Genomic Sequence Comparisons

George Church
Harvard Medical School; Boston, MA 02115
617/432-0503 or -7562, Fax: -7266
http://arep.med.harvard.edu

The first objective of this project is completion of an automated system to sequence DNA using electrophore mass-tag (EMT) primers for dideoxy sequencing. The prototype machine will contain a 60 capillary array with 400 EMT-labeled sequence ladders per capillary. The system is designed to use 100-fold less reagent and have 500-fold higher speed (1000 bases per sec per instrument) than current sequencing technology. Cleavage and laser desorption of EMTs from membranes for subsequent detection by EC-TOF mass spectrometry. The second objective is to overcome the limitations of purely hypothetical annotation of the growing number of reading frames in new genome sequences. We measure gene product levels and interactions using DNA microarrays, whole genome in vivo footprinting and crosslinking.

Our approach involves system integration of instrumentation, organic chemistry, molecular biology, electrophoresis and software to the task of increasing sequencing accuracy and efficiency. Likewise we integrate such instruments and others with the needs of acquiring and annotation of large-scale microbial and human genomic sequence and population polymorphisms.

To establish functions for new genes, we use large scale phenotyping by multiplexed growth competition assays, both by targeted deletion and by saturation insertional mutagenesis. We will continue to develop a system to sequence DNA using electrophore mass-tags (EMTs). We will establish genome-scale experimental methods for sequence annotation.

The most significant findings in 1995-1996 were 1) Demonstration of use of electrophore mass-tags in dideoxy sequencing. 2) Development of IR-laser desorption method and model. 3) A novel dsDNA microarray synthesis strategy. 4) A new amplifiable differential display for whole-genome in vivo DNA-protein interactions. 5) Establishment and application of a microbial DNA-protein interaction database.

DOE Grant No. DE-FG02-87ER60565.

A PAC/BAC End-Sequence Data Resource for Sequencing the Human Genome: A 2-Year Pilot Study

Pieter de Jong
Roswell Park Cancer Institute; Buffalo, NY 14263
716/845-3168, Fax: -8849, *pieter@dejong.med.buffalo.edu*
http://bacpac.med.buffalo.edu

Large scale sequencing of the Human genome requires the availability of high-fidelity clones with large genomic inserts and a mechanism to find clones with minimal overlaps within the clone collections. The first need can be satisfied with bacterial artificial chromosome libraries (PACs and BACs) which already exist and further such libraries now being developed. However, a cost-effective way for establishing high-resolution contig maps for the human genome has not yet been established. Recently, a new approach for virtual screening for overlapping clones has been proposed by several research groups and has been discussed eloquently in a manuscript by Venter et al., 1996 (Nature). We will implement this approach for use with our human PAC and BAC libraries and use the first year as a pilot stage. The goal of the one year pilot is to prove the feasibility of large scale end sequencing and to demonstrate usefulness.

The first goal will be met by sequencing the ends for 40,000 clones from our existing PAC library and from BAC libraries currently being developed under NIH funding within our laboratory. The end-sequencing will be based on our new DOP-vector PCR procedure (Chen et al, 1996, Nucleic Acids Research 24, 2614-2616). All sequence data will be made available through public databases (GSDB, GDB, Genbank) and will also become BLAST searchable through the UTSW WWW site from our collaborator, Glen Evans. In view of our current under-developed informatics structure, we do not expect to provide BLAST search access through our own web site during the pilot phase.

To prove the usefulness of available end sequences, we will prepare a chromosome 14-enriched clone collection from our current 20-fold deep PAC library. To detect the chromosome 14 clones, we will use as hybridization probes a set of 1,000 mapped STS markers available from Paul Dear (MRC, Cambridge, UK), the about 600 markers present in the Whitehead map and the in situ mapped BAC and PAC clones available from Julie Korenberg. We will hybridize with these existing markers in probe pools, specific for regions of chromosome 14. Thus we will isolate region-enriched PAC clone collections.

Assuming that the clone collections will be at least 50%-specific for chromosome 14 (50% false positives) and will include most of the chromosome 14 PACs from our library, a collection of about 35,000 clones is expected.