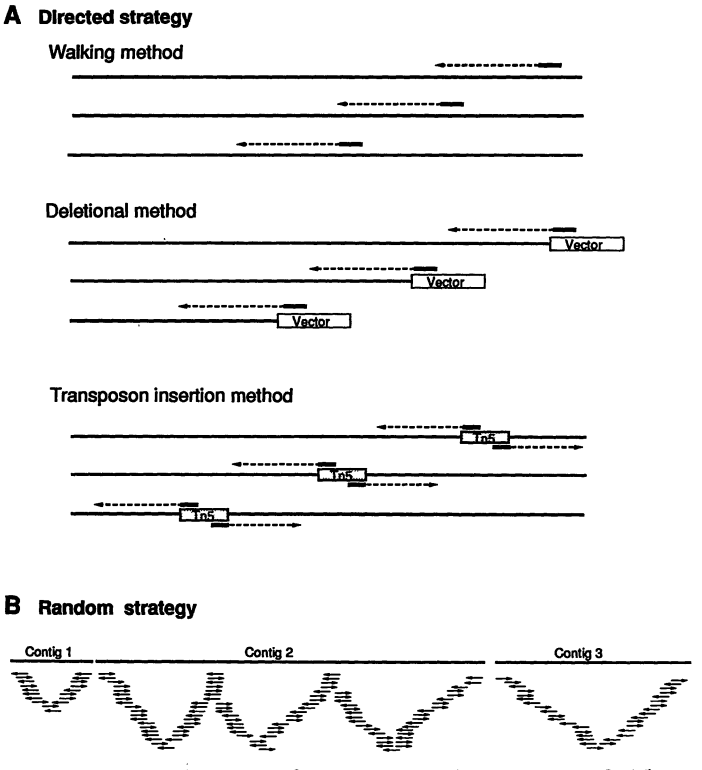

Fig. 3. Two general strategies for DNA sequencing (see text). (**A**) Three directed approaches. The solid bar indicates a primer oligonucleotide and the dotted line actual sequence. (**B**) The random approach generates sequence contigs separated by sequence gaps. The arrows indicate regions obtained from sequenced clones.

Table 1. Statistics associated with the shotgun or random analysis of 94.6 kb of DNA sequence from the mouse T cell receptor α locus (*49*).

Total bases read	Number of clones sequenced	Number of useful clones*	Average bases read per clone	Finished bases	Average redundancy
695,866	2,399	1,577	441	94,647	7.4

*Clones used for sequence assembly.

probes allow for the detection of the sequences from each library. Thus, many different random sequences can be determined from a single gel. The general acceptance of multiplexing for large-scale sequencing is greatly dependent on the development of robust instrumentation for the automatic reading of autoradiographs, since the amount of data that can be obtained per gel cannot easily be analyzed manually in a timely manner.

In an adaptation of the multiplex method (*1, 25*), a genomic DNA sample that has been first digested with an appropriate restriction enzyme is sequenced by chemical cleavage. The complex mixture of reaction products is separated by electrophoresis, and the fragments are transferred from the gel to a blotting membrane. Sequence ladders are then visualized with a labeled known single-copy probe. Cycles of washing and rehybridization with other unique probe sequences allow for the accumulation of large amounts of sequence data from the original genomic sample. The method may be particularly sensitive to the presence of repeat sequences in the genome, which complicates the selection of new unique probe sequences. Currently this approach is being tested in the sequencing of the relatively simple mycoplasm genome (~1,000,000 bp) (*25*).

The polymerase chain reaction and DNA sequencing. The polymerase chain reaction (PCR) is a procedure by which a specific segment of DNA can be amplified by 1 millionfold or more (Fig. 5) (*26*). Two primers of known sequence that flank the region to be amplified initiate double-stranded DNA synthesis in the presence of nucleotides and DNA polymerase. The primers are oriented on opposite strands such that synthesis occurs across the region between the primers. Repeated cycles of heat denaturation (high-temperature separation of the double-stranded DNA into its component single strands), primer annealing (hybridization of the primers to their complementary sequences on the single strands), and extension (new strand synthesis by the polymerase) effect amplification through an ideal doubling of the original target sequence in each cycle. The use of a thermostable DNA polymerase allows the cycling to be carried out automatically without the need for the addition of fresh enzyme after each high-temperature step. PCR has had a significant impact on DNA sequencing in predominantly three ways.

1) Because DNA fragments of length that can be sequenced can be obtained by amplification directly from genomic or large insert cloned DNA, PCR can obviate the need for many of the laborious subcloning and preparative scale procedures currently required to obtain sufficient DNA for sequence analysis. Several procedures have been developed for the direct sequencing of double-stranded PCR products (*27, 28*). Asymmetric PCR (*28, 29*), a variation in which amplification is performed under conditions that result in a preferential amplification of only one of the complementary target strands, has been useful in the generation of single-stranded templates for enzymatic sequencing. Similarly, conventional PCR has been performed with a primer set in which one of the primers has been chemically modified with biotin. The resulting PCR products have one of the two strands specifically end-labeled with biotin that can be immobilized on an avidin-coated solid matrix for strand separation (*30*). Both the free strand and the support-bound strand have been sequenced. The use of magnetic microparticles as the solid support simplifies handling steps and offers possibilities for automation of the procedure (*31*).

2) Sequence contigs can be ordered (see Fig. 3B) by performing PCR reactions with all or a subset of the pairwise combinations of primers generated from the end sequences of the contigs. Once the contigs are ordered and the gaps between contigs are determined, the appropriate primer pairs can be used to generate the physical DNA necessary to fill in the gap sequences.

