O

COMPUTER ASSISTED MULTIPLEX SEQUENCING

Performance Report

for the period of August 1, 1992 to July 15, 1993

George M. Church

Department of Genetics
Alpert Building
Harvard Medical School
200 Longwood Avenue
Boston, MA 02115

July 1993

Prepared for

THE U.S. DEPARTMENT OF ENERGY
GRANT NUMBER: DE-FG02-87ER60565

MAR 23 1994

OSTI

NOTICE:

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product or process disclosed or represents that its use would not infringe privately-owned rights.

MASTER

Abstract:

The objectives of this project are automation for optimization of multiplex sequencing. We have integrated direct transfer electrophoresis, automated multiplex hybridizations and automated film reading and applied this toward sequencing of *E. coli* and human DNA. Primers for the directed dideoxy sequence walking and sequence confirmation steps are synthesized to include DNA tags complementary to an alkaline phosphatase conjugate. A higher throughput synthesis device is well along in testing as are new automated hybridization devices. We have developed software for automatically annotating ORFs and databases of precise termini of proteins and RNA.

Technical progress during the current budget period:

This report covers the activities of the PI (George Church) and the following 7 people: Alec Mian, a visiting research scholar since July 1990, Margret Hoehe, a Visiting Scientist (in our group since 1992) with a joint appointment from NIMH/NIH Clinical Neurogenetics, Keith Robison, a graduate student (in our group since 1992), Dereth Phillips, a graduate student (in our group since 1992), Richard Chin, a Harvard/MIT Health Sciences Technology student (in our group since 1991), Andrew Link, graduate student (in our group since 1988), Laura Jaehn, a Research Assistant II (in our group since 1988).

The objectives of this project are automation and optimization of multiplex sequencing. These objectives often fit in well with the needs of other non-multiplex based sequencing methods including automated gel electrophoresis based and hybridization based sequencing methods.

System integration and testing. A few variations on the original multiplexing have been integrated into a full sequencing assembly line with automated and manual components. The system has been tested by seven sequential projects: *Escherichia* and *Salmonella* scattered sequence tags, 1 large *Salmonella* operon, 2 *Arabidopsis* cosmids, human and mouse simple sequence repeats, 3 *Escherichia* cosmids, human DNA polymorphism studies, and 19 *Mycobacteria* cosmids. For the *Escherichia* cosmids, we integrated direct transfer electrophoresis and automated membrane probing into the system. Sequence patterns for two cosmids were detected using chemiluminescence with oligonucleotide probes

directly conjugated to alkaline phosphatase. Primers for the directed dideoxy sequence walking and sequence confirmation steps were synthesized with a tag component complimentary to an alkaline phosphatase conjugate. Sequence patterns for the third cosmid were detected using radiolabeled oligonucleotide probes. Film data were automatically read and assembled using the programs REPLICA and GTAC. Based on protein matches we estimate that our consensus sequence indel error rate is less than 0.004%

We have worked with Collaborative Research Inc. to upgrade large-scale multiplex sequencing system using duplicates of our instrumentation and software to sequence *Mycobacteria leprae* and tuberculosis genomes (NIH NCHGR grant). In the past 12 months, they have sequenced 15 cosmids (bringing the total to 19) and are well along on another set of cosmids which includes DNA from human chromosome 10. This is all the more significant considering that electrophoretic compressions are encountered in the DNA from these organisms about every 100 bp, in great contrast to their relative absence in *S. cerevisiae*, *H. sapiens*, and *C. elegans* DNA. It should also be noted that this project has had to simultaneously cover some of the instrumentation, application, and software tasks that would normally be accomplished by an ABI-type instrumentation group and the 1000 users/application developers.

Data and technology availability: Our clones, copies of X-ray films, software, and digitized images have been made available to several laboratories prior to publication. Sequence data from 4 *Mycobacteria* and 2 *Escherichia* cosmids have been sent to NCBI prior to publication (a total of 19 and 3 respectively should be at NCBI soon). The *Salmonella* cob project has been deposited and published. Our protein databases of 400 *Escherichia* 2D spots and 700 bacterial ORF homologs are available by ftp from rascal.med.harvard.edu and arrangements are being discussed with NCBI for making our ASN.1 versions more generally available.

Pooled DNA preparations and sequencing reactions. We have developed reproducible Taq thermal cycle sequencing reactions for use with multiplex custom 38 mer primers containing 20 bases of tag attached to 18 bases of sequencing-walking primer. These are probed with same alkaline phosphatase conjugated probes as the recombinant-plasmid-based multiplex tags. Margaret Hoehe has developed multiplex PCR sequencing. Multiplex PCR from human genomic DNA using several

pooled primer pairs (one member of each pair biotinylated and one digoxigeninylated) followed by magnetic bead affinity selection of denatured strands (Stahl, et al. 1988) allows standard T7/Mn sequencing of the PCR DNA pool. The subsequent multiplex hybridizations use probes adjacent to each primer.

