PCT

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification
C12Q 1/68, B01L 3/00

(22) International Filing Date:
15 March 2000 (15.03.00)

(30) Priority Data:
60/125,098 15 March 1999 (15.03.99) US
09/526,070 14 March 2000 (14.03.00) US

(63) Related by Continuation (CON) or Continuation-in-Part (CIP) to Earlier Application
US 09/526,070 (CIP)
Filed on 14 March 2000 (14.03.00)

(71) Applicant (for all designated States except US):
VISTAGEN, INC. [US/US]; 325 East Middlefield Road, Mountain View, CA 94043 (US).

(72) Inventor; and
(75) Inventor/ Applicant (for US only):

(54) Title:
BAR CODING AND IDENTIFYING NUCLEIC ACIDS USING A LIMITED NUMBER OF PROBES AND LOW STRINGENCY CONDITIONS

(57) Abstract

The invention relates generally to methods for identifying and quantitating the presence of nucleic acids of interest. The methods involve the hybridization of nucleic acids of interest to probes under relatively non-stringent conditions and use patterns of relative binding intensity to provide useful information. The methods allow for the use of far fewer probes than previous techniques, and do not require prior knowledge of the sequence of a nucleic acid of interest to detect its presence in a sample. The methods of the invention allow for the comparative assessment of the expression of genes in samples from different sources (for example, from different tissues or cell types, disease states, or development stages).
FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL Albania
AM Armenia
AT Austria
AU Australia
AZ Azerbaijan
BA Bosnia and Herzegovina
BB Barbados
BE Belgium
BF Burkina Faso
BG Bulgaria
BJ Benin
BR Brazil
BY Belarus
CA Canada
CF Central African Republic
CG Congo
CH Switzerland
CI Côte d’Ivoire
CM Cameroon
CN China
CU Cuba
CZ Czech Republic
DE Germany
DK Denmark
EE Estonia
ES Spain
FI Finland
FR France
GA Gabon
GB United Kingdom
GE Georgia
GH Ghana
GN Guinea
GR Greece
HU Hungary
IE Ireland
IL Israel
IS Iceland
IT Italy
JP Japan
KE Kenya
KG Kyrgyzstan
KP Democratic People’s Republic of Korea
KR Republic of Korea
KZ Kazakhstan
LC Saint Lucia
LI Liechtenstein
LK Sri Lanka
LR Liberia
LS Lesotho
LT Lithuania
LU Luxembourg
LV Latvia
MC Monaco
MD Republic of Moldova
MG Madagascar
MK The former Yugoslav Republic of Macedonia
ML Mali
MN Mongolia
MR Mauritania
MW Malawi
MX Mexico
NE Niger
NL Netherlands
NO Norway
NZ New Zealand
PL Poland
PT Portugal
RO Romania
RU Russian Federation
SD Sudan
SE Sweden
SG Singapore
SI Slovenia
SK Slovakia
SN Senegal
SZ Swaziland
TD Chad
TG Togo
TJ Tajikistan
TM Turkmenistan
TR Turkey
TT Trinidad and Tobago
UA Ukraine
UG Uganda
US United States of America
UZ Uzbekistan
VN Viet Nam
YU Yugoslavia
ZW Zimbabwe
BAR CODING AND IDENTIFYING NUCLEIC ACIDS USING A LIMITED NUMBER OF PROBES AND LOW STRINGENCY CONDITIONS

CROSS-REFERENCES TO RELATED APPLICATIONS

This application claims the benefit of U.S. Provisional Application Serial No. 60/125,098, filed March 15, 1999, the disclosure of which is incorporated herein by reference.

TECHNICAL FIELD

The invention relates generally to methods for identifying and quantitating the presence of nucleic acids of interest. The methods involve the hybridization of nucleic acids of interest to probes under relatively non-stringent conditions and use patterns of relative binding intensity to provide useful information. The methods allow for the use of far fewer probes than previous techniques, and do not require prior knowledge of the sequence of a nucleic acid of interest to detect its presence in a sample. The methods of the invention allow for the comparative assessment of the expression of genes in samples from different sources (for example, from different tissues or cell types, disease states, or development stages).

BACKGROUND ART

The determination of whether and which genes are expressed in a cell is useful in a number of contexts. The pathology of many diseases, for example, involves differences in gene expression; indeed, normal tissue and diseased tissue can often be distinguished by the types of active genes and their expression levels. As but one example, cancer cells evolve from normal cells to invasive, metastatic malignancies, which are frequently induced by activation of oncogenes or inactivation of tumor suppressor genes. Altered expression patterns of these genes is then followed by dramatic changes in the expression levels of numerous other genes. Differently expressed gene sequences can serve as markers of the transformed state and are, therefore, of potential value in the diagnosis, classification, and treatment of tumors. Further, differences in the expression profile of genes can facilitate the identification of new genes encoding products having a function of interest or involved in the disease process.
Differences in gene expression can also be of value in screening chemical compositions for development as potential drugs. Almost one third of all prospective human therapeutics fail in the first phase of human clinical trials because of unexpected toxicity. Exposing cells in a culture to a chemical composition and then comparing the gene expression pattern of the exposed cells to that of cells exposed to other chemical agents permits one to detect patterns of expression similar to that of the test compound, and thus to predict that the toxicities of the chemical compositions will be similar. See, e.g., Service, R., Science 282:396-399 (1998).

Detecting differences in gene expression is also valuable in genetics research. Mutant phenotypes often signal the expression of a previously unexpressed gene, or the failure to express one normally expressed in a cell. Determining the changes in gene expression can help identify the source of the mutant phenotype.

In addition to differences in gene expression, detection of differences in genomic DNA can be important. Such differences are the basis for differences between species, and for much of the individual variations among members of the same species. Such variations can be exploited to detect, for example, pathological conditions driven by chromosomal mutations, such as inherited disorders. Such variations can even be used in environmental monitoring, to determine, for example, whether microorganisms which are hard to classify morphologically are of species which are toxic or non-toxic (or pathogenic or non-pathogenic) members of their genus. Another area in which differences in genomic DNA is of importance is in the analysis of forensic specimens.

Detecting differences in gene expression or in genomic DNA poses daunting technical challenges. In the case of gene expression, the number of copies of mRNA of a particular transcript present in a cell may range from one to about 5,000. Thus, not only does any method need to be sensitive, but it must retain sensitivity over a large range. In the case of genomic DNA, detection of alterations in a gene might require detecting that alteration against a background of the 3 billion base pairs in the human genome. Further, the availability of genomic DNA in some applications, particularly in forensic applications, may be fairly small.

The art teaches a number of methods for determining whether differences in gene expression or genomic DNA exist. For example, oligonucleotide probes have long been used to detect complementary nucleic acid sequences in a nucleic acid of interest (the
“target” nucleic acid) and have been used to detect expression of particular genes (for example, by a Northern blot). In these embodiments, crude RNA or mRNA is separated by gel electrophoresis and transferred to a nitrocellulose membrane or filter. Immobilized on the filter, the RNA or mRNA is hybridized with a probe corresponding to sequences of interest. See, e.g., Sambrook, et al., Molecular Cloning, A Laboratory Manual. Cold Spring Harbor Press. Cold Spring Harbor. NY (1989) (hereafter “Sambrook”).

In some assay formats, the oligonucleotide probe is tethered (e.g., by covalent attachment) to a solid support. Arrays of oligonucleotide probes immobilized on solid supports have been used to detect specific nucleic acid sequences in a target nucleic acid. See, e.g., PCT International Publication Nos. WO 89/10977 and 90/11548. Others have proposed the use of large numbers of oligonucleotide probes to provide the complete nucleic acid sequence of a target nucleic acid. See, U.S. Patent Nos. 5,202,231 and 5,002,867, and PCT International Publication No. WO 93/17126. These embodiments have involved the use of up to 1,000,000 probes in a high density array of oligonucleotides. See, e.g., Lockhart, PCT International Publication No. WO 97/10365. Kamb, PCT International Publication No. WO 98/26098, uses oligonucleotides bound to beads to measure the relative levels of nucleic acids present in a sample and the retrieval of specific sequences. Pinkel, PCT International Publication No. WO 96/17958, uses methods of determining relative copy number of target nucleic acids by immobilizing the target nucleic acids on a solid support and hybridizing to them two sets of nucleic acids with distinguishable labels, such as separate colors.

