(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 January 2001 (18.01.2001)

PCT

(10) International Publication Number
WO 01/04355 A1

(51) International Patent Classification[7]: C12Q 1/68, C12P 19/34, C07H 21/04

(21) International Application Number: PCT/US00/18210

(22) International Filing Date: 30 June 2000 (30.06.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/142,816 8 July 1999 (08.07.1999) US
Not furnished 27 June 2000 (27.06.2000) US

(71) Applicant and
(72) Inventor: LIU, Shaorong [—/US]; 303 Ridgeview Drive, Tracey, CA 95376 (US).

(74) Agent: NAKASHIMA, Richard, A.; Fulbright & Jaworski, L.L.P., Suite 2400, 600 Congress Avenue, Austin, TX 78701 (US).

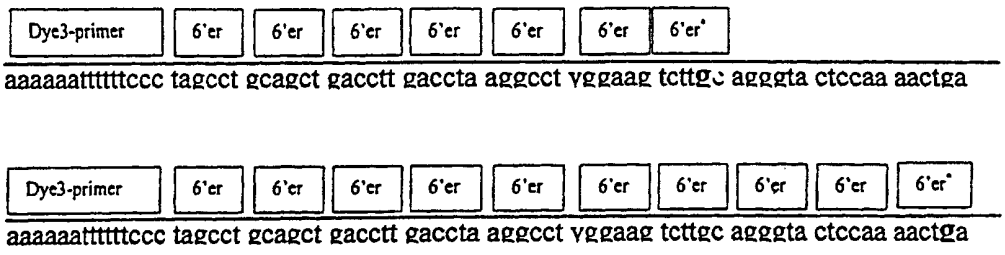
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *With international search report.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: A METHOD FOR SEQUENCING LONG-LENGTH DNA



(57) Abstract: The present invention concerns compositions and methods for DNA sequencing and SNP detection. The methods comprise use of ligation reactions, using one or more primers, sets of n-mer oligonucleotides and n-mer terminators. DNA sequence information is generated by analysis of the ligation products of primers, n-mers and n-mer terminators. Use of multiple primer reactions allows the generation of sequence information with higher throughput, while retaining the longer read-length that is obtained by avoiding the requirement for single-base resolution of DNA sequencing reaction products. Use of specially designed primer(s) and n-mer oligonucleotide(s) allows the generation of ligation reaction products corresponding to SNP(s). Identification of the ligation reaction products allows the determination of the SNP(s).

WO 01/04355 A1

DESCRIPTION

A METHOD FOR SEQUENCING LONG-LENGTH DNA

BACKGROUND OF THE INVENTION

The present patent application claims the benefit under 35 U.S.C. §119(e) of provisional U.S. Patent Application Serial No. 60/142,816, filed July 8, 1999 and U.S. Regular Application Serial No. --/---,---, filed June 27, 2000.

1. Field of the Invention

The present invention concerns the field of DNA analysis. More particularly, the present invention concerns novel compositions and methods for use in DNA sequencing.

2. Description of Related Art

Long read-length sequencing is very important for completion of the Human Genome Project, especially at its later stages when long repeat DNA segments are assembled. DNA repeats of greater than 1000 bases exist in human DNA. It is challenging to sequence these repeats without methods and compositions for long read-length sequencing. Typically, investigators sequence shorter, overlapping fragments of DNA and assemble these fragments in order by matching the identical sequences at the overlapping ends. It is difficult to assemble fragments consisting of repetitive DNA sequences, since there can be identical sequences present in many different fragments. For this purpose, a long read-length sequencing technology is required.

DNA sequencing chemistry was first developed by Sanger *et al.* (1977) and by Maxam and Gilbert (1977). Sanger's dideoxy chain termination method is the most widely used for high-volume sequencing, due to the development of automated fluorescence sequencing based on labeled primers or terminators (Smith *et al.*, 1986; Prober *et al.*, 1987; Tabor *et al.*, 1990; Ansorge *et al.*, 1987). The average read-length from presently available commercial sequencers using this technology is about 600-700 bases, although sequencing read-length over a thousand bases has been reported (Salas-Solano *et al.*, 1998).

The limitation of Sanger dideoxy sequencing to a read-length of less than about 1000 bases is related to the requirement for single-base resolution. That is, to generate DNA sequence data, it is necessary to separate DNA fragments that differ in size by a single nucleotide base. Viewed from the perspective of analytical separation, if it were possible to pre-separate the sequencing fragments

into, for example, two groups, with all fragments possessing an odd number of bases in one group and all fragments having an even number of bases in the other group, then double-base resolution would be satisfactory for sequence determination. With the decreased stringency requirements for double-base resolution, it would be possible to generate longer read-length sequence data.

Actual separation of sequencing fragments into such two groups is presently difficult, if not impossible. The present invention resolves this difficulty by providing novel compositions and methods for DNA sequence determination through ligation of oligomers.

Ligation of hexamers to generate sequential elongation primers was proposed for DNA sequencing by Szybalski (1990). Elongation through ligation works well (Kaczorowski and Szybalski, 1994; Kaczorowski and Szybalski, 1998). In Szybalski's method, study conditions need to be adjusted such that false priming caused by single hexamers does not occur. Using the methods described herein, the annealing conditions can be more vigorous to avoid such false priming but still allow elongation through ligation. The chemistry and separation protocols of the present application also differ from those previously reported.

SUMMARY OF THE INVENTION

The present invention provides compositions and methods of use for long read-length DNA sequencing. In preferred embodiments, DNA sequencing is performed by a method comprising the steps of obtaining a nucleic acid to be sequenced, adding a 5' primer to which one of four distinguishable fluorescent dyes has been attached, adding a set of n-mers that bind by hybridization to the nucleic acid to be sequenced, adding an n-mer terminator, ligating the primer, n-mers and n-mer terminator to form DNA sequencing products, and separating the DNA sequencing products by size to provide DNA sequence information. As used herein, the terms "DNA sequencing products," "ligation reaction products," and "ligation products" are used synonymously to mean the products of ligation of primer, one or more n-mers and an n-mer terminator. An "n-mer" is an oligonucleotide that is "n" bases long, where "n" is an integer between 2 and 10, more preferably between 3 and 8, even more preferably 5 or 6. An "n-mer terminator" is an oligonucleotide that is "n" bases long, wherein the 3' end of the molecule is modified to prevent further ligation or chain elongation at the 3' end, for example, a dideoxynucleotide that is missing a 3'-OH group.

In preferred embodiments, sets of n-mer terminators are designed, so that each member of the set contains a single identified nucleotide at one of the n-positions of the n-mer, while the remaining n-1 positions of the n-mer contain a random mixture of all four nucleotides (A, G, C

and T). Each set as a whole contains terminators with the same identified nucleotide at all of the n-positions of the n-mers. For example, for the set of n-mer terminators where n = 6 and the residue of interest is an A, the set would comprise 6-mer terminators of sequences: 5'-dN-dN-dN-dN-dN-ddA-3'; 5'-dN-dN-dN-dN-dA-ddN-3'; 5'-dN-dN-dN-dA-dN-ddN-3'; 5'-dN-dN-dA-dN-dN-ddN-3'; 5'-dN-dA-dN-dN-dN-ddN-3'; and 5'-dA-dN-dN-dN-dN-ddN-3', where N indicates a mixture of A, T, G and C, dN indicates a deoxynucleotide and ddA indicates a dideoxyadenosine nucleotide. The remaining sets of n-mer terminators would be identical, except for the substitution of a C, G or T in place of the A residue.

In preferred embodiments, the sets of n-mer terminators would be used in the practice of the claimed methods to generate the complete DNA sequence of the target strand. For example, a first reaction would use the set of terminators: 5'-dN-dN-dN-dN-dN-ddA-3'; 5'-dN-dN-dN-dN-dN-ddT-3'; 5'-dN-dN-dN-dN-dN-ddG-3'; 5'-dN-dN-dN-dN-dN-ddC-3'. Separation of the ligation products would identify all nucleotides that are multiples of 6 bases away from the 3' end of the primer. A second reaction would use the set of terminators: 5'-dN-dN-dN-dN-dA-ddN-3'; 5'-dN-dN-dN-dN-dT-ddN-3'; 5'-dN-dN-dN-dN-dG-ddN-3'; and 5'-dN-dN-dN-dN-dC-ddN-3' to identify all nucleotides that are located [(multiples of 6 bases) – 1] away from the 3' end of the primer, and so on until all nucleotides in the nucleic acid to be sequenced have been identified.

