US 20070042388A1

(19) United States

(12) Patent Application Publication (10) Pub. No.: US 2007/0042388 A1

Wong et al. (43) Pub. Date: Feb. 22, 2007

(54) METHOD OF PROBE DESIGN AND/OR OF NUCLEIC ACIDS DETECTION

(76) Inventors: **Christopher W. Wong**, Singapore (SG); **Wing-Kin Sung**, Singapore (SG); **Charlie Lee**, Singapore (SG); **Lance D. Miller**, Singapore (SG)

Correspondence Address:
DICKSTEIN SHAPIRO LLP
1825 EYE STREET NW
Washington, DC 20006-5403 (US)

(21) Appl. No.: **11/202,023**

(22) Filed: **Aug. 12, 2005**

Publication Classification

(51) **Int. Cl.**
C12Q 1/68 (2006.01)
G06F 19/00 (2006.01)

(52) **U.S. Cl.** **435/6**; 702/20

(57) **ABSTRACT**

It is provided a method of designing oligonucleotide probe(s) for nucleic acid detection comprising the following steps in any order: (i) identifying and selecting region(s) of a target nucleic acid to be amplified, the region(s) having an efficiency of amplification (AE) higher than the average AE; and (ii) designing oligonucleotide probe(s) capable of hybridizing to the selected region(s). It is also provided a method of detecting at least one target nucleic acid comprising the steps of: (i) providing a biological sample; (ii) amplifying the nucleic acid(s) of the biological sample; (iii) providing at least an oligonucleotide probe capable of hybridizing to at least a target nucleic acid, if present in the biological sample; and (iv) contacting the probe(s) with the amplified nucleic acids and detecting the probe(s) hybridized to the target nucleic acid(s). In particular, the method indicates the presence of at least a pathogen, for example a virus, in a human biological sample. The probes may be placed on a support, for example a microarray or a biochip.

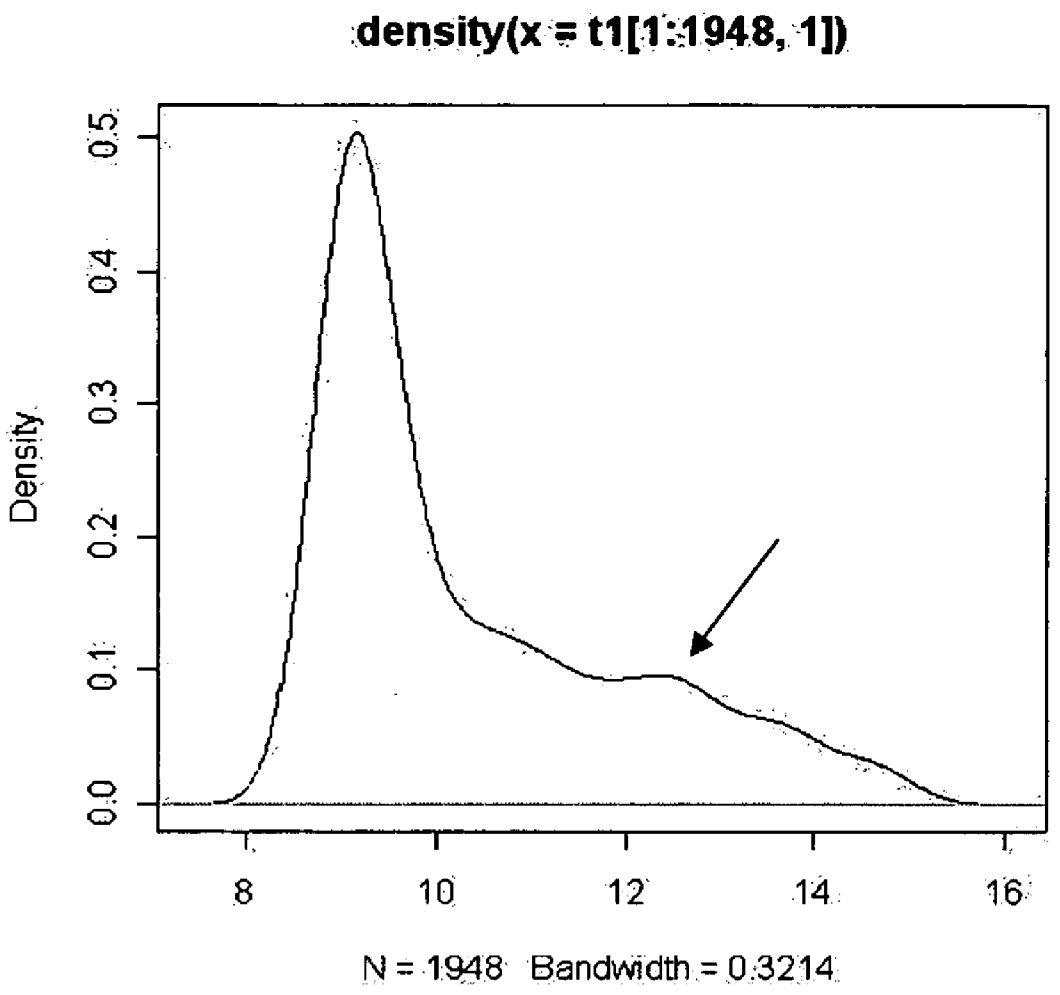


FIGURE 1

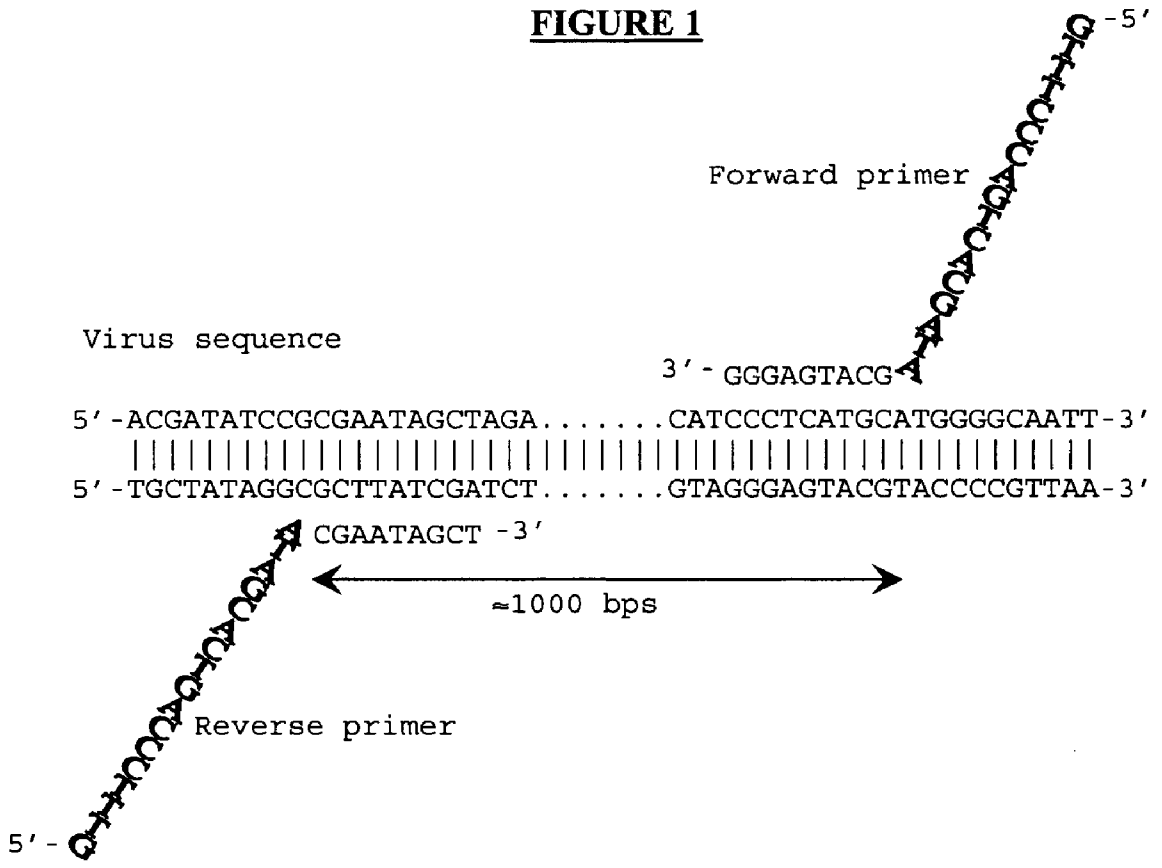

Forward primer
-5'
Virus sequence
3'- GGGAGTACG
5'-ACGATATCCGCGAATAGCTAGA.....CATCCCTCATGCATGGGGCAATT-3'
5'-TGCTATAGGCGCTTATCGATCT.....GTAGGGAGTACGTACCCCGTTAA-3'
CGAATAGCT -3'
≈1000 bps
Reverse primer
5'-

FIGURE 2

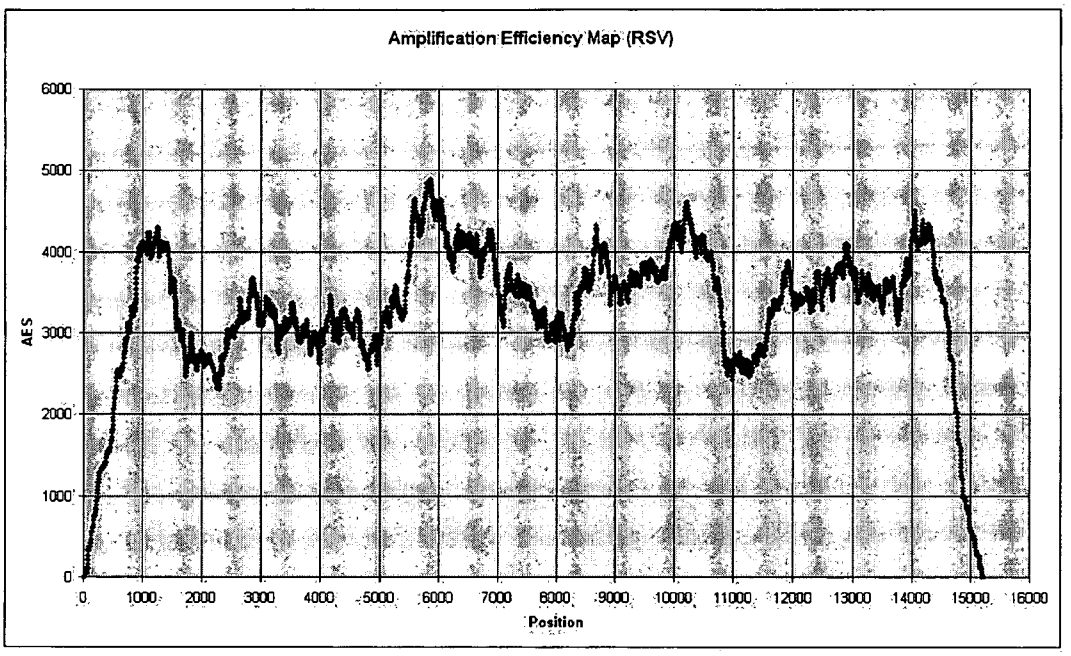

Amplification Efficiency Map (RSV)
AES
Position
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
16000