Most of these techniques are intended to exploit precise matches between the nucleic acid of interest and the oligonucleotide used to capture it, and require the use of high stringency conditions (either at the outset or during washing) to eliminate imperfect matches. They also require at least some knowledge of the sequence of the nucleic acid of interest so that probes can be designed to capture that nucleic acid. Further, as is evident even from the brief descriptions above, these techniques rely on the use of as many as a million separate oligonucleotide probes to capture a nucleic acid of interest. The synthesis of these large numbers of probes requires substantial effort and cost, and is necessarily complex.

A technique which does not require prior knowledge of a sequence of an oligonucleotide of interest is disclosed in Beattie, PCT International Publication No.
WO 97/22720. Beattie relates to “arbitrary sequence oligonucleotide fingerprinting,” in which genomic DNA and other nucleic acid mixtures are compared for variations by observing the respective patterns of binding to an array of “arbitrary” oligonucleotides bound to a solid substrate. Differences in the patterns of binding to the arrays are stated to reflect differences in the oligonucleotide sequences of the samples hybridized to the arrays. Beattie typically employs probes in arrays ranging between several hundred to several thousand in order to create patterns disclosing differences between samples. For example, Beattie uses oligonucleotides nine nucleotides in length in arrays of 20 x 20 (for a total of 400 probes) or 50 x 50 (or 2500 probes).

Although the art has largely not addressed the question, additional methods which permit determination of the expression or level of expression of different genes without the need to first know the sequence of the genes would be useful. Even more useful would be methods which permit doing so without the cost, effort, and complexity needed to synthesize hundreds to tens of thousands of oligonucleotide probes. Moreover, what is needed in the art is a way to gather information pertaining to gene expression in a form easily stored and manipulated. The present invention addresses these and other needs which will be apparent upon complete review of this disclosure.

**DISCLOSURE OF THE INVENTION**

This invention relates to the use of “relative intensity of hybridization” between nucleic acids to provide information concerning which genes are expressed, and the level of their expression. The methods are essentially the reverse of those which have been used in the art. Thus, the invention relates to the discovery that information can be gathered by using many fewer probes, under relatively non-stringent conditions, than methods using high stringency conditions.

This invention provides novel methods for identifying whether a test nucleic acid is identical to a known nucleic acid. In one group of embodiments, the test nucleic acid is placed, or arrayed, on a solid support. The test nucleic acid is hybridized to a set of nucleic acids (such as a set of probes bearing at least one detectable label) comprising at least three members which differ in their degree of complementarity to the test nucleic acid under low or relatively low stringency conditions. The members of the set differentially hybridize to the test nucleic acid in relation to their respective degrees of complementarity to the test nucleic acid, thereby providing a first set of label patterns, and the first set of label patterns
are compared to a second set of label patterns comprising a plurality of hybridization patterns produced by low stringency hybridization of the known nucleic acid to the set of nucleic acids, provided that the members of the set of nucleic acids are not degradation products of a single nucleic acid, that the set of nucleic acids are not chosen to be exactly complementary to the known nucleic acid, and that the hybridizations are not performed *in situ*.

One of the sets of nucleic acids can comprise a molecular library, such as a genomic library or a cDNA library. The first set of hybridization patterns and the second set of hybridization patterns can come from separate hybridization reactions. The second set of nucleic acids can comprise fifty or fewer probes, thirty or fewer probes, twenty or fewer probes or even fifteen or fewer probes.

The molecular library can comprise at least ten million clones, at least a million clones, between about fifty thousand and about one million clones, or between about one thousand and about fifty thousand clones. A plurality of members of the molecular library can be identified by their hybridization to repetitive sets of the first set of nucleic acids (e.g., the probes). The methods can also include a second hybridization step under stringent conditions. The probes or the members of the molecular library can be affixed to a surface, and can be in an ordered arrangement to permit correlation of the hybridizations observed with the members of the molecular library. Replica plating or duplicate sets of the members of the molecular library can also be made.

The methods further include methods of identifying a test nucleic acid in a sample as identical to a known nucleic acid, wherein the method comprises spatially arraying a set of labeled nucleic acids on a solid support in separate regions, wherein the set comprises at least three members which differ in their degree of complementarity, and hybridizing the test nucleic acid to the labeled set under low or relatively low stringency conditions. The members of the set differentially hybridize to the test nucleic acid in relation to their respective degrees of complementarity to the test nucleic acid, thereby providing a first set of label patterns. The first set of label patterns is compared to a second set of label patterns comprising a plurality of hybridization patterns produced by low stringency hybridization of the known nucleic acid to the set of nucleic acids, provided that the members of the set of nucleic acids are not degradation products of a single nucleic acid, that the set of nucleic
acids are not chosen to be exactly complementary to the known nucleic acid, and that the hybridizations are not performed in situ.

The invention further provides methods of identifying a test nucleic acid in a sample, the method comprising hybridizing the test nucleic acid to a set of nucleic acids comprising at least three members, which differ in their degree of complementarity to the test nucleic acid, under low or relatively low stringency conditions. The members of the set differentially hybridize to the test nucleic acid in relation to their respective degrees of complementarity to the test nucleic acid. The relative proportion of hybridization of the test nucleic acid to the set of nucleic acids is measured, thereby providing a first set of hybridization patterns, and the first set of label patterns to a second set of hybridization patterns produced by low stringency hybridization of the known nucleic acid to the set of nucleic acids, provided that the members of the set of nucleic acids are not degradation products of a single nucleic acid, that the set of nucleic acids are not chosen to be exactly complementary to the known nucleic acid, and that the hybridizations are not performed in situ.

Additionally, the members of a molecular library can be bar coded. Members of the library are hybridized to a limited set of probes under low stringency conditions, the resulting hybridization profiles for each member of the library are recorded, and the profiles for hybridization are compiled to provide a bar code for each member of the library. The probe set can comprise fifty or fewer probes, thirty or fewer probes, twenty or fewer probes or even fifteen or fewer probes. The bar code (or the profile) can include differences in the intensity of labeling of at least one probe. The bar codes can be digitized, or a graphical representation of which probes hybridized and at what intensity. Either the probes or the members of the molecular library can be affixed to a surface.

The invention further covers an array reader adapted to read the hybridization patterns of labels on an array, operably linked to a digital computer comprising a data file having a set of at least 500 low-stringency hybridization patterns in a digital format. The array reader can be part of an integrated system for comparing hybridization patterns. The system can include, for example, a robotic armature for fluid delivery to an array. The system can be capable of reading 500 or more labels an hour on an array. The system can further be operably linked to an optical detector for reading the hybridization patterns of labels on the array.
BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a figure and series of eight graphs. The figure presents exemplar hybridization patterns for a series of genes hybridized to oligonucleotide probes.

MODE(S) FOR CARRYING OUT THE INVENTION

A. DEFINITIONS

Terms used herein are in general as typically used in the art. The following terms as used herein are intended to have the following meanings:

An "array" of nucleic acids is an ordered spatial arrangement of one or more nucleic acids on a physical substrate. Row and column arrangements are preferred due to the relative simplicity in making and assessing such arrangements. The spatial arrangement can, however, be essentially any form selected by the user, and preferably is, but need not be, in a pattern. The nucleic acids in the array can be DNA, RNA, or an analog of either; DNA is generally preferred due to the ease of recombinant manipulation and physical stability.

The terms "background" or "background signal intensity" refer to hybridization signals resulting from non-specific binding between labeled oligonucleotides and components of the testing apparatus, such as the array substrate. Background signals may also be produced by components of the testing apparatus. For example, if a fluorescent label is used for the oligonucleotides discussed herein, an intrinsic fluorescence by a plastic substrate would be considered a background signal. It is desirable for the signal from the specific binding (or hybridization) is distinguishable from any non-specific binding. The signal from any non-specific binding should be at least 10% less than the signal for the lowest hybridization to a target of interest, and is more preferably at least 50% less. It is usually determined as the average of the hybridization that occurs on the supporting matrix or substrate.

The term "bar code," as used herein, refers to a hybridization pattern. Most conveniently, the hybridization pattern can be described by assigning a numeric value to represent the intensity with which a probe hybridizes to a nucleic acid sequence of interest, and the "bar code" will therefore be a series of numbers representing the binding of the probes to the nucleic acids of interest.
The term “complement” refers to a nucleic acid sequence to which a second nucleic acid sequence specifically hybridizes to form a perfectly matched duplex.

“Complementarity” and “similarity” refer herein to the degree to which one nucleic acid sequence is more or less complementary to a second nucleic acid sequence. As used in the art, “complementarity” usually implies strandedness, whereas “similarity” implies the sequences are the same. (A gene is composed of two DNA strands, so a sequence the same as one of the strands (high similarity) is complementary to the other strand. Since a sequence can only be complementary to one of the two strands, however, “complementarity” refers to a specific one of the two strands of DNA. As used herein, “complementarity” and “similarity” are generally used interchangeably unless otherwise required by context.

