GATA 8(1): 1–7, 1991

DNA SEQUENCING

Editor's Note: These articles are the first in a series on DNA Sequencing edited by Lloyd M. Smith. Articles in future issues of **GATA** will cover DNA sequencing using mass spectrometry and fluorescence-based automated sequence analysis of DNA.

Rapid DNA Sequencing Based Upon Single Molecule Detection

LLOYD M. DAVIS,
FREDERIC R. FAIRFIELD,
CAROL A. HARGER,
JAMES H. JETT,
RICHARD A. KELLER,
JONG HOON HAHN,
LETITIA A. KRAKOWSKI,
BABETTA L. MARRONE,
JOHN C. MARTIN,
HARVEY L. NUTTER,
ROBERT L. RATLIFF,
E. BROOKS SHERA,
DANIEL J. SIMPSON, and
STEVEN A. SOPER

We are developing a laser-based technique for the rapid sequencing of 40-kb or larger fragments of DNA at a rate of 100 to 1000 bases per second. The approach relies on fluorescent labeling of the bases in a single fragment of DNA, attachment of this labeled DNA fragment to a support, movement of the supported DNA fragment into a flowing sample stream, and detection of individual fluorescently labeled bases as they are cleaved from the DNA fragment by an exonuclease. The ability to sequence large fragments of DNA will significantly reduce the amount of subcloning and the number of overlapping sequences required to assemble megabase segments of sequence information.

From the Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, New Mexico.

Address correspondence to: Dr. R. A. Keller, MS M888, Los Alamos National Laboratory, Los Alamos, NM 87545.

Received August 28, 1990; revised and accepted November 16, 1990.

Introduction

We are developing a laser-based technique for the rapid sequencing of 40-kb or larger fragments of DNA at a rate of 100 to 1000 bases per second [1]. The technique is based upon our ability to detect single chromophores by laser-induced fluorescence in flowing sample streams [2, 3]. This capability is the result of extensive research to improve the sensitivity of fluorescence detection in flowing sample streams [4–6]. The approach relies on fluorescent labeling of the bases in a single fragment of DNA, attachment of this labeled DNA fragment to a support, movement of the supported DNA fragment into a flowing sample stream, and detection of individual fluorescently labeled bases as they are cleaved from the DNA molecule by an exonuclease (Figure 1). The projected sequencing rate is based upon our experience with single molecule detection and known enzymatic cleavage rates. Just as important as the high rate is the ability to sequence large fragments of DNA. This will significantly reduce the amount of subcloning and the number of overlapping sequences required to assemble megabase segments of sequence information. An additional advantage of our approach is the elimination of problems associated with the disposal of radioactive materials and other hazardous wastes, such as acrylamide, which plague current sequencing efforts.

A worldwide effort is now in progress to determine the base sequence of a human genome. When complete, this sequence will consist of $\sim 3 \times 10^9$ bases put into 23 continuous subsequences, one for each human chromosome. With current technology, a continuous sequence of 50,000 bases or an ordered set of DNA fragments covering 1,000,000 bases is considered state of the art [7].

Currently, a sequence of DNA must be built up from many overlapping subsequences, 200–600 bases long. With the overlap requirement, a minimum of 9×10^9 bases must be sequenced. In practice, raw sequence data contains ambiguities that must be resolved by more sequencing reactions. Thus, for a finished, believable sequence of 3×10^9 base pairs, $\sim 5 \times 10^{10}$ base pairs of raw sequence must be accumulated. Even at the rate of 50,000 bases of finished sequence per day (projected for new automated sequencers employing fluorescent detection of electrophoretically separated fragments [8, 9]), sequencing the human genome is a formidable task.