3) Enzymatic sequencing reactions can be thermally cycled with a thermostable DNA polymerase in a manner analogous to PCR amplification, but only one primer is used (*32*). This procedure, termed "cycle sequencing," has several advantages. First, it requires greatly reduced quantities of starting template because of the linear amplification achieved in the temperature-cycling process. Second, repeated cycles of denaturation, primer annealing, and strand synthesis afford an efficient method for the enzymatic sequence analysis of double-stranded templates, a usually inefficient procedure with conventional approaches. Third, it renders the enzymatic sequencing method easily automatable, since no reagent additions are required after the initial assembly of the reaction mixtures.

Automation and Robotics

In attempts to increase both the throughput and consistency of sequencing efforts, as well as decrease the overall cost, many of the constituent operations of sequencing have recently been automated. Primarily, these include "front end" operations such as subclone selection, template preparation, and performance of sequencing reactions, as well as the tasks of separation, raw data acquisition, and sequence determination (base calling). The assembly of overlapping fragments of data into contigs has also been partially automated through the use of computer software (*10*) and special-purpose hardware (*33*).

Clone-plaque selection. Various efforts (*34*) have used some combination of commercial and custom-built robotics and vision-simulation systems to automatically survey and select individual elements (colonies or plaques) from clone libraries. These have proven to be workable but not necessarily useful in the context of smaller scale efforts. The impact of these technologies should increase significantly as larger projects proceed, particularly large shotgun projects. Their ability to regenerate a clone library in microtiter format should also facilitate the construction of arrayed clone libraries of chromosomes, aiding the efforts of physical mapping, clone selection, and library maintenance. Currently, none are commercially available as integrated functional systems.

Template preparation. Existing robotic workstations have been used to automate the isolation of source DNA as well as the preparation of sequencing templates by adapting conventional procedures to the workstation format (*35*). Special purpose devices for DNA isolation have also been introduced (*36*).

Sequencing reactions. Automation of the DNA sequencing reactions is critical to the success of any large-scale sequencing effort because the long-term reproducibility and throughput required are generally beyond the capability of most individuals to maintain.

Seiko developed the first robotic system designed to automate the performance of chemical sequencing reactions (*37*). However, the instrument was never generally available. Others have modified general purpose laboratory robots to perform at least a subset of the redundant pipetting tasks of sequencing. For example, the Beckman Instruments Biomek 1000 robotic workstation has been used extensively for the automation of Sanger sequencing reactions (*38*). The Biomek uses a standard 96-well microtiter plate format and single- and multitip pipetting heads fitted with disposable pipette tips to dispense reagents. Modifications to the basic instrument, including a temperature-controlled reaction block, have been made to improve its usefulness in DNA sequencing.

An instrument designed to address the specific requirements of DNA sequencing reactions has recently been announced [Applied Biosystems, Inc. (ABI)] (*39*). The instrument is designed to perform template preparations as well as a variety of DNA sequencing methods, including cycle sequencing. It is characterized by highly precise control of temperature, evaporation, and small-volume reagent delivery and is designed to ensure long-term reproducibility and accuracy. An even more ambitious system has been described by researchers in Japan (*39*). It is an array of integrated robotic modules that has been designed to automate the entire process of enzymatic sequencing, from clone purification to base calling. However, there are no current plans to make the system generally available.

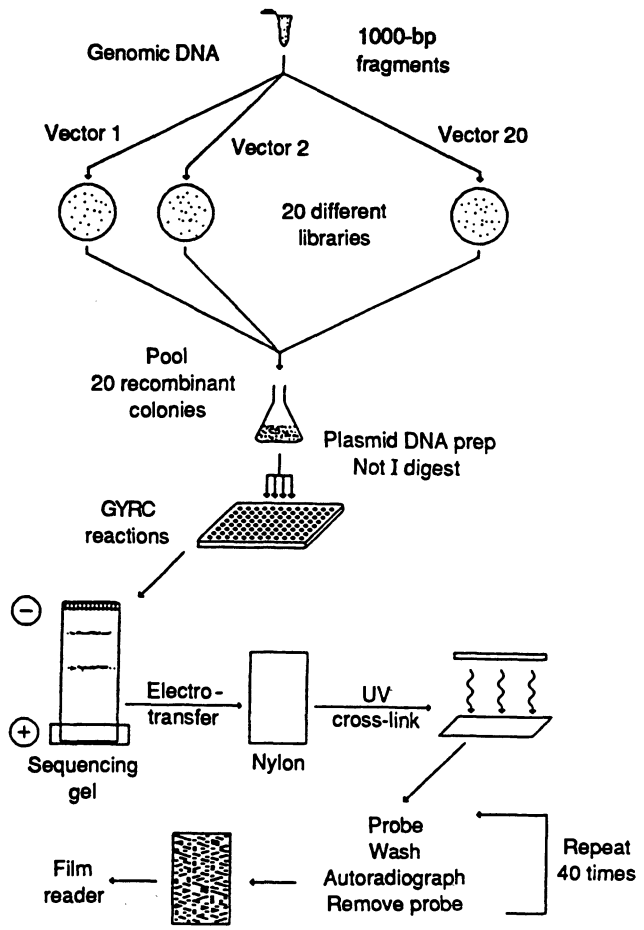

Fig. 4. The steps of multiplex sequencing. Genomic DNA is used to construct 20 libraries, each using a distinct vector (1 through 20). One clone from each library is added to form a pool. This process is repeated to fill a 96-well microtiter plate in which chemical sequencing reactions are carried out. The fragments in each well are separated by gel electrophoresis and transferred to a membrane. Forty autoradiographs are obtained by sequentially probing the membrane with 40 vector-specific labeled probes. Y represents pyrimidine. R represents purine. Not I is a restriction enzyme that cleaves DNA infrequently. Other rare-cutting enzymes might also be used. From (*24*).