In other preferred embodiments, two or more primers labeled with the same dye are used in one reaction. All primers start with the same sequence. The shortest primer is x (where x is an integer number, 1, 2, 3, 4, 5, and so on) nucleotides shorter than the next shortest primer which is x nucleotides shorter than the next one, and so on. For example, three primers (18-mer, 20-mer, and 22-mer), associated with a set of terminators: 5'-dN-dN-dN-dN-dN-ddA-3'; 5'-dN-dN-dN-dN-dN-ddT-3'; 5'-dN-dN-dN-dN-dN-ddG-3'; 5'-dN-dN-dN-dN-dN-ddC-3', may be used for a ligase sequencing reaction. Separation of the ligation products would identify all nucleotides that are multiples of 2 bases away from the 3' end of the 18-mer primer. A second reaction would use the set of terminators: 5'-dN-dN-dN-dN-dA-ddN-3'; 5'-dN-dN-dN-dN-dT-ddN-3'; 5'-dN-dN-dN-dN-dG-ddN-3'; and 5'-dN-dN-dN-dN-dC-ddN-3' to identify all nucleotides that are located [(multiples of 2 bases) – 1] away from the 3' end of the 18-mer primer. All nucleotides in the nucleic acid are identified using two reactions and two separations.

In additional embodiments, when n is other than 6, m (m is an integer number, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, and so on) primers labeled with the same dye are added to one reaction. All primers start with the same sequence. The shortest primer is x (x is an integer number, 1, 2, 3, 4,

5, and so on) nucleotides shorter than the next shortest primer which is x nucleotides shorter than the next one, and so on. The value of m equals to the number of the integer numbers that n can be evenly divided into. For example, when n = 3, n can be evenly divided into 3, or 1, 1 and 1; thus m = 1, or 3 and x = 3, or 1, respectively. When n = 4, n can be evenly divided into 4, or 2 and 2, or 1, 1, 1 and 1; thus, m = 1, or 2, or 4 and x = 4, or 2, or 1, respectively. When n = 5, n can be evenly divided into 5, or 1, 1, 1, 1 and 1; thus m = 1, or 5 and x = 5, or 1, respectively. When n = 7, n can be evenly divided into 7, or 1, 1, 1, 1, 1, 1 and 1; thus m = 1, or 7 and x =7, or 1, respectively. When n = 8, n can be evenly divided into 8, or 4 and 4, or 2, 2, 2 and 2, or 1, 1, 1, 1, 1, 1, 1 and 1; thus, m =1, 2, 4 or 8 and x = 8, 4, 2 or 1, respectively. When n = 9, n can be divided into 9, or 3, 3 and 3, or 1, 1, 1, 1, 1, 1, 1, 1 and 1; thus, m =1, or 3, or 9 and x = 9, or 3 or 1, respectively. When n = 10, n can be divided into 10, or 5 and 5, or 1, 1, 1, 1, 1, 1, 1, 1, 1, 1 and 1; thus, m =1, or 2, or 10 and x = 10, or 5, or 1, respectively. And so on.

In particular embodiments, separation of one reaction mixture generates sequence information of every x bases of a DNA or RNA and this information is useful to check the addition, deletion, and/or mutation of a DNA or RNA. Compared to sequencing separation data, results out of ligation mixture separation are very reliable since high resolution separations are achieved. If there is an addition, or a deletion, or a mutation at the every x base sites, this change will be apparent as a one base shift to the right, or one base shift to the left, or a mismatch in the separation, when the mutant sequence is compared to the wild-type sequence.

Although a "four-color" DNA sequencing method is preferred, in other embodiments the methods disclosed herein may be used with primers labeled with the same fluorescent or other detectable marker. In these embodiments, separate ligation reactions would be performed for each of the individual n-mer terminators, followed by separation of the ligation products and assembly of DNA sequence information.

In alternative embodiments of the present invention, the detectable label may be incorporated into other moieties besides the primer. For example, the n-mer terminators could be labeled with a fluorescent or other detectable marker.

In other preferred embodiments, the primer molecule is between 3 and 35, more preferably between 5 and 30, more preferably between 8 and 25, more preferably between 10 and 20, more preferably between 15 and 18 nucleotides in length.

In preferred embodiments, the nucleic acid to be sequenced may be purified from other cellular constituents, including other nucleic acids. In other embodiments, the nucleic acid to be sequenced may be incorporated into a vector such as a plasmid, virus, YAC, BAC or cosmid or

other vectors known in the art. More preferably, the vector is one that produces single-stranded nucleic acids for use as sequencing templates, for example the vector M13.

In other embodiments, the number of ligation reactions and separations may be reduced to increase the throughput of the system. In preferred embodiments, multiple primers are used for each reaction, for example, by using primers that will hybridize to the same DNA sequence, but whose 3' ends will be offset by a known number of bases.

In additional embodiments, the ligation method is used for single nucleotide polymorphorism (SNP) detection. In preferred embodiments, a primer is designed to hybridize with a sample DNA in which a known SNP is 1 to n bases away from the 3' end of the primer, more preferably n/2 (n/2 is rounded into an integer number if it is not an integer number) ± 3 bases away from the 3' end of the primer, more preferably n/2 ± 2 bases away from the 3' end of the primer, more preferably n/2 ± 1 bases away from the 3' end of the primer, and more preferably n/2 bases away from the 3'-end of the primer. A specially designed labeled n-mer oligonucleotide, complementary to a wild type (or a mutant) DNA sequence that is known to occur on the template strand downstream from the 3'-end of the primer, is allowed to hybridize with the sample DNA and ligated with the primer. If the sample DNA is a wild type (or a mutant) DNA, a ligated nucleic acid incorporating the labeled n-mer is produced. Identification of this ligation reaction product identifies the presence or absence of the SNP.

In particular embodiments, multiple SNPs can be detected if they are separated by less than n bases. In these embodiments, multiple n-mer oligonucleotides are designed, each complementary to the sequence of a particular SNP and each labeled with a distinguishable fluorescent dye or other detectable marker. Identification of the ligation products and their labels allows the identification of multiple SNPs, using a single ligation reaction.

In additional embodiments, the identification occurs by electrophoresis, such as slab gel electrophoresis, capillary gel electrophoresis, or gel electrophoresis on microfabricated chips. These methods are not meant to be exclusive and the skilled artisan will realize that any method known in the art for the detection and identification of labeled nucleic acids may be used within the scope of the present invention.

In certain embodiments, the identification method is based on the attachment of the primer to a solid-phase support. The primer is labeled with the fluorescent dye or the other detectable markers if ligation products are formed. Otherwise, the primer is unlabeled. Through monitoring to determine whether or not the primer is labeled with an n-mer, the presence of an SNP is identified.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1. DNA sequencing products that terminate in an A residue.

FIG. 2. DNA sequencing products that terminate in an C residue.

FIG. 3. DNA sequencing products that terminate in an G residue.

FIG. 4. DNA sequencing products that terminate in an T residue.

FIG. 5. DNA sequencing products that contain an A reside at the next to last base.

FIG. 6. DNA sequencing products that contain an C reside at the next to last base.

FIG. 7. DNA sequencing products that contain an G reside at the next to last base.

FIG. 8. DNA sequencing products that contain an T reside at the next to last base.

FIG. 9. DNA sequencing products that terminate in an A residue, using primer 1.

FIG. 10. DNA sequencing products that terminate in an A residue, using primer 2 that is 3 bases longer than primer 1.

DESCRIPTION OF ILLUSTRATIVE EMBODIMENTS

1. Compositions and Methods of Use of Nucleic Acids

As used herein, the term "nucleic acid" refers to a polymer of DNA, RNA or a derivative or mimic thereof, of two or more bases in length. The term "oligonucleotide" refers to a polymer of DNA, RNA or a derivative or mimic thereof, of between about 3 and about 100 bases in length. The term "polynucleotide" refers to a polymer of DNA, RNA or a derivative or mimic thereof, of greater than about 100 bases in length. Thus, it will be understood that the term "nucleic acid" encompass the terms "oligonucleotide" and "polynucleotide". These definitions generally refer to at least one single-stranded molecule, but in specific embodiments will also encompass at least one double-stranded molecule. Within the scope of the invention, it is contemplated that the terms "oligonucleotide", "polynucleotide" and "nucleic acid" will generally refer to at least one polymer comprising one or more of the naturally occurring monomers found in DNA (A, G, T, C) or RNA (A, G, U, C).