1050-3862/91/$03.50

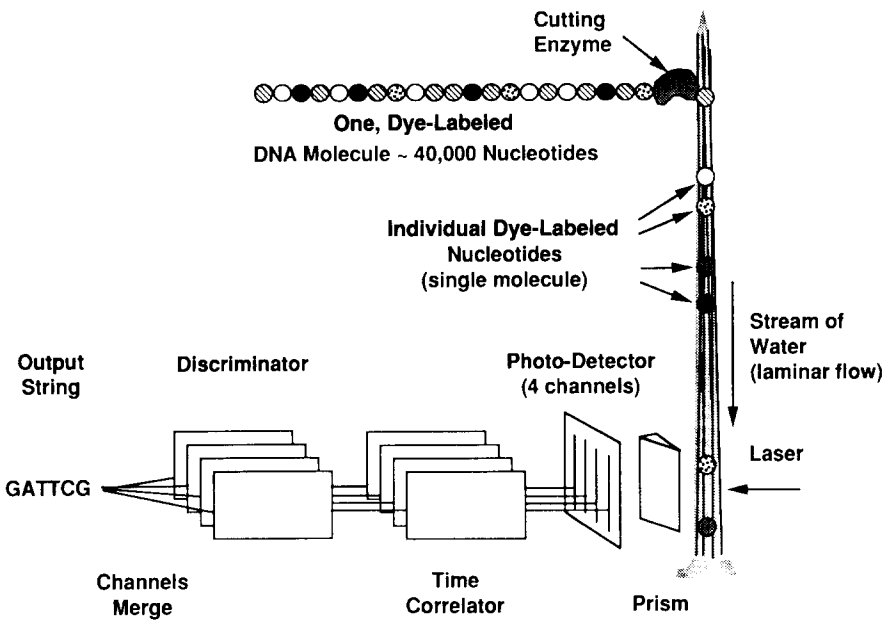


Figure 1. Pictorial representation of a DNA rapid sequencer.

We are attempting to decrease the effort required for sequencing DNA by developing a technique that increases the raw sequencing rate by a factor of ~1000 over the new automated sequencers. Our approach should increase the finished sequencing rate even more because we expect to sequence entire 40-kb DNA fragments in one reaction, thereby reducing the number of overlapping sequence fragments required.

Approach

A summary of our approach is outlined in Figure 2. Details of each step are given below.

Fluorescent Labeling of Bases

Individual free and bound bases found in DNA have intrinsic fluorescence quantum yields of $<10^{-3}$ at room temperature. To detect bases efficiently by a fluorescence technique, modified bases with large fluorescence quantum yields and distinguishable spectral properties are required. Nucleotides will be modified by the covalent addition of a fluorescent dye to the base through a linker arm that enables DNA replication. The labeling will be accomplished by an enzymatic synthesis of a complementary strand of DNA using fluorescently labeled nucleotides in the reaction mixture. Use of a correctly designed linker arm enables the enzymatic synthesis of the complementary strand to proceed without interference from the linker arm or the dye. For each of the four bases, fluorescent, modified nucleotides exist. Examples of various types and lengths of linker arms are shown in Figure 3. Each nucleotide type will be labeled with a characteristic dye. The optimum set of dyes is still under investigation, but a reasonable initial choice might be the dyes used in fluorescent tagging for automated sequencing [8]. As a precedent for this labeling approach, DNA fragments several kilobases in length have been synthesized using biotin-labeled nucleotides [10].

Selection and Suspension of the DNA Fragment to be Sequenced

Large cloned fragments of human DNA will be obtained from the National Gene Library Project [11]. Following excision and denaturation of the cloned DNA fragment, a biotinylated primer will be annealed to one of the 3′ ends, and the complete complementary DNA strand will be synthesized using fluorescently labeled nucleotides (Figure 2A). These fluorescent DNA fragments will then be attached to an avidin-coated microsphere through the biotinylated primer (Figure 2B).

Once the fluorescently tagged DNA fragment is attached to the bead, it may either be sequenced in its double-stranded form (since only the fluorescently modified nucleotides will be detected) or it may be denatured prior to the sequencing (since

