TITLE HIGH-SPEED DNA SEQUENCING: AN APPROACH BASED UPON
FLUORESCENCE DETECTION OF SINGLE MOLECULES

AUTHOR(S) J. H. Jett, R. A. Keller, J. C. Martin, B. L. Marrone,
R. K. Moyzis, R. L. Ratliff, N. K. Seitzinger, E. B. Shera,
and C. C. Stewart

DISCLAIMER

High-Speed DNA Sequencing: An Approach Based Upon
Fluorescence Detection of Single Molecules

James H. Jett, Richard A. Keller, John C. Martin,
Babetta L. Marrone, Robert K. Moyzis, Robert L. Ratliff,
Newton K. Seitzinger, E. Brooks Shera, and Carleton C. Stewart[*]

Los Alamos National Laboratory
Los Alamos, New Mexico 87545

ABSTRACT

We are developing a laser based technique for the rapid sequencing of large fragments (~40 kb) of DNA based upon the detection of single, fluorescently tagged nucleotides cleaved from a single DNA fragment. We have demonstrated significant progress on several of the important steps of this technique. The projected rate of sequencing is several hundred bases per second which is orders of magnitude faster than existing methods. Once developed, this technology could be utilized by investigators for rapid sequencing of genetic material from virtually any source.

[*] Current address - - Director, Laboratory of Flow Cytometry, Roswell Park Memorial Institute, Elm and Carlton Streets, Buffalo, New York 14263

INTRODUCTION

A world wide effort is now in progress with the ultimate goal of determining the entire nucleotide base sequence of the human genome. The magnitude of this task is apparent when one considers that there are 3×10^9 bases in the human genome and current sequencing rates are only a few hundred to a few thousand bases per person per day. Standard procedures, using either the Sanger dideoxy method (1) or the Maxam and Gilbert chemical method (2) for the generation of fragments, and electrophoresis for the separation and identification of the fragments are slow and labor intensive. New automated sequencing techniques are projected to yield sequencing rates of 2,000-30,000 bases per day (3-7). At 1 million bases per day, it would take ten years to sequence the genome of a single human.

All of the above approaches are labor intensive methods requiring relatively long analysis times. Using a completely different approach, we envision a single instrument that can sequence a single fragment of DNA, several tens of kilobases (kb) or more in length, at the rate of several hundred bases per second. The ability to work with large fragments will reduce significantly the amount of subcloning and mapping. Development of this method may lead eventually to a clinical instrument capable of sequencing large fragments of DNA rapidly for diagnosis of genetic diseases.

Our technique is based upon a projected ability to detect single chromophores by laser-induced-fluorescence in flowing sample streams (8-11). The technique involves: (1) labeling the nucleotides with base specific tags suitable for fluorescence detection, (2) selecting a desired fragment of DNA, (3) suspending the *single* DNA fragment in a flowing sample stream, (4) sequentially cleaving labeled bases from the free end of the DNA fragment using an exonuclease, and (5) detecting and identifying the cleaved, labeled bases as they flow through a

2

focused laser beam (12). The steps are illustrated in Fig. 1.

The rate that bases can be sequenced is determined by the kinetics of the exonuclease cleavage reaction and the time required for detection and identification of the labeled bases. Based upon our results for the detection of rhodamine-6G and studies of cleavage rates, we anticipate sequencing rates of several bases per second on a single strand of DNA tens of kb in length.

METHODOLOGY

Fluorescence Labeling of Bases

Individual free and bound bases found in DNA have intrinsic fluorescence quantum yields $<10^{-3}$ at room temperature. In order to detect these bases efficiently by a fluorescence technique, it is desirable to modify (label) them to form species with large fluorescence quantum yields and distinguishable spectral properties. A labeled, complementary strand of DNA will be prepared enzymatically, incorporating fluorescently labeled bases using the DNA strand of unknown sequence as the template. Sequencing will then be done on the complementary, labeled strand.

