(12) EUROPEAN PATENT APPLICATION

(71) Applicant: **PRESIDENT AND FELLOWS OF HARVARD COLLEGE**
17 Quincy Street
Cambridge Massachusetts 02138 (US)

(72) Inventor: **Church, George M.**
**400 Brookline Avenue Apt.21G**
Boston Massachusetts 02215 (US)

Kieffer-Higgins, Stephen
30 Beaumont Street
Dorchester Massachusetts 02124 (US)

(74) Representative: **Field, Roger Norton et al**
Brewer & Son Quality House 5-9, Quality Court Chancery Lane
London WC2A 1HT (GB)

(54) Multiplex sequencing.

(57) This invention features vectors and a method for sequencing DNA. The method includes the steps of:

a) ligating the DNA into a vector comprising a tag sequence, the tag sequence includes at least 15 bases, wherein the tag sequence will not hybridize to the DNA under stringent hybridization conditions and is unique in the vector, to form a hybrid vector,

b) treating the hybrid vector in a plurality of vessels to produce fragments comprising the tag sequence, wherein the fragments differ in length and terminate at a fixed known base or bases, wherein the fixed known base or bases differs in each vessel,

c) separating the fragments from each vessel according to their size,

d) hybridizing the fragments with an oligonucleotide able to hybridize specifically with the tag sequence, and

e) detecting the pattern of hybridization of the tag sequence, wherein the pattern reflects the nucleotide sequence of the DNA.

FIG. I

EP 0 303 459 A2
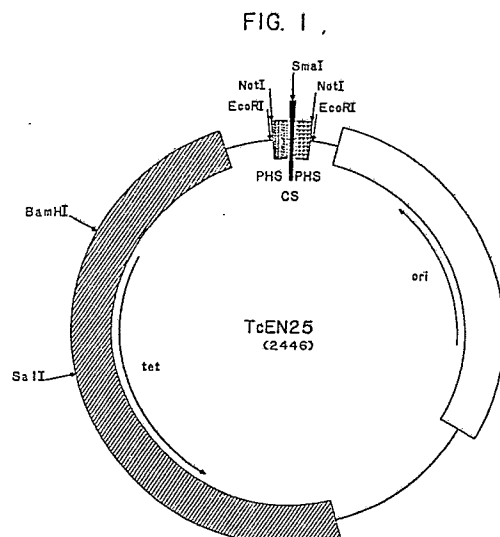
Description

MULTIPLEX SEQUENCING

This invention relates to sequencing of DNA.

The sequence of nucleotide bases in DNA is generally determined using the methods described by Maxam and Gilbert (65 Methods Enzymol. 497, 1980) or by Sanger et al. (74 Proc. Natl. Acad. Sci. USA 5463, 1977). These methods generally involve the isolation of purified fragments of DNA prior to sequence determination.

Church et al. (81 Proc. Natl. Acad. Sci. 1991, 1984) described a method of sequencing directly from genomic DNA. Unlabelled DNA fragments are separated in a denaturing gel after complete restriction endonuclease digestion and partial chemical cleavage of the genome. After binding these fragments to a nylon membrane the DNA is hybridized with a probe comprising RNA homologous to a region near to the region to be sequenced. The membrane can be reprobed with other probes to sequence other regions of interest.

Summary of the Invention

In a first aspect, the invention features a set comprising at least two vectors. Each vector of the set has a DNA construct having at least one restriction endonuclease site. Further, each vector of the set differs from each other vector of the set only at a tag sequence. The tag sequence includes at least 15 base pairs and is located within 50 bases of a restriction endonuclease site. Further, each tag sequence in the set will not hybridize under stringent hybridization conditions to another tag sequence in the set. By stringent hybridization conditions is meant low enough salt counterion concentration and high enough temperature to melt mismatched duplexes, to form single stranded molecules (e.g., at 42°C in 500-1000 mM sodium phosphate buffers). By vector is meant any fragment of DNA, whether linear or circular, which can be ligated to DNA to be sequenced. Such vectors may include an origin of DNA replication.

In a related aspect the invention features a vector for sequencing DNA. The vector comprises a DNA construct having two restriction endonuclease sites, each site being recognised by a first restriction endonuclease, wherein treatment of the vector with this first endonuclease produces a discrete DNA fragment consisting of the DNA extending between the sites. The vector further includes two tag sequences both being located on the DNA fragment, and separated from each other by a second restriction endonuclease site. The tag sequences include at least 15 base pairs and both ends of the tag sequences are located within 50 bases of the nearest first restriction site and within 50 bases of the second restriction endonuclease site. The tag sequences are unique in the vector. By unique is meant that there are no other nucleotide sequences in the vector which exactly correspond to the tag

sequences.

In preferred embodiments of these aspects the tag sequences have two strands, one of which is free from cytosine residues; the vectors are produced from a parental vector having randomly formed tag sequences ligated into it; and the randomly formed sequences are formed in a DNA synthesizer supplied with at least two nucleotides at each addition step.

In another related aspect, the invention features a set of tag oligonucleotides which bind to the tags in the vector, wherein each tag oligonucleotide of the set is unique in the set and neither tag oligonucleotide nor an oligonucleotide homologous to the tag oligonucleotide will hybridize under stringent hybridization conditions to DNA in the set. Each such oligonucleotide is at least 15 bases in length.

In preferred embodiments the set of vectors described above is produced by ligation into a parental vector of this set of tag oligonucleotides.

In a second aspect, the invention features a method for sequencing a DNA specimen, termed multiplex sequencing. The method includes the steps of:

a) ligating the DNA specimen into a vector comprising a tag sequence, to form a hybrid vector. The tag sequence includes at least 15 bases, will not hybridize to the DNA specimen under stringent conditions, and is unique in the vector;

b) treating separate aliquots of the hybrid vector in a plurality of vessels to produce fragments each comprising the tag sequence. These fragments in each vessel differ in length from each other and all terminate at a fixed known base or bases (e.g., A, T, C or G for Sanger dideoxy sequencing; G, A + G, T + C, T or C for Maxam and Gilbert sequencing). The fixed known base or bases differ from these in other vessels;

c) separating the fragments from each vessel according to their size;

d) hybridizing the fragments with an oligonucleotide able to hybridize specifically with the tag sequence; and

e) detecting the pattern of hybridization of the oligonucleotide; this pattern reflects the nucleotide sequence of the DNA specimen.

