DOE GRANT NO. DE-FG02-88ER60699

FINAL TECHNICAL PROGRESS REPORT

P.I. Radomir CRKVENJAKOV

GRANT TITLE: SEQUENCING OF MEGABASE PLUS DNA BY
HYBRIDIZATION: METHOD DEVELOPMENT ENT

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of 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. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

January 31, 1991

MASTER

Note: This report is taken out of the Grant Application "An Integrated Approach for Complex Genome Studies" submitted by Biological and Medical Research Division, Argonne National Laboratory (Radomir Crkvenjakov and Radoje Drmanac, Co-Principal Investigators) to OHER, DOE in October 1990.

PRINCIPLE OF SBH

A linear sequence can be unambiguously represented (written) in two ways. One is defined with (P,E) data set and the other as oligosequences. The former set consists of data in which a sequence element E is defined for each position P, and the latter is a minimal set of oligosequences which define original sequence by their unique overlap. These theoretical possibilities give only two different ways of sequence determination of nucleic acids. Either one experimentally determines which base is on which position, or the minimal content of oligonucleotide sequences. The principle of the process of sequence regeneration from oligosequences data is via the unique overlap of oligomer sequences (10). In the common case, information about the presence in DNA of the majority of oligomer sequences of a particular length and on a small number of longer oligosequences is required to obtain its complete sequence. The numbers and lengths of oligosequences show an approximately quadratic dependence on the sequence length (10).

SBH is the only sequencing method based on the experimental determination of the content of oligonucleotide sequences. The data acquisition relies on the natural process of base pairing. It is possible to determine the content of complementary oligosequences in the target DNA by the process of hybridization with oligonucleotide probes of known sequences (14,15). The number of probes is 4^L , where L is the length of the probe. SBH requires large number of probes for complete sequences of even short fragments. The sequence determination requires hybridization data with the biochemically and technologically most convenient probes (16,000 7-mers to 250,000 9-mers).

OTHER EFFORTS

The first use of multiple oligomer hybridization in genomics was proposed by Lehrach and co-workers in 1986 (16) for linking up cosmids in genomic library. Since then his group obtained the proof of concept of cosmid link-up on the 350-kb genome of the herpes simplex virus, whose sequence is known (17). The delay in applying the oligomer link-up method to more complete genomes (caused by unreliable hybridization of oligomers shorter than 12 bases) was overcome when R. Drmanac brought Belgrade hybridization conditions (15) to Lehrach's lab (H. Lehrach, private communication). This group has worked out robotics as applied to hybridization data collection from membranes. The efficient methods of picking individual cosmid or λ phage clones into microtiter plates and their propagation and transfer onto membranes were developed for both link-up and the cosmid reference library work (H. Lehrach, private communication). Up to 10,000 cosmid colonies from microtiter plates can be spotted per 20 x 20 cm membrane into ordered arrays by a robot developed in-house and efficiently interrogated by hybridization. The wet filters are autoradiographed by Phosphorimager apparatus (Molecular Dynamics, Sunnyvale, California) to transfer hybridization results directly into computer compatible output. This setup currently can produce up to 100,000 hybridization data bits per day.

E. Southern has been working on oligomer synthesis and hybridization since 1988 (18) with the basic idea of simultaneous synthesis of various oligomers attached to a glass surface for purpose of fingerprinting DNA by hybridization. The position of specific oligomers is known from the manner of

synthesis using changeable partial masking of the surface at each elongation step. A glass plate harboring about 4000 oligomers 16 nucleotides in length was made. However, the hybridization is not sufficiently predictable yet (E. Southern, private communication).

Mirzabekov's group in Moscow has proposed sequencing 200-bp DNA fragments by hybridizing them one at a time to a glass surface carrying all 65,000 possible octamers (19). The oligos would be placed on a thin film of acrylamide able to strongly bind them which overlays the glass (20). The branching point ambiguities would be resolved by hybridization of obtained hybrids with a mixture of oligomers and scored under conditions when only duplexes having the covalently noncontinuous oligomer strand survive. We are aware of their study of this model on the level of several oligomers only. Recently, but after us, they have advanced ideas of miniaturization, fluorescence labeling, and the sequencing chip as the future of SBH (21,22). Pevzner has extended their claim of about 80% of random 200-bp fragments not requiring the second round of hybridization to 96% in a computer simulation study (23).