Automated autoradiograph readers. The ladders of bands on an autoradiograph must be translated into DNA sequence. Although traditionally this has been done manually, a number of attempts have been directed at the automation of gel reading. A number of commercial and academic autoradiograph readers have been developed (*40*). The ability to determine the DNA sequence from four reactions separated in independent electrophoretic lanes is limited by the ability to establish the correct register among the lanes. Current variations and temperature gradients across the plane of a gel can lead to significant mobility artifacts, which warp the lanes and bands and obscure the actual order of the fragments. Also, detectable resolution between bands decreases with length of the fragments because of the inverse logarithmic dependence of the separation. Physical distortion of the gel is also likely after electrophoresis during the manipulations required for autoradiography. Unfortunately, because of these complications the development of an instrument and the necessary software that can consistently read gels beyond ~300 bases has proven to be very difficult. Hence, the autoradiograph readers developed to date have required significant subjective input by the operator. Much work has been focused on improving the consistency of the gels themselves to circumvent at least some of these problems. Others have eliminated the physical manipulation of the gel by electrophoresing the DNA fragments directly onto a blotting membrane (*41*). The membrane is continually moving across the bottom of the gel to maintain the band separation. The bands are then visualized either with standard autoradiography or through the use of chemiluminescent reagents.

Automated sequencers. Several automated DNA sequencers have been developed (*42–46*) that have been designed to automate gel electrophoresis, raw data acquisition, and the base-calling steps. Those instruments designed for greater throughput use fluorochrome labeling and laser-activated fluorescence detection to identify the separated products of the sequencing reactions in real time as they migrate through a gel. Two basic approaches have been implemented: use of single fluorescent label with four-lane separation and use of four fluorescent labels with single-lane separation. Sufficient experience has been obtained to date from three commercial instruments to allow comment on their performance.

1) Single-label, four-lane approach. The use of a single fluorescent label with four-lane separation was pioneered by workers at the European Molecular Biology Laboratory (EMBL) (*44*) and also by researchers at Hitachi Instruments in Japan (*46*). In this approach, the primer used in enzymatic sequencing reactions is labeled with a highly sensitive fluorescent reporter. Four reactions are performed with this primer, and the products are separated in four adjacent lanes, directly analogous to conventional methods with autoradiography. As the fluorochrome-labeled fragments migrate through the gel, they are excited by a laser and the emitted light is detected to generate an image of the gel and the bands. A computer is used to correlate the raw data from the four lanes and to determine sequence. However, automated sequence determination with this approach is sensitive to electrophoretic artifacts arising from gel heterogeneity that distort lanes in exactly the same manner as traditional autoradiograms.

Pharmacia currently markets the Automatic Laser Fluorescent (A.L.F.) DNA sequencer (*47*) based on the prototype EMBL instrument. The laser beam of the A.L.F. is directed through the gel across its width (between the glass gel plates). Fluorescence detection is done with 40 identical, fixed detectors, one per lane, that are positioned across the face of the gel. Thus, no moving parts are involved. The 40 detectors allow for the simultaneous recording of ten sequences. Since this detection is done in parallel, the run time

of the A.L.F. is relatively short (~6 hours). Active temperature control of the gel is provided, which results in more consistent electrophoretic separations. The A.L.F. is capable of consistently generating up to 500 bp of raw sequence per reaction.

The optical system of the A.L.F. has been optimized to produce a strong fluorescence signal with the single dye, thus providing for very sensitive detection. Additionally, the labeling chemistry used in connection with the A.L.F. allows for the use of conventional automated DNA synthesis procedures. Furthermore, crude labeled primers can be used effectively in sequencing without the need for lengthy purification. These two considerations make the A.L.F. strategy attractive for projects that use a significant specific-primer–directed sequencing component. The labeling chemistry is also compatible with chemical DNA sequencing.

2) Four-label, single-lane approach. Based on work originated in this laboratory (*48*), ABI has developed chemistry and instrumentation that allow for all four DNA sequencing reactions to be electrophoresed together in a single gel lane (*42*) (Fig. 6). This technology labels each of the four base reaction sets with a different fluorescent dye. All four reactions can be combined and separated in the same lane, with the products of each reaction distinguished by their respective colors. The order of colors passing the fluorescent detector is translated directly into sequence data by a computer. Since all of the fragments of one reaction set migrate through the same gel path, lane-to-lane distortions in the gel have no impact on the ability to determine the order of the different colored fragments in a given lane.

The original chemistry used with the ABI instrument (the 373A) labels different aliquots of the primer used for enzymatic sequencing with the four different fluorescent tags. Each labeled primer is then used to generate the nested fragments for one of the bases. The separate A-, C-, G-, and T-specific reactions are then combined for co-electrophoresis. The beam of an argon ion laser is mechanically scanned across the gel near its bottom to excite the labeled fragments undergoing electrophoresis. The emitted light is collected and focused through a four-wavelength selectable filter wheel onto a single, mobile photomultiplier to give a continuous four-point spectrum of the detected radiation. Each fluorescent label gives rise to a characteristic emission spectrum. After the application of various signal filters to separate the dye signatures and an adjustment for known mobility variations associated with the four different fluorochromes, the color order of the passing bands is translated directly into sequence data by a computer. The mechanical operation of the four-color detection optics results in relatively long run times on the 373A (12 to 14 hours). However, it also allows for the analysis of at least 24 separate sequences per gel and is not sensitive to lane shifting in the gel.