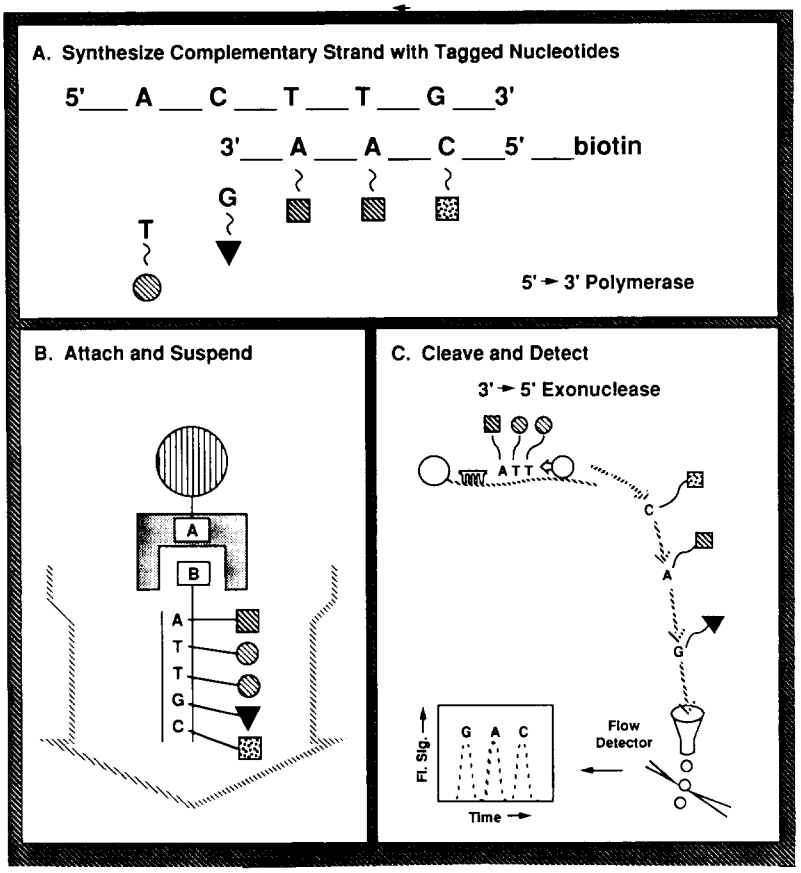


Figure 2. Steps in sequencing a single DNA fragment.

the fluorescent strand is attached to the bead through the biotin–avidin complex). This choice depends upon the stability of the DNA complex and on the exonuclease.

The sequencing technique relies on the detection of the fluorescently labeled nucleotides as they are released from the DNA fragment (Figure 2C). Only one DNA fragment will be attached to the microsphere. If multiple strands of DNA were present, the cleavage of the individual strands would get out of register and it would be difficult, if not impossible, to decipher the data.

Microspheres containing only one DNA fragment will be identified by measuring the total fluorescence, and then selected from the microsphere sample. Standard mechanical micromanipulation techniques [12] may be used in the selection process, but if this method is not successful for the introduction of the microsphere into the flow chamber, we will use optical trapping and manipulation techniques [13]. Optical manipulation could be carried out in an integrated processing chamber similar to that described by Buican et al. in which selection, storage, and access to the flow stream are carried out in separate, interconnecting compartments in a single manipulation chamber [14].

Enzymatic Cleavage of Labeled Nucleotides

An exonuclease added to the flow stream will attach to the 3′ end of the labeled DNA fragment and cleave bases sequentially from the 3′ end. While the presence of the linker arm and the fluorescent dye may inhibit the enzymatic activity of some exonucleases, suitable exonucleases will cleave with only a slight reduction in rate (D. C. Ward, Department of Biochemistry, Yale University, personal communication). The rate of cleavage can be adjusted by varying the exonuclease concentration (nonprocessive only), the cofactor concentration, the temperature, the solvent conditions, or by the use of inhibitors. The normal rate of exonuclease activity is ~100–1000 bases/second. For example, exonuclease I, a highly processive exonuclease, cleaves single-stranded DNA at a rate of 275 bases/second [15]. T4 exonuclease is reported to cleave approximately ten times faster than exonuclease I [16].