In preferred embodiments, the method further includes the step of ligating the DNA specimen to a plurality of vectors to form a plurality of hybrid vectors; each vector differs from each of the others at a tag sequence and each tag sequence is unable to hybridize under stringent conditions to other tag sequences; the method further includes the step of rehybridizing the fragments with each oligonucleotide corresponding to each tag sequence of each vector; each vector includes two tag sequences; the method further comprises binding the fragments to a solid support prior to the hybridizing step; the method further includes, prior to the separating

step, providing a molecular weight marker in one vessel and detecting this marker after the separating step; and the solid support comprises the vector or the oligonucleotide, wherein the vector or oligonucleotide acts as an identifying marker for the support.

In a third aspect, the invention features a method for repeatedly hybridizing a solid support, comprising nucleic acid, with a plurality of labels, including the steps of:

a) enclosing the support within a container comprising material thin enough to allow detection of hybridization of nucleic acid with the labels,

b) inserting hybridization fluid comprising a first label into the container,

c) removing the hybridization fluid from the container, and

d) detecting hybridization of the first label to the solid support, while the solid support remains in the container.

In preferred embodiments the method further comprises, after the detecting step, repeating steps b, c and d using a second label; and most preferably comprises repeating steps b, c and d at least 20 times with at least 20 labels.

In a fourth aspect, the invention features a method for determining the DNA sequence of a DNA specimen, including the steps of:

a) performing a first sequencing reaction on the DNA specimen;

b) performing a second sequencing reaction on a known DNA specimen;

c) placing corresponding products of the first and second sequencing reactions in the same lanes of a DNA sequencing gel;

d) running the products into the gel;

e) detecting the location of the products in the gel, and

f) using the location of the known DNA products to aid calculation of the DNA sequence of the DNA specimen.

In a fifth aspect, the invention features an automated hybridization device, for repeatedly hybridizing a solid support containing nucleic acid, with a plurality of labels. The device includes a container enveloping the solid support, formed of material thin enough to allow detection of hybridization of the nucleic acid with a label by placement of a label-sensitive film adjacent the container. Also included is an inlet and outlet for introducing and withdrawing hybridization fluid and means for automatically regulating introduction and removal of hybridization fluid.

In preferred embodiments the automated hybridization device includes inflatable means for contacting the container with the film; a lightproof box enveloping the container; and a plurality of containers, each separated by a label-opaque shelf.

Multiplex sequencing significantly increases the speed at which DNA sequencing can be performed. Specifically, it reduces the experimental time for preparation of the DNA for sequencing, for the base specific chemical reactions (when using the Maxam and Gilbert methodology) and for the Sanger

dideoxy reactions, for the pouring and running of sequencing gels, and for the reading the image and sequence data from autoradiographs of these gels.

The above vectors are specifically designed for multiplex sequencing having unique DNA sequences positioned appropriately such that each sequence provides a specific probe region for DNA inserted into the vector, and indeed for each strand of the inserted DNA. Thus, any DNA inserted into these vectors is readily sequenced, as a pool of such inserts.

Other features and advantages of the invention will be apparent from the following description of the preferred embodiments and from the claims.

Description of the Preferred Embodiments

The drawings wil first briefly be described.

Drawings

Figure 1 is a diagrammatic representation of a multiplex vector.

Figure 2 is the nucleotide sequence of part of a multiplex vector, including the tag sequences.

Figure 3 is the nucleotide sequences of a set of tag sequences lacking G residues.

Figure 4 is a schematic diagram representing the steps in the method of multiplex sequencing.

Figure 5 is a diagrammatic representation showing data transformations from raw data to ideal data.

Figure 6 is an isometric view of a container for automated probing of membranes.

Figure 7 is a diagrammatic representation of components of an automated probing device.

Multiplex Vectors

Multiplex vectors are used to form a set of vectors suitable for multiplex sequencing. The vectors are provided with a DNA sequence having a) a cloning site, b) at least one tag sequence and c) optional removal sites. The cloning site is a position at which the DNA specimen to be sequenced is placed, usually it is a restriction endonuclease site. The vectors are provided with at least one, and preferably two, tag sequences, or probe hybridization sites. A tag sequence is an unique DNA sequence of between about 15-200 base pairs, preferably greater than 19 base pairs in length, which will act to specifically identify each vector in hybridization tests. These sequences preferably have no homology elsewhere on the vector and no homology to the genomic DNA to be sequenced. Thus, positive hybridization using a part or the whole of a tag sequence as a probe for a mixture of multiplex vectors specifically identifies the multiplex vector containing this tag sequence. Further, since it is preferable that no two tag sequences are the same on a single vector, and since each stand of any one tag sequence is unique, a specific strand of the

vector can also be specifically identified. Removal sites are useful when using the vectors for Maxam and Gilbert type sequencing. Generally, these sites are restriction endonuclease sites, preferably acted upon by the same restriction enzymes, and allow removal of a fragment of DNA, including the tag sequences and the DNA specimen to be sequenced.

The tag sequences are positioned close (preferably within 50 bases) to the cloning site, such that the DNA specimen to be sequenced can be placed between the two tag sequences, or at least downstream from one tag sequence. Further, the tag sequences are preferably positioned between two identical restriction endonuclease sites (removal sites) so that the cloned DNA and tag sequences can be readily isolated as a single fragment by digestion with the specific endonuclease.

Those skilled in the art will recognize that vectors having a single tag sequence may be used in this invention, or even vectors having more than two tag sequences. Such multi-tag vectors will preferably have multiple cloning sites for accepting different DNA fragments to be sequenced.

Example 1: TcEN vectors

An example of such a set of vectors is shown in Fig. 1. In this set of vectors, termed TcEN vectors, each vector has its own pair of unique tag sequences (labelled PHS) separated by Smal site (the cloning site) and surrounded by Notl sites (removal sites). The unique tag sequences are generated by synthesizing oligonucleotides approximately 20 nucleotides in length in a DNA synthesizer, with all 4 nucleotides provided at each polymerization step. All 4^{20} possible 20-mer nucleotide sequences are produced by this procedure. The oligomer mixture was ligated into a parent TcEN plasmid to form constructs containing the sequence (Notl restriction site) - (oligomer tag sequence 1) - (Smal site) - (oligomer tag sequence 2) - (Notl site) using standard procedures. The parent plasmid and each TcEN plasmid contain a tetracycline resistance gene which may be used for selection of clones containing these vectors. Also provided is an origin or replication (ori) which functions in Esherichia coli.

Many recombinant vectors from the original mass ligation were cloned and sequenced, and forty-six different vectors (containing 92 unique oligomer tag sequences) were chosen for initial experiments on the basis of their having the complete tag sequence-containing construct and sufficient sequence diversity to insure uniqueness. Probe oligomers (tag probes) complementary to each of the tag sequences have been prepared synthetically and radioisotopically labelled. Other labels are equally suitable; e.g., fluorescent, luminescent and enzymatic (colorimetric) labelled probes.