Use of dye-labeled primers has proven effective with all of the DNA polymerases commonly used in standard sequencing protocols. However, the time, labor, and expense required to generate a set of four primers each labeled with a different fluorescent reporter exceeds that of conventional primer synthesis. Therefore, the 373A is much better suited for use in random sequencing strategies that use a limited set of primers for all reactions than in directed strategies that require routine primer synthesis. New dye-labeled terminator chemistry may eventually overcome these limitations.

We have sequenced several thousand subclones to date with the 373A and have found that it is capable of consistently resolving from 450 to 500 bp from the priming site per sequencing reaction with an error plus ambiguity rate on the order of 1 to 5%. Of course, the actual results depend on the quality of the template DNA and the sequencing methodology.

DuPont also developed and marketed an automated DNA sequencer, the Genesis 2000, which relies on the four-color, single-lane approach (*43*). This approach relies on the use of labeled chain terminators and allows all four base-terminating reactions to be carried out in the same tube. Its detection system uses two photomultiplier tubes and two filters for data acquisition. Detection and identification of the colored bands is based on the ratio of emitted light intensities detected by each filter plus photomultiplier pair, a value characteristic of each fluorochrome label. Typical throughput is on the order of 250 to 350 bp per reaction with ten lanes on each gel. The Genesis 2000 is no longer commercially available.

Large-scale sequencing and automated sequencers. Our laboratory has had extensive experience with both manual autoradiographic and automated sequencing methods and has recently determined nearly 250,000 bp of finished sequence with a combination of random and directed approaches (*49*). From this experience, we have chosen to conventional enzymatic sequencing. Small quantities of template can be sequenced and of sequence can be obtained simply and efficiently from double-stranded templates.

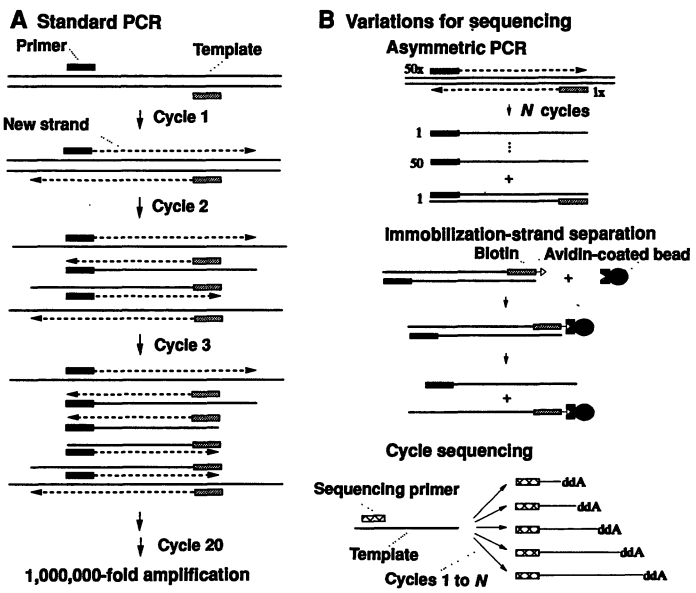

Fig. 5. Schematic diagram of the polymerase chain reaction (PCR, see text) and some variations useful in DNA sequence determination. (**A**) The standard PCR procedure. Each cycle consists of the following three steps. First, the double-stranded template DNA is denatured (separated into its component single strands). Second, oligonucleotide primers flanking the target region to be amplified are annealed (hybridized) to the denatured template. Third, DNA polymerase and nucleotide triphosphates are used to synthesize two new complementary strands from the annealed primers. In principle, each cycle doubles the number of copies of the target region, such that after 20 cycles, a greater than 1 millionfold amplification of the original target sequence has been achieved. (**B**) In asymmetric PCR one primer is used in excess of the other. Initial amplification is exponential as with standard PCR until the quantity of the limiting primer is greatly reduced. At this point, continuing amplification is essentially linear, resulting in the production of single strands originating from the excess primer. These single strands can then be sequenced with the enzymatic method. In immobilization-and strand separation, biotin can be chemically incorporated into either of the two primers used for standard PCR amplification, thus resulting in a PCR product in which one of the two strands carries a covalently attached biotin. The biotinylated PCR products can be efficiently immobilized on a solid matrix coated with avidin, a protein having a very high affinity for biotin. The captured DNA can be denatured, allowing the non–biotin-containing strand to be separated from its immobilized complement. Both strands can then be separately sequenced with the enzymatic method. Cycle sequencing is a modification of asymmetric PCR in which a single primer is used to linearly amplify a region of template DNA in the presence of chain-terminating dideoxynucleotide triphosphates. After many cycles, an amplified nested set of fragments is produced, similar to that obtained with