Nucleic acid sequences that are "complementary" are those that are capable of base-pairing according to the standard Watson-Crick complementary rules. As used herein, the term "complementary sequences" means nucleic acid sequences that are substantially complementary, as may be assessed by the same nucleotide comparison set forth above, or as defined as being capable

of annealing to the nucleic acid segment being described under relatively stringent conditions such as those described herein.

A nucleic acid may be purified on polyacrylamide gels, cesium chloride centrifugation gradients, or by any other means known to one of ordinary skill in the art (see for example, Sambrook *et al.* 1989, incorporated herein by reference).

Nucleic acid molecules having sequence regions consisting of contiguous nucleotide stretches of about 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, 55, 60, 70, 80, 90, 100, 125, 150, 175, 200, 250, 300, 350, 400 or more basepairs, complementary to the target DNA sequence, are particularly contemplated for use in embodiments of the instant invention.

Hybridization Conditions

Hybridization is understood to mean the forming of a double stranded molecule and/or a molecule with partial double stranded nature. Stringent conditions are those that allow hybridization between two homologous nucleic acid sequences, but precludes hybridization of random sequences. For example, hybridization at low temperature and/or high ionic strength is termed low stringency. Hybridization at high temperature and/or low ionic strength is termed high stringency. Low stringency is generally performed at 0.15 M to 0.9 M NaCl at a temperature range of 20°C to 50°C. High stringency is generally performed at 0.02 M to 0.15 M NaCl at a temperature range of 50°C to 70°C. It is understood that the temperature and/or ionic strength of a desired stringency are determined in part by the length of the particular probe, the length and/or base content of the target sequences, and/or to the presence of formamide, tetramethylammonium chloride and/or other solvents in the hybridization mixture. It is also understood that these ranges are mentioned by way of example only, and/or that the desired stringency for a particular hybridization reaction is often determined empirically by comparison to positive and/or negative controls.

Accordingly, the nucleic acids of the disclosure may be used for their ability to selectively form duplex molecules with complementary stretches of DNA and/or RNA. Depending on the application envisioned, it is preferred to employ varying conditions of hybridization to achieve varying degrees of selectivity of probe towards target sequence.

As used herein "stringent condition(s)" or "high stringency" are those conditions that allow hybridization between or within one or more nucleic acid strand(s) containing complementary sequence(s), but precludes hybridization of random sequences. Stringent conditions tolerate little, if any, mismatch between a nucleic acid and a target strand. Such

conditions are well known to those of ordinary skill in the art, and are preferred for applications requiring high selectivity. For applications requiring high selectivity, it is preferred to employ relatively stringent conditions to form the hybrids. For example, relatively low salt and/or high temperature conditions, such as provided by about 0.02 M to about 0.10 M NaCl at temperatures of about 50°C to about 70°C. Such high stringency conditions tolerate little, if any, mismatch between the probe and/or the template and/or target strand. It is generally appreciated that conditions may be rendered more stringent by the addition of increasing amounts of formamide.

In particular embodiments, a nucleic acid may be tagged or labeled with a substance that emits a detectable signal. Non-limiting examples of labels that could be used include fluorescent, luminescent, radioactive, phosphorescent, chemiluminescent or enzymatic. In the case of enzymatic labels, the enzymatic activity is used to catalyze a reaction that is detected, for example, by the formation of a colored or luminescent product. In other embodiments, the binding of label to nucleic acid may occur via the use of a binding pair, for example avidin-streptavidin. In such embodiments, one member of the pair would be covalently or otherwise attached to the nucleic acid and the other member of the pair attached to the label to be detected.

Immobilization of Nucleic Acids on Solid Supports

In certain embodiments, the nucleic acid to be sequenced may be immobilized onto a membrane, filter, magnetic bead or other solid support to facilitate separation of the products of a DNA sequencing reaction from primers, polymerases and other contaminants. The ability to directly attach nucleic acids to solid substrates is well known in the art. See U.S. Patent Nos. 5,837,832 and 5,837,860 both of which are expressly incorporated by reference. A variety of methods have been utilized to either permanently or removably attach nucleic acids to the substrate. Exemplary methods include: the immobilization of biotinylated nucleic acid molecules to avidin/streptavidin coated supports (Holmstrom, (1993)), the direct covalent attachment of short, 5'-phosphorylated nucleic acids to chemically modified polystyrene plates (Rasmussen, *et al.*, (1991)), or the precoating of the polystyrene or glass solid phases with poly-L-Lys or poly L-Lys, Phe, followed by the covalent attachment of either amino- or sulfhydryl-modified nucleic acids using bi-functional crosslinking reagents. (Running, *et al.*, (1990); Newton, *et al.* (1993)). When immobilized onto a substrate, the nucleic acids are stabilized and therefore may be used repeatedly.

In general terms, hybridization and ligation may be performed on an immobilized nucleic acid template attached to a solid surface such as nitrocellulose, nylon membrane or glass. Numerous other matrix materials may be used, including reinforced nitrocellulose membrane,

activated quartz, activated glass, polyvinylidene difluoride (PVDF) membrane, polystyrene substrates, polyacrylamide-based substrate, other polymers such as poly(vinyl chloride), poly(methyl methacrylate), poly(dimethyl siloxane), photopolymers (which contain photoreactive species such as nitrenes, carbenes and ketyl radicals capable of forming covalent links with target molecules.

Binding of the nucleic acid to a selected support may be accomplished by any of several means. For example, DNA is commonly bound to glass by first silanizing the glass surface, then activating with carbodimide or glutaraldehyde. Alternative procedures may use reagents such as 3-glycidoxypropyltrimethoxysilane (GOP) or aminopropyltrimethoxysilane (APTS) with DNA linked *via* amino linkers incorporated either at the 3′ or 5′ end of the molecule during DNA synthesis. DNA may be bound directly to membranes using ultraviolet radiation. With nitrocellose membranes, the DNA is spotted onto the membranes. A UV light source (Stratalinker, from Stratagene, La Jolla, Ca.) is used to irradiate DNA spots and induce cross-linking. An alternative method for cross-linking involves baking the spotted membranes at 80°C for two hours in vacuum.

Ligation Reactions

The methods of the present invention are directed in part to DNA sequencing by ligation of short oligonucleotide fragments. Non-limiting examples of ligation reactions are disclosed, for example, in U.S. Patent Nos. 5,403,708, 5,998,175, 6,001,614, 6,013,456 and 6,020,138, each of which is incorporated herein by reference in its entirety.

Amplification of Nucleic Acids

In particular embodiments, it may be desirable to amplify a nucleic acid of interest. The technique of "polymerase chain reaction," or "PCR," as used herein generally refers to a procedure wherein minute amounts of a specific piece of nucleic acid, RNA and/or DNA, are amplified as described in U.S. Pat. Nos. 4,683,195; 4,683,202; and 4,683,194, which are herein expressly incorporated by reference. Generally, sequence information from the ends of the region of interest or beyond needs to be available, such that oligonucleotide primers can be designed. These primers will be identical or similar in sequence to opposite strands of the template to be amplified. The 5' terminal nucleotides of the two primers may coincide with the ends of the amplified material.

PCR can be used to amplify specific RNA sequences, specific DNA sequences from total genomic DNA, and cDNA transcribed from total cellular RNA, bacteriophage or plasmid sequences, etc. See generally Mullis *et al.*, (1987). As used herein, PCR is considered to be one,

but not the only, example of a nucleic acid polymerase reaction method for amplifying a nucleic acid sample, comprising the use of a known nucleic acid (DNA or RNA) as a primer and utilizes a nucleic acid polymerase to amplify or generate a specific piece of nucleic acid or to amplify or generate a specific piece of nucleic acid that is complementary to a particular nucleic acid.

2. Nucleic Acid Analogs

A nucleic acid may comprise, or be composed entirely of, a derivative or analog of a nucleobase, a nucleobase linker moiety and/or backbone moiety that may be present in a naturally occurring nucleic acid. As used herein a "derivative" refers to a chemically modified or altered form of a naturally occurring molecule, while the terms "mimic" or "analog" refer to a molecule that may or may not structurally resemble a naturally occurring molecule or moiety, but possesses similar functions. As used herein, a "moiety" generally refers to a smaller chemical or molecular component of a larger chemical or molecular structure. Nucleobase, nucleoside and nucleotide analogs or derivatives are well known in the art.

Nucleobases

As used herein a "nucleobase" refers to a heterocyclic base, such as for example a naturally occurring nucleobase (*i.e.,* an A, T, G, C or U) found in at least one naturally occurring nucleic acid (*i.e.,* DNA and RNA), and naturally or non-naturally occurring derivative(s) and analogs of such a nucleobase. A nucleobase generally can form one or more hydrogen bonds ("anneal" or "hybridize") with at least one naturally occurring nucleobase in manner that may substitute for naturally occurring nucleobase pairing (*e.g.,* the hydrogen bonding between A and T, G and C, and A and U).