We have agreed to collaborate with a group of scientists from ORNL led by Dr. B. Jacobsen on application of our oligo-on-based hybridization in flow-cytometry sorting and mass spectroscopy detection for DNA sequencing purposes. They have obtained a SBIR grant for this work from DOE. Finally, we have begun collaboration with Dr. R. Overbeek, senior computer scientist from ANL, on the logical programming of the software for the SBH, with encouraging results.

PROGRESS REPORT

Our theoretical proposal for SBH (10) was the basis for the grant application to DOE. The grant was approved and funded on the basis of United States-Yugoslav treaty on scientific and technological cooperation in 1988. This required an equal share of DOE and Yugoslav funds, with DOE funding materials, equipment and travel and the Yugoslav side salaries and overhead. All proposed goals could not be achieved due to the five-fold reduction of the requested DOE funding. This report summarizes the achievements of the first two years of the grant. The work has proceeded along three main lines: 1) biochemical experiments; 2) computer software development; and 3) theoretical work. Steady progress has been made in all.

Before proceeding with the enumeration of the scientific results, a note about personnel. The Genome Structure Unit of the Institute for Molecular Genetics and Genetic Engineering in Belgrade has a staff of eight people and some temporary personnel. In the four-year existence of the Unit, the staff obtained 3 Ph.D.'s and 4 M.A.'s. The latter required the first authorship on a published paper. The staff turnover was entirely in the direction of young students replacing Ph.D.'s who have gone to the United States and the United Kingdom. Dr. Maja Bucan is now Asst. Prof. at the Department of Psychiatry, Pennsylvania University; Dr. Dean Nizetic is ready to start his own lab in the United States or Europe after a 3-year postdoc in Lehrach's lab at ICRF, London; and Dr. Radoje Drmanac is the co-principal investigator on this proposal. In the last year, seven people (the whole lab except one)

worked on SBH full time. There is a high value in maintaining this talent-training tradition and the productivity of the Genome Structure Unit; it should remain open when co-P.I.'s move to the United States.

BIOCHEMISTRY

A summarized version of our achievements is presented below. Further details and the evidence are to be found in published papers and manuscripts prepared for publication (see Appendix).

Conditions for performing hybridizations with short oligomers have been delineated in support of a strategy for genome sequencing by hybridization. A theoretical analysis defined the optimal conditions and they have been implemented in practice. Valid discrimination of perfectly paired and imperfectly paired hybrids of probe/target DNAs has been made. This capacity has been used in a proof of concept SBH demonstration. Oligonucleotide hybridization capable of discriminating between perfectly and imperfectly mismatched hybrids is the biochemical basis of SBH. The most difficult case to differentiate is a single end-mismatched target. None of the previously described oligohybridization methods necessarily addresses this crucial case. Theoretical and practical solutions for the discriminative hybridization of short oligo probes (15) allows the development of SBH and partial SBH and has importance in many other nucleic acid analyses.

The problem of discriminative hybridization with very short probes can be divided into two parts. One part is the discrimination between the fully matched and the end-mismatched (or other mismatched) duplexes; the other is the detection of the positive signal over background. As the probe shortens, it forms a less stable hybrid duplex. The less stable the duplex, the easier it is to discriminate the perfect from imperfect hybrids, but it is also more difficult to obtain a sufficient amount of hybrid to be able to overcome the background. Using the "all or none" kinetic expressions for the process of hybrid formation and the process of hybrid melting, a discrimination (D) equation is defined as:

$$D = [H_p(o)/B] [B/H_i(o)]^{(A_p/A_i)} e^{(E_{a_i} - E_{a_p})/RT}$$

where:

H(o) = amount of hybrid entering into the differential melting process;
 B = background;
 A = pre-exponential factor in the Arrhenius equation;
 E_a = activation energy; and

the subscripts p and i denote perfectly matched end-mismatched duplex, respectively.

The analysis of the equation leads to the two unexpected conclusions:

1. During hybridization it is better to optimize the hybrid yield at the expense of the obtained discrimination.

2. Better discrimination is obtained by longer melting at lower temperature than by shorter melting at restrictive, stringent temperature.