Each labeled base consists of a linker arm terminated with a fluorescent dye. A correctly designed linker arm will permit the enzymatic synthesis of the complementary strand to proceed without interference from the linker arm or the dye. Modified nucleotides containing a linker arm terminated by biotin are routinely used to label DNA at specific sites (13). Biotinylated nucleotides complementary to all four of the bases with various types and lengths of linker arms can be prepared. An example of a biotinylated nucleotide is shown in Fig. 2. In our approach, highly fluorescent dyes will be substituted for the biotin. Each of the four nucleotide types will be labeled with a different characteristic dye, for example those used for automated sequencing

3

(3). The optimum set of dyes is still under investigation. We anticipate that the linker arm dye combination will have to be carefully designed to optimize the activities of the polymerase and the exonuclease.

Although one might expect that steric effects, due to the linker arm and the biotin, could preclude the synthesis of a DNA fragment in which every base is labeled, our initial experiments have shown that the presence of a linker arm and biotin on the nucleotide did not interfere with replication. A strand of DNA, complementary to $d(A,G)_{2100}$, was prepared by adding biotinylated dCTP and biotinylated dUTP (see Fig. 2) with DNA polymerase I, *E .coli* (Klenow fragment), to a solution containing the template. Tracer amounts of α-^{32}P-dTTP were added in order to follow the replication reaction. The biotinylated nucleotides were incorporated into acid-insoluble, double stranded DNA to the same extent as unlabeled nucleotides in a control reaction. We anticipate that replication of longer biotinylated fragments will not be a problem.

Selection and Suspension of the DNA Fragment to be Sequenced

The selection and suspension of a single DNA fragment in a flowing sample stream can oe accomplished in the following manner. The fragment to be sequenced will be obtained, for example, from one of the large insert clones available from the National Gene Library Project (14). Alternatively, hybridization techniques can be used to select a desired fragment from an inhomogeneous mixture of DNA. A primer, labeled with biotin at the 5' end, will be hybridized to the selected DNA fragment and fluorescent nucleotides and polymerase will be added to form the complementary labeled strand. The labeled strand will then be linked, at the biotin end, to avidin that has been attached to a microsphere. If it is desired, the labeled strand can be separated from the template before attachment to the microsphere. A similar procedure has been

4

described for immobilizing DNA, modified at the 5' terminus, to microspheres (15). The presence of only a single labeled fragment of DNA on a microsphere can be verified by measuring the total fluorescence. The selection and suspension of the DNA fragment to be sequenced is summarized in Fig. 3.

A selected microsphere containing a single DNA fragment will be transferred and suspended in the flow stream. There are several ways to manipulate the microsphere DNA carriers. The simplest method for placing and holding the microsphere in position for sequencing uses microinjection pipette procedures (16). An attractive alternative for microsphere manipulation is to use converging laser beam optical trapping techniques similar to those demonstrated by Ashkin and Dziedzic (17) and used by Buican *et al.* (18) for transporting and sorting biological cells. Microsphere selection and manipulation has been demonstrated in a computer-controlled optical trapping apparatus(19). Precision manipulation of particles has been demonstrated by using the optical trap to pick up a single small human chromosome and place it on a small microsphere (7 μm diameter) (19).

Enzymatic Cleavage of Labeled Nucleotides

Exonuclease will be added to the flow stream to cleave bases sequentially from the end of the labeled DNA strand suspended in the flow stream. While the presence of the linker arm and the fluorescent dye may inhibit the enzymatic activity of some exonucleases, we expect to find suitable exonucleases that will cleave with only a slight reduction in rate. The rate of cleavage can be adjusted by varying the cofactor concentration, the temperature, or by the use of poisoning agents. It appears that a highly processive exonuclease (20) would be more desirable but non-processive exonucleases are also under consideration.