The term “digitizing,” as used herein, refers to the process of assigning numeric values to the hybridization of the probes and the nucleic acids of interest, i.e., converting analog data into digital data.

“Differentially hybridizes” means that a member of a first set of nucleic acids (such as a probe) will bind with a different degree of affinity to a member of a second set of nucleic acids (such as a gene of interest), compared to the binding of other probes (or other members of the first set of nucleic acids) to the same member of the second set of nucleic acids. In the present invention, for example, a group of probes will tend to have different degrees of complementarity to a gene of interest, and will thus bind to it with degrees of affinity which will reflect whether they are more or less complementary to sequences in the gene. Typically, for the differential hybridization to be detectable, the intensity of the binding of at least one of the members of the first set of nucleic acids to a member of the second set of nucleic acids will usually have to be higher than background level.

“Identity” refers to a high degree of similarity between two nucleic acid sequences. Preferably, the similarity is at least about 90%. More preferably, the similarity is about 95% to 100%.

“Intensity” refers to the number of molecules of a probe or other set of nucleic acids which have bound to a second set of nucleic acids. In general, the number of molecules is inferred from the amount of a signal from labeled members of one of the sets of nucleic acids, and is optimally based upon a constant ratio of target to capture probe. The units by which the intensity is described will depend on the nature of the label, and could be, for example, in counts per minute for a radioactive label read in a scintillation counter, in brightness, or in arbitrary units of emitted light for a fluorophore.

The term “label” refers to a composition detectable by spectroscopic, photochemical, biochemical, immunochemical, electrochemical, physiochemical, or chemical means. For example, useful nucleic acid labels include radioisotopes, such as $^{32}$P and $^{35}$S, fluorescent dyes, electron-dense reagents, enzymes (e.g., as commonly used in an ELISA), biotin, digoxigenin, modified nucleic acids, such as bromodeoxyuridine (“BrdU”), for which monoclonal antibodies are available, or haptens and proteins for which antisera or monoclonal antibodies are available.

The term “labeled member” refers to a labeled molecule used to detect the nucleic acid or other molecule of interest. Most commonly in the context of the invention, the molecule will be an oligonucleotide, although it can be, for example, a protein, an antibody or another molecule capable of interacting with a nucleic acid or other target molecule. In the context of the invention, it refers to a probe, when the probe is labeled, or to the nucleic acid of interest, when a nucleic acid, such as a particular clone, is labeled and is being hybridized to a probe.

The term “label pattern” means the hybridization pattern of two groups of molecules, such as that of probes hybridized to a nucleic acid of interest, wherein one of the sets of nucleic acids is labeled. For example, if a labeled probe is hybridized to a set of nucleic acids, the “label pattern” would reflect which of the nucleic acids the probe hybridized to, and the intensity of the hybridization to each such nucleic acid.

The term “library” is used herein in two senses. First, it can be used to refer to the total cDNA or genomic DNA of a eucaryotic cell from a particular species. In this meaning, the term is used as part of the phrases “cDNA library” or “genomic DNA
library.” Second, the term can refer to a plurality of hybridization profiles which permit a comparison of differences in gene expression between two samples. Which meaning is intended will be clear in context.

“Low stringency conditions” are conditions which permit the hybridization of oligonucleotides which are not perfectly complementary, or “matched.” Persons of skill in the art are aware that what constitutes a “low stringency” condition varies according to the degree of complementarity of the particular pair of nucleic acids in question, as well as the length of the oligonucleotides being hybridized. Accordingly, while certain conditions (such as adding buffer at room temperature and washing at room temperature) are usually indicative of low stringency, particular conditions are usually defined empirically by standard measures and by standard formulas, some of which are set forth below. Guides to adjusting conditions for greater or lesser degrees of stringency are also available in standard works, such as Ausubel, et al., Current Protocols in Molecular Biology, John Wiley & Sons, Inc., New York (1988), and its supplements, all of which are incorporated herein by reference.

Functionally, one knows when low stringency conditions are being used in the method of the invention when mismatched nucleic acids hybridize and the intensity of their binding is above background, as measured by, for example, the amount of label present. Typically, low stringency conditions involve the use of temperatures which are 9 to about 25°C, about 10 to about 25°C, about 11 to about 25°C, about 12 to about 25°C, or about 13 to about 25°C below the calculated T_M. (The T_M is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe.) The actual temperature used will take into account the pH and concentrations of salts and organic compounds, and can be tested empirically by methods described within to determine if a satisfactory signal is obtained from the desired hybridization while permitting discrimination of that signal from any background which may also be present.

The term “molecular library” means a cDNA library or a genomic DNA library.

The terms “nucleic acid” or “nucleic acid molecule” refer to a deoxyribonucleotide or ribonucleotide polymer in either single- or double-stranded form, and, unless otherwise limited, would encompass known analogs of natural nucleotides that can function in a manner similar to naturally occurring nucleotides.
An “oligonucleotide” is a single-stranded nucleic acid ranging in length from 2 to about 500 or more monomer units (e.g., nucleotides).

“Operably linked” refers to a connection between two components which allows at least a transfer of data in at least one direction between the two components.

“Profile” means a pattern of hybridization of a molecular library or clone to an array of probes, or of probes to a molecular library or clone.

The term “probe” refers herein to an oligonucleotide selected to hybridize to one or more portions of the genome of interest.

The terms “stringent hybridization conditions” or “stringent conditions” refer to conditions under which a nucleic acid sequence will hybridize to its complement, but not to other sequences in any significant degree. Stringent conditions in the context of nucleic acid hybridizations are sequence dependent and are different under different environmental parameters. Longer sequences hybridize specifically at higher temperatures. An extensive guide to the hybridization of nucleic acids is found in Tijssen, Laboratory Techniques in Biochemistry and Molecular Biology -- Hybridization with Nucleic Acid Probes, Part I, Chapter 2 “Overview of principles of hybridization and the strategy of nucleic acid probe assays.” Elsevier, New York, (1993) (the entirety of Tijssen is hereby incorporated by reference). Very stringent conditions are selected to be equal to the Tm point for a particular probe. Less stringent conditions, by contrast, are those in which a nucleic acid sequence will bind to imperfectly matched sequences. Stringency can be controlled by changing temperature, salt concentration, the presence of organic compounds, such as formamide, or all of these. For convenience and reversibility, stringency is usually controlled by temperature.

A “test nucleic acid” refers to a nucleic acid of interest. A test nucleic acid may, for example, be a particular cDNA.

The term “unique identifier” refers to a hybridization pattern or bar code which is peculiar to the particular nucleic acid sequence to which it pertains for a given set of hybridization conditions.

B. DESCRIPTION OF THE INVENTION

1. Overview

The invention provides a convenient way to monitor gene expression or to screen for and identify new genes. The invention accomplishes these objectives without the need
to have prior knowledge of the sequences of the genes or genome being tested. Moreover, the invention provides a convenient means of managing and storing data relating to gene expression.

As noted in the background section, the art has typically monitored gene expression and detected polymorphisms, within genes by hybridizing samples of interest to arrays of hundreds of, thousands of, tens of thousands of, or even more, nucleotides. As also noted, prior techniques have typically employed stringent hybridization conditions to reduce hybridization of imperfectly matched sequences, which in those techniques is considered to be background “noise” which reduces the ability to discriminate positive results.

While these methods provide information which can be very useful, the synthesis of the numerous oligonucleotide probes used imposes significant costs in time, expense, and effort and demands elaborate synthetic processes. The present invention, in sharp contrast, stems from the discovery that meaningful information can be extracted by taking the opposite tack. Instead of relying on precise hybridization of samples to members of a large set of probes under stringent conditions, the invention compares the patterns of hybridization of samples to a small set of probes, under relatively non-stringent hybridization conditions. The use of a small set of probes sharply reduces the cost and effort needed to conduct the testing.

The present invention derives from the recognition that the exact matches of nucleotide and probe used by current techniques gives only a single piece of information (that is, that a sequence complementary to the probe is present in the sample). In contrast, the present invention stems from the realization that a small set of probes hybridized under less stringent conditions can yield two pieces of information: the nucleic acids to which the probes bind, and the intensity with which the probes bind to those nucleic acids.