In both cases, the effect is greater the larger the difference in activation energies (E_a) between fully matched and mismatched duplexes. The practical consequence of these results is that the same conditions of both hybridization and melting can be used for probes giving hybrids with different stabilities. Maximally, a few temperatures and lengths of hybrid wash (melting) can be employed for all possible octamers. Using the developed conditions of hybridization, the pilot study on 28 oligomers ranging from 12- to 6-mers (but mainly octamers and septamers) has shown that sufficient discrimination can be obtained between targets containing the fully matched versus end-mismatched targets (15). This allows the correct sequence call with all the probes.

The exception are DNAs which have an excessive number of end-mismatched targets. The sum of the weak signals coming from mismatched targets in extreme cases becomes larger than the positive signal coming from a single matched target. This limits the length of DNA that can be sequenced by hybridization with a given probe length. The signal intensity varies as a function of probe sequence. Therefore, positive hybridization controls of known sequence are indispensable for calibrating hybridization intensities in three groups in order of decreasing signal intensity: those caused by a full match, multiple end mismatches, and one or no end mismatches, respectively.

To demonstrate sequencing by octamer and nonamer hybridization, we have chosen a 100-base segment of the 921-bp human b-interferon gene fragment, the sequence of which is known. The selected segment dictated the choice of 93 fully matched oligomers, 73 octamers and 20 nonamers in consecutive overlapping frames displaced by one or two bases. No more than two single end-mismatched targets were present in the rest of the interferon DNA. An additional group of 12 probes were designed to have at least one end mismatch in the 100-bp segment while having a single full match in one of the four control DNAs of known sequence. The DNAs having imperfect targets had the single end-mismatch type for 26 out of 105 probes. Without exception, the DNA with a full match hybridized more strongly than control DNA. Using the control probes having a perfect match in both test and control DNA showed that this discrimination was not due to a higher amount of test DNA compared to control DNA. In many cases there was more DNA in the control than in the interferon test dot.

For 102 probes, the discrimination values, D , defined as ratio of signal intensities in the two dots, were between 2 and 40. Only for three probes was the difference in signal intensity not significant, i.e., below 2. The variations in discrimination are primarily the result of the hybridization efficiency, which allows the visualization of the signal over background. Another important factor is the kind of mismatched target present in the control DNA.

The overall results on these 105 probes are in agreement with those already obtained (15) and demonstrate the efficient determination of presence of both full and end-mismatched targets in test DNA. In summary, the results

IFs sequences are fully regenerated. In view of the informatic richness of partial sequences (see section on partial sequencing), even the segments with uncertain order of the small number of correct SFs (which is what these 20% are) are as useful for the comparative analysis of sequences as are the complete sequences. In summary, the outline of a powerful algorithm is in place for the regeneration of sequences from hybridization data even when the data is partially incorrect. The development of the final version of the algorithm is now a legitimate problem for computer science.

THEORETICAL WORK

Complete Sequencing by SBH

There is ambiguity when a number of sequence sub-fragments (SFs) with overlapping ends can be ordered in several ways with equal probability (10). SFs are obtained due to the multiple occurrences of L-1 oligosequences in a target sequence. If octamers are used as probes and 100-bp fragments are sequenced, one expects only 1% of the 100-bp fragments to have an ambiguous order of SFs. Based on this property there has been proposal to use SBH on short fragments only (24,19).

However, in the sequencing of very long DNA molecules, the main advantages of SBH become apparent. By using convenient libraries of overlapping genomic fragments, it is possible to use longer fragments and to significantly decrease the number of independent samples (clones). The solution allowing the sequencing of DNA of genome length consists of the use of IFs. Also, instead of using of the solitary short 100- to 200-bp DNA fragments, it is possible to use random pools 10-kb in length (10). By these improvements, the amount of necessary experimental work is reduced up to 50-fold, at the expense of an increase in the computational work. Genome sequencing by SBH requires determination of the number of information bits represented by a matrix of N clones (or genomic fragments) by K probes.