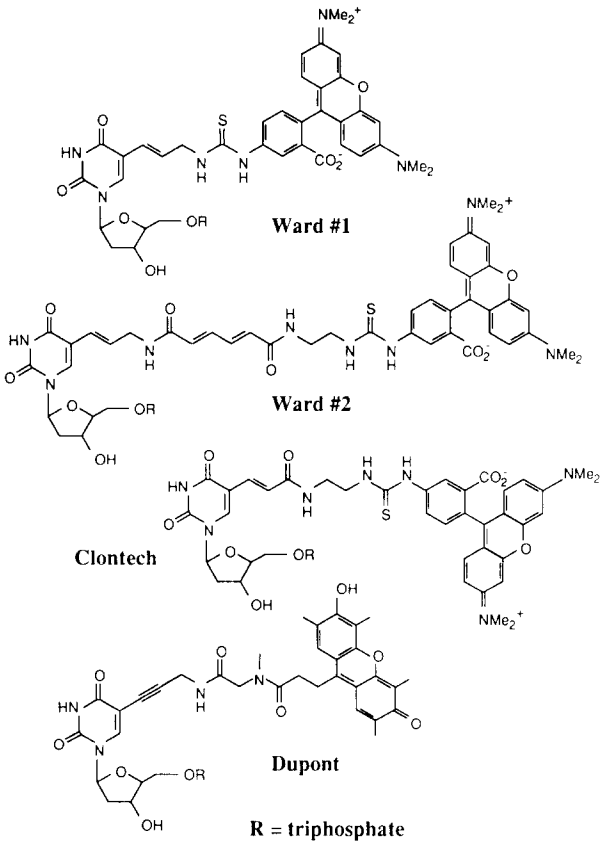


Figure 3. Nucleotide-dye combinations with different types of linker arms.

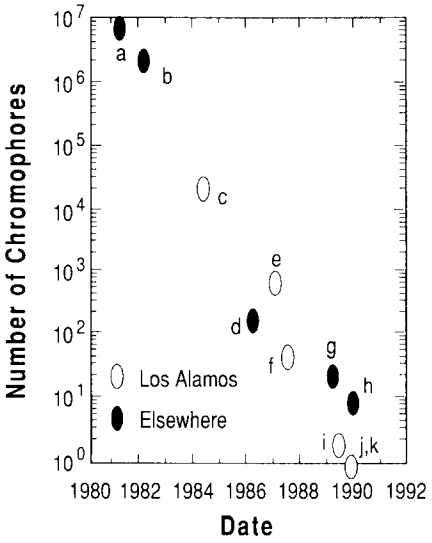
Although we believe that both processive and nonprocessive exonucleases will work, a processive exonuclease appears to be more desirable because of the potentially more rapid cleavage rate and the need to maintain minimum concentrations of fluorescent materials in our sample stream.

Single Base/Tag Detection

The cleaved, labeled bases will be detected by laser-induced fluorescence as they flow through the laser beam in much the same manner that we detect single molecules of rhodamine-6G [2, 3]. Fluorescence from individual molecules is superimposed on the background emission from the solvent. The background consists of Raman, Raleigh, and spectral scattering plus fluorescence arising from solvent impurities. The detection limit is determined by the ability to distinguish fluorescence from the background. A history of the approach to single molecule detection is given in Figure 4. Our fluorescence detection scheme is based on photon-burst detection [17]. As a molecule passes through a focused laser beam, it is repeatedly cycled from the ground electronic state to an excited electronic state with the emission of a photon on each cycle (Figure 5). Under optical saturation conditions, the number of photons emitted (the burst size) can approach the transit time of the molecule across the laser beam divided by the fluorescent lifetime. The burst size can be thousands or hundreds of thousands of photons, ultimately limited by the photostability of the molecule. The photon bursts are correlated in time which enable them to be distinguished from the background. To identify the four bases, it will be necessary to distinguish among the emission spectra of the four fluorescent tags.

The ability to sequence rapidly enables us to work, if necessary, with a system with a relatively high random error rate because identical fragments can be sequenced many times to reach a consensus. Nonrandom errors associated with particular sequences or secondary structure in the DNA can be reduced by sequencing both the selected fragment and its complementary strand. We are investigating algorithms for arriving at a consensus sequence in the presence of random error.