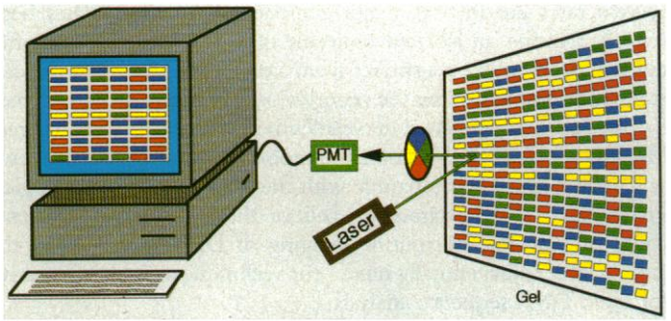

Fig. 6. A schematic illustration of the principle of operation of the ABI 373A automated sequencer (see text); PMT, photomultiplier tube. [Adapted from (*42*) with permission © Eaton]

rely on the automated approach for further efforts of the large-scale project under way. We believe these instruments perform more consistently on a routine basis and have a lower long-term error rate than manual sequencing with isotopic labeling. Although particular individuals can manually obtain a much greater number of bases from a single reaction set through the use of multiple loadings, wedge gels, and ^{35}S labeling, the average number of total bases and finished sequence length generated is no more and is usually less than that obtained through the use of the automated instruments. The manual methods are also very difficult to adequately automate for truly large-scale efforts. We are using the 373A for massive shotgun sequencing and are developing methods to increase its utility for the directed sequencing needed for closure. Because of the simplicity of the primer labeling chemistry for the A.L.F., at least one other group involved in large-scale sequencing is implementing a strategy whereby they use the 373A for bulk random sequencing and the A.L.F. for specific-primer–based sequencing (*50*).

Future DNA Sequencing Technologies

In order to achieve the goals of the Human Genome Project in a timely and cost-effective manner, it is obvious that the current throughput for DNA sequencing must be increased over the next 10 years by 100- to 1000-fold. Such improvement can be obtained in several ways. In the short term, investigations into the means of improving existing sequencing strategies and technologies are well under way. We believe that the throughput of existing systems can be improved at least 100-fold through the clever integration and application of existing technologies. Alternatively, in the longer term, a new and different technology could conceivably replace present methods and increase the rate of DNA sequencing by many orders of magnitude.

Improvement of current approaches. There are a number of ways in which current automated sequencing technology can be significantly improved. It is possible to use more lanes per gel, thus doubling or tripling the number currently analyzed on the automated sequencers. Improvements in the software for the analysis of the raw fluorescence data are likely to extend both the number of lanes as well as the length of sequence that is readable even with the current separation techniques. It is also possible that new gel and electrophoresis protocols (such as the use of field inversion techniques) can provide much sharper resolution of bands, further extending the amount of data from a single reaction. Modifying optical detection systems as well as developing alternative fluorescent labels could improve detection sensitivity, with a concomitant reduction in both the quantities of template required and the degradation of resolution from overloading effects. This in turn would affect methods of template purification, allowing the use of faster, more easily automatable procedures despite their lower yield. The added sensitivity would also facilitate the use of more lanes and enable the use of ultrathin gels, affording the positive advantages in speed and resolution discussed above. These are relatively straightforward developments that do not require a significant new technology or alteration in the basic strategies to provide, together, significant improvement in throughput. For example, if the number of lanes were to double, the read length was increased by 25%, and the speed of electrophoresis increased only three to four times, the potential throughput would be improved by a factor of 7 to 10. We believe that even this value may prove modest.

Additional compatible automation of the numerous steps in the DNA sequencing process (Fig. 1) would not only enormously facilitate the overall process, but would be required if the potential throughput of the automated sequencers is to be matched. Loading of the sequencing reactions onto the gels is currently an extremely tedious and painstaking manual procedure. Moreover, the quality of loading significantly affects the band shapes and resolution and thus can simplify or complicate the automated sequence determination. Automation of this procedure could have a significant impact on the quality and the reproducibility of the gel separations as well as facilitate more accurate sample tracking. With careful consideration and reevaluation of this and other front end processes in the context of large-scale efforts, it should be possible to develop techniques and instrumentation whereby one can readily go from a single clonal plaque or colony on a culture plate to a DNA sequence within 1 day in a completely automated series of modular steps. PCR amplification is likely to be crucial to this process.

New technologies. Although early efforts are promising, it is unclear at present what cost, throughput, and accuracy improvements can ultimately be obtained through the current efforts at automating the classical enzymatic and chemical approaches to DNA sequencing. Therefore, there has been a recent drive to explore alternative sequencing technologies, which has resulted in the emergence of a wide variety of new methodological strategies. Four are of particular interest.

1) Mass spectrometry. Early attempts at the application of mass spectrometric methods to nucleic acid analysis used negative ion fast atom bombardment and plasma desorption to produce suitable ionized DNA species for mass determination (*51*). Unfortunately, only relatively small intact oligonucleotides have been successfully analyzed with this method. However, recent advances in ionization technology, notably electrospray ionization (*52*) and matrix-assisted laser desorption (*53*), offer promise for the analysis of large nucleic acid fragments. Use of the quadrupole ion trap, Fourier transform mass spectrometer, or time-of-flight mass spectrometer should provide adequate sensitivity for DNA sequencing.

In general, three potential approaches for the application of mass spectrometry to DNA sequencing can be envisioned. First, the mass spectrometer can be used to detect and analyze DNA fragments produced in conventional DNA sequencing reactions and separated by gel electrophoresis. The mass spectrometer can be used to detect different mass labels attached to fragments from each of the four sequencing reactions, analogous to the four-color fluorescence approach, or perhaps to analyze the masses of the separated fragments themselves and subsequently deduce the sequence. Second, the mass spectrometer can be used to directly analyze the ionized components of an unfractionated mixture of sequencing reactions, bypassing the need for electrophoretic separation. Finally, a large DNA fragment can be ionized and its sequence assembled from data obtained from repeated cycles of fragmentation and mass analysis in a multiple mass spectrometer experiment. Successful application of mass spectrometers to DNA sequence analysis could provide an enormous increase in throughput potential (*54*).