Because of the requirement for 10-20 genomic equivalents in the randomly generated libraries, the small allowed ratio of clone-to-probe length imposed by the informatic rules (SF formation and ordering), and the biochemical requirements for discriminative hybridization, the libraries to be analyzed must have 100-fold genome redundancy. This means that the required number of hybridization bits is 100-fold larger than the total number of base pairs in the sequence to be determined. For the sequencing of the entire human genome this amounts to 300 billion bits. Theoretically, the number of probes can be reduced at the expense of the number of clones or vice versa, but the number of bits remains approximately the same. The optimal ratio is determined by the practical parameters. The two extremes can be defined, one with the number of clones prevailing and the other with a dominant number of probes. Since the number of hybridization reactions depends on the number of probes, ultimately it will be more efficient to have the oligomers fixed to a support and the DNA fragments being sequenced in solution. Lysov et al. proposed this strategy using 65,000 bound octamers for sequencing 200-bp fragments (19). However, we believe that for genome sequencing in the gigabase range, the optimal threshold requires at least all 1,024,000 decamers to be fixed to the hybridization surface. Making and using this huge number of probes in the inverse oligo hybridization approach are described in the next section.

Sequencing Chip

SBH is based on determining the content and not the positions of oligosequences in target DNA. The method, therefore, does not require macromolecular separation or fractionation of subject DNAs. This allows samples to be applied in microvolumes, and high sample densities can thus be achieved. The samples do not have to be in the ordered positions. These characteristics allow a scheme of microhybridization, an implementation we believe is necessary for an efficient genome technology. Micropositioning of separated samples or oligo probes on defined coordinates can be done by precise robotics or by electromagnetic filing techniques. This gives a miniaturization of the hybridization surface but still requires a handling of separated clones or probes. It is anticipated that the preparation of millions of DNA samples will be the rate-limiting step in genomic sequencing. As mentioned, the number of required genomic fragments in SBH can be reduced at the expense of increasing the number of probes. If a billion 15-mer probes are used, only 1000 overlapping 10-megabase fragments are sufficient for sequencing of the human genome. In this case, the problem of preparation of a larger number of genomic fragments is converted to the preparation of an even larger number of oligo probes.

To solve this latter problem, a scheme for a "sequencing chip" has developed. Different oligos are synthesized attached to support beads (25). In this scheme, only 1000 synthetic reactions are sufficient to make all possible 15-mers. Following the synthesis, one forms a monolayer of beads fixed to the surface. In a "chip" manufactured in this way, every bead carries a sufficient number of molecules of the same oligo, and each is functionally akin to a microhybridization dot on the membrane. In hybridizations with the chip, 20- to 50-bp DNA pieces carrying fluorescent label obtained sharing genomic DNA fragments would be used as probes. The scoring of positive hybridization events would be performed by laser-excited fluorescence microscopy.

The main problem in making such a chip is determining which oligomer is on which bead. Additional combinations of marker oligos could be used to mark beads during oligo synthesis. Through a hybridization of 40 oligomer probes complementary to the marker oligomers after monolayer formation, a billion different beads each carrying a different 15-mer can be distinguished. If 1- μ m beads are employed, the ultimate sequencing chip would have the size of a 3 x 3 cm microscope slide. The manufacture of chips of different level of complexity (containing from 6- to 15-mers) can provide an efficient technological answer for the different levels of genome research, including sequencing, diagnostics, populational and ecological surveys, and others.

Partial Sequencing

Some 100,000 oligomers and highly automated systems will be necessary to fully implement SBH. However, there are highly beneficial uses of SBH which can be pursued now with much smaller oligomer collections. The ordering of a cosmid library through SBH fingerprinting requires only 100 oligomer probes (16). Here, it is shown how a several thousand probes can suffice for a

highly informative comparison of related genomes or their smaller parts up to the cDNA level. We call this procedure partial sequencing, or partial SBH (Table 1 and Fig. 1, pp. 29-32). Table 1 presents the status of partial sequencing in comparison to mapping and complete sequencing using the SBH for all three tasks. A graphical representation of the definition of partial sequence is shown in Fig. 1. Data of the complexity required for partial SBH allow the recognition of gene and genome structure by measuring the homology of unknown sequences to known sequences or sequences constrained by some informatic rule. Another important application is using that set of some 1000 probes to locate important sequence motifs within genomic or cDNAs clones or DNAs amplified by PCR.