Example 2: NoC vectors

A major problem with both dideoxy and chemical sequencing is the formation of hairpins during electrophoretic separation of the sequence reaction products. The hairpins often cause two or more base positions to comigrate on the gels causing an artifact which can be hard to notice or decipher. One partial solution, sequencing of the opposite nucleotide strand, can fail if there are tandem or alternative hairpins, or if the opposite strand has any other artifacts. Other methods, including the use of hot formamide gels, dITP, or dc[7] GTP.

Cytosines can be chemically modified so that they will not base pair with guanines. Although cytosine chemical modification eliminates hairpins, (since at least two GC base pairs are required to stabilize a hairpin in sequencing gels containing 7M urea at 50° C). They also prevent the GC base pair formation required for most tag probes to be able to bind to identify specific vectors later in the sequencing process. This problem is overcome by preparing a set of NoC vectors having tag sequences which completely lack cytosines on one strand (any guanines on the complementary strand) for at least 15 nucleotides in a row (Figs. 2 & 3). These vectors generally have two such oligonucleotide tag sequences flanking each cloning site.

Referring to Fig. 2, the NoC vectors were derived from the TcEN vectors by substituting the sequence shown at the EcoRI sites. H26 represents various combinations of 26 A, C, and T nucleotides present on the strand at this point. D25 represents 25 A, G, and T's. These were chemically synthesized as a mixture (as described above in example 1) and specific examples were cloned and sequenced (see Fig. 3). The object is to have No Cytosines on the appropriate strand in the tag sequence which can be modified by chemical hairpin suppression reactions.

Referring to Fig. 3, the sequences of the tag sequence of the NoC multiplex vectors of Fig. 2 are given. The first twenty nucleotides have also been synthesized as probes. The sequences are all oriented 5′ to 3′, left to right. The 5′ terminal nucleotides shown are the 3′-most cytosines of the Notl sites. The letter P indicates that the tag sequence is closest to the Pstl site; the letter E that the tag sequence is closest to the EcoRI site. The numbers are the vector and tag sequence (or tag probe) numbers; number 00 is the standard plasmid used to provide an internal control of known sequence. The Notl site adjacent to the Pstl site has one C deleted so that Notl cleavage of this vector and probing with tag sequence E00 gives a 2500 base long set of sequence markers (3′ end-labeled as with the unknown sequences).

These vectors also incorporate bacterial transcription terminator sequences flanking the tag sequences, to suppress excessive transcription across the boundary between foreign DNA and the vector DNA and thus promote efficient replication.

Multiplex sequencing method

Multiplex sequencing is a method for keeping a large set of DNA fragments as a precise mixture throughout most of the steps of DNA sequencing. Each mixture can be handled with the same effort as a single sample in previous methods, and so a greater total number of fragments can be handled

within a fixed time period. The mixture must be deciphered at the end of the process. This is done by tagging the fragments at the beginning of the process with unique DNA sequences (tags) and then, at the end of the process, hybridizing complementary nucleic acid (probes) to the sequencing reactions which have been spread out by size, and immobilized on large membranes. Many different sequences are obtained from each pool by hybridizing with a succession of end-specific, strand-specific probes. Thus, in this procedure DNA purifications, base specific reactions and gel loadings normally done on individual fragments are replaced by operations on pools of fragments (e.g., about 46 fragments per pool). In general, the steps in the process are cloning of genomic DNA into a set of multiplex vectors, mixing one clone derived from each vector to form a pool of vectors, performing sequencing reactions on each pool of such vectors, running these reactions on a sequencing gel, binding the DNA fragments in the gel to a solid support, and probing this support with tag probes specific for each vector to identify the DNA sequence of any one cloned fragment.

This methodology can be used to sequence DNA in combination with any standard sequencing procedure, for example, nested deletions (Henikoff, 28 Gene 351, 1984), cDNA (Okayama et al. 2 Mol. Cell. Biol. 161, 1981), shotgun (Sanger et al. 162 J. Mol. Biol. 729, 1982) or DNA hybridization techniques; and can be combined with amplification methodology, for example, the polymerase chain reaction system of Saiki et al. (230 Science 1350, 1985.)

Example 3: Sequencing Escherichia coli genomic DNA

In this example, the steps of the multiplex sequencing process are described below, with reference to Fig. 4.

Step 1

Genomic DNA (or equivalent large DNA fragments) of Escherichia coli was sonically fragmented and electrophoretically separated out for further processing. The fraction containing fragments of 800-1200 bp in length was isolated, since this size is optimal for sequencing. This treatment assures that virtually all possible DNA sequences from the original DNA are represented in a fraction of convenient molecular size for further processing.

Step 2.

The sonically fragmented and sized fraction was treated with the enzymes Bal31, and T4 polymerase, to produce blunt-ended fragments. Aliquots of the blunt-ended fragment mixture were then ligated separately into each of the set of 46 TcEN multiplex vectors and cloned to produce 46 different libraries. Each library contains clones of a large representative sample from the original sonic fraction. All fragments in any library contain the tags derived from the cloning vector used to produce the library,

one at each end of each strand.

Each library was then spread or "plated" onto an agar plate under limiting dilution conditions (conditions such that individual bacteria will be isolated from each other on the plate, and will form separate clonal colonies). The plates contained tetracycline so that only cells which have acquired a TcEN plasmid, with its tetracycline-resistance gene will survive. There were 46 such plates, and about 100 colonies were established per plate. After a period to allow the individual bacteria to grow into colonies, one clone (i.e., one colony, which will contain offspring of a single bacterium) from each of the 46 plates was then mixed together to form a pooled sample. This pooled sample contained a single clone, containing a single approx. 1 kb fragment, from each of the 46 libraries --that is, 46 different 1 kb fragments of the original DNA sample. Each fragment is labeled at each end with the tag oligomers unique to the library from which it was produced. 96 such pooled samples were made.

Step 4.

Each of the pooled samples were treated via the Maxam-Gilbert nucleotide sequencing procedure to produce a large number of DNA fragments of differing length. Each fragment will contain one of the tag sequences at one end. Some of these chemical sequencing reactions were performed in standard 96 well microtiter plates. The usual alcohol precipitation steps were eliminated. This was done by simply adding 10µl of dilute dimethyl sulfate (1 mM), acetic acid (15 mM) or 0.1 mM potassium permanganate (for $G+C$, $A+G$, and T reactions, respectively) to 5µl of DNA followed by 50µl of 1.3M piperidine. The reactions, lyophilization, resuspension, and loading were all done in the 96-well plates using multi-tipped pipets special designed for the 8x12 well format.