The binding of oligonucleotides such as probes to any other nucleic acid sequence is an equilibrium reaction. When non-perfectly matched probes and non-stringent conditions are used, the reaction will therefore continue until the number of molecules of probe dissociating from the nucleic acid of interest (the sample) is the same as the number hybridizing to it. As is well known, equilibrium reactions shift to the right (go further to completion) if the affinity of the reactants is higher, and to the left (more molecules dissociate) if the affinity of the reactants is lower. Probes with higher degrees of complementarity will bind to a sample with a higher degree of affinity, and at equilibrium.
more molecules of such probes will be bound to a sample than will probes with lesser
degrees of complementarity.

Thus, if a nucleic acid of interest (the sample) is hybridized with several different
probes which vary in their degree of complementarity, probes with higher degree of
complementarity will tend to hybridize to the sample in greater numbers than probes with a
lesser degree of complementarity and therefore, will have a higher percentage of the
molecules bound rather than in solution. If the probes are labeled, the differences in the
number of probes binding to the sample at equilibrium will be detectable by differences in
the intensity of the signal from the label. For example, a fluorescent label will result in a
brighter signal for a better matched probe compared to a less well matched probe even
under low stringency conditions. Thus, each of the oligonucleotides which constitute each
probe gives information not only from whether the oligonucleotide binds to a particular
sample nucleic acid at all, but also from the intensity with which the probe binds.

It should be noted that, while labeled probes are a preferred embodiment, it is not
necessary for the probes to be labeled to determine the number of copies hybridized to a
target, such as a nucleic acid of interest. For example, the copies of a probe hybridizing to
a nucleic acid can be eluted from the nucleic acid and run through a mass spectrometer,
which can then determine the number of molecules involved (since the size of the probe is
known and usually much smaller than the nucleic acid of interest, it is simple to determine
that the molecules being counted are that of the probe). Thus, the number of copies can be
measured in ways other than those based on emissions of light or radioactive particles. In
another embodiment, the nucleic acids of interest are labeled, rather than the probes. The
probes are bound to a substrate or other surface, and incubated with the nucleic acids of
interest. Any unbound nucleic acids of interest are then removed (for example, by washing
with buffer), and the amount of signal from the bound, labeled nucleic acids is determined.

If the probes are labeled, it is preferred that the labels for each copy of the same
probe be the same, but that the label for at least one of the probes be distinguishable from
that of other members of the probe set. For example, each probe can be labeled with a
different fluorophore so that each probe has fluoresces at a different color. This permits
simultaneous hybridization with multiple probes and thereby reduces both the number of
individual hybridization steps and the time required to determine the patterns of all the
hybridizations of the members of the probe set. Conveniently, the colors associated with
each different fluorophore can be isolated by an appropriate optical filter to permit quantitation of the binding of each probe.

Optionally, one or more nucleic acid targets can be included in an array as controls to reflect degrees of hybridization. For example, a target perfectly matched to a probe can be placed on one position of an array to give a signal reflecting 100% hybridization (to serve as a positive control), and a nucleic acid completely mismatched to a probe can be placed at a known position on an array to give a signal reflecting 0% hybridization (that is, to serve as a negative control).

The invention uses non-stringent conditions, which permit hybridization of sequences which are not perfectly matched. Under these conditions, the sample may hybridize to some degree to one, two, three, or more members of the probe set. The particular probes which hybridize to the sample, and the relative intensity, or degree, of their hybridization, provides useful information. Since the probes do not need to be an exact match for the nucleic acid of interest, the number of probes needed to detect a sample, such as a gene, can be sharply reduced. For example, if the system has the capability to detect five levels of hybridization, theoretically eight oligonucleotides with 15 nucleotides in length could detect all $10^5$ genes in the human genome.

The pattern of hybridization to the probes can then be compared to the patterns of hybridization to the probe set of other nucleic acids to determine whether the hybridization pattern seen is similar to, or different than, those hybridization patterns. If, for example, the sample of interest is a cDNA, a pattern of hybridization different from that previously seen for a particular organism may indicate the expression of a previously unknown gene. For example, if the hybridization patterns have been determined for several members of a family of genes, a newly generated pattern can be readily compared to see if it is a member of the family or if it represents a novel member. Even the absence of a signal, indicating that no hybridization has occurred, is informative under these circumstances, since it indicates either that the gene is one previously seen (that showed no hybridization) or, if no absence of pattern has previously been seen, that the gene is not one previously known. Thus, the invention can detect the presence of previously unknown genes without prior awareness of their existence or information on their sequence.

It is anticipated that the invention will often be used to investigate samples of cDNA or of genomic DNA wherein each sample represents a single gene. In some
applications, however, it is anticipated that the invention will be applied to samples which may constitute mixtures of genes. For example, an investigator may wish to confirm if two samples contain identical mixtures of genes or whether, instead, additional genes have been expressed. The samples can be hybridized to identical probe sets. Identical hybridization patterns would indicate that the two samples contain the same genes; by contrast, different patterns would indicate that the two samples represented a different mixture of genes. The invention can thus be used as a shortcut to determine whether more precise analysis of the samples is necessary, and as a tool analogous to subtractive hybridization or differential display for comparison of expression profiles of biological samples.

The invention becomes more and more useful as the number of nucleic acids probed with a particular probe set increases, increasing the number of hybridization patterns available for comparison to the hybridization pattern of a current nucleic acid of interest. Once a satisfactory set of probes has been designed, it will therefore usually be desirable to continue to use the same set of probes in hybridizations to further nucleic acids of interest. For convenience, this repetitive use of a given set of probes in hybridizations to additional nucleic acids of interest is referred to herein as the use of “duplicate sets” of probes.

The enhanced ability to compare hybridization patterns as the number of nucleic acids hybridized to a probe set increases can be exploited in many ways. For example, an unknown cDNA with a hybridization pattern almost identical to, for example, the tumor suppressor gene p53 can be identified as encoding a protein in the p53 family, without having known that such a gene was in the sample and without having had to design a probe specific for a sequence encoding a p53 gene. Similarly, hybridization patterns for members of the plant pathogen resistance gene family PR-1 can be used to detect the presence of members of the gene family in plants not previously tested for the presence of pathogen resistance genes. Typically, either the nucleic acids of interest or the probes will be labeled to facilitate monitoring of the hybridizations.

The hybridization pattern of the large sets of molecules to the probes can be represented by numbers. Numeric values, such as 1 to 5, can be assigned to the relative intensity with which each of the probes used label the nucleic acid target, and each nucleic acid can therefore be assigned a number for the binding of each probe in a probe set. The number representing the intensity of binding of each probe to the sample can then be stated in sequence to generate a multi-digit number which represents the hybridization of the
probes to that sample. In general, it is advantageous to select and maintain a particular order of stating the numbers for the individual probes to facilitate comparisons of the numbers generated by the binding of the probe set to different samples.

In addition to numeric representations, the hybridization patterns can be represented by graphical representations, which can look not dissimilar to the "bar codes" seen on items for sale and the like. In this embodiment, each slot or position of the bar code represents a specific oligonucleotide, and the height or width of the "bar" represents the degree, or intensity, of hybridization. For computers a "bit" is a basic unit that can store 2 distinct "values." A "byte" contains 8 bits and can store 256 "values." Whereas storage of the nucleotide sequence comprising a gene can require listing thousands of nucleotides, the barcode representing that gene requires only three bytes (24 bits). Since three bytes can represent at least 16 million distinct "values," such bar codes can present unique identifiers for all of the genes and their spliced versions. This represents a very significant increase in efficiency of manipulation and storage over current methods.

The generation of these numbers or graphical representations permits the ready storage of information about large sets of samples. Moreover, since each number or bar code reflects a physical interactions between a sample and a set of probes under known conditions of hybridization, a comparison of two numbers or bar codes provides information on the degree to which the two molecules of interest share sequence homology.

The combination of advantages provided by the invention are not available in the prior art. Only one of the prior art techniques discussed in the background section, the "arbitrary sequence oligonucleotide fingerprinting" technique taught by Beattie in WO 97/22720, for example, has the capability to detect the presence of nucleic acids for which no prior sequence information is available, yet it still requires the uses of hundreds to thousands of oligonucleotide probes. Further, most of the prior techniques do not permit the ready detection of differentially spliced variants of genes. The invention is the simplest and most cost effective approach yet available for the detection of new genes, the analysis of polymorphisms, and the quantitation of all genes present in a sample.

2. Uses of the Invention

The invention is useful in a variety of contexts. One use contemplated for the invention is in the identification of genes responsive to drugs. For example, duplicate cultures of cells, tissues, or embryoid bodies can be created and one of the cultures
contacted with a chemical composition. The two cultures can then be screened using the invention to identify genes in the test culture which were expressed in response to the chemical composition. Similarly, the same procedure can be used to identify genes responsive to the presence of environmental contaminants, such as PCBs, pesticides, fertilizers, and industrial waste, as well as potential food additives such as dyes and artificial sweeteners.