The high value of this latter application deserves elaboration. Comparisons of nucleic acid sequences and their encoded functions are revealing an increasing number of sequence motifs, including those for transcriptional promotion, translation initiation, transcription termination, RNA processing, active sites of serine proteases, zinc fingers, and other functional domains within proteins. Many of the motifs are so short and degenerate that they could not previously be sought physically through nucleic acid hybridization procedures. Thus the recognition of further examples of known motifs has most commonly required the sequencing of a gene or its corresponding cDNA, so that sequence comparisons could be made. Contemporary procedures for sequencing contiguous segments of DNA are painfully slow for unearthing the nuggets of information represented by functional motifs and their DNA neighborhoods. The total information value will increase much faster than linearly with the number of probes because of the combinatorial utilization of the short sequence information. The motif knowledge thus gained can aid in setting priorities for refined genome locus or cDNA analyses, pending the maturation of technologies for intensive and economical genome sequencing. Importantly, the modest material requirements of partial SBH allow for genomic analysis on the sequence level with the currently available methodology.

The full strategy for genome SBH is very robust in respect to tolerating loss of oligomer binding data. The correct sequence is still computationally obtained, even if many of the experimental oligomer hybridization data are eliminated from the input to the sequence assembly algorithms. With continual elimination of input data, incorrect sequences are not generated. Rather the initial single sequence is broken into sub-fragments. Such SFs may overlap one another, but there is not enough confidence to declare their joining. Breakage into smaller SFs continues up to the level of single probes as more of the input data is eliminated, but each of the SFs is a correct segment of the total sequence. The clone origin of SFs under the map of overlapping clones remains known.

This robustness in the generation of valid SFs stimulated a question. What is the minimal amount of oligomer data that would allow comparison of related DNAs and recognition of motifs? Subsequent computational simulations revealed that even with 1000 of the 100,000 probes required for complete SBH, useful output would be obtained. With each additional probe, longer output SFs are generated. The total sequence of the target chromosome/genome is progressively filled in and confidently linked up as data from more and more probes are entered. There is thus value in the incomplete sequence data early in the SBH process and its value rapidly increases with the number of probes. These are the features which motivate the concept of partial SBH.

When analyzing individual cDNAs or polymerase chain reaction (PCR)-amplified fragments, the positions of oligomers within a target DNA cannot be obtained. However, for known motifs, a relevant combination of short oligomers is chosen. The probability of a motif's presence in target DNAs is computed from the binding data obtained. Thus DNAs which may encode sought motifs are identified for more detailed analyses. Physically, the partial sequence is represented by a population of SFs having various lengths with a nearly random distribution over the genome. Depending on the amount and randomness of the information gained, some sequence segments will not be represented in the SF population. A statistically adequate SF population can be very simply specified. Every positively hybridizing oligomer represents a "quantum" of information. A minimal desirable sampling of oligomer sequences for obtaining most of the biological data is achieved when most base pairs (over 75%) are covered at least with one oligomer probe. This number of probes for a random sequence is defined as $4/2L$, where L is the length of the probes. It amounts to about 350 6-mers, 1200 7-mers, or 4100 8-mers. The corresponding maximal lengths of double-stranded DNA fragments which can be partially sequenced are 1, 4, and 16 kb, respectively. Uniformly distributed information can be obtained with a minimal possible overlap of probes. Conversely, for determination of longer oligomer sequence segments, the groups of oligos with overlap provide the required data.

Applications of Partial SBH

Since partial sequencing is a methodological concept which does not have an established field of application, here we briefly explain its uses. More rigorously defined partial sequencing comprises the amount and type of sequence information necessary for:

- 1) Measuring biologically relevant homology of unknown sequences;
- 2) Recognition of gene and genome structures; and
- 3) Physical isolation of a partially sequenced DNA fragment.

In addition to primary tasks of measuring homology and recognizing gene and genome structures, the partial sequence should provide in most cases a possibility for a simple physical reisolation of sequenced DNA fragments via definition of STSs. The recovery can be accomplished either by reprobng with a specific combination of a few oligomer probes or by PCR with primers synthesized on the basis of partial sequences. This is an important advantage of sequence information over mapping data which makes it unnecessary to permanently store original clones. For the prediction of the biochemical function(s) of a unit of genomic DNA, it is possible to use the following sequence characteristics: 1) global content of subsequences, i.e., differences from a random sequence such as G+C content, CpG/TpA ratio, triplet content, or simplicity; 2) total homology; 3) content of functional subsequences (promoters, enhancers, protein motifs, etc.); 4) rate of changes in the evolution (variable and conserved regions, allelic differences); and 5) position effects (precise distances, reading frames, etc.). It is possible to recognize coding regions by measuring sequence characteristics delineated in items 1 and 4 (26). The information from item 1 is sufficient to recognize isochore structures or HTF islands. The biochemical functions of genes could

be predicted by the determination of homology and content of protein motifs. New members of a repetitive family can be recognized by measuring homology with known sequences.