The pooled, reacted samples were then electrophoretically separated on sequencing gels by conventional means. Typically, the four sequencing reaction mixtures of twenty-four different pool samples were applied in separate lanes on each gel. The separated nucleotide patterns were then transferred to a nylon membrane, and fixed to the membrane (other solid media are equally suitable), as described by Church & Gilbert, supra.

Step 5.

Each nylon membrane, with each lane containing overlapping sequencing-fragments of 46 different 1 kb fragments, was then hybridized with one of the 92 probe oligonucleotides, each one of which is complementary of one of the tag sequences. Unreacted tag probe was then washed off with a 7-fold dilution of the hybridization buffer at 23°C, and the autoradiographic pattern of the treated gel recorded. Only those sequencing fragments containing the complementary tag sequence will hybridize with any given probe oligonucleotide, and thus be detectable by autoradiography. Hybridized probe was then removed, and the procedure

repeated with another tag probe oligonucleotide. The nucleotide pattern fixed on the nylon membrane can be repeatedly hybridized in 7% SDS 770 mM sodium phosphate, pH 7.2 at 42°C, with a tag probe oligonucleotide, excess probe washed off, auto-radiographed, then washed free of probe (by melting the 20 base pair DNA duplexes at 65°C in low salt (2mM tris-EDTA-SDS)), and reacted with another tag probe. Over 45 successive probings have been achieved. There is no reason to expect that this is the limitation on the number of successive probings possible on one membrane. Each tag probe oligonucleotide will detect only those sequencing fragments derived from whichever of the 46 original 1 kb fragments was cloned with the complementary tag probe oligonucleotide.

The hybridizations, washes and exposures were performed in an automated probing device (see below) with the membrane sealed in a container made of thin plastic (Scotchpak® #229, about 60 μm thick). The solutions were introduced through a tube to the inside and the X-ray film pressed tightly against the plastic layer by applying pressure (e.g., 3mm Hg air pressure above atmospheric pressure) to the film via a 4 kg flat weight on a constrained inflatable element, such as a plastic bag.

The membranes were marked with vector DNA used in a manner like ink on paper to automatically identify the membrane and the probe used by the hybridization properties of the DNA markings. Markings to identify the probe used consisted of one vector per dot or line corresponding to each probe, or any appropriate mixture of vectors or fragments thereof. Markings to identify the membrane were a mixture of all vectors, this will hybridize with all the probes used in sequencing procedures.

Internal standards were also provided in at least one lane in the gel. These known internal standards gave sequencing patterns which are useful for interpretation of unknown sequences. Specifically, during digital image processing of the film sequence data the prior knowledge of lane and band positions and shapes for the sequencing film can speed processing for all subsequent films prepared from the same membrane. Prior knowledge of reaction chemistry deviations and local gel artifacts seen in a standard sequencing reaction also helps in accurate estimation of errors for all subsequent films. For example, internal standards can be applied to each lane and discerned prior to probing with oligonucleotides. One such standard is the product of a sequencing reaction on a vector having no insert, thus providing a known DNA sequence. Such known DNA sequences provide ideal internal standards since the band shapes, lane shapes, and reaction chemistry visible from one probing are congruent with those in all subsequent probings. This aids the data reduction steps and the quantative recognition of trouble spots. This technique can also be used for mapping of restriction enzyme sites in DNA molecules.

Interpretation of the X-ray films generated by the above-process includes six steps to transform the raw data to an idealized form. The resulting data is interpreted to provide a DNA sequence and esti-

mates of the liklihood of alternative interpretations made. The steps are: a) to find the boundaries of the lanes and straighten them if necessary; b) to adjust the interband distances; c) to straighten the bands across each lane; d) to adjust the interlane displacements; e) to adjust the band thicknesses; and f) to adjust the reaction specificities (i.e., to ensure that each band in each of the lanes in a set of reactions is a true band). Some of these steps are discussed by Elder et al. 14 Nuc. Acid. Res. 417, 1986.

Referring to Figure 5, a schematic example of band interpretation is provided. DNA sequencing reactions from an internal standard (left panel) and an unknown DNA sample (second panel from left) were placed in the same lanes of a gel and probed with appropriate probes. Figure 5 is a schematic of the resulting X-ray films. As can be seen from the left two panels both the standard and the unknown deviate from an ideal result (shown in the right panel) in an identical manner. The DNA sequence of the standard is provided on the left of Figure 5. Using this sequence it is possible to determine which bands on the film represent true bands, rather than artifacts. From this analysis the DNA sequence of the unknown can then also be determined by making allowance for artifacts detected in the standard sample. Thus, referring to Figure 5, the transformation steps described above can be applied. In the first step lane curvature is straightened in all four lanes; in the second step interband distances are made equal; in the third step band curvature is straightened; in the fourth step severe interlane displacement in the C lane relative to the other three lanes is adjusted; in the fifth step thick bands in the R lane are made thinner; and in the sixth step false bands are removed. After these steps the ideal result is determined and thence the DNA sequence of the unknown (shown on the right of Figure 5).

Each adjustment that involves manual or automatic pattern recognition on the entire film data set can take about an hour. By using the above multiplexing sequencing method with an internal standard of known sequence each of the adjustments can be more accurately determined than for an unknown sequence, and then applied to each set of unknown film data obtained from subsequent probings of the same membrane. Thus, one standard pattern can be used for a large number of later probings of the same membrane. The transformation requirements can be memorized by the computer and recalled as necessary.

Automated Probing Device

An automated probing device is useful for automation of the visualization of latent multiplexed DNA sequences immobilized on nylon membranes, or for any other process where multiple cycles of probing are necessary.

In general, after the DNA to be probed has been transferred and crosslinked to a membrane support, eg., a nylon membrane, the nylon membrane is heat-sealed into a polyester/polyethylene laminate bag such as Scotchpak®, the automated probing device performs the following steps: a) 100ml of

prehybridization buffer (7% SDS, 10% PEG (MW 6000), 0.13M sodium phosphate pH 7.2, and 0.25 M NaCl) is introduced into the bags and the membranes equilibrated; b) this buffer is removed and 90ml of hybridization buffer with probe (5pM radiolabeled oligonucleotide in prehybridization buffer) is introduced; followed by a 10ml chase of prehybridization buffer; c) the membranes are then incubated 4 to 16 hours at 42°C; d) wash buffer (1% SDS, 70 mM sodium phosphate pH 7.2) is introduced, incubated at room temperature for 45 minutes and removed; the wash is repeated a minimum of 5 times; e) after the last wash, X-ray films are placed over the membranes and exposed. To ensure good contact between the film and membrane an air bag is inflated directly over the film; f) after exposure the air bag is deflated by use of a vacuum and the films processed; g) the probe is removed with hot stripping buffer (2 mM EDTA, TRIS base pH 7.5, 0.2% SDS at 80°C) and the cycle repeated with a new probe.