FIGURE 3

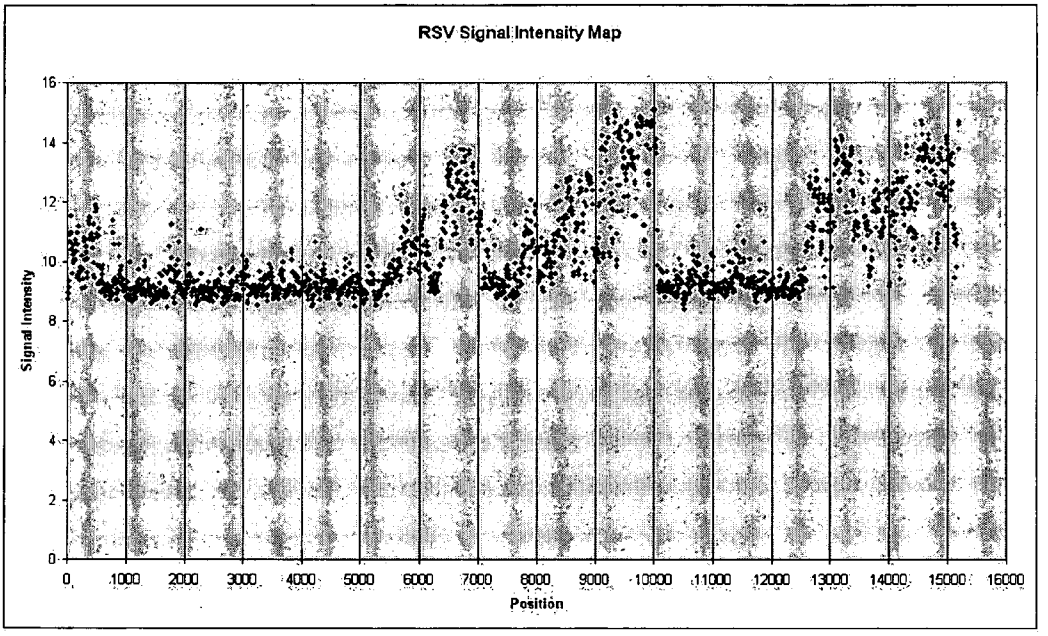
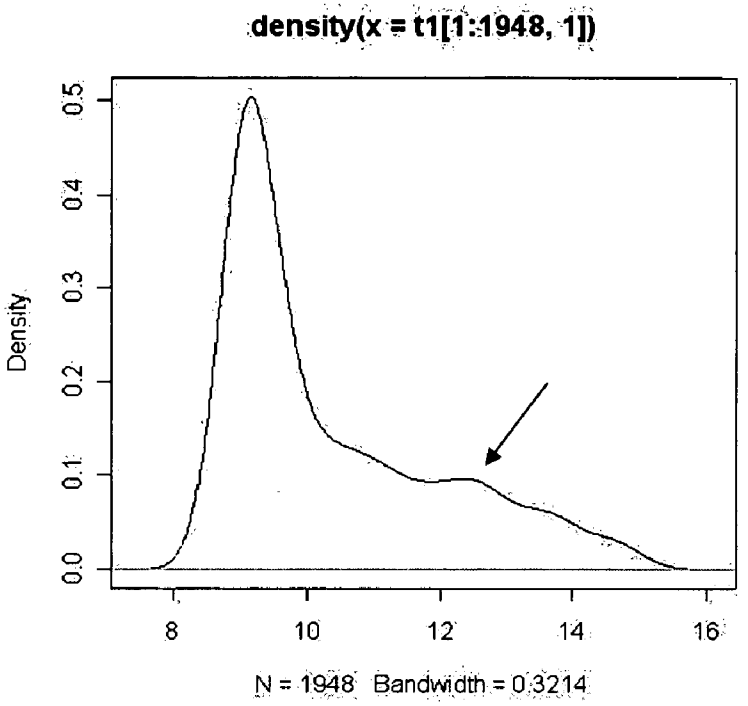
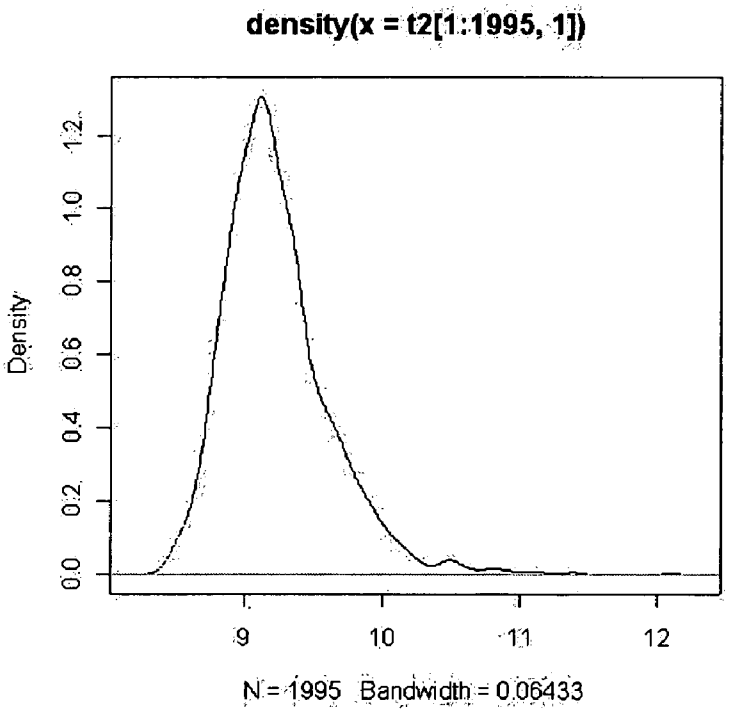

RSV Signal Intensity Map
Signal Intensity
0
2
4
6
8
10
12
14
16
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
16000
Position

FIGURE 4(A, B)



(A)



(B)

FIGURE 5

Given PDC data D with virus set V and probe set P,

Let $V' = \Phi$

For every $v_a \in V$,

- Compute one-tailed t-test with significance level 0.05 of probes in v_a

If(p-value of $t_a < 0.05$)

- Accept if $KL(P_a \| P) \geq 0.1$; $V' = V' \cup \{ v_a \}$
- Reject otherwise;

Reject otherwise;

Return V'

METHOD OF PROBE DESIGN AND/OR OF NUCLEIC ACIDS DETECTION

FIELD OF THE INVENTION

[0001] The present invention relates to the field of oligoucleotide probe design and nucleic acids detection. The method according to the invention may be used for the detection of pathogens, for example viruses.

BACKGROUND OF THE INVENTION

[0002] The accurate and rapid detection of viral and bacterial pathogens in human patients and populations is of critical medical and epidemiologic importance. Historically, diagnostic techniques have relied on cell culture passaging and various immunological assays or staining procedures. More recently PCR-based assays have been implemented, allowing for more rapid diagnosis of suspected pathogens with higher degree of sensitivity of detection. In clinical practice, however, the etiologic agent often remains unidentified, eluding detection in myriad ways. For example, some viruses are not amenable to culturing. At other times, a patient's sample may be of too poor quality or of insufficient titre for pathogen detection by conventional techniques. Moreover, both PCR- and antibody-based approaches may fail to recognize suspected pathogens simply due to natural genetic diversification resulting in alterations of PCR primer binding sites and antigenic drift.

[0003] While the concept of using oligonucleotide hybridization microarrays as a tool for determining the presence of pathogens has been proposed, significant hurdles remain, preventing the use of these microarrays routinely (Striebel, H. M., 2003). These hurdles include probe design and data analysis (Striebel, H. M., 2003; Bodrossy, L. & Sessitsch, A., 2004; Vora, G. J., et al., 2004).

[0004] Accordingly, there is a need in this field of technology for alternative and improved methods of detection of nucleic acids. In particular, there is a need for alternative and/or improved diagnostic methods for the detection of pathogens.

SUMMARY OF THE INVENTION

[0005] The present invention addresses the problems above, and in particular provides an alternative and/or improved method of probe design and/or for nucleic acid detection.

[0006] According to a first aspect, the present invention provides a method of designing oligonucleotide probe(s) for nucleic acid detection comprising the following steps in any order:

[0007] (i) identifying and selecting region(s) of a target nucleic acid to be amplified, the region(s) having an efficiency of amplification (AE) higher than the average AE; and

[0008] (ii) designing oligonucleotide probe(s) capable of hybridizing to the selected region(s).

[0009] In particular, in step (i) a score of AE is determined for every position i on the length of the target nucleic acid or of a region thereof and subsequently, an average AE score is obtained. Those regions showing an AE score higher than the average may be selected as the region(s) of the target nucleic acid to be amplified. In particular, the AE of the selected region(s) may be calculated as the Amplification Efficiency Score (AES), which is the probability that a forward primer r_i can bind to a position i and a reverse primer r_j can bind at a position j of the target nucleic acid, and |i-j| is the region of the target nucleic acid desired to be amplified. In particular, the region |i-j| may be $\leqq$10000 bp, more in particular $\leqq$5000 bp, or $\leqq$1000 bp, for example $\leqq$500 bp. In particular, the forward and reverse primers may be random primers.

[0010] According to another aspect, the step (i) comprises determining the effect of geometrical amplification bias for every position of a target nucleic acid, and selecting the region(s) to be amplified as the region(s) having an efficiency of amplification (AE) higher than the average AE. For example, the geometrical amplification bias is the PCR bias.

[0011] The step (ii) of designing oligonucleotide probe(s) capable of hybridizing to the region(s) selected in step (i) may be carried out according to any probe designing technique known in the art. In particular, the oligonucleotide probe(s) capable of hybridizing to the selected region(s) may be selected and designed according to at least one of the following criteria:

[0012] (a) the selected probe(s) has a CG-content from 40% to 60%;

[0013] (b) the probe(s) is selected by having the highest free energy computed based on Nearest-Neighbor model;

[0014] (c) given probe s_a and probe s_b substrings of target nucleic acids v_a and v_b, s_a is selected based on the hamming distance between s_a and any length-m substring s_b and/or on the longest common substring of s_a and probe s_b;

[0015] (d) for any probe s_a of length-m specific for the target nucleic acid v_a, the probe s_a is selected if it does not have any hits with any region of a nucleic acid different from the target nucleic acid, and if the probe s_a length-m has hits with the nucleic acid different from the target nucleic acid, the probe s_a length-m with the smallest maximum alignment length and/or with the least number of hits is selected; and

[0016] (e) a probe p_i at position i of a target nucleic acid is selected if p_i is predicted to hybridize to the position i of the amplified target nucleic acid.

[0017] Accordingly, two or more of the criteria indicated above may be used for designing the oligonucleotide probe(s). For example, the probe(s) may be designed by applying all criteria (a) to (e). Other criteria not explicitly mentioned herein but which are within the knowledge of a skilled person in the art may also be used.

[0018] In particular, under the criterion (e), a probe p_i at position i of a target nucleic acid v_a is selected if $P(p_i|v_a)>\lambda$, wherein λ is 0.5 and $P(p_i|v_a)$ is the probability that p_i has to hybridize to the position i of the target nucleic acid V_a. More in particular, λ is 0.75.

[0019] In particular,

$$P(p_i \mid v_a) \approx P(X \le x_i) = \frac{c_i}{k},$$

wherein X is the random variable representing the amplification efficiency score (AES) values of all probes of v_a, k is the number of probes in v_a, and c_i is the number of probes whose AES values are $\leqq x_i$.

[0020] According to another aspect of the invention, the method of designing the oligonucleotide probe(s) as described above further comprises a step of preparing the selected and designed probe(s). The probe may be prepared according to any standard method known in the art. For example, by chemical synthesis.

[0021] According to another aspect, the present invention provides a method of detecting at least one target nucleic acid comprising the steps of:

[0022] (i) providing a biological sample;

[0023] (ii) amplifying the nucleic acid(s) of the biological sample;

[0024] (iii) providing at least one oligonucleotide probe capable of hybridising to at least a target nucleic acid, if present in the biological sample, wherein the probe(s) is prepared by using a method according to any aspect of the invention described herein; and

[0025] (iv) contacting the probe(s) with the amplified nucleic acids and detecting the probe(s) hybridised to the target nucleic acid(s).

[0026] The amplification step (ii) may be carried out in the presence of random primers. For example, the amplification step (ii) may be carried out in the presence of more than two random primers. Any amplification method known in the art may be used. For example, the amplification is a RT-PCR.

[0027] In particular, a forward random primer binding to position i and a reverse random primer binding to position j of a target nucleic acid v_a are selected among primers having an amplification efficiency score (AES_i) for every position i of a target nucleic acid v_a of

$$AES_i = \sum_{j=i-Z}^{i} \left\{ P^f(j) \times \sum_{k=i+1}^{j+Z} P^r(k) \right\},$$

wherein

$$\sum_{k=i+1}^{j+Z} P^r(k) = P^r(i+1) + P^r(i+2) + \ldots P^r(j+Z),$$

P^f (i) and P^r (i)are the probability that a random primer r_i can bind to position i of v_a as forward primer and reverse primer, respectively, and $Z \leqq 10000$ bp is the region of v_a desired to be amplified. More in particular, Z may be $\leqq 5000$ bp, $\leqq 1000$ bp, or $\leqq 500$ bp.