The invention can also be used to determine differences in gene expression between healthy persons and persons suffering from disease, or persons who have been exposed to agents such as those described in the previous paragraph. In these cases, biopsied tissue or autopsy specimens can be compared to specimens from persons not known to have suffered from the condition, or exposed to the agent in question, to determine differences in gene expression between those cells or tissues.

While these embodiments contemplate the application of the techniques to human cells, the invention can also be used to find drug responsive genes in other species. For example, malarial parasites, helminths, fungal parasites, and other organisms which cause disease, can also be exposed to chemical agents and their gene expression compared to that of control populations to determine which ones might be potential agents against one or more of those organisms. Nor do the organisms need to be eucaryotic. Prokaryotes can also be screened using the methods of the invention to determine whether various agents have the potential to inhibit or prevent their growth. Thus, pathogenic bacteria can be screened for the induction of genes indicating whether they are harmed or benefited by contact with various agents.

Similarly, plant species can be screened for the effect of chemical compositions to determine if the compositions can be used as therapeutics, as fertilizers, or as growth retardants. For example, a desirable plant can be screened for compositions which induce expression of genes associated with protection of plants from plant pathogens, such as the PR family of genes, and plants from species considered to be weeds can be screened for expression of genes associated with decreased lifespan or increased mortality.

In another use, polymorphisms can be detected in genes isolated from two or more individuals. The change in one or more nucleotides in a sequence will likely not change the hybridization of all the probes to the sample, but is likely to change the hybridization to one or more of them. This change will of course alter the hybridization pattern, and
indicate a change in the gene compared to a control. It should be noted that variations in
gene splicing will also be reflected by altered hybridization patterns. This is appropriate,
since such splicing changes the gene product expressed and can result in phenotypic
differences. Accordingly, the ability of the invention to discriminate among spliced
versions of a gene is a useful one. It should be noted that one artificial variant can be
introduced which is not useful. cDNAs are created by the action of reverse transcriptase on
mRNA. On occasion, the reverse transcriptase stops short of completing the reverse
transcription of the entire mRNA. This shorter sequence will create a different
hybridization pattern, which in this case is an artifact. This artifact can be minimized by
following known procedures for increasing the percentage of full-length molecules. The
artifact cannot be entirely eliminated, and represents a price paid for the convenience of not
having to sequence all the genes in the sample.

Gene expression can also reveal relationships between members of genera and
evolutionary relationships between members of different genera. The invention can be
helpful in elucidating the degree of homology and relationship between species and
between members of different genera.

In many embodiments, duplicates will be made of the samples to undergo
hybridization (for example, by pressing a plate over a surface spotted with bacterial
samples to be hybridized to form a replica plate, or by dispensing cloned DNA molecules
into duplicate wells of microtiter plates, or the like). If the hybridization patterns reveal the
presence of a gene of interest, a duplicate sample can be used to obtain and sequence the
gene of interest. Once the gene is sequenced, the hybridization pattern observed can be
correlated to it and the next occurrence of that pattern recognized as indicating the presence
of the gene.

In a common embodiment, the invention is used to study the biology of “lead
compounds” (that is, a limited universe of compounds which are considered by a company
to be the best candidates for development as pharmaceuticals) and, in particular, their effect
on both known and unknown genes. The ability of the invention to detect the effects of the
compounds on unknown genes, or genes which would not have been expected to be
affected by a particular compound, is a significant advantage of the invention.

By standard statistical practices it is possible to calculate the number of cDNAs
required to have representation of mRNAs expressed in low copy numbers. In most cases,
creating a library of approximately 100,000 cDNAs gives a reasonable assurance that genes expressed at low levels are represented in the cDNA population. The invention provides a convenient way of screening these large numbers of cDNAs quickly and effectively.

C. SELECTING PROBES

As noted, the invention contemplates the use of probe sets far smaller than those used by other methods. The number of probes used in practicing the invention will range from about 10 to about 75, more preferably from about 10 to about 50, even more preferably from about 10 to about 30, and still more preferably from about 10 to about 25. Most preferably, the number of probes used will range from about 15 to about 20.

Typically, the probes will be from about 12 to about 30 nucleotides in length. Probes about 20 nucleotides or shorter are preferred since longer probes tend to be more specific in their hybridizations.

Since the number of probes used is so small, their design is important to the practice of the invention. Fortunately, many of the criteria for selection of oligonucleotides suitable for use as probes in the invention are the same as the criteria for selecting oligonucleotides for performing the polymerase chain reaction (“PCR”). Thus, selection of oligonucleotides for use as probes can begin with a software program for selecting oligonucleotide primers for PCR. A number of such programs are available commercially and on-line, including programs found at the web sites maintained by: chemie.uni-marburg.de; genome.wi.mit.edu; alces.med.umn.edu; and gcg.com.

Probe design can also be performed manually. As with the programs for selecting PCR primers, probe design typically begins with the selection of a “seed” set of genes. That is, a group of genes, usually selected to be different, is chosen for consideration as to whether they contain sequences useable as probes. Conveniently, genes known to be different can be selected simply by selecting genes from the human, mouse, and rat “UniGene” (“unique gene”) libraries. The UniGene human library currently contains clones of over 20,000 human genes chosen as representing unique, individual genes. Some 4,000 clones of this library have validated sequences and are of named genes (16,000 more clones in the library are also of unique human genes, but have not yet been identified).

Additional thousands of unique mouse and rat genes are available from the UniGene libraries of those species. Clones from the human, mouse, and rat UniGene libraries are
commercially available from Research Genetics, Inc. (Huntsville, AL), Genome Systems Inc. (St. Louis, MO), and the American Type Culture Collection (Manassas, VA).

Once a set of genes has been selected as a seed set, the sequences of the genes are first examined in the same manner as for selecting PCR primers. For example, standard texts teach with respect to PCR that typical primers are 15 to 28 nucleotides in length and have 50 to 60% G and C composition. See, e.g., Innis, et al., eds., PCR Protocols, Academic Press, San Diego (1990) at page 9 (the entirety of Innis et al. is hereby incorporated by reference). Further, Innis et al. teach that complementarity within primer pairs should be avoided as this promotes the formation of primer-dimer artifacts, as should palindromic sequences within primers. Id. All of these considerations can be used in designing probes for use in the invention.

Oligonucleotides can be tested for self-complementarity and for complementarity among themselves using annealing tests described by Hillier and Green (PCR Methods and Applications. 1:124-128 (1991)), with slight modifications. For tests of self-complementarity, a sequence in the 5′ to 3′ orientation is compared with the same sequence in the 3′ to 5′ orientation. For tests of complementarity between two different oligonucleotides, the two sequences are placed in opposing orientations (that is, one of placed in the 5′ to 3′ orientation and the other in the 3′ to 5′ orientation).

The sequences are compared in every register of comparison using a scoring matrix containing values of complementarity for every pair of nucleotide symbols. For each register of comparison, the score of each base pair comparison is determined. The scores of contiguous base pairs with positive comparison values are summed. The maximum score of all such contiguous segments, taken over all registers of comparison between the sequences, determines the total “oligo-oligo” annealing score. Complementarity at the 3′ ends of the oligonucleotide sequences has a particularly large influence on PCR-induced oligonucleotide-dimer formation. Therefore, the maximum score of all contiguous segments that include the position of either oligonucleotide sequence, taken over all registers of comparison, is separately determined as the 3′ oligo-oligo annealing score. The lower the score, the more desirable the oligonucleotide for use as a probe. All other factors being equal, the oligonucleotides are ranked for use as probes in order of their score, from lowest to highest.
Once an initial group of oligonucleotides for potential use of probes has been
5 generated, either manually or by use of a program for selecting PCR primers, and has met
the criteria set forth above, they can be further tested to select a group of oligonucleotides
particularly suitable for use in the invention. First, it is desirable that the oligonucleotides
selected for the probes have melting temperatures within a few degrees of one another.
Thus, it is preferable that the oligonucleotides have melting temperatures within about 8°C
of one another, and more preferable that the melting temperatures be within about 6°C of
each other. Even more preferably, the oligonucleotides should have melting temperatures
within about 4°C of each other. Most preferably, the melting temperatures will be within
about 2°C of each other. It will often occur that it is difficult to obtain sets of
oligonucleotides which meet all the other criteria and in which all of the Tms are within
2°C or 4°C of each other. Typically, however, the Tms of the oligonucleotides under
consideration will cluster in two or more groups, with each group consisting of
oligonucleotides with Tms which are close to each other. These groups of oligonucleotides
with closely related Tms will frequently be employed.