The overall rate of cleavage is determined by the rate of diffusion of the exonuclease to

5

the DNA strand, the rate of attachment to the strand, the rate of cleavage once on the strand, and the processivity. The rate determining step is usually the diffusion rate and leads to observed cleavage rates of several bases per second. A detailed kinetic analysis is necessary to extract the cleavage rate of a processive exonuclease. To our knowledge, this analysis has been applied only to exonuclease I, a highly processive exonuclease, and it is encouraging that the measured cleavage rate after attachment is 275 bases/sec (21).

We have studied the kinetics of the cleavage reactions of various exonucleases on duplex, biotinylated DNA fragments ($d(C,U)_{2100}$, prepared as described above). The data obtained using the exonuclease activities of T4 and T7 polymerase are shown in Fig. 4. The extent of the cleavage reaction for the biotinylated fragment was less than for the unlabeled fragments. The smaller extent of cleavage of the biotinylated fragment may reflect uneven rates of cleavage of the biotinylated and unmodified strands of double stranded DNA by exonucleases. We are currently investigating cleavage rates of other exonucleases, the effect of reaction conditions and added reagents, and the use of single stranded substrates.

Single Base/Tag Detection

The cleaved, labeled bases will be detected by laser-induced fluorescence as they flow through the focused laser beam. The detection technique will be an improved version of one developed for sensitive detection of rhodamine-6G (11) and phycoerythrin (10, 22).

Our detection capability is limited by the signal-to-noise ratio. With our current apparatus, an average of 1 photoelectron from rhodamine-6G fluorescence and 14.5 photoelectrons from background emission are detected in the 80 μs that it takes a molecule to transit the focused laser beam (11). The noise observed is related to statistical fluctuations in the background. The signal-to-noise ratio is given by $S/N \sim S/\sqrt{(background)} = 1/\sqrt{(14.5)} = 0.26$,

6

which is insufficient to permit detection of single molecules whose fluorescence is comparable to rhodamine-6G. For sequencing it will be necessary to detect and distinguish the emission spectra of four separate tags. Improvements in the fluorescence detection as outlined below will enable us to achieve the required sensitivity.

Since the noise is proportional to the square root of the background counts, the S/N ratio can be improved by increasing the total amount of light collected (both fluorescence and background). Our present detection optics collect approximately 5% of the total emitted light. The flow cell and the optics have been redesigned to collect approximately 50% of the emitted light. A further increase in collected light can be obtained by reducing the flow velocity and increasing the observation length which correspondingly increases the viewing time of the molecules as they pass through the laser beam.

A further improvement in the S/N ratio can be made by suppressing background. The majority of our background originates from the weak Raman bending band of the water solvent. We are investigating the use of a frequency doubled, modelocked Nd:YAG laser to reduce this background. This laser operates at 82 MHz with a pulse width of less than 100 picoseconds. The Raman scattered background light occurs only during the laser excitation pulse while the emitted fluorescence light decays with a lifetime of a few nanoseconds. Hence, gating on the photon counter only between laser pulses will greatly reduce the Raman background.

To discriminate further against background sources which are stationary with respect to the moving sample stream (e.g. reflections) or emissions which occur randomly in pace (e.g. Raman scattering or fluorescence from solvent impurities), a position-sensitive optical detector will be used to record the fluorescent light emitted as tagged bases pass through the flow cell (23). This device will furnish the arrival time and point of origin in the sample stream for each detected photon. Computer analysis of these data will permit us to discriminate against

7

background photons that do not originate from a molecule traveling through the flow cell at the known flow velocity. This method will also allow us to avoid sequencing errors that may occur when two or more tagged bases are present simultaneously in the flow cell, as may occur occasionally during rapid sequencing.

As a demonstration of this technique, we have analyzed data which were acquired using a moving attenuated lase·beam to simulate the passage of a molecule and diffused light from a tungsten lamp to provide the random background. The data analysis demonstrated that time correlation can provide a large discrimination against stationary or random background signals. For example, we simulated a typical sequencing experiment where one might expect to detect, on average, 26 photoelectrons from a passing molecule, against a total background of 10,000 photoelectrons. Under these conditions, we measured a random error rate of approximately 2% (1% missed molecules plus 1% false identifications from background). It is a property of Poisson statistics ·hat small improvements in signal-to-noise ratio can result in a large reduction in error rate. For example, when the the number of simulated fluorescence photoelectrons was increased to 30, the total error rate dropped to < 0.1%.