[0028] The amplification step may comprise forward and reverse primers, and each of the forward and reverse primers may comprise, in a 5'-3' orientation, a fixed primer header and a variable primer tail, and wherein at least the variable tail hybridizes to a portion of the target nucleic acid v_a. In particular, the amplification step may comprise forward and/or reverse random primers having the nucleotide sequence of SEQ ID NO:1 or a variant or derivative thereof.

[0029] The biological sample may be any sample taken from a mammal, for example from a human being. The biological sample may be tissue, sera, nasal pharyngeal washes, saliva, any other body fluid, blood, urine, stool, and the like. The biological sample may be treated to free the nucleic acid comprised in the biological sample before carrying out the amplification step. The target nucleic acid may be any nucleic acid which is intended to be detected. The target nucleic acid to be detected may be at least a nucleic acid exogenous to the nucleic acid of the biological sample. Accordingly, if the biological sample is from a human, the exogenous target nucleic acid to be detected (if present in the biological sample) is a nucleic acid which is not from human origin. According to an aspect of the invention, the target nucleic acid to be detected is at least a pathogen genome or fragment thereof. The pathogen nucleic acid may be at least a nucleic acid from a virus, a parasite, or bacterium, or a fragment thereof.

[0030] Accordingly, the invention provides a method of detection of at least a target nucleic acid, if present, in a biological sample. The method may be a diagnostic method for the detection of the presence of a pathogen in the biological sample. For example, if the biological sample is obtained from a human being, the target nucleic acid, if present in the biological sample, is not from human.

[0031] The probe(s) designed and prepared according to any method of the present invention may be used in solution or may be placed on an insoluble support. For example, the probe(s) may be applied, spotted or printed on an insoluble support according to any technique known in the art. The support may be a solid support or a gel. The support with the probes applied on it may be a microarray or a biochip.

[0032] The probes are then contacted with the nucleic acid(s) of the biological sample, and, if present, the target nucleic acid(s) and the probe(s) hybridize, and the presence of the target nucleic acid is detected. In particular, in the detection step (iv), the mean of the signal intensities of the probes which hybridize to v_a is statistically higher than the mean of the probes $\notin v_a$, thereby indicating the presence of v_a in the biological sample.

[0033] More in particular, in the detection step (iv), the mean of the signal intensities of the probes which hybridize to v_a is statistically higher than the mean of the probes $\notin v_a$, and the method further comprises the step of computing the relative difference of the proportion of probes $\notin v_a$ having high signal intensities to the proportion of the probes used in the detection method having high signal intensities, the density distribution of the signal intensities of probes v_a being more positively skewed than that of probes $\notin v_a$, thereby indicating the presence of v_a in the biological sample.

[0034] For example, in the detection step (iv), the presence of a target nucleic acid in a biological sample is given by a

value of t-test ≦0.1 and/or a value of Weighted Kullback-Leibler divergence of ≧1.0, preferably ≧5.0. In particular, the t-test value is ≦0.05.

[0035] According to another aspect, the present invention provides a method of determining the presence of a target nucleic acid v_a comprising detecting the hybridization of a probe (the probe being selected and designed according to any known method in the art and not necessary limited to the methods according to the present invention) to a target nucleic acid v_a and wherein the mean of the signal intensities of the probes which hybridize to v_a is statistically higher than the mean of the probes $\notin v_a$, thereby indicating the presence of v_a. In particular, the mean of the signal intensities of the probes which hybridize to v_a is statistically higher than the mean of the probes $\notin v_a$, and the method further comprises the step of computing the relative difference of the proportion of probes $\notin v_a$ having high signal intensities to the proportion of the probes used in the detection method having high signal intensities, the density distribution of the signal intensities of probes v_a being more positively skewed than that of probes $\notin v_a$, thereby indicating the presence of v_a. More in particular, the presence of a target nucleic acid in a biological sample is given by a value of t-test ≦·0.1 and/or a value of Weighted Kullback-Leibler divergence of ≧1.0, preferably ≧5.0. For example, the t-test value may be ≦0.05.

[0036] According to another aspect, the present invention provides a method of detecting at least a target nucleic acid, comprising the steps of:

[0037] (i) providing a biological sample;

[0038] (ii) amplifying the nucleic acid(s) of the biological sample;

[0039] (iii) providing at least one oligonucleotide probe capable to hybridize to at least a target nucleic acid, if present in the biological sample; and

[0040] (iv) contacting the probe(s) with the amplified nucleic acids and detecting the probe(s) hybridized to the target nucleic acid(s), wherein the mean of the signal intensities of the probes which hybridize to v_a is statistically higher than the mean of the probes $\notin v_a$, thereby indicating the presence of v_a in the biological sample.

[0041] In step (iv), the mean of the signal intensities of the probes which hybridize to v_a is statistically higher than the mean of the probes $\notin v_a$, and the method further comprises the step of computing the relative difference of the proportion of probes $\notin v_a$ having high signal intensities to the proportion of the probes used in the detection method having high signal intensities, the density distribution of the signal intensities of probes v_a being more positively skewed than that of probes $\notin v_a$, thereby indicating the presence of v_a in the biological sample. In particular, in step (iv) the presence of a target nucleic acid in a biological sample is given by a value of t-test ≦0.1 and/or a value of Weighted Kullback-Leibler divergence of ≧1.0, preferably ≧5.0. The t-test value may be ≦0.05. The nucleic acid to be detected is nucleic acid exogenous to the nucleic acid of the biological sample. The target nucleic acid to be detected may be at least a pathogen genome or fragment thereof. The pathogen nucleic acid may be at least a nucleic acid from a virus, a parasite, or bacterium, or a fragment thereof. In particular, when the sample is obtained from a human being, the target nucleic acid, if present in the biological sample, is not from the human genome. The probes may be placed on an insoluble support. The support may be a microarray or a biochip.

BRIEF DESCRIPTION OF THE FIGURES

[0042] FIG. **1** shows a RT-PCR binding process of a pair of random primers on a virus sequence.

[0043] FIG. **2** shows an Amplification Efficiency Scoring (AES) Map for the RSV B genome.

[0044] FIG. **3** shows signal intensities for 1 experiment for RSV B.

[0045] FIGS. **4**(A, B). FIG. **4**A shows the density distribution of signal intensities of a virus that is the sample tested. An arrow indicates the positive skewness of the distribution. This indicates that although there is noise, there is significant amount of real signals as well. FIG. **4**B shows the density distribution of signal intensities of a virus not in the sample. It is noise dominant.

[0046] FIG. **5** shows an analysis framework of pathogen detection chip data.

DETAILED DESCRIPTION OF THE INVENTION

[0047] Bibliographic references mentioned in the present specification are, for convenience, listed in the form of a list of references and added to the end of the examples. The whole content of such bibliographic references is herein incorporated by reference.

[0048] The present invention addresses the problems of the prior art, and in particular provides an alternative and/or improved method of probe design and/or of nucleic acids detection.

[0049] While the concept of using oligonucleotide hybridization microarrays as a tool for determining the presence of pathogens has been proposed, significant hurdles remain, thus preventing the use of these microarrays routinely (Striebel, H. M., 2003). These hurdles include probe design and data analysis (Striebel, H. M., 2003; Bodrossy, L. & Sessitsch, A., 2004; Vora, G. J., et al., 2004). The present inventors observed in a pilot microarray that despite meticulous probe selection, the best in silico designed probes do not necessarily hybridize well to patient samples. However, purified samples from cell-culture hybridized well to the probes on the microarray. The inventors realized that to generate probes which would hybridize consistently well to patient material, it would be necessary to develop a new and/or improved method of probe design so as to determine the optimal design predictors. In particular, as described in the Example section, the present inventors created a microarray comprising overlapping 40-mer probes, tiled across 35 viral genomes. However, the invention is not limited to this particular application, probe length and type of target nucleic acid.

[0050] According to a particular aspect of the invention, the present inventors describe how a support, in particular a microarray platform, is optimized so as to become a viable tool in target nucleic acid detection, in particular, in pathogen detection. The inventors also identified probe design

predictors, including melting temperature, GC-content of the probe, secondary structure, hamming distance, similarity to human genome, effect of PCR primer tag in random PCR amplification efficiency, and/or the effect of sequence polymorphism. These results were considered and/or incorporated into the development of a method and criteria for probe design. According to a more particular aspect, the inventors developed a data analysis algorithm which may accurately predict the presence of a target nucleic acid, which may or may not be a pathogen. For example the pathogen may be, but not limited to, a virus, bacteria and/or parasite(s). The algorithm may be used even if probes are not ideally designed. This detection algorithm, coupled with a probe design methodology, significantly improves the confidence level of the prediction (see Tables 1 and 2).

[0051] According to a particular aspect, the method of the invention may not require a prediction of the likely pathogen, but may be capable of detecting most known human viruses, bacteria and/or parasite(s), as well as some novel species, in an unbiased manner. Genome or a fragment thereof is defined as all the genetic material in the chromosomes of an organism. DNA derived from the genetic material in the chromosomes of a particular organism is genomic DNA. A genomic library is a collection of clones made from a set of randomly generated overlapping DNA fragments representing the entire genome of an organism. The rationale behind this detection platform according to the invention is that each species of virus, bacteria and/or parasite(s) contains unique molecular signatures within the primary sequence of their genomes. Identification of these distinguishing regions allows for rational oligonucleotide probe design for the specific detection of individual species, and in some cases, individual strains. The concomitant design and/or preparation of oligonucleotide (oligo) probes that represent the most highly conserved regions among family and genus members, will enable the detection and partial characterization of some novel pathogens. Furthermore, the inclusion of all such probes in a single support may allow the detection of multiple viruses, bacteria and/or parasite(s) that simultaneously co-infect a clinical sample. The support may be an insoluble support, in particular a solid support. For example, a microarray or a biochip assay.

[0052] According to a particular aspect, the invention may be used as a diagnostic tool, depending on the way in which oligonucleotide probes are designed, and/or how the data generated by the microarray is interpreted and analyzed.

Determination of Efficiency of Amplification

[0053] According to a first aspect, the present invention provides a method of designing oligonucleotide probe(s) for nucleic acid detection comprising the following steps in any order:

[0054] (i) identifying and selecting region(s) of a target nucleic acid to be amplified, the region(s) having an efficiency of amplification (AE) higher than the average AE; and

[0055] (ii) designing oligonucleotide probe(s) capable of hybridizing to the selected region(s).

[0056] In particular, in step (i) a score of AE is determined for every position i on the length of the target nucleic acid or of a region thereof and an average AE is obtained. Those regions showing an AE higher than the average may be selected as the region(s) of the target nucleic acid to be amplified. In particular, the AE of the selected region(s) may be calculated as the Amplification Efficiency Score (AES), which is the probability that a forward primer r_i can bind to a position i and a reverse primer r_j can bind at a position j of the target nucleic acid, and |i-j| is the region of the target nucleic acid desired to be amplified. In particular, the region |i-j| may be ≦10000 bp, more in particular ≦5000 bp, or ≦1000 bp, for example ≦500 bp. In particular, the forward and reverse primers may be random primers.

[0057] According to another aspect, the step (i) of identifying and selecting region(s) of a target nucleic acid to be amplified comprises determining the effect of geometrical amplification bias for every position of a target nucleic acid, and selecting the region(s) to be amplified as the region(s) having an efficiency of amplification (AE) higher than the average AE. The geometrical amplification bias may be defined as the capability of some regions of a nucleic acid to be amplified more efficiently than other regions. For example, the geometrical amplification bias is the PCR bias.

Modeling of Amplification Efficiency

[0058] Since it is not known what target nucleic acid (for example a pathogen) exists within the patient sample, random primers may be used during the amplification step and/or the reverse-transcription (RT) process to ensure unbiased reverse-transcription of all RNA present into DNA. Any random amplification method known in the art may be used for the purposes of the present invention. In the present description, the random amplification method will be RT-PCR. However, it will be clear to a skilled person that the method of the present invention is not limited to RT-PCR. In particular, the RT-PCR approach may be susceptible to signal inaccuracies caused by primer-dimer bindings and poor amplification efficiencies in the RT-PCR process (Bustin, S. A., et al, 2004). To overcome this hurdle, the inventors have modeled the RT-PCR process by using random primers.