Referring to Figs. 6 and 7, automated prober 10 is formed of a light-tight aluminum housing 11, having a series of grooves 20 for slideably accepting a series of shelves 18. Shelves 18 are spaced apart about 0.4" to allow a re-usable pouch 12 formed from plastic of thickness 0.65 micron, an X-ray film (not shown) and an air bag 16 to be sandwiched between each shelf 18. Housing 11 has a height A of 10.75", a width B of 20" and a depth C of 24". Each pouch 12 is provided with an inlet pipe 22 and an outlet pipe 24, and the housing is heated by heater 26. A door (not shown) of thickness 0.25" is provided with a 0.125" neoprene light-tight gasket seal and a latch. Automated prober 10 is assembled and manufactured by standard procedure. Light-tight construction of automated prober 10 allows exposures to be performed outside of a photographic darkroom; and aluminum housing 11 provides efficient attenuation of beta particles from ^{32}P decay, with minimal secondary X-ray formation.

Re-usable pouches 12 allow tight contact with the X-ray film to be exposed, without need for pouch disassembly. Membranes 14 are fixed (by heat tacking of corners) inside each pouch 12 to prevent movement and creasing of membranes 14. In addition to membranes for multiplex sequencing the automated probing device is capable of accommodating membranes from standard restriction digest transfers, and dot and slot blots used in diagnostics. The horizontal format allows uniform spatial distribution of a probe, without need for mesh supports. The rigid bookshelf type structure allows air bags 16 to expand between shelves 18 to ensure tight contact between each X-ray film and an adjacent membrane 14. Each shelf 18 easily slides out of grooves 20 for membrane replacement.

A computer (not shown) controls temperature, liquid, air, and vacuum flow. Only the X-ray films and vessels containing liquid input and output have to be manually changed.

Also provided is a heated, ventilated output waste vessel 30 which allows 100-fold concentration of liquid wastes, thereby reducing disposal costs.

Automated prober 10 is constructed as repeating shelf units 42 of ten membranes, each repeat of ten membranes having dedicated reservoirs and valves for service buffers including hybridization buffer, probe, wash solution and stripping buffer. Each individual membrane has its own input valve. The waste output valves are placed in series with a 12VDC-controlled Gorman-Rupp reciprocating pump; the buffer input valves are also placed in series via a 12VDC-controlled Gorman-Rupp reciprocating pump. Air input is used to inflate air bags for X-ray film exposures. These air bags may be covered by conductive foam to enhance membrane contact. A house vacuum is used to evacuate the airbags after exposure. During probing the temperature is maintained at 42°C by heating mats 26 placed inside the top and bottom of the apparatus.

Computer control is by TRS-80 Model 102 which controls the valves via a CIP/35 serial controller (SIAS Engineering). A BASIC program controls all buffer inputs and output by timing loops; thus it is important to calibrate the flow rates before setting parameters. Parameters are set by creating a text file that can be opened and read by the BASIC program. The input file has lines that have a number as the first element followed by tab before any descriptors are used. The line order is:

1. - membrane number (1-10)
2. - repeat units (1-13)
3. - minimum membrane number
4. - maximum membrane number
5. - prehyb input time (min)
6. - prehyb incubate time (min)
7. - prehyb output time (min)
8. - probe input time (min)
9. - probe chase time (min)
10. - probe incubation time (hours)
11. - probe output time (min)
12. - wash input time (min)
13. - wash incubation time (min)
14. - wash output time (min)
15. - number of washes (integer)
16. - strip input time (min)
17. - strip incubate time (min)
18. - strip output time (min)

Other embodiments are within the following claims.

Claims

1. A set comprising at least two vectors, each said vector of said set characterised by having a DNA construct having at least one restriction endonuclease site and having at least one tag sequence comprising at least 15 base pairs, each tag sequence being located within 50 bases of one said site, each said tag sequence being incapable of hybridizing with another of said tag sequences under stringent hybridization conditions,
each said vector of said set differing from each other said vector of said set only at said tag sequence.

2. A vector for sequencing genomic DNA of an organism, said vector characterised by a DNA construct having at least two restriction endonuclease sites, each said site being recognised by a first restriction endonuclease, wherein treatment of said vector with said first endonuclease produces a discrete DNA fragment consisting of the DNA extending between said sites comprising two tag sequences separated from each other by a second restriction endonuclease site, wherein said tag sequences each comprise at least 15 bases and each end of said tag sequences is located within 50 bases of the nearest said first and second restriction site, said tag sequences being unique in said vector.

3. A set as claimed in claim 1, wherein said tag sequences have no cytosine residues in one strand.

3. A vector as claimed in claim 2, wherein said tag sequences have no cytosine residues in one strand.

5. A set as claimed in claim 1 wherein said vectors are produced from a parental vector having randomly formed tag sequences ligated into said parental vector.

6. A set as claimed in claim 5 wherein said randomly formed sequences are formed in a DNA synthesizer supplied with at least two nucleotides at each addition step.

7. A set of tag oligonucleotides, characterised in that each said tag oligonucleotide of said set is unique in said set and neither said tag oligonucleotide nor oligonucleotide homologous to said tag oligonucleotide will hybridize under stringent hybridization conditions to other said tag oligonucleotides in said set, wherein each tag oligonucleotide is at least 15 bases in length.

8. A set of vectors produced by ligation into a parental vector of the set of tag oligonucleotides as claimed in claim 7.

9. A set as claimed in any one of claims 1, 3, 5, 6 or 7 wherein said hybridization conditions comprise conditions in which mismatched duplexes are melted to form single stranded molecules.

10. A set as claimed in claim 9 wherein said hybridization conditions comprise heating at 42° C in 500-1000 mM sodium phosphate buffer.

11. A method for sequencing a DNA specimen, said method characterised by the steps of:

a) ligating said DNA specimen into a vector comprising a tag sequence, said tag sequence comprising at least 15 bases, wherein said tag sequence will not hybridize to said DNA specimen under stringent hybridization conditions and is unique in said vector, to form a hybrid vector,

b) treating separate aliquots of said hybrid vector in a plurality of vessels to produce fragments comprising said tag sequence, wherein said fragments in each vessel differ in length from each other and all terminate at a fixed known base or bases, wherein said fixed known base or bases differ from those in other said vessels,

c) separating said fragments from each said vessel according to their size,

d) hybridizing said fragments with an oligonucleotide able to hybridize specifically with said tag sequence, and

e) detecting the pattern of hybridization of said tag sequence, wherein said pattern reflects the nucleotide sequence of said DNA specimen.

12. A method as claimed in claim 11 further comprising ligating said DNA to a plurality of said vectors to form a plurality of said hybrid vectors.