Figure 4. History of single molecule detection.



a. Kelley T.A., Christian G.D., Anal. Chem. 53, 2110 (1981).
b. Folestad S., Johnson L., Josefsson B., Galle B., Anal. Chem. 54, 925 (1982).
c. Dovichi N.J., Martin J.C., Jett J.H., Keller R.A., Anal. Chem. 56, 348 (1984).
d. Watson J.V., Walport M.J., J. Immunol. Meth. 93, 171 (1986).
e. Nguyen D.C., Keller R.A., Trukula M., J. Opt. Soc. Am. B 4, 138 (1987).
f. Nguyen D.C., Keller R.A., Jett J.H., Martin J.C., Anal. Chem. 59, 2158 (1987).
g. Peck K., Stryer L., Glazer A.N., Mathies R.A., Proc. Natl. Acad. Sci. 86, 4087 (1989).
h. Ramsey J.M., Whitten W.B., in preparation.
i. Hahn J.H., Soper S.A., Nutter H.L., Martin J.C., Jett J.H., Keller R.A., Appl. Spect. (in press).
j. Shera E.B., Seitzinger N.K., Davis L.M., Keller R.A., Soper S.A., Chem. Phys. Lett. 174, 553 (1990).
k. Soper S.A., Hahn J.H., Nutter H.L., Martin J.C., Jett J.H., Keller R.A., Anal. Chem. (in press).

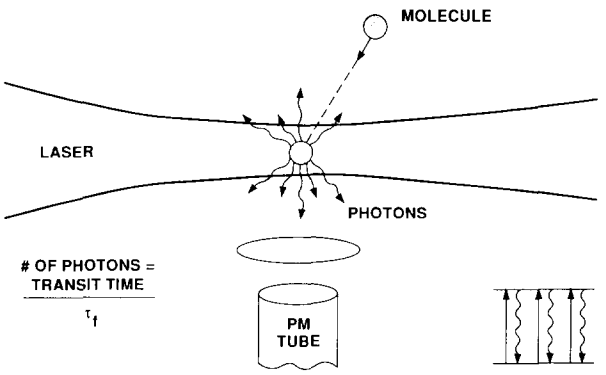


Figure 5. Photon-burst detection of single molecules.

Status

Fluorescent Labeling of Bases

Steric effects of the linker arms and the dye might preclude the synthesis of a DNA fragment in which every base is linked to a fluorescent dye. Since all four nucleotides, each containing a linker arm and fluorescent dye, were not available to us, nucleotides containing linker arms terminated in biotin were used as model compounds. We have demonstrated that steric effects are not a problem with biotinylated nucleotides by synthesizing a completely biotinylated strand of DNA complementary to the template $d(A,G)_{2100}$. This template was used because only biotinylated dUTP and dCTP were available at the time. The biotinylated strand of DNA was prepared by reacting bio-11-dCTP, bio-11-dUTP (Enzo Biochem, New York, NY), a primer, and the template with *Escherichia coli* DNA polymerase I (Klenow large fragment). Analysis of the resulting products by electrophoresis demonstrated that the reaction went to completion and that a totally biotinylated complementary fragment, $d(C,U)_{2100}$, was formed [1]. These experiments demonstrated that this DNA polymerase could recognize modified nucleotides that are similar to the nucleotides that we plan to use for the fluorescent tagging.

To discern whether DNA synthesis using modified nucleotides would be more difficult with natural DNAs, the modified bacteriophage M13mp18, 7250 bases in length, was selected as a model system. This template contains all four bases, has a known sequence and length, and is well characterized. We studied four commercially available biotin-modified nucleotides: bio-11-dCTP, bio-11-dUTP, bio-7-dATP, and bio-14-dATP (Life Technologies, Grand Island, NY). Using *E. coli* DNA polymerase I (Klenow large fragment), bio-11-dCTP or bio-11-dUTP, with the other three nucleotides unmodified, the complete replication of the M13mp18 bacteriophage was demonstrated. With both bio-11-dCTP and bio-11-dUTP, the replication is almost, but not quite, complete. Only partial replication occurs when either bio-7-dATP or bio-14-dATP are included in the reaction mixture. We attribute this to the fact that the linker arm is attached to the 6-amino position of the dATP, which is involved in the base pairing with thymine, thereby weakening the bond between the base pairs. These replications have been confirmed by various methods, including agarose gel sizing and restriction enzyme digests (L. A. Krakowski and R. A. Ratliff, unpublished results).