[0059] According to a particular aspect of the invention, the amplification step comprises forward and reverse primers, and each of the forward and reverse primers comprises, in a 5'-3' orientation, a fixed primer header and a variable primer tail, and wherein at least the variable tail hybridizes to a portion of the target nucleic acid v_a. The size of the fixed primer header and that of the variable primer tail may be of any size, in mer, suitable for the purposes of the method according to the present invention. The fixed header may be 10-30 mer, in particular, 15-25 mer, for example 17 mer. The variable tail may be 1-20 mer, in particular, 5-15 mer, for example 9 mer. An example of these forward and reverse primers is shown in FIG. **1**. More in particular, the amplification step may comprise forward and/or reverse random primers having the nucleotide sequence 5'-GTTTCCCAGTCACGATANNNNNNNNN-3', (SEQ ID NO:1), wherein N is any one of A, T, C, and G or a derivative thereof.

[0060] In particular, the present inventors have modeled the random RT-PCR process as follows. Let v_a be the actual virus in the sample. The random primer used in the RT-PCR process has a fixed 17-mer header and a variable 9-mer tail of the form (5'-GTTTCCCAGTCACGATANNNNNNNN-3')(SEQ ID NO:1). To get a RT-PCR product in a region between positions i and j of v_a, the inventors required (1) a forward primer binding to position i, (2) |i-j|≦1000, and (3)

a reverse primer binding to position j. However, even if in the above particular example |i-j|, which is the region of the target nucleic acid desired to be amplified, is ≦1000, the region |i-j| may be ≦10000 bp, more in particular ≦5000 bp, or also ≦500 bp. The quality of the RT-PCR product depends on how well the forward primer and the reverse primer bind to v_a. Some random primers can bind to v_a better than others. The identification of such primers and where they bind to v_a gives an indication of how likely a particular region of v_a will be amplified. Using this approach, an amplification efficiency model may be proposed which computes an Amplification Efficiency Score (AES) for every position of v_a.

[0061] For a particular position i of a target nucleic acid v_a, P^f (i)and P^r (i) are the probabilities that a random primer r_i can bind to position i of v_a as forward primer and reverse primer respectively. For simplicity, it is assumed that a random primer can only bind to v_a if the last 9 nucleotides of the random primer is a substring of the reverse complement of v_a (forward primer) or a substring of v_a (reverse primer). This is shown in FIG. **1**. Based on well-established primer design criteria (Wu, D. Y., et al., 1991), the P^f (i) was estimated to be low if r_i forms a significant primer-dimer or has extreme melting temperature. On the other hand, if r_i does not form any significant primer-dimer and has optimal melting temperature, then P^f (i) will be high. Note that if the header of the random primer is similar to v_a, it may also aid in the binding and thus result in a higher P^f (i). Similarly, the P^r (i) was computed.

[0062] The binding of the random primer r_i at position i of v_a as a forward primer affects the quality of the RT-PCR product for at least 1000 nucleotides upstream of position i. Similarly, the binding of the random primer r_i at position i of v_a as a reverse primer affects the quality of the RT-PCR product for at least 10000 nucleotides downstream of position i. Thus, an amplification efficiency score, AES_i, for every position i of v_a can be computed by considering the combined effect of all forward and reverse primer-pairs that amplifies it:

$$AES_i = \sum_{j=i-Z}^{i} \left\{ P^f(j) \times \sum_{k=i+1}^{j+Z} P^r(k) \right\}$$

[0063] wherein

$$\sum_{k=i+1}^{j+Z} P^r(k) = P^r(i+1) + P^r(i+2) + \ldots P^r(j+Z)$$

[0064] P^f (i) and P^r (i) is the probability that a random primer r_i can bind to position i of v_a as forward primer and reverse primer, respectively, and Z ≦10000 bp is the region of v_a desired to be amplified.

[0065] Z may be ≦5000 bp, ≦1000 bp or ≦500 bp. For example, Z is ≦10000 bp.

[0066] To verify if the variation in signal intensities displayed by different regions of a virus has direct correlation with their corresponding amplification efficiency scores, a total of five microarray experiments were performed on a common pathogen affecting human, the human respiratory syncytial virus B (RSV B).

Probe Design

[0067] The step (ii) of designing oligonucleotide probe(s) capable of hybridizing to the selected region(s) may be selected to any one of the probe designing techniques known in the art.

[0068] For example, given a set of target nucleic acids (for example, viral genomes) $V=\{v_1, v_2, \ldots, v_n\}$, for every $v_i \in V$, a set of length-m probes (that is a substring of v_i) which satisfies the following conditions may be designed taking into consideration, for example, at least one of the following:

[0069] (a) established probe design criteria of homogeneity, sensitivity and specificity (Sung, W. K. & Lee, W. H., 2003);

[0070] (b) no significant sequence similarity to human genome; and

[0071] (c) efficiently amplified, for example by RT-PCR, as herein described.

Homogeneity, Sensitivity and Specificity

[0072] Homogeneity requires the selection of probes which have similar melting temperatures. It was found that probes with low CG-content did not produce reliable hybridization signal intensities, and that probes with high CG-content had a propensity to produce high signal intensities through non-specific binding. Thus, it could be established that the CG-content of probes selected should be from 40% to 60%.

[0073] Accordingly, the present invention provides a method of designing oligonucleotide probe(s) for nucleic acid detection, comprising selecting the probes having a CG-content from 40% to 60%.

[0074] The term "hybridization" refers to the process in which the oligo probes bind non-covalently to the target nucleic acid, or portion thereof, to form a stable double-stranded. Triple-stranded hybridization is also theoretically possible. Hybridization probes are oligonucleotides capable of binding in a base-specific manner to a complementary strand of target nucleic acid. Hybridizing specifically refers to the binding, duplexing, or hybridizing of a molecule substantially to or only to a particular nucleotide sequence or sequences under stringent conditions when that sequence is present in a complex mixture (e.g., total cellular) of DNA or RNA. Hybridizations, e.g., allele-specific probe hybridizations, are generally performed under stringent conditions. For example, conditions where the salt concentration is no more than about 1 Molar (M) and a temperature of at least 25° C., e.g., 750 mM NaCl, 50 mM NaPhosphate, 5 mM EDTA, pH 7.4 (5 times SSPE) and a temperature of from about 25° C. to about 30° C. Hybridization is usually performed under stringent conditions, for example, at a salt concentration of no more than 1 M and a temperature of at least 25° C. For stringent conditions, see also for example, Sambrook and Russel, Molecular Cloning: A Laboratory Manual, Cold Springs Harbor Laboratory, New York (2001) which is hereby incorporated by reference in its entirety for all purposes above.

[0075] Sensitivity requires that probes that cannot form significant secondary structures be selected in order to detect low-abundance mRNAs. Thus, probes with the highest free energy computed based on Nearest-Neighbor model are selected (SantaLucia, J., Jr., et al., 1996).

[0076] Accordingly, the present invention provides a method of designing oligonucleotide probe(s) for nucleic acid detection, wherein the probe(s) are selected by having the highest free energy computed based on Nearest-Neighbor model.

[0077] Specificity requires the selection of probes that are most unique to a viral genome. This is to minimize cross-hybridization of the probes with other non-target nucleic acids (for example, viral genomes). Given probe s_a and probe s_b substrings of target nucleic acids v_a and v_b, s_a is selected based on the hamming distance between s_a and any length-m substring s_b from the target nucleic acid v_b and/or on the longest common substring of s_a and probe s_b. In particular, let s_a and s_b be length-m substrings from viral genome v_a and v_b respectively, where ($v_a \neq v_b$).

[0078] The length of the probe(s) to be designed may be of any length useful for the purposes of the present invention. The probes may be less than 100 mer, for example 20 to 80 mer, 25 to 60 mer, for example 40 mer. The hamming distance and/or longest common substring may also vary.

[0079] According to Kane's criteria (Kane, M. D., et al., 2000), s_a is specific to v_a if:

[0080] (a) the hamming distance between s_a and any length-m substring s_b from viral genome v_b, is more than 0.25 m;

[0081] (b) the longest common substring of s_a and s_b is less than 15.

[0082] The cutoff value(s) for the hamming distance may be chosen according to the stringency desired. It will be evident to any skilled person how to select the hamming distance cutoff according to the particular stringency desired. According to a particular example of the herein described probe design, the inventors used hamming distance cutoffs of >10 for specific probes, and <10, preferably <5 for conserved probes. With a specific probe, it indicates a probe which only hybridizes to a specific target nucleic acid, while with a conserved probe it indicates a probe which may hybridize to any member of the family of the target nucleic acid.

[0083] Accordingly, the present invention also provides a method of designing oligonucleotide probe(s) for nucleic acid detection, wherein given probe s_a and probe s_b substrings of target nucleic acids v_a and v_b comprised in the biological sample, s_a is selected if the hamming distance between s_a and any length-m substring s_b from the target nucleic acid v_b is more than 0.25m, and the longest common substring of s_a and probe s_b is less than 15.

Sequence Similarity to Human Genome

[0084] In case the target nucleic acid to be detected is of human origin (for example, human samples containing viral genomes), probes with high homology to the human genome should also be avoided. Accordingly, for any probe s_a of length-m specific for the target nucleic acid v_a, the probe s_a is selected if it does not have any hits with any region of a nucleic acid different from the target nucleic acid, and if the probe s_a length-m has hits with the nucleic acid different from the target nucleic acid, the probe s_a length-m with the smallest maximum alignment length and/or with the least number of hits is selected. In particular, for any length-m probe s_a, hits of s_a with the human genome are found with the BLAST algorithm (Altschul, S. F., et al., 1997). A BLAST word size of (W=15) and an expectation value of 100 was used to find all hits. s_a is selected if it does not have any hits with the human genome, that is, it is specific to v_a. However, if all length-m substrings of v_a have hits with the human genome, those with the smallest maximum alignment length and with the least number of hits was selected.

[0085] Accordingly, the present invention provides a method of designing oligonucleotide probe(s) for nucleic acid detection, wherein for any probe s_a of length-m specific for the target nucleic acid v_a, the probe s_a is selected if it does not have any hits with any region of a nucleic acid different from the target nucleic acid, and if the probe s_a length-m has hits with the nucleic acid different from the target nucleic acid, the probe s_a length-m with the smallest maximum alignment length and/or with the least number of hits is selected.

[0086] Further, the design of the oligonucleotide probe(s) may be also carried out by AES according to the invention. In particular, the invention provides a method of selecting and designing probes wherein a probe p_i at position i of a target nucleic acid is selected if p_i is predicted to hybridize to the position i of the amplified target nucleic acid.

[0087] In particular, the oligonucleotide probe(s) capable of hybridizing to the selected region(s) may be selected and designed according to at least one of the following criteria:

[0088] (a) the selected probe(s) has a CG-content from 40% to 60%;

[0089] (b) the probe(s) is selected by having the highest free energy computed based on Nearest-Neighbor model;

[0090] (c) given probe s_a and probe s_b substrings of target nucleic acids v_a and v_b, s_a is selected based on the hamming distance between s_a and any length-m substring s_b from the target nucleic acid v_b and/or on the longest common substring of s_a and probe s_b;

[0091] (d) for any probe s_a of length-m specific for the target nucleic acid v_a, the probe s_a is selected if it does not have any hits with any region of a nucleic acid different from the target nucleic acid, and if the probe s_a length-m has hits with the nucleic acid different from the target nucleic acid, the probe s_a length-m with the smallest maximum alignment length and/or with the least number of hits is selected; and

[0092] (e) a probe p_i at position i of a target nucleic acid is selected if p_i is predicted to hybridize to the position i of the amplified target nucleic acid.

[0093] According to a particular aspect of the invention, two or more of the criteria indicated above may be used for designing the oligonucleotide probe(s). For example, the probe(s) may be designed by applying all criteria (a) to (e). Other criteria, not explicitly mentioned herein but which are evident to a skilled person in the art may also be used.