"Purine" and/or "pyrimidine" nucleobase(s) encompass naturally occurring purine and/or pyrimidine nucleobases and also derivative(s) and analog(s) thereof, including but not limited to, those a purine or pyrimidine substituted by one or more of an alkyl, carboxyalkyl, amino, hydroxyl, halogen (*i.e.,* fluoro, chloro, bromo, or iodo), thiol or alkylthiol moiety. Preferred alkyl (*e.g.,* alkyl, carboxyalkyl, etc.) moieties comprise of from about 1, about 2, about 3, about 4, about 5, to about 6 carbon atoms. Other non-limiting examples of a purine or pyrimidine include a deazapurine, a 2,6-diaminopurine, a 5-fluorouracil, a xanthine, a hypoxanthine, a 8-bromoguanine, a 8-chloroguanine, a bromothymine, a 8-aminoguanine, a 8-hydroxyguanine, a 8-methylguanine, a 8-thioguanine, an azaguanine, a 2-aminopurine, a 5-ethylcytosine, a 5-methylcyosine, a 5-bromouracil, a 5-ethyluracil, a 5-iodouracil, a 5-chlorouracil, a 5-propyluracil, a thiouracil, a 2-methyladenine, a methylthioadenine, a N,N-diemethyladenine, an

azaadenines, a 8-bromoadenine, a 8-hydroxyadenine, a 6-hydroxyaminopurine, a 6-thiopurine, a 4-(6-aminohexyl/cytosine), and the like. A table listing non-limiting purine and pyrimidine derivatives and analogs is provided below.

Table 1 – Purine and Pyrmidine Derivatives or Analogs

Abbr.	Modified base description	Abbr.	Modified base description
ac4c	4-acetylcytidine	Mam5s2 u	5-methoxyaminomethyl-2-thiouridine
chm5u	5-(carboxyhydroxylmethyl)uridine	Man q	Beta,D-mannosylqueosine
Cm	2′-O-methylcytidine	Mcm5s2 u	5-methoxycarbonylmethyl-2-thiouridine
Cmnm5s 2u	5-carboxymethylaminomethyl-2-thioridine	Mcm5u	5-methoxycarbonylmethyluridine
Cmnm5u	5-carboxymethylaminomethyluridine	Mo5u	5-methoxyuridine
D	Dihydrouridine	Ms2i6a	2-methylthio-N6-isopentenyladenosine
Fm	2'-O-methylpseudouridine	Ms2t6a	N-((9-beta-D-ribofuranosyl-2-methylthiopurine-6-yl)carbamoyl)threonine
gal q	beta,D-galactosylqueosine	Mt6a	N-((9-beta-D-ribofuranosylpurine-6-yl)N-methyl-carbamoyl)threonine
Gm	2'-O-methylguanosine	Mv	Uridine-5-oxyacetic acid methylester
I	Inosine	o5u	Uridine-5-oxyacetic acid (v)
I6a	N6-isopentenyladenosine	Osyw	Wybutoxosine
m1a	1-methyladenosine	P	Pseudouridine
m1f	1-methylpseudouridine	Q	Queosine
m1g	1-methylguanosine	s2c	2-thiocytidine

Table 1 – Purine and Pyrmidine Derivatives or Analogs			
Abbr.	Modified base description	Abbr.	Modified base description
m1I	1-methylinosine	s2t	5-methyl-2-thiouridine
m22g	2,2-dimethylguanosine	s2u	2-thiouridine
m2a	2-methyladenosine	s4u	4-thiouridine
m2g	2-methylguanosine	T	5-methyluridine
m3c	3-methylcytidine	t6a	N-((9-beta-D-ribofuranosylpurine-6-yl)carbamoyl)threonine
m5c	5-methylcytidine	Tm	2'-O-methyl-5-methyluridine
m6a	N6-methyladenosine	Um	2'-O-methyluridine
m7g	7-methylguanosine	Yw	Wybutosine
Mam5u	5-methylaminomethyluridine	X	3-(3-amino-3-carboxypropyl)uridine, (acp3)u

A nucleobase may be incorporated into a nucleoside or nucleotide, using any chemical or natural synthesis method described herein or known to one of ordinary skill in the art.

Nucleosides

As used herein, a "nucleoside" refers to an individual chemical unit comprising a nucleobase covalently attached to a nucleobase linker moiety. A non-limiting example of a "nucleobase linker moiety" is a sugar comprising 5-carbon atoms (*i.e.*, a "5-carbon sugar"), including but not limited to a deoxyribose, a ribose, an arabinose, or a derivative or an analog of a 5-carbon sugar. Non-limiting examples of a derivative or an analog of a 5-carbon sugar include a 2'-fluoro-2'-deoxyribose or a carbocyclic sugar where a carbon is substituted for an oxygen atom in the sugar ring.

Different types of covalent attachment(s) of a nucleobase to a nucleobase linker moiety are known in the art. By way of non-limiting example, a nucleoside comprising a purine (*i.e.*, A or G) or a 7-deazapurine nucleobase typically covalently attaches the 9 position of a purine or a 7-deazapurine to the 1'-position of a 5-carbon sugar. In another non-limiting example, a nucleoside comprising a pyrimidine nucleobase (*i.e.*, C, T or U) may covalently attach a 1 position of a pyrimidine to a 1'-position of a 5-carbon sugar.

Nucleotides

As used herein, a "nucleotide" refers to a nucleoside further comprising a "backbone moiety". A backbone moiety generally covalently attaches a nucleotide to another molecule

comprising a nucleotide, or to another nucleotide to form a nucleic acid. The "backbone moiety" in naturally occurring nucleotides typically comprises a phosphorus moiety, which is covalently attached to a 5-carbon sugar. The attachment of the backbone moiety typically occurs at either the 3'- or 5'-position of the 5-carbon sugar. However, other types of attachments are known in the art, particularly when a nucleotide comprises derivatives or analogs of a naturally occurring 5-carbon sugar or phosphorus moiety.

3. Separation and Quantitation Methods

It will be desirable to separate the products the DNA sequencing reaction from each other and from the template and the excess primer and oligonucleotides for the purpose of DNA sequence analysis. A variety of techniques for size-separation of nucleic acids are known in the art. The skilled artisan will realize that the practice of the methods claimed herein do not rely upon any specific technique for size separation of nucleic acids, but rather any method that is suitable for one-base or two-base resolution of nucleic acids may be used within the scope of the present invention.

Gel electrophoresis and Chromatographic Techniques

In one embodiment, DNA sequencing products are separated by agarose, agarose-acrylamide or polyacrylamide gel electrophoresis using methods commonly known to one of ordinary skill in the art. (Sambrook *et al.*, 1989).

Alternatively, chromatographic techniques may be employed to effect separation. There are many kinds of chromatography which may be used: adsorption, partition, ion-exchange and molecular sieve, and many specialized techniques for using them including column, paper, thin-layer and gas chromatography (Freifelder, 1982). The only requirement for the practice of the present invention is that the technique used must be capable of separating, according to size, DNA fragments differing in size by one base or two bases.

Microfluidic Techniques

Microfluidic techniques include separation on a platform such as microcapillaries, designed by ACLARA BioSciences Inc., or the LabChipTM "liquid integrated circuits" made by Caliper Technologies Inc. These microfluidic platforms require only nanoliter volumes of sample, in contrast to the microliter volumes required by other separation technologies. Miniaturizing some of the processes involved in genetic analysis has been achieved using microfluidic devices. For example, published PCT Application No. WO 94/05414, to Northrup and White, incorporated herein by reference, reports an integrated micro-PCR™ apparatus for

collection and amplification of nucleic acids from a specimen. U.S. Patent No. 5,856,174 describes an apparatus which combines the various processing and analytical operations involved in nucleic acid analysis and is incorporated herein by reference.

Capillary Electrophoresis

Microcapillary array electrophoresis generally involves the use of a thin capillary or channel which may be filled with a particular separation medium. Electrophoresis of a sample through the capillary provides a size based separation profile for the sample. The use of microcapillary electrophoresis in size separation of nucleic acids has been reported in, for example, Woolley and Mathies, 1994. Microcapillary array electrophoresis generally provides a rapid method for size-based sequencing product analysis. The high surface to volume ratio of these capillaries allows for the application of higher electric fields across the capillary without substantial thermal variation across the capillary, consequently allowing for more rapid separations. Furthermore, when combined with confocal imaging methods, these methods provide sensitivity in the range of attomoles, which is comparable to the sensitivity of radioactive sequencing methods.