The ability to sequence rapidly allows one to work, if necessary, with a system with a relatively high random error rate because sequencing can be repeated on identical fragments to reach a consensus. Non-random, sequence dependent errors can be reduced by sequencing separately the complements to both strands of the selected fragment.

SUMMARY

We have outlined a technique that has the projected capability to sequence large fragments of DNA at a rate of several hundred bases per second. We have demonstrated significant progress on several important steps in this technique. This is a challenging project with several

difficult steps, but we foresee no fundamental obstacle to its success. While our discussion concentrated on sequencing DNA, the technique also has the potential to sequence RNA.

We envision that other scientific and technological spinoffs will emerge as the techniques described here are developed. For example, the ability to work at the single molecule/strand level should do much to reveal the mechanisms of the action of polymerases and exonucleases at the molecular level. Improved sensitivity for fluorescence detection in flowing sample streams will have applications in flow cytometry and bioanalytical diagnostics (see, for example, reference 24).

Finally, the concept and method of base-specific tagging of the nucleotides followed by sequencing the tags has application for other possible sequencing techniques. For example, metal atoms could be used as tags in electron microscopy and x-ray holography, or particles of differing sizes used in scanning tunneling microscopy.

ACKNOWLEDGEMENTS

REFERENCES

1. F. Sanger, S. Nicklen, and A. R. Coulson, "DNA Sequencing With Chain-Terminating Inhibitors", Proc. Nat. Acad. Sci. U.S.A. 74, 5463 (1977).

2. A. M. Maxam and W. Gilbert, "Sequencing End-Labeled DNA with Base-Specific Chemical Cleavages", Meth. Enzym. 65, 499 (1980).

3. L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. H. Kent, and L. E. Hood, "Fluorescence Detection in Automated DNA Sequence Analysis", Nature 321, 674 (1986).

4. J. M. Prober, G. L. Trainor, R. J. Dam, F. W. Hobbs, C. W. Robertson, R. J. Zagursky, A. J. Cocuzza, M. A. Jensen, and K. Baumeister, "A System for Rapid DNA Sequencing with Fluorescence Chain-terminating Dideoxynucleotides", Science 238, 336 (1987).

5. W. Ansorge, B. Sproat, J. Stegmann, C. Schwager, and M. Zenke, "Automated DNA Sequencing: Ultrasensitive Detection of Fluorescent Bands During Electrophoresis", Nuc. Acid Res. 15, 4593 (1987).

6. W. Ansorge. B. Sproat, J. Stegemann, and C. Schwager, "A Non-radioactive Automated Method for DNA Sequence Determination", J. Biochem. and Biophy. Meth. 13, 315 (1986).

7. J. A. Braumbaugh, L. R. Middendorf, D. L. Grone, and J. L. Ruth, "Continuous, on-line DNA Sequencing Using Oligodeoxynucleotide Primers with Multiple Fluorophores", Proc. Natl. Acad. Sci. USA 85, 5610 (1988).

8. N. J. Dovichi, J. C. Martin J. H. Jett, and R. A. Keller, "Attogram Detection Limit for Aqueous Dye Samples by Laser-Induced Fluorescence", Science 219, 845 (1983).

9. N. J. Dovichi, J. C. Martin J. H. Jett, M. Trkula, and R. A. Keller, "Laser-Induced Fluorescence of Flowing Samples as an Approach to Single-Molecule Detection in Liquids", Anal. Chem. 56, 348 (1984).

10. D. C. Nguyen, R. A. Keller, J. H. Jett, and J. C. Martin, "Detection of Single Molecules of Phycoerythrin in Hydrodynamically Focused Flows by Laser-Induced Fluorescence", Anal. Chem. 59, 2158 (1987).