[0094] In particular, under the criterion (e), a probe p_i at position i of a target nucleic acid v_a is selected if $P(p_i|v_a)>\lambda$, wherein λ is 0.5 and $P(p_i|v_a)$ is the probability that p_i has to hybridize to the position i of the target nucleic acid v_a. More in particular, λ is 0.75.

[0095] According to another aspect, the invention provides a method as above described wherein $P(p_i|v_a)\approx P(X\leqq x_i)=c_i/k$, wherein X is the random variable representing the amplification efficiency score (AES) values of all probes of v_a, k is the number of probes in v_a, and c_i is the number of probes whose AES values are $\leqq x_i$.

Synthesis of Oligonucleotide Probes on a Support

[0096] According to another aspect of the invention, the method of selecting and designing the oligonucleotide probe(s) as described above further comprises a step of preparing the selected and designed probe(s). Designing a probe comprises understanding its sequence and/or designing it by any suitable means, for example by using a software. The step of preparing the probe comprises the physical preparation of it. The probe may be prepared according to any standard method known in the art. For example, the probes may be chemically synthesized or prepared by cloning. For example, as described in Sambrook and Russel, 2001.

[0097] The probe(s) designed and prepared according to any method of the present invention may used in solution or may be placed on an insoluble support. For example, may be applied, spotted or printed on an insoluble support according to any technique known in the art. The support may be a solid support or a gel. The support with the probes applied on it, may be a microarray or a biochip.

[0098] More in particular, the present invention provides an oligo microarray hybridization-based approach for the rapid detection and identification of pathogens, for example viral and/or bacterial pathogens, from PCR-amplified cDNA prepared from primary tissue samples. In particular, from random PCR-amplified cDNA(s).

[0099] In the following description, the preparation of probes is made with particular reference to a microarray. However, the support, as well as the probes, may be prepared according to any description across the whole content of the present application. In particular, an "array" is an intentionally created collection of molecules which can be prepared either synthetically or biosynthetically. The molecules in the array can be identical or different from each other. The array can assume a variety of formats, e.g., libraries of soluble molecules; libraries of compounds tethered to resin beads, silica chips, or other solid supports. Array Plate or a Plate is a body having a plurality of arrays in which each array is separated from the other arrays by a physical barrier resistant to the passage of liquids and forming an area or space, referred to as a well.

Sample Preparation and Hybridization onto the Microarray

[0100] The biological sample may be any sample taken from a mammal, for example from a human being. The biological sample may be blood, a body fluid, saliva, urine, stool, and the like. The biological sample may be treated to free the nucleic acid comprised in the biological sample before carrying out the amplification step. The target nucleic acid may be any nucleic acid which is intended to be detected. The target nucleic acid to be detected may be at least a nucleic acid exogenous to the nucleic acid of the biological sample. Accordingly, if the biological sample is from a human, the exogenous target nucleic acid to be detected (if present in the biological sample) is a nucleic acid which is not from human origin. According to an aspect of the invention, the target nucleic acid to be detected is at least a pathogen genome or fragment thereof. The pathogen nucleic acid may be at least a nucleic acid from a virus, a parasite, or bacterium, or a fragment thereof.

[0101] According to an aspect of the present invention, there is provided a method of target nucleic acid detection analysis. The target nucleic acid(s) from a biological sample desired to be detected may be any target nucleic acid, RNA and/or DNA. For example, mRNA and/or cDNA. More in particular, the target nucleic acid to be detected may be a pathogen or non-pathogen. For example, it may be the genome or a fragment thereof of at least one virus, at least one bacterium and/or at least one parasite. The probes selected and/or prepared may be placed, applied and/or fixed on a support according to any standard technology known to a skilled person in the art. The support may be an insoluble support, for example a solid support. In particular, a microarray and/or a biochip.

[0102] According to a particular example, RNA and DNA was extracted from patient samples e.g. tissues, sera, nasal pharyngeal washes, stool using established protocols and commercial kits. For example, Qiagen Kit for nucleic acid extraction may be used. Alternatively, Phenol/Chloroform may also be used for the extraction of DNA and/or RNA. Any technique known in the art, for example as described in Sambrook and Russel, 2001 may be used. RNA was reverse-transcribed to CDNA using tagged random primers, based on a protocol described by Bohlander et. al., 1992 and Wang et. al., 2003. The cDNA was then amplified by random PCR. Fragmentation, labeling and hybridization of sample to the microarray were carried out as described by Wong et. al., 2004.

Microarray Synthesis

[0103] According to a particular experiment described in the Examples section, the present inventors selected 35 viral genomes representing the most common causes of viral disease in Singapore. Using the complete genome sequences downloaded from Genbank, 40-mer probes which tiled across the entire genomes and overlapping at five-base resolution were generated. Seven replicates of each virus probe were synthesized directly onto the microarray using Nimblegen technology (Nuwaysir, E. F., et al., 2002). The probes were randomly distributed on the microarray to minimize the effects of hybridization artifacts. To control the non-specific hybridization of sample to probes, 10,000 oligonucleotide probes were designed and synthesized onto the microarray. These 10,000 oligonucleotides did not have any sequence similarity to the human genome, or to the pathogen genomes. They were random probes with 40-60% CG-content. These probes measured the background signal intensity. As a positive control, 400 oligonucleotide probes to human genes which have been known or inferred functions in immune response were synthesized on the array. A plant virus, PMMV, was included as a negative control, for a total of approximately 380,000 probes. In the following description, the invention will be described in more particu-

larity with reference to a pathogen detection chip analysis (also referred to as PDC). However, the analysis (method) is not limited to this particular embodiment, but encompasses the several aspects of the invention as described across the whole content of the present application.

Non-Specific Hybridization Controls

[0104] To control the non-specific hybridization of sample to probes, 10,000 oligonucleotide probes were designed and synthesized onto the microarray. These 10,000 oligonucleotides did not have any sequence similarity to the human genome, or to the pathogen genomes. They were random probes with 40-60% CG-content. These probes measured the background signal intensity.

Method of Detecting Target Nucleic Acid(s)

[0105] According to another aspect, the present invention provides a method of detecting at least one target nucleic acid comprising the step of:

[0106] (i) providing a biological sample;

[0107] (ii) amplifying the nucleic acid(s) of the biological sample;

[0108] (iii) providing at least one oligonucleotide probe capable of hybridizing to at least a target nucleic acid, if present in the biological sample, wherein the probe(s) is prepared by using a method according to any aspect of the invention herein described;

[0109] (iv) contacting the probe(s) with the amplified nucleic acids and detecting the probe(s) hybridized to the target nucleic acid(s).

[0110] The amplification step (ii) may be carried out in the presence of random primers. For example, the amplification step (ii) may be carried out in the presence of more than two random primers. Any amplification method known in the art may be used. For example, the amplification is a RT-PCR.

[0111] In particular, the present inventors developed a method of detecting the probe(s) hybridized to the to the target nucleic acid based on the amplification efficiency score (AES). This may herein also be referred to as the algorithm according to the present invention. In particular, a forward random primer binding to position i and a reverse random primer binding to position j of a target nucleic acid v_a are selected among primers having an amplification efficiency score (AES_i) for every position i of a target nucleic acid v_a of:

$$AES_i = \sum_{j=i-Z}^{i} \left\{ P^f(j) \times \sum_{k=i+1}^{j+Z} P^r(k) \right\},$$

wherein

$$\sum_{k=i+1}^{j+Z} P^r(k) = P^r(i+1) + P^r(i+2) + \ldots P^r(j+Z)$$

[0112] P^f (i)and P^r (i)are the probabilities that a random primer r_i can bind to position i of v_a as forward primer and reverse primer, respectively, and Z≦10000 bp is the region of v_a desired to be amplified. More in particular, Z may be ≦5000 bp, ≦1000 bp, or ≦500 bp.

[0113] The amplification step may comprise forward and reverse primers, and each of the forward and reverse primers may comprise, in a 5'-3' orientation, a fixed primer header and a variable primer tail, and wherein at least the variable tail hybridizes to a portion of the target nucleic acid v_a. In particular, the amplification step may comprise forward and/or reverse random primers having the nucleotide sequence of SEQ ID NO:1 or a variant or derivative thereof.

[0114] The biological sample may be any sample taken from a mammal, for example from a human being. The biological sample may be tissue, sera, nasal pharyngeal washes, saliva, any other body fluid, blood, urine, stool, and the like. The biological sample may be treated to free the nucleic acid comprised in the biological sample before carrying out the amplification step. The target nucleic acid may be any nucleic acid which is intended to be detected. The target nucleic acid to be detected may be at least a nucleic acid exogenous to the nucleic acid of the biological sample. Accordingly, if the biological sample is from a human, the exogenous target nucleic acid to be detected (if present in the biological sample) is a nucleic acid which is not from human origin. According to an aspect of the invention, the target nucleic acid to be detected is at least a pathogen genome or fragment thereof. The pathogen nucleic acid may be at least a nucleic acid from a virus, a parasite, or bacterium, or a fragment thereof.

[0115] Accordingly, the invention provides a method of detection of at least a target nucleic acid, if present, in a biological sample. The method may be a diagnostic method for the detection of the presence of a pathogen into the biological sample. For example, if the biological sample is obtained from a human being, the target nucleic acid, if present in the biological sample, is not from human.

[0116] The probe(s) designed and prepared according to any method of the present invention may used in solution or may be placed on an insoluble support. For example, may be applied, spotted or printed on an insoluble support according to any technique known in the art. The support may be a solid support or a gel. The support with the probes applied on it, may be a microarray or a biochip.

[0117] The probes are then contacted with the nucleic acid of the biological sample, and if present the target nucleic acid(s) and the probe(s) hybridize, and the presence of the target nucleic acid is detected. In particular, in the detection step (iv), the mean of the signal intensities of the probes which hybridize to v_a is statistically higher than the mean of the probes $\notin v_a$, thereby indicating the presence of v_a in the biological sample.

[0118] More in particular, in the detection step (iv), the mean of the signal intensities of the probes which hybridize to v_a is statistically higher than the mean of the probes $\notin v_a$, and the method further comprises the step of computing the relative difference of the proportion of probes $\notin v_a$ having high signal intensities to the proportion of the probes used in the detection method having high signal intensities, the density distribution of the signal intensities of probes v_a being more positively skewed than that of probes $\notin v_a$, thereby indicating the presence of v_a in the biological sample.

[0119] For example, in the detection step (iv), the presence of a target nucleic acid in a biological sample is given by a

value of t-test ≧0.1 and/or a value of Weighted Kullback-Leibler divergence of ≧1.0, preferably ≧5.0. In particular, the t-test value is ≧0.05.

[0120] According to another aspect, the present invention provides a method of determining the presence of a target nucleic acid v_a comprising detecting the hybridization of a probe to a target nucleic acid v_a and wherein the mean of the signal intensities of the probes which hybridize to v_a is statistically higher than the mean of the probes $\notin v_a$, thereby indicating the presence of v_a. In particular, the mean of the signal intensities of the probes which hybridize to v_a is statistically higher than the mean of the probes $\notin v_a$, and the method further comprises the step of computing the relative difference of the proportion of probes $\notin v_a$ having high signal intensities to the proportion of the probes used in the detection method having high signal intensities, the density distribution of the signal intensities of probes v_a being more positively skewed than that of probes $\notin v_a$, thereby indicating the presence of v_a. More in particular, the presence of a target nucleic acid in a biological sample is given by a value of t-test ≦0.1 and/or a value of Weighted Kullback-Leibler divergence of ≧1.0, preferably, ≧5.0. For example, the t-test value may be ≦0.05.

[0121] According to another aspect, the present invention provides a method of detecting at least one target nucleic acid, comprising the steps of:

[0122] (i) providing a biological sample;

[0123] (ii) amplifying the nucleic acid(s) of the biological sample;

[0124] (iii) providing at least one oligonucleotide probe capable of hybridizing to at least a target nucleic acid, if present in the biological sample;

[0125] (iv) contacting the probe(s) with the amplified nucleic acids and detecting the probe(s) hybridized to the target nucleic acid(s), wherein the mean of the signal intensities of the probes which hybridize to v_a is statistically higher than the mean of the probes $\notin v_a$, thereby indicating the presence of v_a in the biological sample.