2) Sequencing by hybridization. Any DNA sequence can be thought of as an assembly of overlapping shorter subsequences. Sequencing by hybridization (SBH) is a technology in which the set and order of subsequences is determined by oligomer hybridization (*55*). It has been demonstrated mathematically that the sequence of a relatively short DNA fragment can be assembled from data obtained by hybridizing a library of all possible *N*-mer oligonucleotide sequences to the fragment and recording which oligomers provide perfectly complementary hybrids (*55*). In general, *N* is in the range 8 to 10 to provide manageable library sizes while affording reasonable hybridization parameters (for example, there are 65,536 octamers). Such an approach should be simple, low cost, and reasonably rapid. However, sequence assembly from the hybridization data is computationally quite demanding, and errors arising from difficulty in distinguishing between perfect matches and single-base mismatches may prove too numerous to afford completely objective sequence determination, although the volume of sequence information afforded by the technique tolerates significant error rates. Also, repeated sequences cannot be determined unambiguously by SBH. The strategy has recently been successfully tested on a 100-bp template of known sequence with a limited library of oligomers (*56*). Automation of the SBH procedure should benefit from advances in the technology of assembling high-density probe arrays in a microformat (*57*), such as the light-directed spatially addressable biopolymer synthesis techniques recently reported by Affymax, Inc. (*58*). It is possible that the most productive future uses of this technology would be in the rapid screening of DNA samples for the presence of mutations and for resolving ambiguities or verifying the sequence information obtained by other means.

3) Single-molecule sequencing. Workers at the Los Alamos National Laboratories (LANL) have developed a technology for the detection of fluorescence from single molecules and have proposed a strategy for rapid DNA sequencing (*59*). Template DNA molecules would first be enzymatically modified (or synthesized) to contain only bases possessing distinguishing fluorochromes. A single labeled molecule would be selected and immobilized with light pressure from two laser beams ("laser tweezers"). The DNA molecule would then be suspended in a flow system and subjected to treatment with a processive exonuclease that sequentially removes bases from one end of the immobilized nucleic acid. The flow carries the excised labeled bases through the single-molecule fluorescence detector, the order of detected labels specifying the DNA sequence. The LANL group is currently developing the required biochemical and molecular manipulation technologies necessary to implement this potentially powerful sequencing approach.

4) Atomic probe microscopy. Recent developments in atomic probe microscope technology such as scanning tunneling microscopes (STM) and atomic force microscopes (AFM) have led to vigorous attempts to develop reliable methods for the rapid, high-resolution direct imaging of DNA molecules (*60*). In STM, a conductive tip of near-atomic dimensions is used to scan the surface of a sample. Continuous measurement of the electronic tunneling current between the moving tip and the surface gives rise to a three-dimensional image. Numerous research groups have attempted to image DNA fragments by STM with varying degrees of success. Individual nucleotides in single-stranded DNA samples have been resolved with the STM. The STM images at atomic resolution of a double-stranded DNA clearly show the double-helical structure and individual base pairs and correlate well with dimensions obtained by x-ray crystallographic determinations. Other groups are attempting to use AFM (*61*), which measures the force between a scanning tip and a surface, for the atomic-scale imaging of DNA. Although results from scanning tip technologies have been promising, the procedures are far from reliable at present and the success rates are low. A major impediment to success has been reliably locating an isolated molecule of DNA to be imaged after random deposition on a surface from a dilute solution. This process is further complicated by the observation that the moving tip tends to displace the deposited molecules before they can be imaged. Much work is devoted to robust immobilization of DNA fragments on surfaces in ways compatible with the scanning tip technologies. Proponents of the technology believe that future improvements could provide for the routine imaging of DNA molecules at the atomic level and eventually render the technique a viable alternative for rapid DNA sequence analysis.

Discussion

The potential impact of automation on large-scale DNA sequencing is clear. For example, the ABI 373A is capable of producing 12,000 bp of raw data per day, 60,000 bp per working week, and 3,120,000 bp per year. Despite this potential, however, there have been no sequences greater than 30,000 bp deposited in the public databases that have made exclusive use of one of the commercial automated DNA sequencers, despite their having been on the market for several years. [One large sequence was done with the prototype of the A.L.F. (*62*).] There are several reasons for this. First, the automated sequencers and the protocols and strategies to use them efficiently have only relatively recently become reliable for large-scale projects. Second, our own experience indicates that the major bottlenecks to large-scale sequencing have not been the automation of the data acquisition and base calling, but rather the inability of conventional methods to provide consistently reliable DNA templates and sequencing reactions. Also, the computational methods for handling, assembling, and analyzing enormous amounts of raw sequence are just being developed. Third, the scale of sequencing has generally been much less in the past. However, several laboratories, including our own, have now undertaken large-scale projects relying on automated procedures and have begun to generate hundreds of kilobases of data. These data should begin to appear in the national databases during the next year.

Although there is debate on the issue, we believe that large-scale DNA sequencing requires an organizational structure quite distinct from those of various research efforts commonly found in academia.

1) A sequencing production line must be established that uses the most powerful current technologies and robust and stable protocols. The key is to generate sequence continually at high throughput independent of the other demands of large-scale sequencing.

2) Separate groups within the sequencing project should be established to govern quality control, troubleshooting, sequence closure and assembly, and protocol development and testing. This latter group should focus on identifying and eliminating bottlenecks in the overall procedure (Fig. 1). For example, automated sequencing instruments with higher throughput would not add to the overall rate of sequence generation if serious bottlenecks exist in the front end operations or sequence assembly and analysis procedures.