To recognize most of the mentioned sequence characteristics (1 to 4), it is not necessary to know the complete sequence. The more different the sequences are from the random case or the higher the homology, the amount of required information is less. For a given DNA segment, the determination of only a part of the complete oligosequences content is sufficient to recognize many sequence characteristics and to predict some biochemical and biological functions. A main requirement is that the oligomer content must be determined for fragments of a length which approximately coincides with length of functional segments. For example, partial SBH may serve for recognition of most exons if the content of oligosequences is determined in fragments 100-200 bp in length, because that is close to their average length (27). As we proposed previously (10), it is not necessary to actually work with such short fragments. Short physical fragments can be replaced by informational fragments. It is possible to get an overview of genome organization in a fast and economical way by partial SBH of appropriate genomic libraries. For partial SBH of the human genome one needs to hybridize 5 million clones of 1-3 kb with 1000 6 to 8-mer probes. The job can be accomplished on a chromosome-by-chromosome basis. Since this analysis allows the recognition of the major structural elements of genomic sequence, it can give very important genetic information for molecular biology. The elements are: i) isochore organization (28); ii) disposition of both known and yet undiscovered repetitive sequence families; iii) global placement of genes and the specific localization of most of them; and iv) global pattern of known and potential regulatory elements. A deeper understanding of the structural organization detected in this way can be obtained by comparisons of partial genomic sequence with partial sequences of cDNA clones from different tissues, as well as with the comparative analysis of partial sequences of individual genomes within a species and genomes from related species.

Partial SBH has a significant potential in the field of molecular epidemiology by measuring, among other things, the low rate of mutations in natural populations. Partial SBH that reads most of the bases can discover most polymorphic and mutated sites by comparison of genomic DNAs from several individuals. Direct gene markers are obtained by partial sequencing of cDNA libraries. It is sufficient to identify 1% of several million polymorphic sites in a genome to obtain the desired 10,000 to 100,000 markers. It is necessary to develop the procedures for the isolation of the 10,000 to 100,000 specific fragments of genomic DNA and for their partial sequencing in order to efficiently determine the genetic status of individuals. This is feasible to do now with PCR using a specific human chromosome or with a cDNA part of a genome in the megabase range (easily recoverable from a number of individuals). The problem of the number of required primers and amounts of DNA needed for partial sequencing as future routine work over the whole span of the genome can be resolved in principle by the mixed PCRs and by the miniaturization of hybridization (sequencing chip concept).

These theoretical results give us a clear direction for further work on the development of SBH. Our plan is presented in following sections.

LITERATURE

1. Cook-Deegan, R.M., The Alta Summit, December 1984, Genomics 5 (1989), pp. 661-663.
2. Burke, D.T., G.G. Carle, and M.V. Olson, Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors, Science 236 (1987), p. 806.
3. Olson, M V., L. Hood, C.R. Cantor, and D. Botstein, A common language for physical mapping of the human genome, Science 245 (1989), pp. 1434-1435.
4. Studier, F.W., A strategy for high-volume sequencing of cosmid DNAs: random and directed priming with a library of oligonucleotides, Proc. Natl. Acad. Sci. USA 86 (1989) pp. 6917 -6921.
5. Phandis, S.H., H.V. Huang, and D.E. Berg, Tn5supF, a 264-base-pair transposon derived from Tn5 for insertion mutagenesis and sequencing DNAs cloned in phage λ , Proc. Natl. Acad. Sci. USA, 86 (1989), pp. 5908-5912.
6. Edwards, A., H. Voss, P. Rice, A. Civitello, J. Stegemann, C. Schwager, J. Zimmermann, H. Erfle, C.T. Caskey, and W. Ansorge, Automated DNA sequencing of the human HPRT locus, Genomics 6 (1990), pp. 593-608.
7. Church, G.M., and S. Keifer-Higgins, Multiplex DNA sequencing, Science 240 (1988), pp. 185-188.
8. Cantor, C., cited in McCormick, D., Genome Know-How, Biotechnology 8 (1990), p.5
9. Notice inviting grant applications, Office of Energy Research, DOE, U. S. Federal Register 55 (1990), 10486.
10. Drmanac, R., I. Labat, I. Brukner, and R. Crkvenjakov, Sequencing of megabase plus DNA by hybridization: Theory of the method, Genomics 4 (1989), pp. 114-128.
11. T.P. Beebe, Jr., T.E. Wilson, D.F. Ogletree, J.E. Katz, R. Balhorn, M.B. Salmeron, W.J. Siekhaus, Direct observation of native DNA structures with the scanning tunneling microscope, Science 243, 370-372.
12. Jett, J.H., 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, High Speed DNA sequencing: an approach based upon fluorescent detection of single molecules, J. Biomol. Struct. Dyn. 7 (1989), p. 301.
13. R. Drmanac, In: C.R. Cantor and H.A. Lim, eds., Electrophoresis, supercomputers and the human genome; Proc. of Conference, April 10-13, 1990, Tallahassee, Florida, World Scientific, (Dec. 1990), in press.