[0126] In step (iv), the mean of the signal intensities of the probes which hybridize to v_a is statistically higher than the mean of the probes $\notin v_a$, and the method further comprises the step of computing the relative difference of the proportion of probes $\notin v_a$ having high signal intensities to the proportion of the probes used in the detection method having high signal intensities, the density distribution of the signal intensities of probes v_a being more positively skewed than that of probes $\notin v_a$, thereby indicating the presence of v_a in the biological sample. In particular, in step (iv) the presence of a target nucleic acid in a biological sample is given by a value of t-test ≦0.1 and/or a value of Weighted Kullback-Leibler divergence of ≧1.0, preferably ≧5.0. The t-test value may be ≦0.05. The nucleic acid to be detected is nucleic acid exogenous to the nucleic acid of the biological sample. The target nucleic acid to be detected may be at least a pathogen genome or fragment thereof. The pathogen nucleic acid may be at least a nucleic acid from a virus, a parasite, or bacterium, or a fragment thereof. In particular, when the sample is obtained from a human being, the target nucleic acid, if present in the biological sample, is not from the human genome. The probes may be placed on an insoluble support. The support may be a microarray or a biochip.

Test Using the Template Sequence of RSV B

[0127] To verify if the variation in signal intensities displayed by different regions of a virus has direct correlation with their corresponding amplification efficiency scores, a total of five microarray experiments were performed on a common pathogen affecting human, the human respiratory syncytial virus B (RSV B).

[0128] Next, the probe design criteria, as described above, were applied on the template sequence of RSV B obtained from NCBI (NC_001781). This resulted in 1948 probes spotted onto each microarray. The amplification efficiency map for RSV B was also computed prior to the actual experiments and shown in FIG. **2**. This figure shows the peaks having the AES higher than the average AES and indicating the regions of the RSV B with higher probability of amplification.

[0129] Using 5 samples containing the human respiratory syncytial virus B (RSV B), independent microarray experiments were conducted. The resultant signal intensities for one such experiment is shown in FIG. **3**.

[0130] For each experiment, the signal intensities of the 1948 probes were ranked in decreasing order and were correlated with their corresponding AES value. The p-value was found to be $<2.2e^{-}16$ on the average. This indicates that the correlation between the signal intensity of probe at position i of RSV B with AES_i is not at all random. Further investigations revealed that about 300 probes, which consistently produced high signal intensities in all five experiments, have amplification efficiency scores in the 90^{th} percentile level.

[0131] Having shown that the described amplification efficiency model works well on the RSV B genome, it was desired to show that the model according to the invention may be extended to other viral genomes as well. Another microarray experiment was performed on the human metapneumonia virus (HMPV). This time, there were 1705 probes on the microarray. Again, the amplification efficiency map for HMPV was computed. In this experiment, the correlation test between signal intensities and amplification efficiency scores gave a p-value of $1.335e^{-9}$.

[0132] Thus, there is a strong indication that the amplification efficiency model according to the invention is able to predict the relative strength of signals produced by different regions of a viral genome in the described experiment set-up. Probes from regions with low amplification efficiency scores have a high tendency to produce no or low signal intensities. This would result in a false negative on the microarray. Such probes will complicate the analysis of the microarray data and this is made even more complicated since a probe with a low signal intensity may be due to its target genome not being present or simply that it was not amplified. As such, probes in regions with reasonably high amplification efficiency scores should be selected to minimize inaccuracies caused by the RT-PCR process using random primers.

[0133] The threshold for amplification efficiency scores for probe selection for a virus v_a is determined by the cumulative distribution function of the AES values v_a. Let X be the random variable representing the AES values of all probes of v_a. Let k be the number of probes in v_a. Then, we denote the probability that the AES value is less than or equal to x be $P(X \leqq x)=c/k$, where c is the number of probes which have AES values less than or equal to x. For a probe p_i at position i of v_a, let x_i be its corresponding AES value.

Since the signal intensity of a probe is highly correlated to its AES value, we estimate $P(p_i|v_a)$, the probability that p_i has high signal intensity in the presence of v_a, to be $P(X \leqq x_i)$. Thus,

$$P(p_i \mid v_a) \approx P(X \leq x_i) = \frac{c_i}{k}$$

where c_i is the number of probes whose AES values are less than or equal to x_i.

[0134] For probe selection, probe p_i is selected if $P(p_i|v_a)>\lambda$. In the present experiments, λ was set as $\lambda=0.75$.

[0135] Accordingly, the present invention also provides a method of probe design and/or of target nucleic acid detection wherein a probe p_i at position i of a target nucleic acid v_a is selected if $P(p_i|v_a)>\lambda$, wherein λ is 0.75 and $P(p_i|v_a)$ is the probability that p_i has a high signal intensity in the presence of v_a. More in particular,

$$P(p_i \mid v_a) \approx P(X \leq x_i) = \frac{c_i}{k},$$

wherein X is the random variable representing the amplification efficiency score (AES) values of all probes of v_a, k is the number of probes in v_a, and c_i is the number of probes whose AES values are less than or equal to x_i.

Target Nucleic Acid Detection Analysis

[0136] In the following description, the invention will be described in more particularity with reference to a pathogen detection chip analysis (also referred to as PDC). However, the analysis (method) is not limited to this particular embodiment, but encompasses the several aspects of the invention as described across the whole content of the present application. Therefore, in particular, given a PDC with a set of length-m probes $P=\{p_1, p_2, \ldots, p_l\}$, which is designed for a set of viral genomes $V=\{v_1, v_2, \ldots, v_n\}$, the pathogen detection chip analysis problem is to detect the virus present in the sample based on the chip data. The chip data here refers to the collective information provided by the probe signals on the PDC. Thus, the chip data $D=\{d_i, d_2, \ldots, d_x\}$ is the set of corresponding signals of the probe set P on the PDC.

[0137] Given a sample, it is not known what pathogens are present in the sample, how many different pathogens there are, if present at all. However, if a virus v_a is indeed in the sample, then the signal intensities of the probes of v_a should differ significantly from the signal intensities of probes from other viruses. Specifically, a higher proportion of probes of v_a should have high signal intensities compared to other viruses. Hence, it would be expected that the mean of the signal intensities of the probes in v_a should be statistically higher than that of probes $\notin v_a$.

[0138] Accordingly, the invention provides a method wherein the mean of the signal intensities of the probes which hybridize to v_a is statistically higher than the mean of the probes $\notin v_a$, which may indicate the presence of v_a in the biological sample.

[0139] However, having a statistically higher mean may still be insufficient to conclude that v_a is in the sample. Preferably, an additional step may be required. We need to compute the relative difference of the proportion of probes $\notin v_a$ having high signal intensities to the proportion of probes on the PDC having high signal intensities. This is based on the observation that the distribution of the signal intensities of probes $\in v_a$ is more positively skewed than that of probes $\notin v_a$ (see the arrow in FIG. **4**A. For comparison see FIG. **4**B).

[0140] Based the above observations, the chip data D for the presence of viruses was analyzed as follows. For every virus $v_a \in V$, we used a one-tail t-test (Goulden, C. H., 1956) to determine if the mean of the signal intensities of the probes $\in v_a$ was statistically higher than that of the signal intensities of the probes $\notin v_a$. Thus, the t-statistic was computed:

$$t_i = \frac{\mu_a - \mu_{a'}}{\sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_{a'}^2}{n_{a'}}}}$$

where μ_a, σ_a^2 and n_a is the mean, variance, and size of the signal intensities of the probes $\in v_a$ respectively and μ_s, σ_a^2, and n_a, is the mean, variance, and size of the signal intensities of the probes $\notin v_a$ respectively.

[0141] To test the significance of the difference, the level of significance was set to 0.05. This means that the hypothesis that the mean of the signal intensities of the probes $\in v_a$ is higher than that of the signal intensities of the probes $\notin v_a$ would only be accepted if the p-value of $t_a<0.05$. In this case, v_a is likely to be present in the sample.

[0142] The t-test alone, which allows the inventors to know if the distribution of the signal intensities of a virus is different from that of other viruses, may not be sufficient to determine if a particular virus is in the sample. It is also essential to know how similar or different the two distributions are. A ruler that can be used to measure the similarity between a true distribution and a model distribution is the Kullback-Leiber divergence (Kullback and Leiber, 1951) (also known as the relative entropy). In this application, the probability distribution of the signal intensities of the probes in v_a is the true distribution while the probability distribution of the signal intensities of all the probes in P is the model distribution. Let P_a be the set of probes in v_a. The Kullback-Leibler (KL) divergence of the probability distribution of the signal intensities of P_a and P is:

$$KL(P_a \,||\, P) = \sum_{\mu \leq x \leq \max(D)} f_a(x) \log\left(\frac{f_a(x)}{f(x)}\right)$$

where μ is the mean signal intensity of the probes in P; $f_a(x)$ is the fraction of probes in P_a with signal intensity x; and f(x) is the fraction of probes in P with signal intensity x. It follows that if $KL(P_a||P)=0$ then the probability distribution of P_a is exactly the same as that of P. Otherwise they are different.

[0143] Since a virus that is present in the sample would have signal intensities higher than that of the population, this implies that v_a has a chance of being present in the sample if $KL(P_a||P)>0$. Thus, the larger the value of $KL(P_a||P)$, the

more different are the two probability distributions and the more likely that v_a is indeed present in the sample.

[0144] It is important to note that the Kullback-Leibler divergence is the collective difference over all x of two probability distributions. Thus, while the Kullback-Leibler divergence is good at finding shifts in a probability distribution, it is not always so good at finding spreads, which affect the tails of the probability distribution more. As described in FIG. **4**(A,B), the tails of the probability distribution provides the most information about whether a virus is present in the sample. Hence, the Kullback-Leibler divergence statistic must be improved to reflect more accurately such an observation.

[0145] To increase its sensitivity out on the tails, we introduced a stabilized or weighted statistic to the Kullback-Leibler divergence, the Anderson-Darling statistic (Stephens, M. A. (1974). EDF Statistics for Goodness of Fit and Some Comparisons, Journal of the American Statistical Association, Vol. 69, pp. 730-737). Thus the Weighted Kullback-Leibler divergence (WKL) is:

$$WKL(P_a \,||\, P) = \sum_{\mu \le x \le \max(D)} \frac{f_a(x) \log \frac{f_a(x)}{f(x)}}{\sqrt{Q(x)[1 - Q(x)]}}$$

where Q(x) is the cumulative distribution function of the signal intensities of the probes in P.

[0146] Empirical tests show that in samples where there are no viruses, viruses that pass the t-test with significance level 0.05 have WKL<5.0. In samples where there is indeed a virus present, the actual viruses not only pass the t-test with significance level 0.05 but are also the only viruses to have WKL>5.0. Thus we set the Weighted Kullback-Leiber divergence threshold for a virus to be present in the sample to be 5.0.

[0147] This analysis framework is shown in FIG. **5**.

[0148] Having now generally described the invention, the same will be more readily understood through reference to the following examples, which are provided by way of illustration, and are not intended to be limiting of the present invention.

EXAMPLES

[0149] Standard molecular biology techniques known in the art and not specifically described were generally followed as described in Sambrook and Russel, Molecular Cloning: A Laboratory Manual, Cold Springs Harbor Laboratory, New York (2001).

[0150] The present inventors present 2 sets of experiments to demonstrate the effects of probe design on experimental results and then to show the robustness of the analysis algorithm according to the present invention.

Example 1

[0151] In the first set of experiments, a PDC containing 53555 40-mer probes from 35 viruses affecting human was used for 4 independent microarray experiments. These 53555 probes were chosen based on a 5-bps tiling of each virus and were not subjected to any of our probe design criteria. Thus, we would expect errors arising due to CG-content, cross-hybridization and inefficient amplification to be significantly more than that of a PDC with well-designed probes. We tested our analysis algorithm in such an adverse setting for 4 experiments.

[0152] In each experiment, a human sample with an unknown pathogen was amplified by the RT-PCR process using random probes and then hybridized onto the PDC. We subjected the probes for each of the 35 viruses on our PDC to the one-tailed t-test with significance level 0.05 and computed the Weighted Kullback-Leibler (WKL) divergence of their signal intensities to the signal intensities of all the probes on the chip to determine which virus was in the sample for each experiment. Confirmation of the accuracy of the analysis by our program was done by wet-lab PCR to identify the actual virus in the sample. We present the results of our analysis for the 4 experiments and their corresponding PCR verifications in Table 1.