3) The computational problems in large-scale sequencing are still largely unsolved. It is important to establish the laboratory information management systems needed to track and record all of the details of the sequencing procedures, to identify bottlenecks and suggest better sequencing strategies, to monitor costs, and to facilitate the final assembly process of finished DNA sequence. Quantization and statistics must be brought into sequence analysis so that in the future every base position can be reported with a four-dimensional probability vector (for example, position 507 is 0.97 A, 0.02 G, 0.005 T, and 0.005 C). The question of an acceptable error rate for the finished sequence is a matter of vigorous

debate because of the increasing cost associated with decreasing error rate. We believe that a rate of 0.001 to 0.0001 is acceptable in genomic sequencing. Likewise, new computational techniques must be developed for delineating the features of DNA sequence (for example, coding regions, regulatory elements, and so forth). Advances in informatics and biological computation are critical to the success of large-scale sequencing.

We anticipate that during the next 10 years DNA sequencing technologies will experience a 100-fold increase in throughput, a 100-fold decrease in cost (from current estimates of $2 to $10 per base), and a significant decrease in error rate. We suggest that there is about a 50% chance that large-scale DNA sequencing in 10 years would use greatly improved conventional techniques, and an equal chance that an entirely new DNA sequencing technology would emerge. In either case, the sequence information generated during the next 15 years should revolutionize our understanding of biology and fundamentally alter the practice of medicine.