Microfabrication of microfluidic devices including microcapillary electrophoretic devices has been discussed in detail in, for example, Jacobsen *et al.*, 1994; Effenhauser *et al.*, 1994; Harrison *et al.*, 1993; Manz *et al.*, 1992; and U.S. Patent No. 5,904,824, here incorporated by reference. Typically, these methods comprise photolithographic etching of micron scale channels on a silica, silicon or other crystalline substrate or chip, and can be readily adapted for use in the present invention. In some embodiments, the capillary arrays may be fabricated from the same polymeric materials described for the fabrication of the body of the device, using the injection molding techniques described herein.

In many capillary electrophoresis methods, the capillaries, e.g., fused silica capillaries or channels etched, machined or molded into planar substrates, are filled with an appropriate separation/sieving matrix. Typically, a variety of sieving matrices are known in the art may be used in the microcapillary arrays. Examples of such matrices include, e.g., hydroxyethyl cellulose, polyacrylamide, agarose and the like. Generally, the specific gel matrix, running buffers and running conditions are selected to maximize the separation characteristics of the particular application, e.g., the size of the nucleic acid fragments, the required resolution, and the presence of native or undenatured nucleic acid molecules. For example, running buffers may include denaturants, chaotropic agents such as urea or the like, to denature nucleic acids in the sample.

Separation and quantitation of ligation reaction products may also be performed using the apparatus and methods disclosed in co-pending U.S. Patent Application entitled, "Microfabricated Injector and Capillary Array Assembly for High-Resolution and High Throughput Separations," by Shaorong Liu, the entire text of which is incorporated herein by reference.

EXAMPLES

The following examples are included to demonstrate preferred embodiments of the invention. It should be appreciated by those of skill in the art that the techniques disclosed in the examples which follow represent techniques discovered by the inventor to function well in the practice of the invention, and thus can be considered to constitute preferred modes for its practice. However, those of skill in the art should, in light of the present disclosure, appreciate that many changes can be made in the specific embodiments which are disclosed and still obtain a like or similar result without departing from the spirit and scope of the invention.

Example 1: Chemistry and Separation Protocols

In the practice of the present invention, DNA sequencing is performed by hybridization of dye-labeled primers, an n-mer oligonucleotide library and a sub-library of n-mer terminators to a nucleic acid template molecule to be sequenced. After hybridization, a DNA ligase and ligase reaction mixture are added to covalently link the primer, one or more n-mers and an n-mer terminator together. Following ligation, the products of the ligation reaction are separated by size and the DNA sequence is read.

The 5'-end of the dye-labeled primer is blocked so that extension won't occur in the 3' to 5' direction. The n-mer oligonucleotide library contains a set of random oligonucleotides of "n" bases in length, where each position of the "n" positions within the library contains a random mixture of A, G, T and C bases.

Sets of n-mer terminators are designed so that each member of the set contains a single identified nucleotide at one of the n-positions of the n-mer, while the remaining n-1 positions of the n-mer contain a random mixture of all four nucleotides (A, G, C and T). Each set as a whole contains terminators with the same identified nucleotide at each of the n-positions of the n-mers. For example, for the set of n-mer terminators where n = 6 and the residue of interest is an A, the set would comprise 6-mer terminators of sequences: 5'-dN-dN-dN-dN-dN-ddA-3'; 5'-dN-dN-dN-dN-dA-ddN-3'; 5'-dN-dN-dN-dA-dN-ddN-3'; 5'-dN-dN-dA-dN-dN-ddN-3'; 5'-dN-dA-dN-

dN-dN-ddN-3'; and 5'-dA-dN-dN-dN-dN-ddN-3', where N indicates a mixture of A, T, G and C, dN indicates a deoxynucleotide and ddA indicates a dideoxyadenosine nucleotide. The remaining sets of n-mer terminators would be identical, except for the substitution of a C, G or T in place of the A residue.

In the practice of the invention, the sets of n-mer terminators would be used to generate the complete DNA sequence of the target strand. For example, where the primer is the labeled moiety in the mixture, a first reaction would use the terminator 5'-dN-dN-dN-dN-dN-ddA-3'. A second reaction with a distinguishably labeled primer would use the terminator 5'-dN-dN-dN-dN-dN-ddT-3'. A third reaction with a distinguishably labeled primer would use the terminator 5'-dN-dN-dN-dN-dN-ddG-3'. A fourth reaction with a distinguishably labeled primer would use the terminator 5'-dN-dN-dN-dN-dN-ddC-3'. With a detector capable of identifying each of the four labels, the ligation products of the four reactions could be mixed together and size-separated. Separation of the ligation products would identify all nucleotides that are multiples of 6 bases away from the 3' end of the primer.

Alternatively, if the terminator molecules are distinguishably labeled, it would be possible to conduct all four ligation reactions in a single tube. In this case, the reaction would contain a single unlabeled primer, a library of n-mers, and a mixture (sub-library) of the four distinguishably labeled terminators 5'-dN-dN-dN-dN-dN-ddA-3'; 5'-dN-dN-dN-dN-dN-ddT-3'; 5'-dN-dN-dN-dN-dN-ddG-3'; and 5'-dN-dN-dN-dN-dN-ddC-3'. Separation of ligation products would provide sequence information as recited above.

Complete sequence determination would require additional ligation reactions. A second reaction would use the sub-library of terminators: 5'-dN-dN-dN-dN-dA-ddN-3'; 5'-dN-dN-dN-dN-dT-ddN-3'; 5'-dN-dN-dN-dN-dG-ddN-3'; and 5'-dN-dN-dN-dN-dC-ddN-3' to identify all nucleotides that are located [(multiples of 6) – 1 bases] away from the 3' end of the primer. A third reaction would use the sub-library of terminators: 5'-dN-dN-dN-dA-dN-ddN-3'; 5'-dN-dN-dN-dT-dN-ddN-3'; 5'-dN-dN-dN-dG-dN-ddN-3'; and 5'-dN-dN-dN-dC-dN-ddN-3' to identify all nucleotides that are located [(multiples of 6) – 2 bases] away from the 3' end of the primer and so on until all nucleotides in the nucleic acid to be sequenced have been identified.

The practice of the invention is illustrated in FIG. 1 through FIG. 8. For illustrative purposes, the arbitrary sequence of the DNA template was chosen as: 5'-aaaaaatttttccc tagcct gcagct gacctt gaccta aggcct yggaag tcttgc agggta ctccaa aactga". Using this template with a primer complementary to the 5'-aaaaaatttttccc-3' sequence, a mixture of all possible 6-mers and the terminator sub-library of dN-dN-dN-dN-dN-ddT, the resulting ligation reaction products are shown

in FIG. 1, where 6'er stands for a 6-mer complementary to the indicated template sequence and 6'er* is a terminator complementary to the indicated template sequence. These ligation products, when separated by size, provide sequence information showing the presence of an "A" residue at 24, 48, 54 and 60 bases downstream from the 3' end of the primer. Similarly, use of the same reaction mixture, substituted with a terminator sub-library dN-dN-dN-dN-dN-ddG would give the reaction products shown in FIG. 2, indicating the presence of a "C" residue 42 bases downstream from the 3' end of the primer. Use of the terminator sub-library dN-dN-dN-dN-dN-ddC would give the reaction products shown in FIG. 3, indicating the presence of a "G" residue 36 bases downstream from the 3' end of the primer. Use of the terminator sub-library dN-dN-dN-dN-dN-ddA would give the reaction products shown in FIG. 4, indicating the presence of a "T" residue at 6, 12, 18 and 30 bases downstream from the 3' end of the primer.

The above reactions provide complete sequence information for all residues located at (6 x M) residues downstream from the 3' end of the primer, where "M" is the set of all integers between 1 and [(total number of bases in template)/6] . To obtain complete sequence information, the corresponding nucleotide identifications must be obtained for residues located at (6 x M) – 1; (6 x M) – 2; (6 x M) – 3; (6 x M) – 4 and (6 x M) – 5 bases downstream from the 3' end of the primer.

The next set of reactions would use the same primers and n-mer libraries, but would use a set of terminator sub-libraries where the known base (A, G, T or C) is located at the 5th position of the n-mer instead of at the end.