TABLE 1

Analysis results done on a PDC with no probe design criteria applied. The virus determined by our analysis algorithm to be the actual virus in the sample tested for each experiment is highlighted in light gray colour.

	Experiment 1		Experiment 2		Experiment 3		Experiment 4	
Sample Name	35259_324		35179_122		35253_841		35915_111	
ID	53555		53555		53555		53555	
Viruses (Accession No.)	t-test p-value	WKL	t-test p-value	WKL	t-test p-value	WKL	t-test p-value	KL
NC_001781.1	0	16.391	1	NA	0	10.85635	1	NA
NC_003461.1	1	NA	1	NA	1	NA	1	NA
NC_003443.1	0.999324	NA	0.873017	NA	0.99802	NA	0.999961	NA
NC_001796.2	1	NA	1	NA	1	NA	1	NA
AY283794.1	0	0.5435	0.108141	NA	0	0.775959	0	0.435427
NC_005147.1	0	1.2896	1	NA	0	1.399591	0	1.762912
NC_002645.1	0	1.2943	0.999847	NA	0	1.655888	0	2.079334
NC_004148.2	1	NA	0.002733	5.762907	1	NA	1	NA
NC_002023.1								
NC_002022.1								
NC_002021.1								
NC_002020.1	1	NA	0.579561	NA	1	NA	1	NA
NC_002019.1								
NC_002018.1								
NC_002017.1								

TABLE 1-continued

Analysis results done on a PDC with no probe design criteria applied. The virus determined by our analysis algorithm to be the actual virus in the sample tested for each experiment is highlighted in light gray colour.

NC_002204.1								
NC_002205.1								
NC_002206.1								
NC_002207.1	1	NA	1	NA	1	NA	1	NA
NC_002208.1								
NC_002209.1								
NC_002210.1								
NC_002211.1								
NC_001563.2	1	NA	0.000001	0.537826	1	NA	0.995013	NA
NC_002031.1	1	NA	0.000005	0.758758	0.998873	NA	0.363947	NA
NC_002728.1	1	NA	0.999062	NA	1	NA	1	NA
NC_002617.1	0.999994	NA	0	0.571844	1	NA	0.769098	NA
NC_001802.1	1	NA	0.999966	NA	1	NA	1	NA
NC_003977.1	0	2.7424	0	2.189827	0	3.978747	0	1.490665
NC_001576.1	0.371224	NA	0.004643	0.94841	0.009599	1.257041	0	3.961532
NC_002554.1	0.000062	0.7146	0	1.527292	0.299334	NA	0.000002	0.166239
NC_001545.1	0	1.4545	0	2.438558	0	0.869782	0	0.989592
NC_001489.1	0	1.7088	0.319125	NA	0	2.593065	0	1.510399
NC_005222.1	0.999757	NA	0.646314	NA	0.773912	NA	0.807875	NA
NC_005217.1	0.60477	NA	0.999903	NA	0.354358	NA	0.000871	0.626818
NC_004394.1	0	1.8411	0.000523	0.902399	0	2.43215	0.000007	0.538531
NC_004291.1	0.662386	NA	0.954137	NA	0.255422	NA	0.099148	NA
NC_001437.1	1	NA	0	0.593093	1	NA	1	NA
AB189128.1	1	NA	0.906213	NA	1	NA	1	NA
AF326573.1	1	NA	0.038503	0.539783	1	NA	1	NA
AF489932.1	1	NA	0.899797	NA	1	NA	1	NA
M87512.1	1	NA	0.759668	NA	1	NA	1	NA
NC_001430.1	1	NA	0.912496	NA	1	NA	0.999912	NA
NC_001428.1	0.999988	NA	0.284792	NA	0.999346	NA	0.957164	NA
NC_001612.1	0.970379	NA	0.000001	0.557865	0.998878	NA	0.061226	NA
NC_003986.1	1	NA	0.000012	0.604474	1	NA	0.997945	NA
NC_001472.1	0.999999	NA	0.0046	0.455194	0.999579	NA	0.143404	NA
NC_001617.1	0.721465	NA	0.98373	NA	0.178733	NA	0.414209	NA
NC_001490.1	0.999808	NA	0.995029	NA	0.997369	NA	0.859025	NA
Deduction	NC_001781.1 (RSV)		NC_004148.2 (HMPV)		NC_001781.1 (RSV)		None	
Virus Confirmation Virus (PCR)	NC_001781.1 (RSV)		NC_004148.2 (HMPV)		NC_001781.1 (RSV)		None	

[0153] The present results show that the analysis algorithm accurately deduces the actual virus in the sample tested in the first 3 experiments. Furthermore, we were able to deduce that the sample has no viruses in the last experiment. Note that if we had just used the t-test with level of significance 0.05, then the number of viruses detected to be present for each sample is shown in Table 2.

TABLE 2

False positive detection of viruses using t-test alone

	Sample Name			
	35259_324	35179_122	35253_841	35915_111
Viruses Detected Using T-test	9	14	9	10
False Positives	8	13	8	10
Max KL divergence (>5.0)	16.391	5.76	10.85	—
Viruses Detected Using T-test followed by KL divergence	1	1	1	0

TABLE 2-continued

False positive detection of viruses using t-test alone

	Sample Name			
	35259_324	35179_122	35253_841	35915_111

[0154] By using the Weighted Kullback-Leibler divergence of the viruses that pass the t-test, we were able to remove all false positive viruses and identify the actual virus. Thus, our analysis algorithm can robustly determine the virus under a high level of noise.

[0155] Next, we investigated the effects of using a PDC with probe design criteria applied on our analysis results. Firstly, the amplification efficiency map for each of the 35 viruses was computed. Then, the exact 53555 probes on the original PDC were subjected to probe design criteria. Probes which had extreme levels of CG-conten, high similarity to human and non-target viruses, and low amplifictaion efficiency scores were removed from the chip. A total of 10955 probes were retained for the second set of experiments. Using the samples used if the first set of experiments, we repeated the 4 experiments with the new chip. The experimental results are presented in Table 3.

TABLE 3

Analysis results done on a PDC with probe design criteria applied. The virus determined by our analysis algorithm to be the actual virus in the sample tested for each experiment is highlighted in light gray colour.

	Experiment 1		Experiment 2		Experiment 3		Experiment 4	
Sample Name	35259_324		35179_122		35253_841		35915_111	
ID	10955		10955		10955		10955	
Viruses (Accession No.)	t-test p-value	WKL	t-test p-value	WKL	t-test p-value	WKL	t-test p-value	KL
NC_001781.1	0	18.54859	1	NA	0	11.17914	1	NA
NC_003461.1	1	NA	1	NA	1	NA	1	NA
NC_003443.1	0.548718	NA	0.53727	NA	0.002783	0.837121	0.020436	0.603552
NC_001796.2	1	NA	0.999907	NA	1	NA	1	NA
AY283794.1	0	1.347801	0.024116	0.858364	0	1.523272	0	1.128637
NC_005147.1	0	1.604381	0.999697	NA	0	2.150019	0	2.893555
NC_002645.1	0	2.802742	0.999895	NA	0	4.612482	0	3.635771
NC_004148.2	1	NA	0.000003	9.324785	1	NA	1	NA
NC_002023.1								
NC_002022.1								
NC_002021.1								
NC_002020.1	1	NA	0.124517	NA	1	NA	0.999163	NA
NC_002019.1								
NC_002018.1								
NC_002017.1								
NC_002204.1								
NC_002205.1								
NC_002206.1								
NC_002207.1	1	NA	0.998724	NA	1	NA	1	NA
NC_002208.1								
NC_002209.1								
NC_002210.1								
NC_002211.1	0.986443	NA	0.428418	NA	0.76002	NA	0.112011	NA
NC_001563.2	0.998103	NA	0.003435	2.52162	0.278672	NA	0.409527	NA
NC_002031.1	0.999375	NA	0.30951	NA	0.969492	NA	0.297244	NA
NC_002728.1	0.63418	NA	0.003578	0.965856	0.247148	NA	0.025188	0.861163
NC_002617.1	1	NA	0.998118	NA	1	NA	1	NA
NC_001802.1	0	3.062956	0.000028	3.027442	0	4.574591	0	3.277708
NC_003977.1	0.579342	NA	0.101093	NA	0.155219	NA	0.026417	3.280335
NC_001576.1	0.6722	NA	0	2.289379	0.80654	NA	0.106683	NA
NC_002554.1	0	2.225817	0	2.794877	0.000019	1.674329	0	1.97064
NC_001545.1	0.099427	NA	0.999985	NA	0.000336	1.829543	0.000006	3.023235
NC_001489.1	0.999735	NA	0.294141	NA	0.974031	NA	0.356952	NA
NC_005222.1	0.916186	NA	0.994358	NA	0.600759	NA	0.032616	2.105628
NC_005217.1	0.867625	NA	0.235197	NA	0.100961	NA	0.052759	NA
NC_004294.1	0.992032	NA	0.964128	NA	0.714211	NA	0.206422	NA
NC_004291.1	1	NA	0.001058	1.563913	1	NA	0.857228	NA
NC_001437.1	1	NA	0.732737	NA	0.999997	NA	0.98859	NA
AB189128.1	1	NA	0.435629	NA	0.999986	NA	0.905393	NA
AF326573.1	1	NA	0.322655	NA	0.999996	NA	0.996837	NA
AF489932.1	0.999617	NA	0.057346	NA	0.999758	NA	0.937937	NA
M87512.1	1	NA	0.865038	NA	1	NA	0.882339	NA
NC_001430.1								
NC_001428.1	1	NA	0.522986	NA	0.999351	NA	0.749412	NA
NC_001612.1	0.991708	NA	0.751091	NA	0.990929	NA	0.257635	NA
NC_003986.1	0.999997	NA	0.02014	0.93616	0.937996	NA	0.708985	NA
NC_001472.1	0.99959	NA	0.977242	NA	0.957869	NA	0.692936	NA
NC_001617.1	0.435562	NA	0.474076	NA	0.028549	1.699567	0.079676	NA
NC_001490.1	1	NA	0.90881	NA	0.996231	NA	0.518662	NA
Deduction Virus	NC_001781.1 (RSV)		NC_004148.2 (HMPV)		NC_001781.1 (RSV)		None	
Confirmation Virus (PCR)	NC_001781.1 (RSV)		NC_004148.2 (HMPV)		NC_001781.1 (RSV)		None	

Example 2

[0156] In the second set of experiments, the analysis algorithm correctly detected the actual virus in the 3 samples and also the negative sample. After designing good probes for our chip, the Weighted Kullback-Leibler divergence of the acutal viruses in Experiment 1, 2 and 3 was greater than that of the corresponding experiments without probe design. This means that the signal intensities from the actual virus were relatively higher than the background noise in the PDC. This showed that our probe design criteria had removed some bad probes from the PDC, which resulted in a more accurate analysis.

[0157] Again, we present results of the 4 experiments if we had just used the t-test with a level of significance 0.05. This time, the number of viruses detected to be present for each sample is shown in Table 4:

TABLE 4

False positive detection of viruses using t-test alone in a PDC with probe design.

	Sample Name			
	35259_324	35179_122	35253_841	35915_111
Viruses Detected Using T-test	6	9	9	10
False Positives	5	8	8	10
Max KL divergence (>5.0)	18.54859	9.324785	11.17914	—
Viruses Detected Using T-test followed by KL divergence	1	1	1	0

[0158] From Table 4, it can be seen that probe design has reduced the number of false positive viruses detected by the t-test for samples 35259_324 and 35179_122. A more important observation is that the Weighted Kullback Leiber divergence for the actual virus has increased for all 4 samples. This means that the signals of the actual virus are more differentiated than the background signals when probe design criteria are applied on the PDC.

[0159] In conclusion, we showed that using the one-tailed t-test with significance level 0.05, followed by computing the Weighted Kullback-Leibler divergence for the signal intensities of each virus, we were able to accurately analyze the data on the PDC and determine with high probability the actual pathogen in the sample. Although the analysis algorithm works well even under a high level of noise, we showed that the accuracy of the analysis is improved by using the above-described probe design criteria to select a good set of probes for the PDC.