13. A method as claimed in claim 12 wherein each said vector differs at said tag sequence and each said tag sequence is unable to hybridize under stringent hybridization conditions to other said tag sequences.

14. A method as claimed in claim 11 further comprising rehybridizing said fragments with each said tag sequence of each said vector.

15. A method as claimed in claim 11, wherein each said vector comprises two said tag sequences.

16. A method as claimed in claim 11 further comprising binding said fragments to a solid support prior to said hybridizing.

17. A method as claimed in claim 11 or claim 13 wherein said stringent hybridization conditions comprise conditions in which mismatched duplexes are melted to form single stranded molecules.

18. A method for repeatedly hybridizing a solid support, comprising nucleic acid, with a plurality of labels, characterised by the steps of:

a) enclosing said support within a container comprising material thin enough to allow detection of hybridization of said nucleic acid with said labels,

b) inserting hybridization fluid comprising a first said label into said container,

c) removing said hybridization fluid from said container, and

d) detecting hybridization of said first label to said solid support, while said solid support remains in said container.

19. A method as claimed in claim 18, further comprising after said detecting step, repeating steps b,c and d using a second said label.

20. A method as claimed in claim 18, wherein steps b,c and d are repeated at least 20 times with at least 20 said labels.

21. A method as claimed in claim 11, further comprising prior to said separating step, the steps of providing a molecular weight marker in a said vessel and detecting said markers after said separating step.

22. A method as claimed in claim 16 wherein said solid support comprises said vector or said oligonucleotide, wherein said vector or oligonucleotide act as an identifying marker for said support.

23. A method for determining the DNA sequence of a DNA specimen characterised by the steps of:

    a) performing a first sequencing reaction on said DNA specimen;

    b) performing a second sequencing reaction on a known DNA specimen;

    c) placing corresponding products of said first and second sequencing reactions in the same lanes of a DNA sequencing gel;

    d) running the products into said gel;

    e) detecting the location of said products in said gel, and

    f) using the location of said known DNA products to aid calculation of the DNA sequence of said DNA specimen.

24. An automated hybridization device, for repeatedly hybridizing a solid support comprising nucleic acid, with a plurality of labels, characterised by:
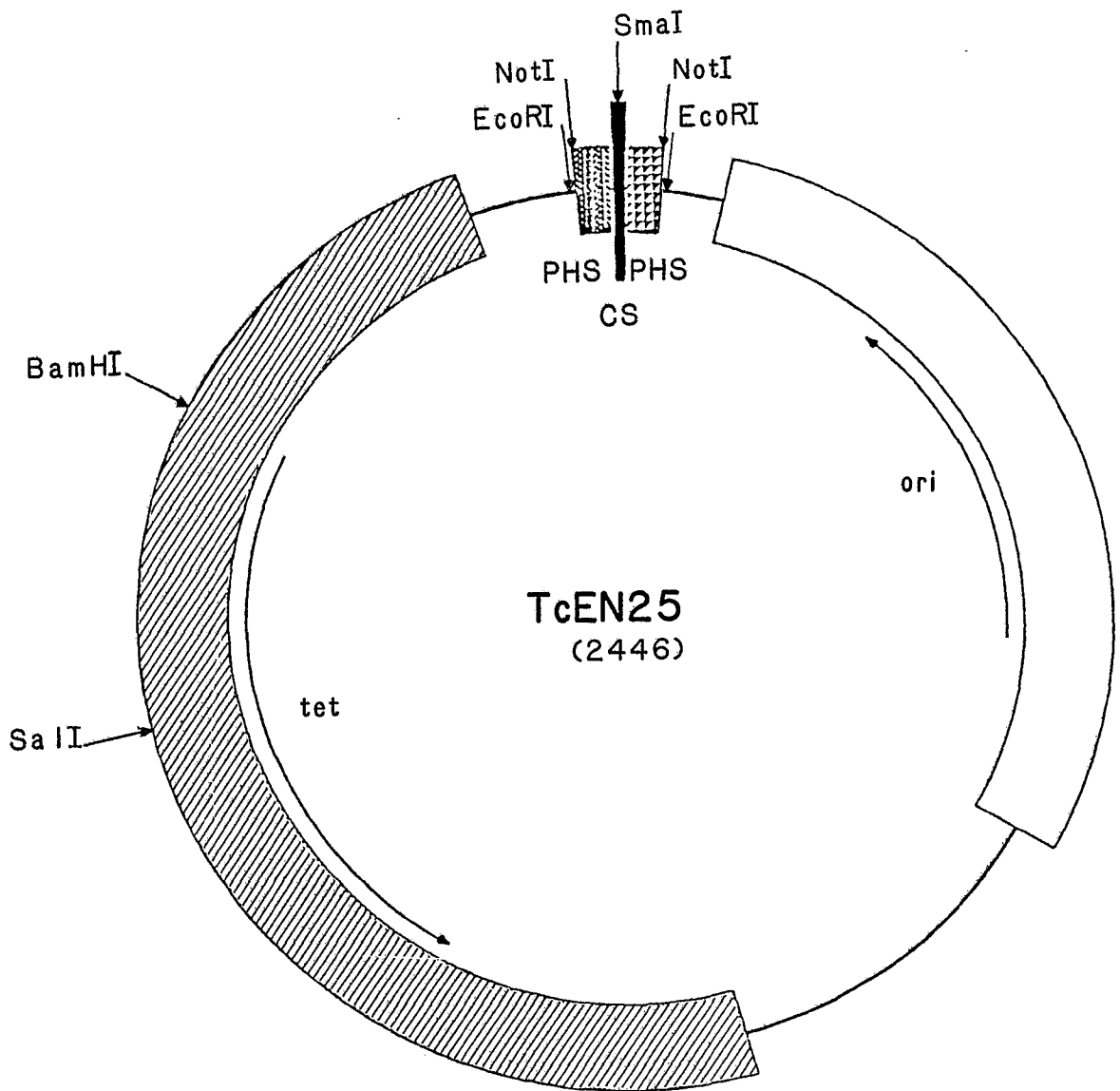
a container enveloping said solid support, comprising material thin enough to allow detection of hybridization of said nucleic acid with a said label by placement of a label-sensitive film adjacent said container,

said container including an inlet and outlet for introducing and withdrawing hybridization fluid, and comprising means for automatically regulating introduction and removal of said hybridization fluid.

25. An automated hybridization device as claimed in claim 24 comprising inflatable means for contacting said container with said film.

26. An automated hybridization device as claimed in claim 24 or 25 comprising a lightproof box enveloping said container.

27. An automated hybridization device as claimed in claim 24 or 25 comprising a plurality of said containers, each separated by a label-opaque shelf.