FIG. 5 shows the ligation products using the terminator sub-library dN-dN-dN-dN-dT-ddN, indicating the presence of an "A" residue at 35 and 53 bases downstream from the 3' end of the primer. FIG. 6 shows the ligation products using the terminator sub-library dN-dN-dN-dN-dG-ddN, indicating the presence of a "C" residue at 5, 11 and 29 bases downstream from the 3' end of the primer. FIG. 7 shows the ligation products using the terminator sub-library dN-dN-dN-dN-dC-ddN, indicating the presence of a "G" residue at 41 and 59 bases downstream from the 3' end of the primer. FIG. 8 shows the ligation products using the terminator sub-library dN-dN-dN-dN-dA-ddN, indicating the presence of a "T" residue at 17, 23 and 47 bases downstream from the 3' end of the primer.

Similarly, one uses a terminator libraries of dN-dN-dN-dA-dN-ddN, dN-dN-dN-dC-dN-ddN, dN-dN-dN-dG-dN-ddN, and dN-dN-dN-dT-dN-ddN for the third set of reactions, dN-dN-dA-dN-dN-ddN, dN-dN-dC-dN-dN-ddN, dN-dN-dG-dN-dN-ddN, and dN-dN-dT-dN-dN-ddN for the fourth set of reactions, dN-dA-dN-dN-dN-ddN, dN-dC-dN-dN-dN-ddN, dN-dG-dN-dN-dN-ddN, and dN-dT-dN-dN-dN-ddN for the fifth set of reactions, and dA-dN-dN-dN-dN-ddN, dC-dN-dN-dN-dN-ddN, dG-dN-dN-dN-dN-ddN, and

dT-dN-dN-dN-dN-ddN for the sixth set of reactions. Completion of the sequencing reactions using the entire set of terminators would provide the entire sequence of the template strand. As indicated above, using four-color labeled terminators instead of labeled primers would reduce the number of separate reactions needed by a factor of four.

The claimed method allows for long read-length DNA sequencing because the requirements for separation of similarly sized molecules is greatly reduced. Even with a four-color reaction scheme, the ligation products to be separated must differ by at least 6 nucleotides in size. This reduced stringency required for the separation of nucleic acids allows sequence determination to be performed for much longer DNA fragments. In principle, if 500 bases of sequences can be obtained using the standard Sanger dideoxy method, which requires single base resolution of nucleic acids, then the use of 6-mers as the units for size separation should allow a read-through length of approximately (6 x 500) or 3,000 bases.

Example 2: DNA Sequencing Using Different Sized Primers

Under certain circumstances, it may be desirable to reduce the number of reactions and separations in order to increase throughput. A simple and elegant solution is to use multiple primers for each reaction, with all other conditions the same.

Consider, for example, two distinguishable dye-labeled primers, primer 1 with a sequence complementary to 5'aaaaaatttttccc-3' and primer 2 with a sequence complementary to 5'-aaaaaatttttccctag-3'. Although they both start at the same position on the template strand, the binding site for primer 2 is 3 bases longer than the binding site for primer 1. If these two primers were used with a terminator sub-library of dN-dN-dN-dN-dN-ddT, the ligation products of the first primer should be as shown in FIG. 9, indicating the presence of an "A" residue at 24, 28, 54 and 60 bases downstream from the 3' end of primer 1. The ligation products incorporating primer 2 would be as shown in FIG. 10, indicating the presence of an "A" residue at 6 bases downstream from the 3' end of primer 2 (or 9 bases downstream from the 3' end of primer 1). By combining sets of primers in each reaction mixture, the number of reactions is correspondingly decreased. In the example given, the sizes of nucleic acids to be separated differ by at least three bases in length. Thus, the increase in read-length will be mainly retained, while allowing a corresponding increase in sample throughput.

If a set of three primers that differ in length from each other by two bases are used in the reaction, with four-color labeled terminators, full sequence information can be obtained with only two separations of ligation products. Double-base resolution is required for these separations to generate accurate sequencing data, but the read-length should be twice as long as for a single-base

determination. The novel methods and compositions of the present invention thus provide significant advantages over the prior art in terms of read-length and high through-put analysis of DNA sequences.

* * *

All of the APPARATUS and/or METHODS disclosed and claimed herein can be made and executed without undue experimentation in light of the present disclosure. While the compositions and methods of this invention have been described in terms of preferred embodiments, it will be apparent to those of skill in the art that variations may be applied to the APPARATUS and/or METHODS and in the steps or in the sequence of steps of the method described herein without departing from the concept, spirit and scope of the invention. More specifically, it will be apparent that certain agents which are both chemically and physiologically related may be substituted for the agents described herein while the same or similar results would be achieved. All such similar substitutes and modifications apparent to those skilled in the art are deemed to be within the spirit, scope and concept of the invention as defined by the appended claims.

REFERENCES

The following references, to the extent that they provide exemplary procedural or other details supplementary to those set forth herein, are specifically incorporated herein by reference.

Ansorge *et al.*, Automated DNA Sequencing: Ultrasensitive Detection of Fluorescent Bands during Electrophoresis. *Nucleic Acids Research* 15:4593, 1987.

Effenhauser *et al.*, *Anal. Chem.*, 66:2949-2953, 1994

Freifelder, *Physical Biochemistry Applications to Biochemistry and Molecular Biology*, 2nd ed. Wm. Freeman and Co., New York, NY, 1982.

Harrison *et al.*, *Science*, 261:895-897, 1993.

Holmstrom *et al.*, *Anal. Biochem.* 209:278-283, 1993.

Jacobsen *et al.*, *Anal. Chem.*, 66:1107-1113, 1994.

Kaczorowski and Szybalski, *Anal. Biochem.*, 221:127-135, 1994.

Kaczorowski and Szybalski, Gene, 223:83-91, 1998.

Manz *et al.*, *J. Chromatogr.*, 593:253-258, 1992.

Maxam and Gilbert, A New Method for Sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.* 74:560-4, 1977.

Mullis *et al.*, *Cold Spring Harbor Symp. Quant. Biol.*, 51:263, 1987.

National Human Genome Research Institute, "Advanced Development of Genomic Technologies", PA number: PAR-99-047, Release data: Jan. 22, 1999.

Newton, *et al.*, *Nucl. Acids Res.* 21:1155-1162, 1993.

PCT Application No. WO 94/05414

Prober *et al.*, A System for Rapid DNA Sequencing with Fluorescent Chain-Terminating Dideoxynucleotides. *Science* 238:336-41, 1987.

Rasmussen, *et al.*, *Anal. Biochem.* 198:138-142, 1991.

Running, *et al.*, *BioTechniques* 8:276-277, 1990.

Salas-Solano *et al.*, *Anal. Chem.*, 70:3996-4003, 1998.

Sambrook *et al.*, "Molecular Cloning," *A Laboratory Manual, 2d Ed.*, Cold Spring Harbor Laboratory Press, New York, 13.7-13.9:1989.

Sanger *et al.*, DNA Sequencing with Chain Termination Inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* 74:5463-7, 1977.

Smith *et al.*, Fluorescence Detection in Automated DNA Sequence Analysis. *Nature* 321:674-9, 1986.

Szybalski, *Gene*, 90:177-178, 1990.

Tabor *et al.*, DNA Sequence Analysis with a Modified Bacteriophage T7 DNA Polymerase effect of Pyrophosphorolysis and Metal Ions. *J. Biol. Chem.* 265:8322-8, 1990.

U.S. Patent No. 4,683,194

U.S. Patent No. 4,683,195

U.S. Patent No. 4,683,202

U.S. Patent No. 5,403,708

U.S. Patent No. 5,837,832

U.S. Patent No. 5,837,860

U.S. Patent No. 5,856,174

U.S. Patent No. 5,871,628

U.S. Patent No. 5,904,824

U.S. Patent No. 5,908,755

U.S. Patent No. 5,998,175

U.S. Patent No. 6,001,614

U.S. Patent No. 6,013,456

U.S. Patent No. 6,020,138

Woolley and Mathies, *Proc. Natl. Acad Sci. U.S.A.*, 91:11348-11352, 1994

CLAIMS

A METHOD FOR SEQUENCING LONG-LENGTH DNA

1. A method for long read-length sequencing of a nucleic acid comprising:

a) obtaining a template nucleic acid to be sequenced;

b) adding a composition comprising (i) one or more primers; (ii) a library of n-mers; and (iii) a sub-library of n-mer terminators, under high stringency hybridization conditions;

c) ligating primers, n-mers and terminators that are hybridized to said template to form ligation reaction products;

d) separating said ligation reaction products by size; and

e) analyzing said separated ligation reaction products to provide nucleic acid sequence data.