14. Wallace, B.R., Shaffer, F.R. Murphy, T. Hirose, and K. Itakura, Hybridization of synthetic oligonucleotides to ϕ X174 DNA: the effect of single base pair mismatch, *Nucleic Acids Research* 6 (1979), pp. 3543-3557.
15. Drmanac, R., Z. Strezoska, I. Labat, S. Drmanac, and R. Crkvenjakov, Reliable hybridization of oligonucleotides as short as 6 nucleotides in length, *DNA and Cell Biology* 9 (1990), pp. 527-534.
16. Poustka, A., T. Pohl, D.P. Barlow, G. Zehetner, A. Craig, F. Michiels, E. Ehrlich, A.M. Frischauf, and H. Lehrach, Molecular approach to mammalian genetics, *Cold Spring Harbor Symp. Quant. Biol.* 51 (1986), pp. 131-139.
17. Craig, A.G., D. Nizetic, J.D. Hoheisel, G. Zehetner, and H. Lehrach, Ordering of cosmid clones covering the Herpes simplex virus type I (HSV-I) genome: a test case for fingerprinting by hybridization, *Nucleic Acids Research* 18 No. 9 (1990), pp. 2653-2660.
18. Southern, E., U. K. patent application GB 8810400 (1988).
19. Lysov, Yu. P., V.L. Florentiev, A.A. Khorlin, V.V. Khrapko, V.V. Shick, and A.D. Mirzabekov, *Dokl. Acad. Sci. USSR* 303, pp. 1508-1511.
20. Khrapko, K., Y.P. Lysov, A.A. Khorlin, V.V. Shick, V.L. Florentiev, and A.D. Mirzabekov, An oligonucleotide hybridization approach to DNA sequencing, *FEBS Letters* 256 (1989), 118-122.
21. Mirzabekov, A.D., S. Khrapko, Yu.P. Lysov, A.A. Khorlin, V.V. Shick, and V. L. Florentiev, An oligonucleotide hybridization approach to DNA sequencing, *Symposium on Human Genome Research: Strategies and Priorities*, January 29-31, 1990, Paris, France.
22. Lysov, Yu.P., K. Khrapko, A. Khorlin, V. Florentiev, and A. Mirzabekov, DNA sequencing by oligonucleotide hybridization, *The First International Conference on Electrophoresis, Supercomputing, and the Human Genome*, April 10-13, 1990, Tallahassee, Florida.
23. Pevzner, P.A., L-tuple DNA sequencing: computer analysis, *J. Biomol. Structure and Dynamics* 7 (1989), pp. 63-73.
24. Bains, W., and G. Smith, A novel method for nucleic acid sequence determination, *J. Theor. Biol.* 135 (1988), pp. 303-307.
25. R. Drmanac and R. Crkvenjakov, Yugoslav patent application No. 767 (1989)
26. J.W. Fickett, Recognition of protein coding regions in DNA sequences, *Nucleic Acids Res.* 10 (1982), pp. 5303-5319.
27. M.W. Smith, Structure of vertebrate genes: a statistical analysis implicating selection, *J. Mol. Evol.* 27 (1988), pp. 45-55.