5

10

15

20

25

30

35

40

45

50

55

60

65

FIG. 1

FIG. 2

```
                                                             T  C
                                                           T     G
                                                           C-G
                                                           G-C
                    T  C                                   C-G
                  G     T                                  C-G
                   C-G                                     C-G
                   G-C                                     A-T
                   T-A                                     A-T
                   C-G                                     A-T
                   G-C                                     A-T
                   C-G                                     A-T
                   C-G                                     A-T
ATTAAAAAA    CCGC-H26-GGTACCC-!-GGGAGCTC-D25-GCGGCCGCGAATTC    TT CTG
                              !
   ter    Pst  Not        Kpn Sma Sac         Not     Eco  rpoC-ter
```

FIG. 3

```
01  P   CCTTTCATTACAACTAATCA  TAACAT.
02  P   CATAAATTATTCTACTTATA  AAAAATA
03  P   CAACTCCTATCAACTCTACA  CTTACTC
04  P   CAATATTTAAACCTCACACT  TCCAATA
05  P   CATAATATAACCTAAACCCT  AAATCTT
06  P   CCACATCCAAAATAATCAAT  CAACATA
07  P   CTACTAAATTTCCTTTATAA  TCCCCAA
08  P   CCCCTCCAATATAATATATA  ATTACAT
09  P   CATTAACAATCATACCACTA  CCAAATA
10  P   CCTAATCATCAATATACTCA  ATACAAC
11  P   CATCACCACACATATTATCC  TTATCCT
12  P   CTAATCTATTAACCACTTAA  AATACTT
13  P   CAACTTACTCTACACCCCTT  TTATCAC