2. The method of claim 1, wherein said n is an integer number selected from the group consisting of 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34 and 35.

3. The method of claim 1, wherein said nucleic acid is DNA.

4. The method of claim 1, wherein said nucleic acid is RNA.

5. The method of claim 1, wherein said primer is labeled.

6. The method of claim 5, wherein said label is fluorescent, luminescent, phosphorescent, spectroscopic, enzymatic or radioactive.

7. The method of claim 1, wherein said n-mer terminators are labeled.

8. The method of claim 7, wherein said labels provide four-color nucleic acid sequencing.

9. The method of claim 1, wherein said n-mer and said n-mer terminator are 6 nucleotides in length.

10. The method of claim 9, wherein said n-mer terminators consist of dN-dN-dN-d-dN-ddA, dN-dN-dN-dN-dN-ddC, dN-dN-dN-dN-dN-ddG, dN-dN-dN-dN-dN-ddT, dN-dN-dN-dN-dA-ddN, dN-dN-dN-dN-dC-ddN, dN-dN-dN-dN-dG-ddN, dN-dN-dN-dN-dT-ddN, dN-dN-dN-dA-dN-ddN, dN-dN-dN-dC-dN-ddN, dN-dN-dN-dG-dN-ddN, dN-dN-dN-dT-dN-ddN, dN-dN-dA-dN-dN-ddN, dN-dN-dC-dN-dN-ddN, dN-dN-dG-dN-dN-ddN, dN-dN-dT-dN-dN-ddN, dN-dA-dN-dN-dN-ddN, dN-dC-dN-dN-dN-ddN, dN-dG-dN-dN-dN-ddN, dN-dT-dN-dN-dN-ddN, dA-dN-dN-dN-dN-ddN, dC-dN-dN-dN-dN-ddN, dG-dN-dN-dN-dN-ddN, and dT-dN-dN-dN-dN-ddN.

11. The method of claim 10, wherein each of said n-mer terminators is added to a separate ligation reaction mixture.

12. The method of claim 10, wherein all n-mer terminators containing an identified nucleotide at the same position of the n-mer are added to the same ligation reaction mixture.

13. The method of claim 1, wherein said primers are blocked at the 5' end to prevent the ligation of oligonucleotides to the 5' end of the primer.

14. The method of claim 13, wherein only one primer is added.

15. The method of claim 9, wherein a first and a second primer s are added, wherein the binding sites for both primers start at the same position on the template strand and wherein the first primer is three nucleotides shorter than the second primer.

16. The method of claim 9, wherein a first, a second and a third primer are added, wherein the binding sites for all three primers start at the same position on the template strand, wherein the first primer is two nucleotides shorter than the second primer and the second primer is two nucleotides shorter than the third primer. .

17. The method of claim 9, wherein 6 primers are added, wherein the binding sites for all six primers start at the same position on the template strand, wherein all 6 primers are of different length and wherein each primer is one nucleotide longer than the next shortest primer.

18. The method of claim 2, wherein two or more primers are added, wherein the binding site for each primer starts at the same position on the template strand, wherein each primer is of a different

length and wherein each primer differs in length by the same number of nucleotides from the next closest primer in size.

19. The method of claim 1, wherein said n-mer and n-mer terminators have a size selected from the group consisting of 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 and 20 nucleotides in length.

20. A composition comprising a library of n-mers and a sub-library of n-mer terminators.

21. The composition of claim 17, wherein said n-mers and said n-mer terminators are 6 nucleotides in length.

22. The composition of claim 18, wherein said n-mer terminators are selected from the group consisting of dN-dN-dN-d-dN-ddA, dN-dN-dN-dN-dN-ddC, dN-dN-dN-dN-dN-ddG, dN-dN-dN-dN-dN-ddT, dN-dN-dN-dN-dA-ddN, dN-dN-dN-dN-dC-ddN, dN-dN-dN-dN-dG-ddN, dN-dN-dN-dN-dT-ddN, dN-dN-dN-dA-dN-ddN, dN-dN-dN-dC-dN-ddN, dN-dN-dN-dG-dN-ddN, dN-dN-dN-dT-dN-ddN, dN-dN-dA-dN-dN-ddN, dN-dN-dC-dN-dN-ddN, dN-dN-dG-dN-dN-ddN, dN-dN-dT-dN-dN-ddN, dN-dA-dN-dN-dN-ddN, dN-dC-dN-dN-dN-ddN, dN-dG-dN-dN-dN-ddN, dN-dT-dN-dN-dN-ddN, dA-dN-dN-dN-dN-ddN, dC-dN-dN-dN-dN-ddN, dG-dN-dN-dN-dN-ddN, and dT-dN-dN-dN-dN-ddN.

23. The composition of claim 19, wherein each of said n-mer terminators is present in a separate composition.

24. The composition of claim 19, wherein all n-mer terminators containing an identified nucleotide at the same position of the n-mer are present in the same composition.

25. The method of claim 1, wherein said separating step occurs by capillary electrophoresis.

26. The method of claim 1, wherein said separating step occurs on a microfabricated electrophoresis chip.

27. The method of claim 1, wherein said separating step occurs on a hybrid microfabricated injector and capillary array assembly.

28. The method of claim 1, further comprising using an automated DNA sequencer.

29. The method of claim 2, wherein the reaction products are used to generate sequence information of every x bases after the 3'-end of the primer.

30. The method of claim 29, wherein said information is used to check the sequence of a DNA or RNA molecule.

31. The method of claim 2, further comprising: (i) using one or more distinguishably labeled n-mers complementary to a known single nucleotide polymorphorism (SNP) sequence; and (ii) detecting one or more SNP sequences in the template strand.

32. The method of claim 31, wherein the primer is designed to hybridize with a sample DNA in which a known SNP location 1 to n bases away from the 3' end of the primer.

33. The method of claim 31, wherein the primer is designed to hybridize with a sample DNA in which a known SNP is n/2 (where n/2 is rounded into an integer number if it is not an integer number) ± 3 bases away from the 3' end of the primer.

34. The method of claim 31, wherein the primer is designed to hybridize with a sample DNA in which a known SNP is n/2 ± 2 bases away from the 3' end of the primer.

35. The method of claim 31, wherein the primer is designed to hybridize with a sample DNA in which a known SNP is n/2 ± 1 bases away from the 3' end of the primer.

36. The method of claim 31, wherein the primer is designed to hybridize with a sample DNA in which a known SNP is n/2 bases away from the 3'-end of the primer.

37. The method of claim 31, wherein the n-mer oligonucleotide is complementary to the sequence of a wild type DNA following the 3'-end of the primer.

38. The method of claim 31, wherein the n-mer oligonucleotide is complementary to the sequence of a mutant DNA following the 3'-end of the primer.

39. The method of claim 31, wherein the n-mer oligonucleotide is labeled with a fluorescent dye or other detectable markers.

40. The method of claim 31, wherein multiple SNPs are detected.

41. The method of claim 38, wherein multiple n-mer oligonucleotides are used.

42. The method of claim 39, wherein each n-mer oligonucleotide is complementary to the sequence of a particular SNP following the 3'-end of the primer.

43. The method of claim 40, wherein each n-mer oligonucleotide is labeled with a unique fluorescent dye or other detectable markers.

44. The method of claims 30 or 40, wherein the identification method is selected from the group consisting of slab gel electrophoresis, capillary gel electrophoresis and gel electrophoresis on microfabricated chips.

45. The method of claims 30 or 40, wherein the identification method is based on the attachment of the primer to a solid-phase support.

46. The method of claim 43, wherein the one or more SNPs are identified by monitoring whether the primer is labeled with a fluorescent dye or other detectable marker.