The question still remains as to whether a fluorescent dye can be substituted for the biotin without interfering with the synthesis. There is a potential problem with the intercalation of the fluorescent dye into the template of the newly synthesized DNA strand. This intercalation could interfere with the synthesis by causing the replicating strand to peel away from the template. We plan to investigate the use of nucleotides with linker arms of high rigidity to prevent intercalation. Ward demonstrated uninhibited DNA synthesis using fluoresceinated nucleotides and linker arms containing conjugated double bonds to provide rigidity (D. C. Wood, personal communication).

Enzymatic Cleavage of the Labeled Nucleotides

The replicated strand, $d(C,U)_{2100}$, prepared as described above with $d(A,G)_{2100}$ template and biotinylated dUTP and dCTP, was cleaved with only a slight reduction in rate using the exonucleases T_4 or T_7. The results of these experiments suggest that the activity of T_4 and T_7 was not significantly reduced with these modified nucleotides.

Selection and Suspension of the DNA Fragment to be Sequenced

In advance of the development of modified fluorescent bases for DNA labeling, we have begun preliminary studies of the DNA–microsphere attachment and manipulation. We have assembled a system for detecting and observing weakly fluorescent DNA fragments based on an epifluo-

rescence microscope, a cooled CCD camera, and video recording. We are characterizing the fluorescence of ethidium-bromide-stained lambda DNA (~48.5 kb). Individual DNA molecules and microspheres are visualized by their images in the field of an epifluorescence microscope coupled to a cooled CCD camera. Hin*d*-III restriction fragments of lambda DNA are being used to determine the lower limits of our microscope detection system. Whereas microspheres are readily seen in the microscope, the fluorescence from single lambda DNA molecules is very dim, requiring the use of a cooled CCD camera for detection. Under these measurement conditions, we have observed objects in dilute solutions of lambda DNA that were roughly 2–3 μm in diameter. Unlike the microspheres, these objects are irregular in shape and in distribution of fluorescence. An equivalent amount of lambda DNA digested with Hin*d* III resulted in a large increase in the number of smaller objects. The objects that we have imaged have the same behavior as lambda DNA reported by Smith et al. [18]. Observation of these objects during low-voltage electrophoresis along with enzymatic digestion will provide conclusive identification.

Single Base/Tag Detection

We are detecting single molecules of rhodamine-6G (R-6G) by two techniques. The first involves cw excitation with the 514.5-nm output from an Ar ion laser irradiating R-6G molecules in a flowing stream of ethanol (EtOH) [2]. The increased photostability of R-6G and its greater fluorescence quantum yield in EtOH, as compared to water, makes single molecule detection easier in this solvent system. We are investigating alternative solvent systems that exhibit increased photostability for various chromophores and are also compatible with the enzymes needed for the cleavage reactions. To demonstrate single molecule detection, it is necessary to work at a very low analyte concentration to minimize the probability of two molecules occupying the probe volume in any one time interval and thereby prevent overlapping bursts. An autocorrelation analysis of the photon counts demonstrates the existence of photon bursts from molecules of R-6G as they pass through the laser beam. Histograms of the photon count data from the blank (pure EtOH) and EtOH with R-6G indicate the presence of a tail in the distribution at high count rates; present in the case of R-6G, but absent in the blank. These high counts can be attributed only to photon bursts from individual molecules traveling through the probe volume.

The second technique has enabled us to detect and count individual fluorescent molecules as they pass through the laser beam [3]. The detection scheme, which involves excitation by a repetitively pulsed laser and time-gated selection of fluorescent photons, has reduced the background against which we must detect passing molecules by approximately two orders of magnitude. For the single chromophore dye, R-6G, we have achieved a detection efficiency of ~85% with only ~0.01 false counts per second. This is a significant achievement since it is the clearest evidence obtained to date that molecules have been individually detected and counted in a liquid environment.