REFERENCES AND NOTES

1. Office of Technology Assessment, *OTA-H-298* (U.S. Government Printing Office, Washington, DC, 1986); C. DeLisi, *Am. Sci.* **76**, 488 (1988); L. Roberts, *Science* **249**, 1497 (1990).
2. D. T. Burke, G. F. Carle, M. V. Olson, *Science* **236**, 806 (1987).
3. N. Sternberg, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 103 (1990).
4. J. Collins and B. Hohn, *ibid.* **75**, 4242 (1978).
5. N. E. Murray and K. Murray, *Nature* **251**, 476 (1974).
6. J. Sambrook, E. F. Fritsch, T. Maniatis, *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1989).
7. A. T. Bankier and B. G. Barrell, in *Nucleic Acids Sequencing: A Practical Approach*, C. M. Howe and C. J. Rawlings, Eds. (IRL Press, Oxford, 1989), pp. 37–73.
8. J. Messing and A. T. Bankier, in *ibid.*, pp. 1–36.
9. L. Roberts, *Science* **252**, 1618 (1991).
10. P. M. Rice, K. Elliston, M. Gribskov, in *Sequence Analysis Primer*, M. Gribskov and J. Devereux, Eds. (Stockton, New York, 1991), pp. 1–59.
11. M. S. Chee *et al.*, *Curr. Top. Microbiol. Immunol.* **154**, 125 (1990).
12. T. Hunkapiller, R. J. Kaiser, B. F. Koop, L. Hood, *Curr. Opin. Biotech.* **2**, 92 (1991).
13. A. Maxam and W. Gilbert, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 560 (1977).
14. F. Sanger, S. Nicklen, A. R. Coulson, *ibid.*, p. 5463.
15. F. Sanger, A. Coulson, B. Barrell, A. Smith, B. Roe, *J. Mol. Biol.* **143**, 161 (1980).
16. B. W. Birren, M. I. Simon, E. Lai, *Nucleic Acids Res.* **18**, 1481 (1990).
17. H. Swerdlow and R. Gesteland, *ibid.* **18**, 1415 (1990); J. A. Luckey *et al.*, *ibid.*, p. 4417; R. L. Bromley, Jr., and L. M. Smith, *ibid.*, in press.
18. R. Tizard *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 4514 (1990).
19. F. W. Studier, *ibid.* **86**, 6917 (1989); W. Szybalski, *Gene* **90**, 177 (1990); D. R. Siemieniak and J. L. Slightom, *ibid.* **96**, 121 (1990).
20. M. Poncz *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 4298 (1982); S. Henikoff, *Methods Enzymol.* **155**, 156 (1987); D. A. Sorge and L. A. Blinderman, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 9208 (1989).
21. T. Adachi *et al.*, *Nucleic Acids Res.* **15**, 771 (1987); A. Ahmed, *Methods Enzymol.* **155**, 177 (1987).
22. J. Messing, R. Crea, P. H. Seaburg, *Nucleic Acids Res.* **9**, 309 (1981); S. Anderson, *ibid.*, p. 3015; A. T. Bankier and B. G. Barrell, *Tech. Nucleic Acid Biochem.* **B5**, 1 (1983); P. Deininger, *Anal. Biochem.* **129**, 216 (1983).
23. R. Staden, in *Nucleic Acid and Protein Sequence Analysis: A Practical Approach*, M. J. Bishop and C. J. Rawlings, Eds. (IRL Press, Oxford, 1987), p. 173.
24. G. M. Church and S. Kieffer-Higgins, *Science* **240**, 185 (1988).
25. W. Gilbert, personal communication.
26. K. Mullis *et al.*, *Cold Spring Harbor Symp. Quant. Biol.* **51**, 275 (1986); R. K. Saiki *et al.*, *Science* **230**, 1350 (1985); R. K. Saiki *et al.*, *ibid.* **239**, 487 (1988).
27. D. R. Engelke, P. A. Hoener, F. S. Collins, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 544 (1988).
28. M. A. Innis, K. B. Myambo, D. H. Gelfand, M. D. Brow, *ibid.*, p. 9436.
29. R. K. Wilson, C. Chen, L. Hood, *BioTechniques* **8**, 184 (1990).
30. L. G. Mitchell and C. R. Merril, *Anal. Biochem.* **178**, 239 (1989).
31. M. T. Uhlen, T. Hultman, J. Whalberg, *J. Cell. Biochem.* **13E**, 310 (1989); T. Hultman, S. Bergh, T. Moks, M. Uhlen, *BioTechniques* **10**, 84 (1991).
32. A. M. Carothers *et al.*, *BioTechniques* **7**, 494 (1989); D. P. Smith, E. M. Johnstone, S. P. Little, H. M. Hsiung, *ibid.* **9**, 48 (1990).
33. ABI inherit DNA analysis software and T. Hunkapiller, unpublished results.
34. S. Lewis, personal communication; P. Jones *et al.*, *Genome Mapping and Sequencing* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1991), p. 107.
35. J. Zimmerman *et al.*, *Methods Mol. Cell. Biol.* **1**, 29 (1989); E. R. Mardis and B. A. Roe, *BioTechniques* **7**, 840 (1989).
36. Autogen Instruments Inc. now markets an instrument for small-scale plasmid, phage, and cosmid preparations.
37. A. Wada, *Nature* **325**, 771 (1987).
38. R. K. Wilson *et al.*, *Biotechnology* **6**, 776 (1988); J. D. Cunha *et al.*, *BioTechniques* 9, 80 (1990); ABI Catalyst is marketed by Applied Biosystems Inc., Foster City, California.
39. D. Swinbanks, *Nature* **351**, 593 (1991).
40. J. K. Elder, D. K. Green, E. M. Southern, *Nucleic Acids Res.* **14**, 417 (1986); J. K. Elder and E. M. Southern, in *Nucleic Acid and Protein Sequence Analysis: A Practical Approach*, M. J. Bishop and C. J. Rawlings, Eds. (IRL Press, Oxford, 1987), p. 165.
41. F. M. Pohl and S. Beck, *Methods Enzymol.* **155**, 250 (1987). Betagen Corp. markets a commercial device.
42. C. Connell *et al.*, *BioTechniques* **5**, 342 (1987).
43. J. M. Prober *et al.*, *Science* **238**, 336 (1987).
44. W. Ansorge *et al.*, *Nucleic Acids Res.* **15**, 4593 (1987).
45. J. A. Brumbaugh, L. R. Middendorf, D. L. Grone, J. L. Ruth, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 5610 (1988).
46. H. Kambara, T. Nishikawa, Y. Katayama, T. Yamaguchi, *Biotechnology* **6**, 816 (1988).
47. M. C. Freeman, C. Baehler, S. Spotts, *ibid.* **8**, 147 (1990).
48. L. M. Smith, S. Fung, M. W. Hunkapiller, T. Hunkapiller, L. E. Hood, *Nucleic Acids Res.* **13**, 2399 (1985); L. M. Smith *et al.*, *Nature* **321**, 674 (1986).
49. R. K. Wilson *et al.*, unpublished results; B. K. Koop *et al.*, unpublished results.
50. R. K. Wilson, personal communication.
51. C. J. McNeal *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 77, 735 (1980); L. Grotjahn, H. Bloecker, F. Frank, *Biomed. Mass Spectrom.* **12**, 514 (1985); L. Grotjahn and L. E. Taylor, *Org. Mass Spectrom.* **20**, 146 (1985).
52. T. R. Covey, R. F. Bonner, B. I. Shushan, J. D. Henion, *Rapid Commun. Mass. Spectrom.* **2**, 249 (1988). J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, C. M. Whitehouse, *Science* **246**, 64 (1989).
53. B. T. Chait and R. C. Beavis, paper presented at the 200th American Chemical Society meeting, Washington, DC, 26 to 31 August 1990; M. Karas, D. Bachman, U. Bahr, X. Hillenkamp, *Int. J. Mass Spectrom. Ion Processes* **78**, 53 (1989).
54. K. B. Jacobson *et al.*, *Genomics* 9, 51 (1991).
55. R. Drmanac, I. Labat, I. Brunker, R. Crkvenjakov, *ibid.* **4**, 114 (1989); R. Drmanac, Z. Stevanovic, R. Crkrenjakou, *DNA Cell Biol.* 9, 527 (1990).
56. P. Crkrenjakou, personal communication.
57. E. M. Southern and U. Maskos, in *Genome Mapping and Sequencing* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1991), p. 195.
58. S. P. A. Fodor *et al.*, *Science* **251**, 767 (1991).
59. J. H. Jett *et al.*, *J. Biomol. Struct. Dyn.* 7, 301 (1989); D. C. Nguyen, R. A. Keller, J. H. Jett, J. C. Martin, *Anal. Chem.* **56**, 348 (1987).
60. S. M. Lindsay and M. Phillip, *Genet. Anal. Tech. Appl.* **8**, 8 (1991); D. P. Allison *et al.*, *Scanning Microsc.* **4**, 517 (1990); R. J. Driscoll, M. G. Youngquist, J. D. Baldeschwieler, *Nature* **346**, 294 (1990); M. Salmeron *et al.*, *J. Vac. Sci. Technol.* **8**, 635 (1990).
61. A. L. Weisenhorn *et al.*, *Scanning Microsc.* **4**, 511 (1990).
62. A. Edwards *et al.*, *Genomics* **6**, 593 (1990); L. Roberts, *Science* **250**, 756 (1990).

63. We thank the NSF, DOE, NIH, and Alfred P. Sloan Foundation for support, Drs. D. Nickerson and J. R. Yates for careful reading of the manuscript, and K. McCarthy for help in preparing the manuscript.