Melting temperatures can be calculated using the nearest-neighbor model of Borer,
15 et al. (J. Mol. Biol. 86:843-853 (1974)) as modified slightly by Rychlik, et al. (Nucleic
Acids Res. 18:6409-6412 (1990)) and the thermodynamic parameters for DNA nearest-
neighbor interactions determined by Breslauer, et al. (Proc. Natl. Acad. Sci. USA. 83:
3746-3750 (1986)); \( T_{(m)}^{( oligo) } = \Delta H / (\Delta S + R x \ln(c/4)) - 273.15 + 16.6 x \log[K^{+}] \) where
\( \Delta H \) is the enthalpy of helix formation, \( \Delta S \) is the entropy of helix formation (including
helix initiation), \( R \) is the molar gas constant (1.987 cal/degree Celsius/mol), and \( c \) is the
oligo concentration. These calculations are also an integral part of most common PCR
oligo primer selection programs, including one or more of those discussed above.

Second, once a group of potential probes has been selected (either by a program or
20 manually), the potential probes should be prioritized by similarity analysis for
oligonucleotides that are similar to many (5-5000) cDNAs, allowing for a mismatch of 1-5
nucleotides. Similarity analysis is a standard method in the art and is routinely performed
using any of several widely available programs, such as the Basic Local Alignment Search
Tool, or “BLAST” program, which is made available by the U.S. National Institutes of
Health and can be used online at, or downloaded from ncbi.nlm.nih.gov/BLAST. Other
such programs include FASTA. (see Pearson and Lipman, Proc Natl Acad Sci (USA) 85:2444-2448 (1988) and Pearson, Meth Enzym 183:63-98 (1990), and BLITZ.

Third, each of the potential probes is preferably hybridized to a small set of cDNAs (about 50) to determine if it displays an adequate range of hybridization, preferably from zero to some cDNAs to high hybridization (80 - 100%) to others. Preferably, the gene to which the sequence is complementary is included among the cDNAs to serve as a positive control for 100% hybridization. Potential probes which show a limited range of hybridization, for example 25% to about 75% (excluding the gene to which they are complementary), or 25% to 50%, are less optimal and are less preferred.

Fourth, potential probes which perform well in this limited hybridization are tested on a more random, larger set (450-500) of cDNA sequences (hereafter, the “test cDNAs”). (If desired, the first, smaller set of hybridizations can be skipped and the practitioner can test the candidate probe set directly against the test cDNAs.). Conveniently, the test cDNAs can be selected from a UniGene library, although they can also be purchased from other sources or cloned directly from different tissues or organisms.

Hybridization is first conducted at stringent conditions. A series of sequential hybridizations is then performed with slightly less stringency in each succeeding hybridization, until non-stringency is achieved, and preferably until optimal non-stringency is achieved. Optimal non-stringency is determined as the temperature at which the broadest range of intensities of hybridizations to the cDNAs is seen, while permitting easy discrimination of the hybridizations of interest from any background which may be present. While stringency can be relaxed by changing temperature, salt conditions, or the concentration of compounds such as formamide or DMSO, it is often most conveniently controlled by changing the temperature. The hybridization patterns of each of the potential probes are then examined and oligonucleotides are selected such that the set of oligonucleotides selected for use as probes detects all of the test cDNAs to which they have been hybridized, at a full range of hybridization intensities. That is, each of the oligonucleotides selected as a probe should detect a number of the cDNAs, but should do so with intensities which vary depending on which cDNA to which it has been hybridized.

Fifth, it is desirable that no oligonucleotide be selected which hybridizes to the majority of cDNAs at the same relative intensity (sometimes referred to as an “intensity band” since in such cases, the oligonucleotide fails to provide a range of intensities to a
majority of cDNAs and therefore fails to permit discrimination among them. Thus, an oligonucleotide is less preferred for use as a probe, if it has, for example, 50% hybridization to 75% of the cDNAs, or 80% hybridization to 100% of the cDNAs, or 10% hybridization to 100% of the cDNAs. If any of the cDNAs are not detected by the set of potential probes, an oligonucleotide is designed to detect that cDNA, tested, and added to the probe set if it meets the other criteria.

Statistical analyses have shown that a probe set of 50 or fewer probes used in the methods of the invention is capable of detecting essentially all cDNAs, and that a probe set of 15 to 20 probes is capable of detecting a high percentage of all cDNAs. The preferred probe set size of 15 to 20 probes represents a balance between using an especially convenient number of probes and detecting a large enough percentage of all cDNAs to be highly useful. Larger sets, such as 25, 30, 35 or even 40, 45 or 50 probes, can, however, be used in applications in which especially high precision is desired, in which it is desired to discriminate among essentially all cDNAs, or in which an especially rare cDNA is sought.

If a particular oligonucleotide fails either of these tests, then an alternative oligonucleotide is selected. Further, if an additional cDNA is added to the test that is not detected by the set of oligonucleotides under consideration, a new oligonucleotide is designed to detect that cDNA which meets the conditions. If the new oligonucleotide does not match or is not similar to the patterns shown by any of the previous oligonucleotides, it is added to the probe set. If, instead, the pattern is a close match for the hybridization patterns of an oligonucleotide already in the proposed probe set, the oligonucleotide with the most similar hybridization pattern to the new probe is replaced with the new one. In this manner, the probe set can be optimized to detect a large number of cDNAs or other nucleic acids of interest while using only a small number of probes.

Although the sequences which ultimately become probes originate as sequences complementary to a limited set of the genes selected to be the “seed” sequences, they end by being able to detect the presence of virtually any gene. Moreover, a large number of genes can serve as the “seed” sequences. Accordingly, the particular genes selected to be “seed” sequences are not critical to the practice of the invention. Moreover, it is anticipated that numerous oligonucleotides will work in the methods of the invention. Once the practitioner understands the concept of the invention as taught herein, a large number of probe sets can be selected by routine methods to function in the methods taught
D. SELECTING LOW STRINGENCY CONDITIONS

The optimum hybridization conditions to use in the methods of the invention will vary depending on the particular set of probes chosen by the practitioner. Accordingly, once a probe set is selected, the hybridization conditions to be used with that probe set will have to be determined.

This determination is conveniently done empirically. Typically, the determination is conducted by hybridizing the probe set multiple times to a group of dissimilar cDNAs. The conditions under which the hybridizations are made less stringent for each successive hybridization, until conditions are reached under which members of the probe set hybridize to a substantial proportion of the cDNAs, and with differing levels of intensity which are above background. The hybridization conditions to be used need only be determined once for any particular probe set.

The determination usually commences by selecting the sample nucleic acids to which the probe set will be hybridized. Typically, the nucleic acids will be cDNAs (for ease of discussion, the text below will refer to the sample nucleic acids as cDNAs, although other nucleic acids can be used). It is desirable that the cDNAs be chosen to be dissimilar so that the ability of the probes to detect a variety of cDNAs can be confirmed. Conveniently, the cDNAs can be chosen from the UniGene collection of libraries so that each one is known to be from a different gene. The number of cDNAs used should not be so great as to be unwieldy, but large enough so that the practitioner can observe differences in binding affinity. Usually, between about 25 and about 100 cDNAs is sufficient for these determinations. The number of cDNAs need not exceed 1000.

Typically, the cDNAs are “spotted” on a surface in an ordered fashion, to form an array with the cDNAs; in known positions. One member of the probe set is selected (this can be done at random or in an order chosen for the practitioner’s convenience) and the melting temperature of the probe determined. Melting temperatures can be estimated or calculated by standard equations based on the purine and pyrimidine composition of the nucleotides comprising the probe. Preferably, the cDNA from which the probe was derived is included in the array to provide a positive control to provide a signal representing 100% hybridization to assist in quantitating hybridizations to other cDNAs.
The probes are initially hybridized to the cDNAs under fairly stringent conditions to provide a starting point at which hybridization is not expected to a significant degree except to the cDNA from which the probe was derived. Typically, the first hybridization is conducted at about 5°C below the melting temperature of the probe, and the hybridization of the probe to the cDNAs is then determined. Since it is unlikely that the probe will closely match the sequence of any of the genes (except for the cDNA from which it was derived and related cDNAs that have identical stretches of sequence), it is expected that it will not hybridize to a significant degree to even one of the cDNAs in the array during this first hybridization, except for the cDNA from which it was derived.