Although single molecules are being detected with reasonable efficiency, our detection capabilities must be improved further in order for this technique to become useful for sequencing DNA. To this end, we are improving our optical collection efficiency, investigating the use of time correlation with position-sensitive detection to discriminate further against the Raman background, and infrared excitation and detection to reduce interferences from solvent impurities and the Raman background.

Error Analysis

An important consideration in any sequencing scheme is the error rate. The error has two components—random and nonrandom. Random errors arise from missing molecules, identifying background as molecules, and misidentification of molecules. Random errors can be reduced by sequencing identical strands and using a consensus scheme to arrive at the final error. Our rapid sequencing makes this approach attractive because we will be able to compare results from several strands. The exact error analysis is difficult but a first approximation is given by (D.C. Torney, Los Alamos National Laboratory, personal communication)

Random error in a single strand	ϵ
Number of strands compared	$2N+1$
Consensus error	ϵ^{N+1}

The following example illustrates how the consensus error reduces rapidly. With a 1% error in a

single strand, the consensus error drops to 1 in 10^4 by comparing three strands and 1 in 10^6 by comparing five strands. This analysis will enable us to make a trade-off between reducing the error in a single strand or sequencing more identical strands at a higher error rate. Nonrandom errors, such as those associated with a particular sequence or secondary structure in the DNA, are more difficult. Hopefully, we can reduce some of these by sequencing the strand of DNA complementary to the original strand which will have a different sequence and different secondary structure.

Summary

We have outlined a technique that is projected to have the capability to sequence large fragments of DNA at a rate of 100–1000 bases/second. This is a challenging project with several difficult steps, but we see no fundamental reason why it cannot be made to work. Major hurdles include enzymatic labeling with fluorescently tagged nucleotides and high processive, rapid exonuclease cleavage of modified nucleotides. We anticipate a demonstration experiment within a year. Successful completion of this approach will have a large impact on the Human Genome Project.

References

1. Jett JH, Keller RA, Martin JC, Marrone BL, Moyzis RK, Ratliff RL, Seitzinger NK, Shera EB, Stewart CC: J Biomol Struct Dynamics 7:301, 1989
2. Soper SA, Shera EB, Martin JC, Jett JH, Hahn JH, Nutter HL, Keller RA: Anal Chem 1990 (submitted for publication)
3. Shera EB, Seitzinger NK, Davis LM, Keller RA, Soper SA: Chem Phys Lett 174:553, 1990
4. Dovichi NJ, Martin JC, Jett JH, Keller RA: Science 219:845, 1983
5. Dovichi NJ, Martin JC, Jett JH, Trkula M, Keller RA: Anal Chem 56:348, 1984
6. Nguyen DC, Keller RA, Trkula M: J Opt Soc Am [B] 4:138, 1987
7. Roberts L: Science 242:1245, 1988
8. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SBH, Hood LE: Nature 321:674, 1986
9. Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA, Baumeister K: Science 238:336, 1987
10. Langer PR, Waldrop AA, Ward DC: Proc Natl Acad Sci USA 78:6633, 1981
11. VanDilla MA, Deaven LL, Albright KL: Bio/Technology 4:537, 1986
12. DeMayo FJ, Bullock DW: In Schrader WT, O'Malley BW (eds): Laboratory Methods Manual for Hormone Action and Molecular Endocrinology. Houston, TX, Houston Biological Association, 1987
13. Buican TN, Smyth MJ, Crissman HA, Salzman GC, Stewart CC, Martin JC: Appl Opt 26:5311, 1987
14. Buican TN, Neagley DL, Morrison W, Upham BD: in Proceedings of the SPIE Conference on New Technologies in Cytometry, 19–20 January 1989
15. Brody RS, Doherty KG, Zimmerman PD: J Biol Chem 261:7136, 1986
16. Thomas KR, Olivera BM: J Biol Chem 253:424, 1978
17. Greenless GW, Clark DL, Kaufman SL, Lewis DA, Tonn JF, Broadhurst JH: Opt Commun 23:236, 1977
18. Smith SB, Aldridge PK, Callis JB: Science 243:203, 1989