14  P   CTCTTCATAATATAATTTAC  TACTCTT
15  P   CAAACCAACATTTAACACAA  TATATCA
16  P   CATAAACACCCATTCATCCA  ACTCTTA
17  P   CTCACCTTCTTAAAACCCAA  ATTCTCT
18  P   CAAATCTACTTCCAACCACT  ATTCAAA
19  P   CAACCAACTCTACTTATCAC  TCCCAAC
20  P   CCCCACCAAATTACAACTAC  AATACTC
00  P   CCTTCCTAAACCACACTCCA  TTTAACC

01  E   CCCCCAATAAAATCATACTA  CCTTAT
02  E   CTTTTACACAATAACTCTTA  CTCAAA
03  E   CTAACAACAAACCTTACTAC  ATCCTA
04  E   CAACACCCATCCACTAAACT  TAAACA
05  E   CACATAACTCAAATCTCAAA  TTCACC
06  E   CACCCCATATCAATTTACAA  TAATCT
07  E   CCATAAACTCCTCATCTTCC  CCACTC
08  E   CCTCTTCCCATTTTTATTAT  TACTTC
09  E   CTATTCACAATACACCACCA  CAATCC
10  E   CACCTAATACCCTATATATA  ATACCA
11  E   CCACCTTTATACTTCTTACA  ACTCAT
12  E   CCCTTCTACCAACCTATATC  ATTTAT
13  E   CAAAAACTAATTCCCAAAAA  ATCCTC
14  E   CTTTCTTTTCTCAACCCCTT  CACCCA
15  E   CTCACTAAATCTTATCTTAC  TTTACA
16  E   CTTACAATCCCCTATCATAA  CATTAT
17  E   CCACCATCTTCACCCCCCAA  TTTAAT
18  E   CAACCACAACCAACCCTACT  CCCCTA
19  E   CCAACCCTACATTAACTTCT  ATCATC
20  E   CCATAACTTTAACCTTTAAC  ATTAAT
00  E   CTTCCTCTTTATCTTTATTA  ACATTT
```
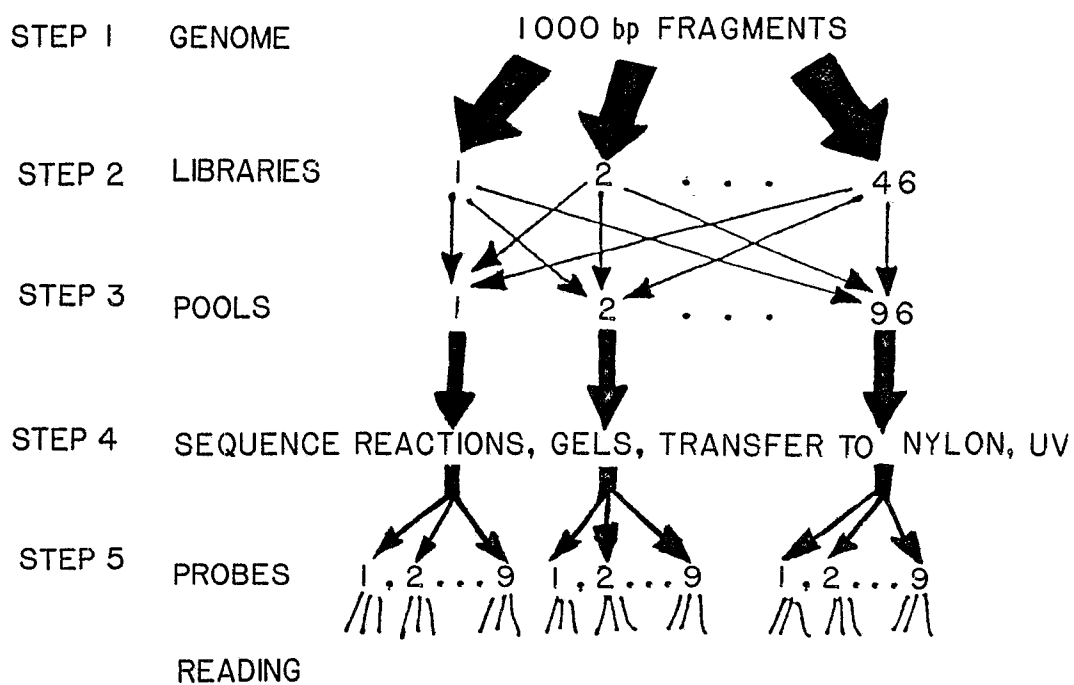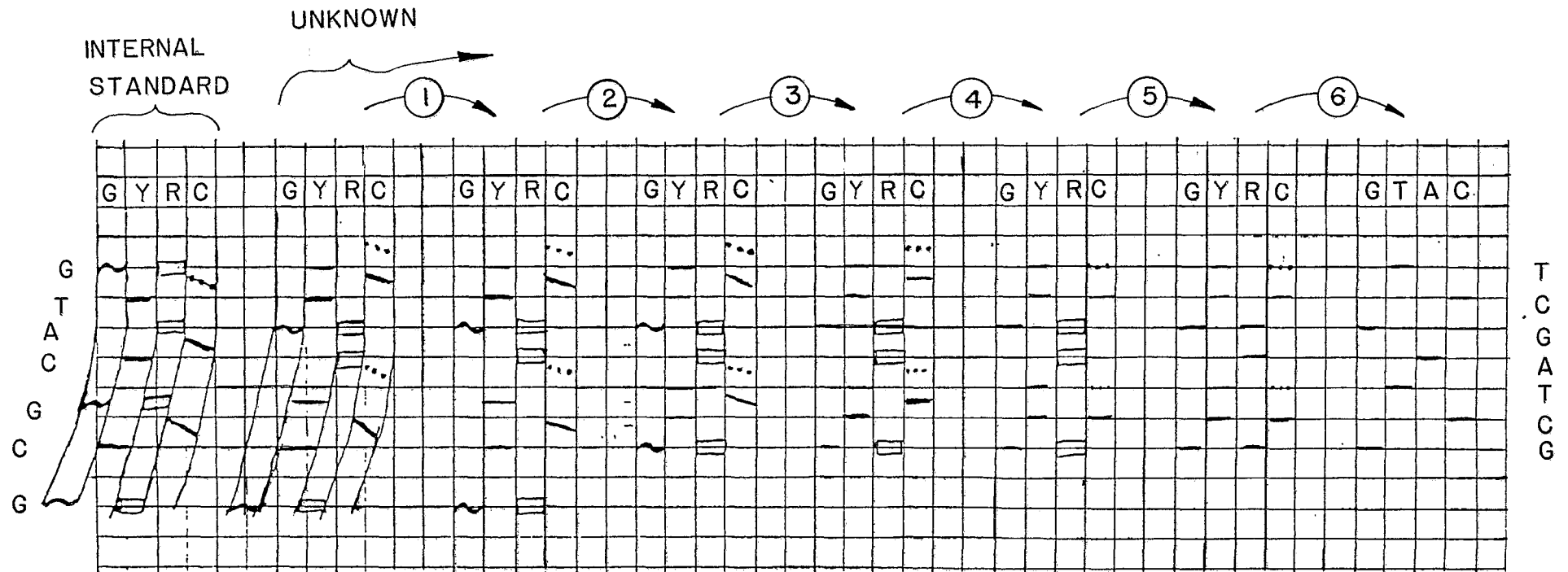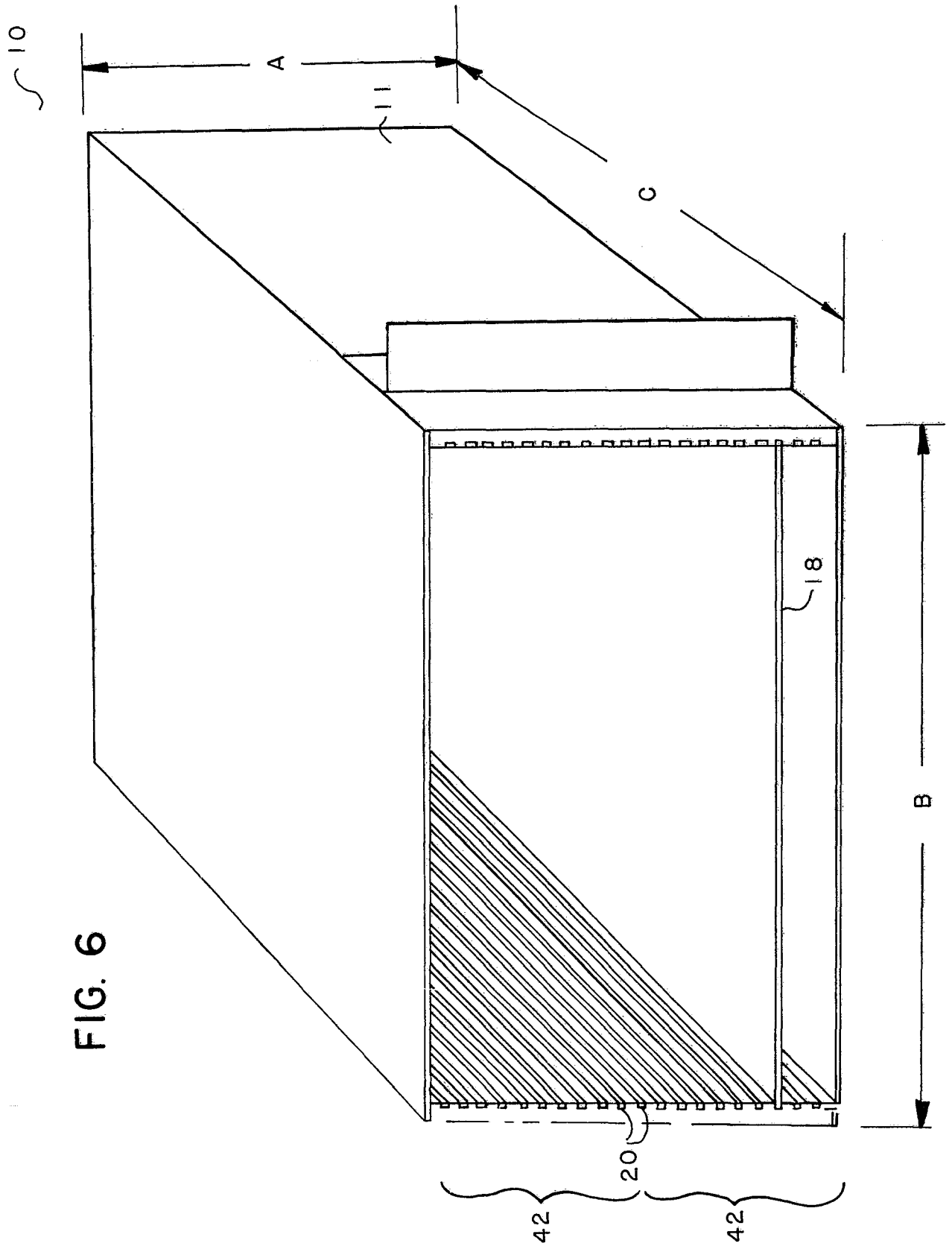
FIG. 4

STEP 1   GENOME   1000 bp FRAGMENTS

STEP 2   LIBRARIES

STEP 3   POOLS

STEP 4   SEQUENCE REACTIONS, GELS, TRANSFER TO NYLON, UV

STEP 5   PROBES   1,2...9 1,2...9 1,2...9

READING

FIG. 5

FIG. 6

FIG. 7

PREHYB

PROBE

WASH

STRIP

16

22

14

12

18

14 16

12

24

10

AIR

VACUUM

26

HEATER 26

WASTE 30

→⊗— PUMP

—⊙— SOLENIOD VALVE