11. D. C. Nguyen, R. A. Keller, and M. Trkula, "Ultrasensitive Laser-induced Fluorescence Detection in Hydrodynamically Focused Flows", J. Opt. Soc. Am. B., 4, 138 (1987).

12. U. S. Patent Application submitted.

13. P. R. Langer, A. A. Waldrop, and D. C. Ward, "Enzymatic Synthesis of Biotin-Labeled Polynucleotides: Novel Nucleic Acid Affinity Probes", Proc. Natl. Acad. Sci. USA 78, 6633 (1981).

14. M. A. Van Dilla et. al., "Human Chrosome-Specific DNA Libraries: Construction and Availability", Bio/Tech. 4, 537 (1986).

15. J. N. Kremsky, J. L. Wooters, J. P. Dougherty, R. E. Meyers, M. Collins and E. L. Brown, "Immobilization of DNA via Oligonucleotides Containing an Aldehyde or Carboxylic Acid Group at the 5' Terminus", Nucl. Aci. Res. 15, 2891 (1987).

16. F. J. DeMayo and D. W. Bullock, "Microinjection and Transfer of Mouse Embryos", In, *Laboratory Methods Manual for Hormone Action and Molecular Endocrinology.* W. T. Schrader and B. W. O'Malley, eds. Houston Biological Assoc., Inc., Houston, TX (1987).

17. A. Ashkin and J. M. Dziedzic, "Optical Trapping and Manipulation of Viruses and Bacteria", Science 235, 1517 (1987).

18. T. N. Buican, M. J. Smyth, H. A. Crissman, G. C. Salzman, C. C. Stewart, and J. C. Martin, "Automated Single-cell Manipulation and Sorting by Light Trapping", Appl. Opt. 26, 5311 (1987).

19. T. N. Buican, D. L. Neagley, W. C. Morrison, B. D. Upham, "Optical Trapping, Cell Manipulation and Robotics", SPIE Proceedings, Vol. 1063 (1989) pp 190-197.

20. K. R. Thomas and B. M. Olivera, "Processivity of DNA Exonucleases", J. Biol. Chem. 253, 424 (1978).

21. R. S. Brody, K. G. Doherty, and P. D. Zimmerman, "Processivity and Kinetics of the Reaction of Exonuclease I from *Escherichia coli* with Polydeoxyribonucleotides", J. Biol. Chem. 261, 7136 (1986).

22. K. Peck, L. Stryer, A. N. Glazer, and R. A. Mathies, "Single-Molecule Fluorescence Detection: Autocorrelation Criterion and Experimental Realization with Phycoerythrin", Proc. Natl. Acad. Sci. USA 86 4087 (1989).