28. G. Bernardi, B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunierrotival and F. Rodier, The mosaic genome of warm-blooded vertebrates, *Science* 228 (1985), pp. 953-958.
29. U. Asseline et al., *Proc. Natl. Acad. Sci. USA* 81 (1984), 3297.
30. Lathe, R., Synthetic oligonucleotide probes deduced from amino acid sequence data; theoretical and practical considerations, *J. Mol. Biol.* 183 (1985) pp. 1-12.
31. Oishi, M., In gel reassociation at DNA: its application for mammalian genome analysis, In: First International Conference on Electrophoresis, Supercomputers, and the Human Genome, April 10-13, 1990, Tallahassee, Florida.
32. Davis, M.M., D.I. Cohen, E.A. Nielsen, M. Steinmetz, W.E. Paul, and L. Hood, *Proc. Natl. Acad. Sci. USA* 81 (1984), pp. 2194-2198.
33. Schweinfest, C.W., K.W. Henderson, J. Gu, S.D. Kottaridis, S. Besbeas, E. Panotopoulou, and T.S. Papas, Subtraction hybridization cDNA libraries from colon carcinoma and hepatic cancer, *Genet. Annal. Techn. Appl.* 7 (1990), pp. 64-70.
34. Aguan, K., T. Kusano, N. Suzuki, and Y. Kitagawa, An improved method for the construction of high efficiency cDNA library in plasmid or lambda vector, *Nucleic Acids Research*, 18 (1990), p. 1071.
35. Carrano, A.V., Lamerdin, Ashworth, Watkins, Branscomb, Slezak, Raff, de Jong, Keith, McBride, Meister, and Kronick, *Genomics* 4 (1989), pp. 129-136.
36. Deininger, P.L., Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis, *Anal. Biochem.* 129 (1983), pp. 216-223.
37. Southern, E., International Application published under the Patent Cooperation Treaty (PCT), (1989), PCT/GB89/01114.
38. Tsurui, H., E. Hara, K. Oda, A. Suyama, S. Nakada, and A. Wada, A rapid and efficient cloning method with a solid-phase DNA probe: application for cloning the 5'-flanking region of the gene encoding human fibronectin, *Gene* 88 (1990), pp. 233-239.
39. Zarytova, V.F., and I.G. Shishkina, Affinity chromatography of DNA fragments and P-modified oligonucleotides, *Anal. Biochem.* 188 (1990), pp. 214-218.
40. 3M Co., from advertisement on Near Commercialization of Novel Support Material, *Bioprocessing Technology*, October 1989.
41. Duncan, C.H., and S. Cavalier, Affinity chromatography of a sequence-specific DNA binding protein using Teflon-linked oligonucleotides, *Anal. Biochem.* 169 (1988), pp. 104-108.

42. Pochet, S., B. Arcangioli, and T. Huynh-Dinh, Solid-supported ligation primer, *Nucleic Acids Research* 16 (1988), p. 1619.
43. Majors, R., and M. Hopper, *J. Chromatic. Sci.* 12 (1974), p. 767.
44. Bayer, E.A., and M. Wilchek, *Methods Biochem. Anal.* 26 (1980), 1-45.
45. Chen, E.Y., Yu-C. Liao, D.S. Smith, H.A. Barbera-Saldana, R.E. Gelinas, and P.H. Seeburg, The human growth hormone locus: nucleotide sequence, biology, and evolution, *Genomics* 4 (1989), pp. 479-497.
46. Yoshitake. S., B.G. Schach, D.C. Foster, E.W. Davie, and K. Kurachi, Nucleotide sequence of the gene for human factor IX (antihemophilic factor B), *Biochemistry* 24 (1985), pp. 3736-3750.

TABLE 1. Human Genome by Oligonucleotide Hybridization

HUMAN GENOME BY OLIGONUCLEOTIDE HYBRIDIZATION

METHOD	No of PROBES	No of HYBRIDIZ. DATA BITS	EXPECTED RESULTS
MAPPING (non-oligo, more complex DNA probe can be used)	100	$3 \cdot 10^8$	ordered clone libraries, no biological information by itself
PARTIAL SEQUENCING	1,000 - 10,000	10^{10} - $5 \cdot 10^{10}$	location and kinds of genes, motifs, other sequences; tissue expression patterns via cDNA; overview of chromosomes organization
COMPLETE SEQUENCING	100,000	$3 \cdot 10^{11}$ - 10^{12}	definitive information for each base in the sequence

Incomplete sequence

Fractional sequence

Partial sequence

**DATE
FILMED**

6/3/94

END