Slab & capillary array transfer electrophoresis. We have developed ultrathin slab direct transfer electrophoresis (DTE) such that we routinely pack 96 lanes per gel and read out to 400 bases. Read lengths out to 850 bases have been obtained, but the longer, heavier gels required have put the production use of this technique off until an automated gel device is in place. (see attached Methods in Enzymology paper). We have constructed a prototype for capillary array transfer electrophoresis (CATE), but only very limited testing has been done. The system already seems to achieve narrow (less than 0.8 mm) lanes with more predictable positioning of the lanes on the membrane than slab DTE. Although speed and resolution has not yet been addressed for the membrane transfer system, our collaborators in the Karger group observe single base resolution at 350 bases in 30 minutes using detection within the capillaries using the same batch of coated capillaries and polyacrylamide (Ruiz-Martinez, et al. submitted). We have shown that 4 capillaries can be filled at once from a common piston pressurized buffer chamber; there is no obvious upper limit for this step.

Automated membrane probing. We have designed new automated membrane array probing (AMAP) which combines aspects of various previous manual and automated protocols. A prototype is being made and is discussed in our grant renewal. The target performance is to probe up to 100 membranes simultaneously at speeds of one batch per 4 hours.

Automated picomole scale DNA synthesis for walking and PCR. Many directed sequencing strategies use large numbers of synthetic oligonucleotides in picomole amounts. Current chemical DNA synthesis devices require the syntheses to occur inside columns, with typically 40 nanomoles or more of solid phase initiator nucleoside is necessary for each synthesis. Scaling-down in volumes generates problems of large dead-volumes and/or the clogging of very small tubing. Scaling-up in column number involves an entire set of complicated plumbing and wiring per column.

We have been developing an approach to synthesize 96 oligonucleotides simultaneously at scales of 100 picomoles or less as determined by the size of the pin tip, eliminating the problems mentioned above by dipping solid phases into shared liquid vessels (troughs). One large region for TCA, oxidizer, and rinses and a set of 5 thin monomer troughs (A,C,G,T, and special) is cut into the top surface of a block of polypropylene. Each trough is connected to reagent dispensing syringes and lies beneath a grid of rods. There can be a maximum of 96 rods, the arrangement of which is the same as the standard spacing for the popular 96 well plates (9 mm). The tip of each rod is coated with controlled pore glass particles containing the initiating 3' nucleoside, and the other end of each rod is attached to a computer controlled actuator (solenoids and pneumatic have been used). The input for the controlling program is a file containing the sequences of up to 96 desired oligonucleotides. Syntheses occur in a small, sealed box in an acetonitrile-saturated anhydrous argon atmosphere.

For a 0.5 mm tip diameter the yields are uniformly 100 picomoles. The products are assayed by priming DNA sequencing, PCR, kinase labeling and gel electrophoresis. By these assays synthetic oligonucleotides made are comparable in quality to those made on commercially available synthesizers. Looking beyond oligonucleotides, the individually addressable pin array represents a very versatile system for sub-microliter manipulations, scaling up in number and down in volume easily.

Automated annotation. Minimal annotation for sequence database contributions should include locations of RNAs, proteins, motifs, and homologies. To help develop and test programs making such inferences we have established a database of *E. coli* proteins, N-terminal sequences, pI, MW, abundance, and subcellular localization. To be maximally useful as a reference protein sequence collection, the *E. coli* genomic database is expected to have few or no completely uncharacterized ORFs. The above protein database was established by Andy Link using sub-cellular and 2D gel fractionations plus automated sequencing to quantitate and establish the 12 N-terminal residues. Blast searches revealed 70% unambiguous hits, with the rest either ORFs or unknowns. Whole genome knock-out libraries have been constructed which combine multiplex sequencing tags as gene mutation, counterselection with *sacB* (Blomfield et al 1991), and

temperature sensitive replication origin replacement vectors (Hamilton, et al. 1989) compatible with many strains (including wild types) of enterics. Alec Mian and Richard Chin have developed methods which can extend the concept of systematic study of protein termini to that of RNA termini and DNA-protein interaction sites. Such a database is vital for testing algorithms for prediction of sites in genomic sequences. In order to avoid ambiguities in the identity of terminal bases, Alec Mian has established a oligomer attachment procedure based on RNA ligase and PCR and has used this procedure to sequence several non-polyadenylated transcripts from repetitive regions.

Keith Robison has developed a C++ object-oriented library for molecular biology which he has applied to automatically finding and graphically annotating ORFs including matches to database sequences. This has turned up 128 ORFs greater than 70 codons long in the E. coli contigs of which 31 were previously known, and a new potential operon(s) covering 13 consecutive long ORFs without homologs in any databases. The latter is by far the longest found in E. coli or Salmonella to our knowledge. Within it are two 1200 codon ORFs with self similarity. In addition about 280 ORFs have been identified in the Mycobacteria data and over 700 new protein homologs found in bacterial DNA previously believed to be or by default non-coding. We have worked with Jim Ostell at NCBI to develop ASN.1 tools for output of sequence motif and error-likelihood annotation for DNA and protein sequences and with Kenn Rudd at NCBI on the E.coli database. Keith will be working at NCBI for a short time this August.

Proposed research for the coming year if different from the original proposal: The research has continuing essentially as originally proposed; new approaches are described in a grant renewal application submitted July 15, 1993.

DATE

FILMED

5/4/94

END