23. U. S. Patent # 4793705 (1988).

24. Y-F. Cheng and N. J. Dovichi, "Subattomole Amino Acid Analysis by Capillary Zone Electrophoresis and Laser-Induced Fluorescence", Science 242, 562 (1988).
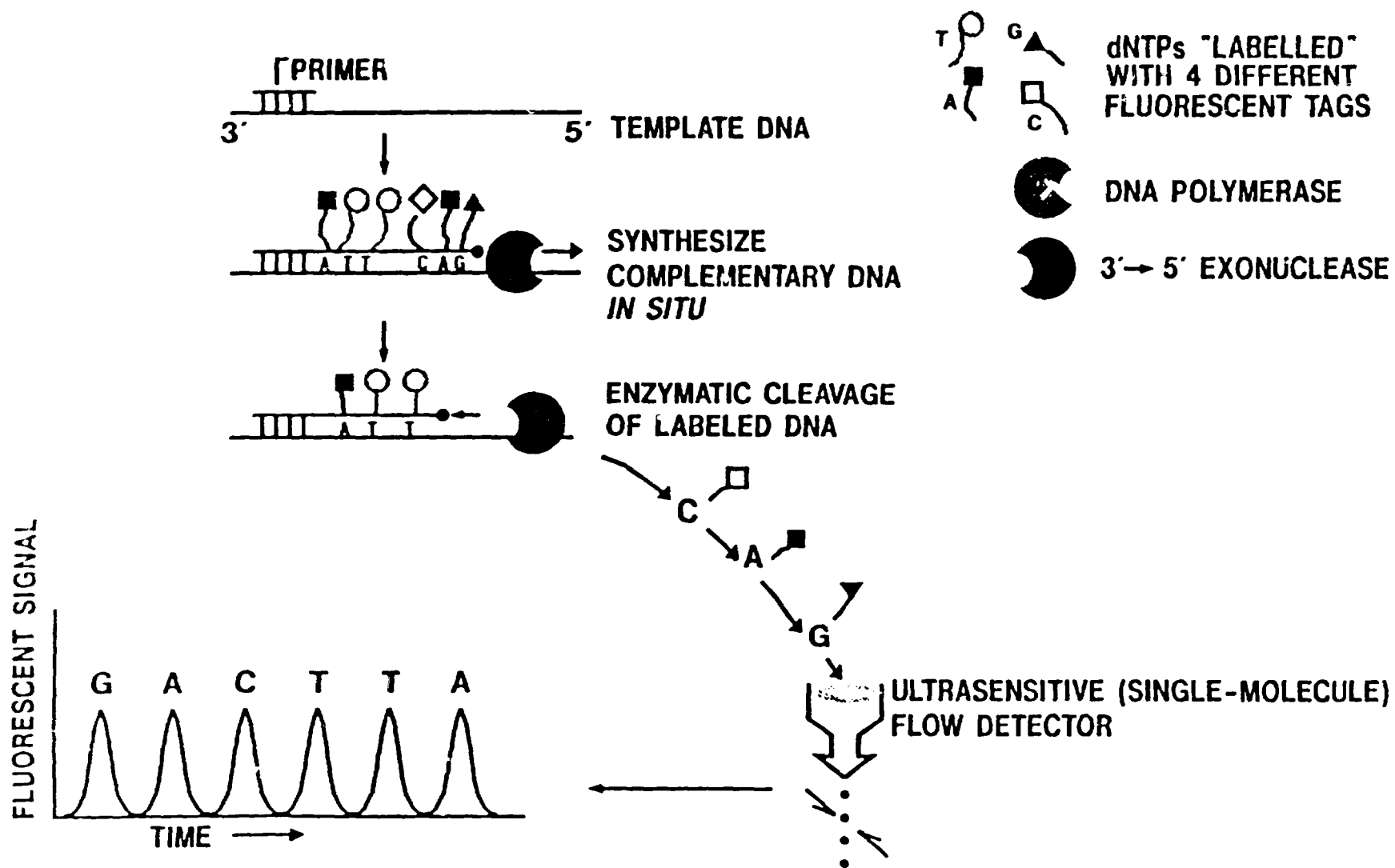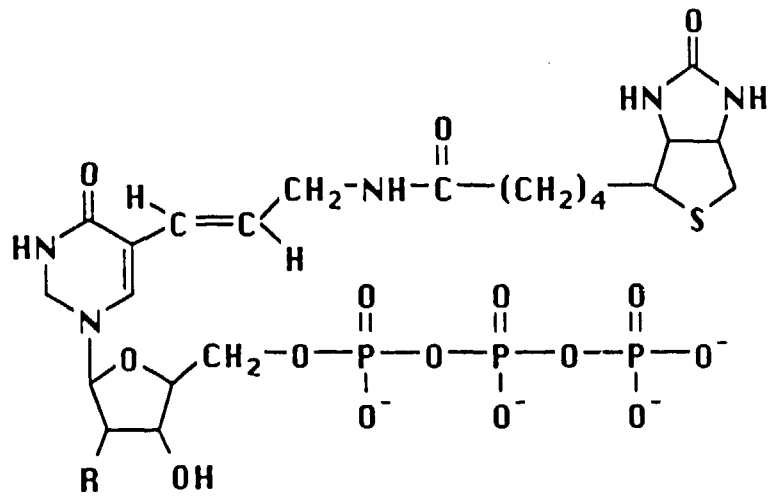
Figure Captions