REFERENCES

[0160] 1. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-3402. (1997).

[0161] 2. Bodrossy, L. & Sessitsch, A. Oligonucleotide microarrays in microbial diagnostics. *Curr Opin Microbiol* 7, 245-254 (2004).

[0162] 3. Bohlander, S. K., Espinosa, I., Rafael, Le Beau, M. M., Rowley, J. D. & Diaz, M. O. A method for the rapid sequence-independent amplification of microdissected chromosomal material. *Genomics* 13, 1322-1324 (1992).

[0163] 4. Bustin, S. A. & Nolan, T. Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction. *J Biomol Tech* 15, 155-166 (2004).

[0164] 5. Goulden, C. H. Methods of Statistical Analysis, Edn. 2nd. (John Wiley & Sons, Inc., New York; 1956).

[0165] 6. Kane, M. D. et al. Assessment of the sensitivity and specificity of oligonucleotide (50 mer) microarrays. *Nucleic Acids Res* 28, 4552-4557 (2000).

[0166] 7. Kullback, S. & Leiber, R. A. On information and sufficiency. *Ann. Math. Stat.* 22, 79-86 (1951).

[0167] 8. Nuwaysir, E. F. et al. Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res* 12, 1749-1755 (2002).

[0168] 9. Sambrook and Russel, Molecular Cloning: A Laboratory Manual, Cold Springs Harbor Laboratory, New York (2001).

[0169] 10. SantaLucia, J., Jr., Allawi, H. T. & Seneviratne, P. A. Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* 35, 3555-3562 (1996).

[0170] 11. Stephens, M. A. (1974). EDF Statistics for Goodness of Fit and Some Comparisons, Journal of the American Statistical Association, Vol. 69, pp. 730-737.

[0171] 12. Striebel, H. M., Birch-Hirschfeld, E., Egerer, R. & Foldes-Papp, Z. Virus diagnostics on microarrays. *Curr Pharm Biotechnol* 4, 401-415 (2003).

[0172] 13. Sung, W. K. & Lee, W. H. Fast and Accurate Probe Selection Algorithm for Large Genomes. CSB (2003).

[0173] 14. Vora, G. J., Meador, C. E., Stenger, D. A. & Andreadis, J. D. Nucleic acid amplification strategies for DNA microarray-based pathogen detection. *Appl Environ Microbiol* 70, 3047-3054 (2004).

[0174] 15. Wang, D. et al. Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol* 1, E2 (2003).

[0175] 16. Wong, C. W. et al. Tracking the Evolution of the SARS Coronavirus Using High-Throughput, High-Density Resequencing Arrays. *Genome Res.* 14, 398-405 (2004).

[0176] 17. Wu, D. Y., Ugozzoli, L., Pal, B. K., Qian, J. & Wallace, R. B. The effect of temperature and oligonucleotide primer length on the specificity and efficiency of amplification by the polymerase chain reaction. *DNA Cell Biol* 10, 233-238 (1991).

SEQUENCE LISTING

<160> NUMBER OF SEQ ID NOS: 7

<210> SEQ ID NO 1
<211> LENGTH: 26
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Description of Artificial Sequence: Synthetic primer
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (18)..(26)
<223> OTHER INFORMATION: a, c, g, or t

<400> SEQUENCE: 1

gtttcccagt cacgatannn nnnnnn 26

<210> SEQ ID NO 2
<211> LENGTH: 26
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Description of Artificial Sequence: Synthetic primer

<400> SEQUENCE: 2

gtttcccagt cacgatagca tgaggg 26

<210> SEQ ID NO 3
<211> LENGTH: 26
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Description of Artificial Sequence: Synthetic primer

<400> SEQUENCE: 3

gtttcccagt cacgatacga atagct 26

<210> SEQ ID NO 4
<211> LENGTH: 22
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Description of Artificial Sequence: Synthetic fragment of virus sequence

<400> SEQUENCE: 4

acgatatccg cgaatagcta ga 22

<210> SEQ ID NO 5
<211> LENGTH: 23
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Description of Artificial Sequence: Synthetic fragment of virus sequence

<400> SEQUENCE: 5

catccctcat gcatggggca att 23

<210> SEQ ID NO 6
<211> LENGTH: 22

-continued

<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Description of Artificial Sequence: Synthetic fragment of virus sequence

<400> SEQUENCE: 6

tgctataggc gcttatcgat ct 22

<210> SEQ ID NO 7
<211> LENGTH: 23
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Description of Artificial Sequence: Synthetic fragment of virus sequence

<400> SEQUENCE: 7

gtagggagta cgtaccccgt taa 23

1. A method of designing oligonucleotide probe(s) for nucleic acid detection comprising the following steps in any order:

(i) identifying and selecting region(s) of a target nucleic acid to be amplified, the region(s) having an efficiency of amplification (AE) higher than the average AE; and

(ii) designing oligonucleotide probe(s) capable of hybridizing to the selected region(s).

2. The method according to claim 1, wherein the AE of the selected region(s) is calculated as the Amplification Efficiency Score (AES), which is the probability that a forward primer r_i can bind to a position i and a reverse primer r_j can bind at a position j of the target nucleic acid, and |i-j| is the region of the target nucleic acid desired to be amplified.

3. The method according to claim 2, wherein |i-j| is ≦10000 bp.

4. The method according to claim 2, wherein |i-j| is ≦500 bp.

5. The method according to claim 1, wherein the oligonucleotide probe(s) capable of hybridizing to the selected region(s) is selected and designed according to at least one of the following criteria:

(a) the selected probe(s) has a CG-content from 40% to 60%;

(b) the probe(s) is selected by having the highest free energy computed based on Nearest-Neighbor model;

(c) given probe s_a and probe s_b substrings of target nucleic acids v_a and v_b, s_a is selected based on the hamming distance between s_a and any length-m substring s_b from the target nucleic acid v_b and/or on the longest common substring of s_a and probe s_b;

(d) for any probe s_a of length-m specific for the target nucleic acid v_a, the probe s_a is selected if it does not have any hits with any region of a nucleic acid different from the target nucleic acid, and if the probe s_a length-m has hits with the nucleic acid different from the target nucleic acid, the probe s_a length-m with the smallest maximum alignment length and/or with the least number of hits is selected; and

(e) a probe p_i at position i of a target nucleic acid is selected if p_i is predicted to hybridize to the position i of the amplified target nucleic acid.

6. The method according to claim 1, wherein the method further comprises a step of preparing the selected and designed probe(s).

7. A method of detecting at least one target nucleic acid comprising the step of:

(i) providing a biological sample;

(ii) amplifying the nucleic acid(s) of the biological sample;

(iii) providing at least one oligonucleotide probe capable of hybridizing to at least a target nucleic acid, if present in the biological sample, wherein the probe(s) is prepared according to the method of claim 13; and

(iv) contacting the probe(s) with the amplified nucleic acids and detecting the probe(s) hybridized to the target nucleic acid(s).

8. The method according to claim 7, wherein the amplification step (ii) is carried out in the presence of at least one random forward primer and at least one reverse random primer.

9. The method according to claim 7, wherein the amplification step is a RT-PCR.

10. The method according to claim 7, wherein the forward random primer binding to position i and the reverse random primer binding to position j of a target nucleic acid v_a are selected among primers having an amplification efficiency score (AES_l) for every position i of a target nucleic acid v_a of:

$$AES_i = \sum_{j=i-Z}^{i} \left\{ P^f(j) \times \sum_{k=i+1}^{j+Z} P^r(k) \right\}$$

-continued

$$\text{wherein } \sum_{k=i+1}^{j+Z} P^r(k) = P^r(i+1) + P^r(i+2) + \ldots P^r(j+Z);$$

P^f (i) and P^r (i)are the probabilities that a random primer r_i can bind to position i of v_a as forward primer and reverse primer respectively, and $Z \leqq 10000$ bp is the region of v_a desired to be amplified.

11. The method according to claim 7, wherein the target nucleic acid to be detected is nucleic acid exogenous to the nucleic acid of the biological sample.

12. The method according to claim 7, wherein the target nucleic acid to be detected is at least a pathogen genome or fragment thereof.

13. The method according to claim 12, wherein the pathogen nucleic acid is at least a nucleic acid from a virus, a parasite, or bacterium, or a fragment thereof.

14. The method according to claim 7, wherein the biological sample is obtained from a human being and the target nucleic acid, if present in the biological sample, is not from human.

15. The method according to claim 7, wherein the probes are placed on an insoluble support.

16. The method according to claim 7, wherein in the detection step (iv), the mean of the signal intensities of the probes which hybridize to v_a is statistically higher than the mean of the probes $\notin v_a$, thereby indicating the presence of v_a in the biological sample.

17. The method according to claim 7, wherein in the detection step (iv), the mean of the signal intensities of the probes which hybridize to v_a is statistically higher than the mean of the probes $\notin v_a$, and the method further comprises the step of computing the relative difference of the proportion of probes $\notin v_a$ having high signal intensities to the proportion of the probes used in the detection method having high signal intensities, the density distribution of the signal intensities of probes v_a being more positively skewed than that of probes $\notin v_a$, thereby indicating the presence of v_a in the biological sample.

18. The method according to claim 7, wherein in the detection step (iv), the presence of a target nucleic acid in a biological sample is given by a value of t-test $\leqq 0.1$ and/or a value of Kullback-Leibler divergence of $\geqq 1.0$.

19. A method of determining the presence of a target nucleic acid v_a comprising detecting the hybridization of a probe to a target nucleic acid v_a and wherein the mean of the signal intensities of the probes which hybridize to v_a is statistically higher than the mean of the probes $\notin v_a$, thereby indicating the presence of v_a.

20. The method according to claim 19, wherein the mean of the signal intensities of the probes which hybridize to v_a is statistically higher than the mean of the probes $\notin v_a$, and the method further comprises the step of computing the relative difference of the proportion of probes $\notin v_a$ having high signal intensities to the proportion of the probes used in the detection method having high signal intensities, the density distribution of the signal intensities of probes v_a being more positively skewed than that of probes $\notin v_a$, thereby indicating the presence of v_a.

21. The method according to claim 19, wherein the presence of a target nucleic acid in a biological sample is given by a value of t-test $\leqq 0.1$ and/or a value of Kullback-Leibler divergence of $\geqq 1.0$.

22. A method of detecting at least one target nucleic acid comprising the steps of:

(i) providing a biological sample;

(ii) amplifying the nucleic acid(s) of the biological sample;

(iii) providing at least one oligonucleotide probe capable of hybridizing to at least a target nucleic acid, if present in the biological sample; and

(iv) contacting the probe(s) with the amplified nucleic acids and detecting the probe(s) hybridized to the target nucleic acid(s), wherein the mean of the signal intensities of the probes which hybridize to v_a is statistically higher than the mean of the probes $\notin v_a$, thereby indicating the presence of v_a in the biological sample.

23. The method according to claim 22, wherein in step (iv) the mean of the signal intensities of the probes which hybridize to v_a is statistically higher than the mean of the probes $\notin v_a$, and the method further comprises the step of computing the relative difference of the proportion of probes $\notin v_a$ having high signal intensities to the proportion of the probes used in the detection method having high signal intensities, the density distribution of the signal intensities of probes v_a being more positively skewed than that of probes $\notin v_a$, thereby indicating the presence of v_a in the biological sample.

24. The method according to claim 22, wherein in step (iv) the presence of a target nucleic acid in a biological sample is given by a value of t-test $\leqq 0.1$ and/or a value of Kullback-Leibler divergence of $\geqq 1.0$.

25. The method according to claim 22, wherein the target nucleic acid to be detected is nucleic acid exogenous to the nucleic acid of the biological sample.

26. The method according to claim 22, wherein the target nucleic acid to be detected is at least a pathogen genome or fragment thereof.

27. The method according to claim 26, wherein the pathogen nucleic acid is at least a nucleic acid from a virus, a parasite, or bacterium, or a fragment thereof.

28. The method according to claim 22, wherein the biological sample is obtained from a human being and the target nucleic acid, if present in the biological sample, is not from human genome.

29. The method according to claim 22, wherein the probes are placed on an insoluble support.

* * * * *