The hybridization is then repeated, but with the annealing temperature reduced by 2 to 5°C, and the hybridization to the cDNAs; at the new temperature determined. The cycle of reducing temperature and determining hybridization continues until the probe has been found to hybridize to between 5-40% (optimally, about 25%) of the cDNAs present, with a range of intensities. The same procedure is then followed for the next probe, and so on, until optimal hybridization conditions have been determined for all the probes.

If the probes have been selected carefully, the conditions for their optimal hybridization should be closely matched. As noted above with respect to the Tm of the oligonucleotides, even if the optimal conditions for the probes are not identical for all of the probes, it will usually be the case that the optimal conditions for the probes will fall into two or three groups. Hybridizations for the probes falling into each individual group can be carried out at the same time, provided that the probes within the group are labeled in a way which permits them to be distinguished from each other.

In the practice of the invention, a test nucleic acid (such as a sample) is hybridized to at least three members of a set of nucleic acids (such as a probe set), including a member having some percentage of complementarity to the test nucleic acid, a member with a lesser percentage of complementarity to the test nucleic acid than does the first member of the set, and a third with a still lower degree of complementarity. The hybridizations are commenced at the higher end of the temperature range (for example, at 10°C below the calculated Tm), and repeated at successively lower temperature until satisfactory low stringency conditions are reached. One of skill knows when satisfactory low stringency conditions are reached when the three nucleic acids bind to the test nucleic acid with a degree of intensity proportional to their respective degrees of complementarity to the test
nucleic acid and above background. That is, the member of the set with the highest degree of complementarity binds with the greatest intensity, the member with the intermediate degree of complementarity binds with an intensity below that of the most complementary member of the set but higher than that of a less complementary member, and the least complementary member binds with less intensity than either of these two members. The lowest intensity binding should still be distinguishable as above background so that the practitioner can confirm all three oligonucleotides have bound to the test nucleic acid.

It is desirable both during the testing and during use of the final probe set to include as one of the test nucleic acids the cDNA from which each probe was derived. This provides a positive control for 100% hybridization for each probe, and permits quantitation of the signal from other hybridizations.

The following are four examples of non-stringent hybridization conditions which can be used to hybridize probes to nucleic acids of interest, illustrating four different ways of varying parameters to obtain non-stringent conditions.

1. The nucleic acids of interest and the probes can be incubated overnight at 42°C, in 6x SSC (that is, 0.9M NaCl, 0.1M NaCitrate). 5x Denhardt’s (that is, 0.1% each Ficoll, polyvinylpyrrolidone, and bovine serum albumin), and 0.1% sodium dodecyl sulphate (SDS). The hybridized nucleic acids and probes are then washed 2 to 3 times with 6x SSC and 0.1% SDS for 30 minutes each time at room temperature. Finally, they are washed again for 20 min, at about 9 to 25°C, at about 10 to 25°C, at about 11 to 25°C, or at about 12 to about 25°C below the calculated Tm of the oligonucleotide, in 6x SSC, and 0.1% SDS. The actual temperature used will take into account the salt and formamide concentrations, and can be tested empirically by the method described above to determine if a satisfactory signal is obtained from the desired hybridization while permitting discrimination of that signal from any background which may also be present.

2. Exemplary formamide conditions. Prehybridization and hybridization are conducted at room temperature (considered to be from about 66°F to about 73°F) in a solution composed 50% of formamide, 5x SSC, 20 mM Tris (pH 7.6), 1% of Denhardt’s, 10% of dextran sulfate, and 0.1% SDS. The wash is conducted in 0.2x SSC, 0.1% SDS, at about 9 to 25°C, at about 10 to 25°C, at about 11 to 25°C, or at about 12 to about 25°C below the calculated Tm. The actual temperature used will take into account the salt and formamide concentrations, and can be tested empirically by the method.
described above to determine if a satisfactory signal is obtained from the desired hybridization while permitting discrimination of that signal from any background which may also be present. See, e.g., Denhardt, Biophys. Res. Comm. 23:641 (1966); Gillespie and Spigelman, J. Mol. Biol., 12:829 (1965).

3. Modified Church’s Procedure. Prehybridization and hybridization are conducted at 45°C, in a solution of 0.25 mM sodium phosphate, pH 7.2, and 0.1% SDS. The wash is conducted in the same solution at about 9 to 25°C, at about 10 to 25°C, at about 11 to 25°C, or at about 12 to about 25°C below the calculated Tm. The actual temperature used will take into account the salt and formamide concentrations, and can be tested empirically by the method described above to determine if a satisfactory signal is obtained from the desired hybridization while permitting discrimination of that signal from any background which may also be present. See, e.g., Church and Gilbert, Proc. Natl. Acad. Sci. (USA) 81:1991 (1984).

4. TMAC conditions. Prehybridization and hybridization are conducted at 45°C, in a solution of 3 M tetramethylammonium chloride (“TMAC”), 0.1 mM sodium phosphate, pH 6.8, 1 mM EDTA, 5x Denhardt’s, and 0.6% SDS. The wash is conducted in a solution containing 3 M TMAC, 50 mM Tris-Cl, pH 8, and 0.2% SDS at about 9 to 25°C, at about 10 to 25°C, at about 11 to 25°C, or at about 12 to about 25°C below the calculated Tm. The actual temperature used will take into account the salt and formamide concentrations, and can be tested empirically by the method described above to determine if a satisfactory signal is obtained from the desired hybridization while permitting discrimination of that signal from any background which may also be present. In this method, the calculation of Tm is based on the length of the oligonucleotide, as defined in Jacobs et al., Nucl. Acids Res. 16:4637-4650 (1988).

Persons of skill in the art will appreciate that numerous permutations can be made in the combination of temperature and other parameters, such as salt concentration and the presence or absence of organic compounds such as formamide or DMSO, can be used to achieve non-stringent conditions for use in the present invention. The effects of changing these parameters are well known in the art. The effect on Tm of changes in the concentration of formamide, for example, is reduced to the following equation: $T_m = 81.5 + 16.6 \left(\log \text{Na}^+\right) + 0.41 \%\text{G+C} - (600/\text{oligo length}) - 0.63\%\text{formamide}$. Reductions in Tm due to TMAC and the effects of changing salt concentrations are also well known.
Changing the temperature is, however, a preferred method of changing the stringency of hybridization due to its simplicity and ease of control.

**E. PERFORMING HYBRIDIZATIONS ON SOLID SUPPORTS OR FLUID ENVIRONMENTS**

The hybridizations can be performed while either the probes or the nucleic acids of interest are attached to solid supports, or while they are in a fluid environment.

In one set of embodiments, the hybridizations are performed on a solid support. For example, the nucleic acids of interest (or “samples”) can be spotted onto a surface. Conveniently, the spots are placed in an ordered pattern, or array, and the placement of where the nucleic acids are spotted on the array is recorded to facilitate later correlation of results. The probes are then hybridized to the array.

The composition of the solid support can be anything to which nucleic acids can be attached. It is preferred if the attachment is covalent. The material for the support for use in any particular instance should be chosen so as not to interfere with the labeling system to be used for the probes or the nucleic acids. For example, if the nucleic acids are labeled with fluorescent labels, the material chosen for the support should not be one which fluoresces at wavelengths which would interfere with reading the fluorescence of the labels.

Preferably, the support is of a material to which the samples and probes bind or one which is substantially non-porous to them, so that the oligonucleotides remain accessible (i.e., to the probes or the samples) at the surface of the support. Membranes porous to the nucleic acids may be used so long as the membrane can bind sufficient amounts of nucleic acid to permit the hybridization procedures to proceed. Suitable materials should have chemistries compatible with oligonucleotide attachment and hybridization, as well as the intended label, and include, but are not limited to, resins, polysaccharides, silica or silica-based materials, glass and functionalized glass, modified silicon, carbon, metals, nylon, natural and synthetic fibers, such as wool and cotton, and polymers.

In some embodiments, the solid support has reactive groups such as carboxy-amino- or hydroxy groups to facilitate attachment of the oligonucleotides (that is, the samples or the probes). Plastics may be used if modified to accept attachment of nucleic acids or oligonucleotides (since plastic usually has innate fluorescence, the use of non-fluorescent labels is preferred for use with plastic substrates. If plastic materials are used with
fluorescent labels. Appropriate adjustments should be made to procedures or equipment, such as the use of color filters, to reduce any interference in detecting results due to the fluorescence of the substrate. Polymers may include, e.g., polystyrene, polyethylene glycol tetraphtalate, polyvinyl acetate, polyvinyl chloride, polyvinyl pyrrolidone, buty rubber, and polycarbonate. The surface can be in the form of a bead. Means of attaching oligonucleotides to such supports are well known in the art, and are set forth, for example, in U.S. Patent Nos. 4,973,493 and 4,569,774 and PCT International Publications WO 98/26098 and WO 97/46313. See also, Pon et al., Biotechniques 6:768-775 (1988); Dammba, et al., Nuc. Acids Res. 18:3813-3821 (1990).