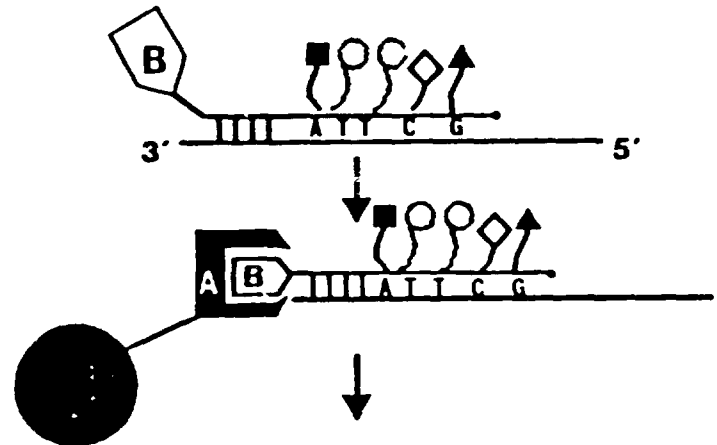Figure 1. Overall view of our proposed rapid DNA sequencing technique.

Figure 2. Biotinylated nucleotide used in the resynthesis of the complementary DNA strand.

Figure 3. Pictorial representation of the selection and suspension of a single labeled DNA fragment in the flowing sample stream. The DNA fragment is attached to the microsphere using biotin (B) and aviden (A) binding.

Figure 4. Labeled and unlabeled 2100 bp fragments of DNA were incubated at 37 C with 5U of T4 polymerase (top) or T7 polymerase with thioredoxin (bottom). At each time point, the amount of acid insoluble DNA remaining was determined. Control contains no biotinylated bases. Bio C, Bio U, and Bio C Bio U were made from the same template as the control DNA except that each of the indicated nucleotide substrates were biotinylated.