FIG. 1

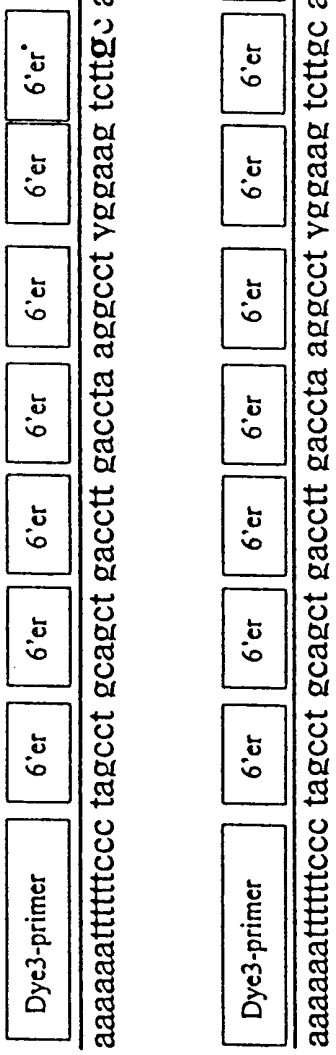

Dye3-primer
6'er
6'er
6'er
6'er
6'er
6'er
6'er
aaaaaatttttccc tagcct gcagct gacctt gaccta aggcct yggaag tcttgc aggta ctccaa aactga
Dye3-primer
6'er
6'er
6'er
6'er
6'er
6'er
6'er
6'er
aaaaaatttttccc tagcct gcagct gacctt gaccta aggcct yggaag tcttgc aggta ctccaa aactga

FIG. 2

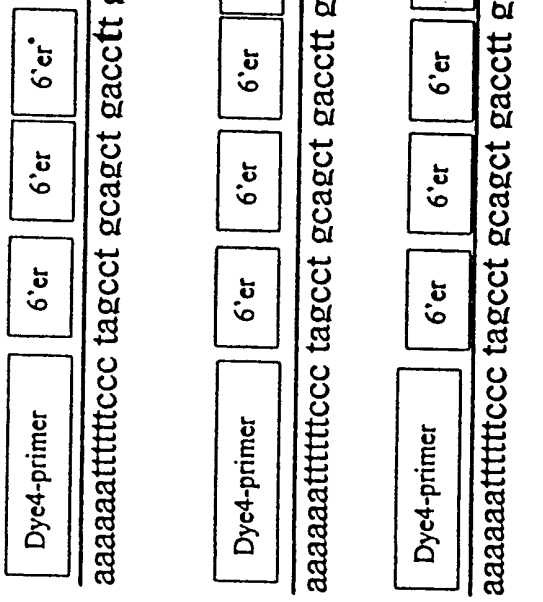

Dye4-primer
6'er
6'er
6'er'
aaaaaatttttccc tagcct gcagct gaccta aggcct yggaag tcttgc agggta ctccaa aactga
Dye4-primer
6'er
6'er
6'er'
aaaaaatttttccc tagcct gcagct gacctt gaccta aggcct yggaag tcttgc agggta ctccaa aactga
Dye4-primer
6'er
6'er
6'er
6'er
6'er
6'er
6'er'
aaaaaatttttccc tagcct gcagct gacctt gaccta aggcct yggaag tcttgc agggta ctccaa aactga

FIG. 3

Dye1-primer
6'er
6'er
6'er
6'er
6'er
6'er'
aaaaaatttttccc tagcct gcagct gacctt gaccta aggcct yggaag tcttgc agggta ctccaa aactga
Dye1-primer
6'er
6'er
6'er
6'er
6'er
6'er
6'er
6'er'
aaaaaatttttccc tagcct gcagct gacctt gaccta aggcct yggaag tcttgc agggta ctccaa aactga

FIG. 4

Dye2-primer
6'er'
åaaaaaattttttccc tagcct gcagct gaccta aggcct yggaag tcttgc agggta ctccaa aactga
Dye2-primer
6'er
6'er'
aaaaaaattttttccc tagcct gcagct gaccta aggcct yggaag tcttgc agggta ctccaa aactga
Dye2-primer
6'er
6'er
6'er
6'er
6'er'
aaaaaaattttttccc tagcct gcagct gaccta aggcct yggaag tcttgc agggta ctccaa aactga

FIG. 5

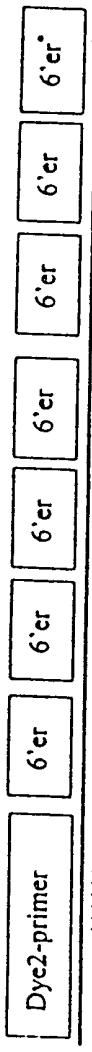

Dye2-primer
6'er
6'er
6'er
6'er
6'er
6'er
6'er*
aaaaaatttttccc tagcct gcagct gacctt gaccta aggcct yggaag tcttgc agggta ctccaa aactga

FIG. 6

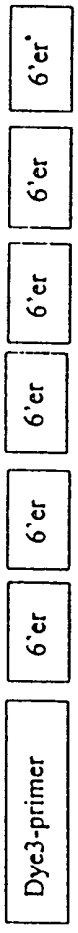

Dye3-primer
6'er
6'er
6'er
6'er
6'er
6'er'
aaaaaatttttccc tagcct gcagct gacctt gaccta aggcct yggaag tcttgc agggta ctccaa aactga

FIG. 7

Dye4-primer
6'er
aaaaaattttttccc tagcct gcagct gacctt gaccta aggcct yggaag tcttgc agggta ctccaa aactga
Dye4-primer
6'er
6'er
aaaaaattttttccc tagcct gcagct gacctt gaccta aggcct yggaag tcttgc agggta ctccaa aactga
Dye4-primer
6'er
6'er
6'er
aaaaaattttttccc tagcct gcagct gacctt gaccta aggcct yggaag tcttgc agggta ctccaa aactga
Dye4-primer
6'er
6'er
6'er
6'er
6'er
aaaaaattttttccc tagcct gcagct gacctt gaccta aggcct yggaag tcttgc agggta ctccaa aactga

FIG. 8

Dye1-primer
6'er
6'er
6'er
6'er*
aaaaaatttttccc tagcct gcagct gaccta aggcct ygggaag tcttgc agggta ctccaa aactga
Dye1-primer
6'er
6'er
6'er
6'er
6'er
6'er
6'er
6'er*
aaaaaatttttccc tagcct gcagct gaccta aggcct ygggaag tcttgc agggta ctccaa aactga
Dye1-primer
6'er
6'er
6'er
6'er
6'er
6'er
6'er
6'er
6'er*
aaaaaatttttccc tagcct gcagct gaccta aggcct ygggaag tcttgc agggta ctccaa aactga
Dye1-primer
6'er
6'er
6'er
6'er
6'er
6'er
6'er
6'er
6'er*
aaaaaatttttccc tagcct gcagct gaccta aggcct ygggaag tcttgc agggta ctccaa aactga

FIG. 9

Dye1-primer1
6'er
6'er
6'er
6'er'
aaaaaatttttccc tagcct gcagct gaccta aggcct ygaaag tcttgc agggta ctccaa aactga
Dye1-primer1
6'er
6'er
6'er
6'er
6'er
6'er
6'er'
aaaaaatttttccc tagcct gcagct gaccta aggcct ygaaag tcttgc agggta ctccaa aactga
Dye1-primer1
6'er
6'er
6'er
6'er
6'er
6'er
6'er
6'er'
aaaaaatttttccc tagcct gcagct gaccta aggcct ygaaag tcttgc agggta ctccaa aactga
Dye1-primer1
6'er
6'er
6'er
6'er
6'er
6'er
6'er
6'er
6'er'
aaaaaatttttccc tagcct gcagct gaccta aggcct ygaaag tcttgc agggta ctccaa aactga

FIG. 10

Dye 1-primer2
6'er*
aaaaaatttttccctag cctgca gctgac cttgac ctaagg cctygg aagtct tgcagg tgcagg gtactc caaaac tga

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/18210

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : C12Q 1/68; C12P 19/34; C07H 21/04

US CL : 435/6, 91.2; 536/24.3

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6, 91.2; 536/24.3

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EAST, DIALOG

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	KACZOROWSKI et al. Assembly of 18-nucleotide primers by ligation of three hexamers: Sequencing of large genomes by primer walking. Analytical Biochemistry. 1994, Vol. 221, pages 127-135, see entire document.	1-9 11-21 23-46
Y	KACZOROWSKI et al. Genomic DNA sequencig by SPEL-6 primer walking using hexamer ligation. Gene. 26 November 1998, Vol. 223, pages 83-91, see entire document.	1-9 11-21 23-46
X	US 5,780,231 A (BRENNER) 14 July 1998 (12.07.1998) see entire document.	1-9 11-21 23-46
Y	US 5,824,481 A (KAMBARA et al) 20 October 1998 (20.10.1998) see entire document.	1-9 11-21 23-46

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search

15 SEPTEMBER 2000

Date of mailing of the international search report

25 OCT 2000

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

BJ FORMAN

Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet) (July 1998) ★

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/18210

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y, P	US 6,007,987 A (CANTOR et al) 28 December 1999 (28.12.1999) see entire document.	1-9 11-21 23-46
Y	Promega. 1993/1994 catalog. Primers, Table 3E.	10, 22