PRIMER

3' —————————————— 5' **TEMPLATE DNA**

SYNTHESIZE
COMPLEMENTARY DNA
IN SITU

ENZYMATIC CLEAVAGE
OF LABELED DNA

T   G   **dNTPs "LABELLED"
WITH 4 DIFFERENT
FLUORESCENT TAGS**
A   C

DNA POLYMERASE

3' → 5' EXONUCLEASE

C

A

G

ULTRASENSITIVE (SINGLE-MOLECULE)
FLOW DETECTOR

FLUORESCENT SIGNAL

G   A   C   T   T   A

TIME ——→

O

HN — C — NH

S

O

H C — C CH₂ — NH — C — (CH₂)₄

HN H

O O O

N CH₂ — O — P — O — P — O — P — O⁻

O O⁻ O⁻ O⁻

R OH

Labeled DNA Strand
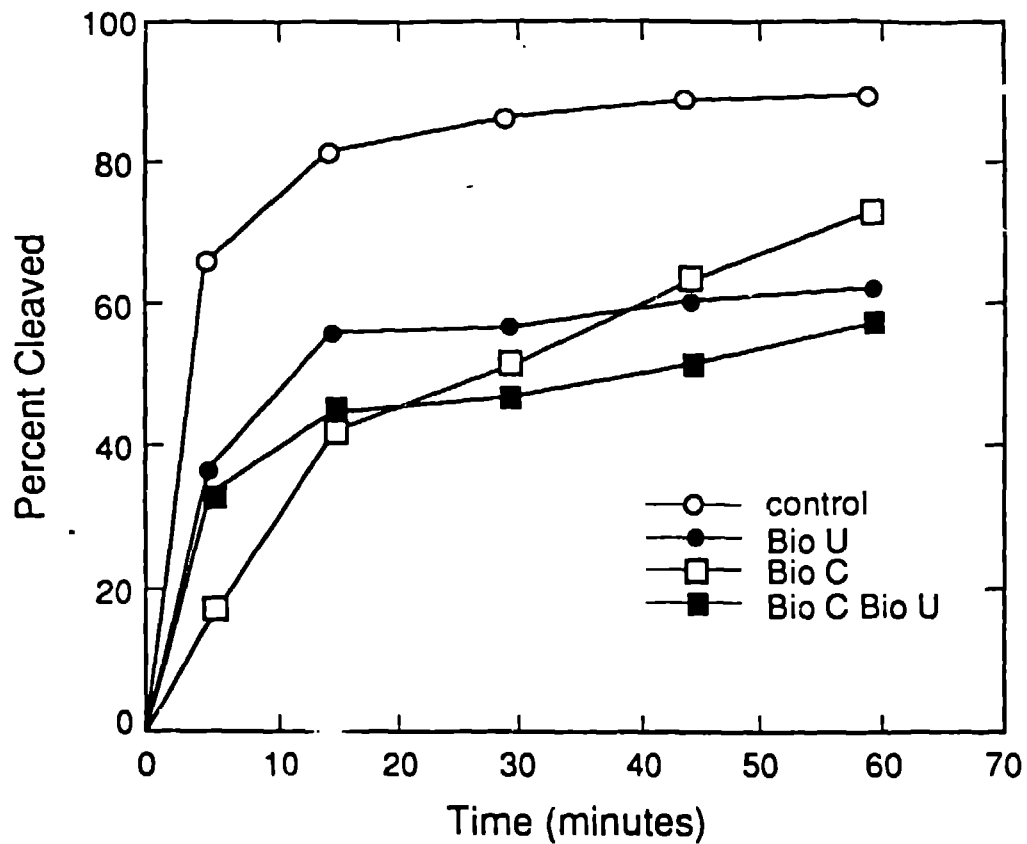With Biotin at 5´ End

Labeled DNA Is Attached
to Avidin–Coated
Microsphere at Biotin End
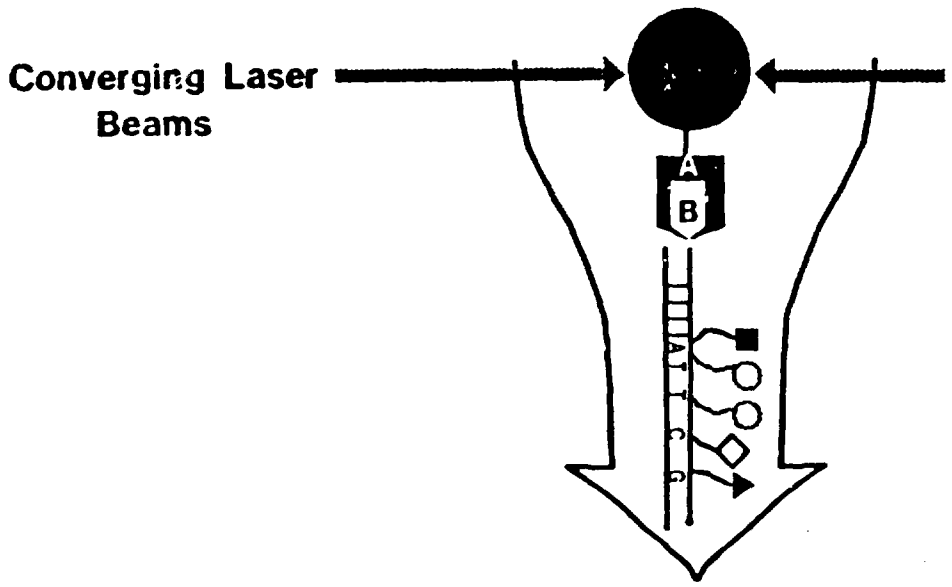
Converging Laser
Beams

Microsphere With Labeled
DNA Attached Is
Suspended by Optical
Manipulator in the
Flow Stream

T4

T7 and Thioredoxin