Alternatively, the samples can be placed in separate wells or chambers and hybridized in their respective well or chambers. The art has developed robotic equipment permitting the automated delivery of reagents to separate reaction chambers, including “chip” and microfluidic techniques, which allow the amount of the reagents used per reaction to be sharply reduced. Chip and microfluidic techniques are taught in, for example, U.S. Patent No. 5,800,690, Orchid. “Running on Parallel Lines” New Scientist, Oct 25, 1997, McCormick, et al., Anal. Chem. 69:2626-30 (1997), and Turgeon, “The Lab of the Future on CD-ROM?” Medical Laboratory Management Report. Dec. 1997, p. 1. Automated hybridizations on chips or in a microfluidic environment are contemplated methods of practicing the invention.

Although microfluidic environments are one embodiment of the invention, they are not the only defined spaces suitable for performing hybridizations in a fluid environment. Other such spaces include standard laboratory equipment, such as the wells of microtiter plates, Petri dishes, centrifuge tubes, or the like can be used.

F. CONDUCTING HYBRIDIZATIONS

1. Sequence of hybridizing

The probes can be hybridized sequentially. When this procedure is followed, a probe is hybridized to the sample and its hybridization intensity noted. The hybridized probe is then subjected to conditions (such as heat or a change in salt concentration) causing the probe to separate from the sample, whereupon the probe is washed off. The process is then repeated until the hybridization of all of the probes to the sample has been conducted and the results recorded.
Alternatively, if the probes are labeled in a way which permits them to be
distinguished from each other, the probes so labeled can be hybridized to the sample at the
same time. Hybridization of multiple probes at the same time is preferred since in the
process of removing one probe so that the next can be introduced, some of the sample of
interest can be lost. By contrast, hybridizing multiple probes at the same time normalizes
the amount of the sample of interest present with respect to those probes and permits
comparisons of relative intensities of hybridization. As noted in previous sections, it is
likely that the probes will fall into several groups, based on the non-selective hybridization
conditions which work well for the members of that group. In these cases, one group of a
set of probes may be hybridized and then removed before the next group of the set is
hybridized. To reduce loss of sample to the extent possible, and permit readier
comparisons of hybridization intensity, it is desirable that as many of the probes be grouped
together as their optimal hybridization conditions permits, and that the number of separate
groups hybridized be minimized.

When more than one probe is hybridized at a time, two or more probes may
compete for the same binding site. In this case, which probe binds to the target will depend
on such variables as the relative affinity of the probes to the target site and the respective
molar concentrations of the probes. The hybridization pattern obtained when a single probe
is hybridized to a target in the absence of competing probes may therefore not be the same
as when the same probe is hybridized to the same target in the presence of other probes.
Thus, if the probes are to be hybridized in groups, once the probes have been placed in
appropriate groups by hybridization conditions, it is desirable for them to continue to be
hybridized as part of the same group of probes thereafter to reduce variations which might
be introduced into the hybridization patterns of the probes by changes in the group of
probes used. Changing the molar ratios of the probes can also introduce variations in the
resulting hybridization patterns. For example, increasing the molar ratio of probe A to
probe B by ten-fold may allow probe A to outcompete probe B for any binding sites they
might share, and thus change the resulting hybridization pattern. It is therefore desirable
that the molar ratios of the respective probes not be varied enough to affect the

2. **Controlling concentration of sample nucleic acids**
It is desirable to control the concentration of the sample nucleic acids (such as cDNAs) which are being probed. Left uncontrolled, variations in the concentration or amount of the nucleic acids will result in variations in the intensity of binding of the various probes and render the results more difficult to interpret.

In many embodiments, the practitioner will validate the system to be used, finding the degree of variation, and reducing it to 5% or less. This can be tested by standard methods. For example, if a membrane is being used, a single species of labeled cDNA, such as one tagged with a radioactive or fluorescent label, is spotted multiple times on the membrane and the amount of the cDNA on each spot is quantitated. Standard statistical analysis of the variation in the amount of cDNA will give the practitioner information on the degree of difference in intensity needed to reflect a real difference in the amount of probe bound to the cDNA. By standard statistical methods, for example, a difference of three times the error rate or more (e.g., 15%, if the error rate is 5%) gives a 98% confidence level that a difference in intensity is due to hybridization of the probe, rather than a difference in the amount of cDNA present on a spot. For many applications of the invention, that will provide sufficient certainty for meaningful comparisons to be made.

In embodiments where the reproducibility of sample placement is not well controlled, where the precision of the system is particularly important (as might be the case, for example, in forensic applications), or where otherwise desired, internal controls can be included to increase the precision of the comparisons of intensity. Such controls can include tagging the samples with, for example, radioactive or fluorescent labels, quantitating the amount of nucleic acid for one or more of the samples of interest, and adjusting the intensity read for each probe to account for the variation in the amount of nucleic acid in the sample. In applications where precision is critical, every sample can be quantitated or normalized.

An alternative method of normalizing the amount of nucleic acid is available particularly with respect to cDNAs. All of the cDNAs used in a particular study will usually be prepared in the same cloning vector. One can design a vector-specific probe that can be hybridized under stringent conditions to the vector, with the amount of hybridization being indicative of the amount of test nucleic acid present. This value can be used to normalize the readings of the intensity of the hybridizations of the probes to the test nucleic acid. In a preferred embodiment, the vector-specific probe is labeled with a label.
distinguishable from any labels used for the test or probe nucleic acids. In a further preferred embodiment, the vector-specific probe is designed to have a Tm sufficiently lower than the Tm of the probe set hybridized to the test nucleic acid so that conditions which are non-stringent for the probes are stringent for the vector-specific probe. This permits the stringent hybridization to the vector to be conducted at the same time as a non-stringent hybridization to the test nucleic acid.

The amount of nucleic acid present can be also normalized using the bar code produced by hybridization of the probes to the cloning vector. Since the bar code of the cloning vector will be the same, it can then be subtracted from the hybridization to the cDNA to obtain an accurate reading of the bar codes for the nucleic acid of interest. If normalization using the cloning vector is not needed or not desired, an additional criterion can be added to the selection of the probes that they do not hybridize to the cloning vector. While this complicates the selection process, it is possible since the cloning vector constitutes a short, defined sequence. Because of the extra difficulty of this selection, as opposed to the relative ease of subtracting out the effect of the cloning vector, and of the possible assistance in normalization from the bar code of the vector, selecting probes which do not hybridize to the vector will not normally be preferred.

**G. OBTAINING MOLECULAR LIBRARIES**

Methods for obtaining genomic and cDNA libraries are well known in the art. Sambrook, for example, sets forth detailed instructions for choosing sources of mRNA, assuring the integrity of the mRNA, determining approximately how many clones are needed have a reasonable probability of obtaining a clone containing the mRNA of interest, appropriate vectors and adapters, and cloning cDNA from the mRNA. *See*, Sambrook, *supra*, at chapter 8. In particular, Sambrook provides detailed instructions on constructing bacteriophage vectors. More recently, techniques such as “CAP traps” and anchoring the 3’ tail have increased the probability of obtaining full-length cDNAs. *See*, e.g., Chenchik, A., *et al.*, Biotechniques, 21(3):526-534 (1996); Hakvoort, T., *et al.*, Nucl Acids Res 24(17):3478-3480 (1996); Shirozu, M., *et al.*, Genomics, 37(3):273-280 (1996); Carninci, P., *et al.*, Genomics, 37(3):327-336 (1996). *See also*, the following Internet sites (all these sites are at clontech.com): /clontech/TechTips/ SMARTTechTip; /clontech/CATALOG/librarytoc; /archive/JAN96UPD/CapFinder.
Bacteriophage vectors and plasmid vectors are preferred for practicing the invention. Preferably, the process generates a full length cDNA. The full length cDNA should not vary appreciably since it contains the coding portion of the gene and therefore should yield a consistent hybridization pattern.

Sambrook further sets forth detailed protocols for obtaining genomic DNA libraries. See, Sambrook, *supra*, at chapter 9. Commonly, the genomic DNA is treated with restriction enzymes to cleave it at desired restriction sites. If necessary, the resulting fragments can be subjected to size fractionation through, for example, gel electrophoresis or HPLC, prior to cloning and hybridization.

Genomic and cDNA libraries can also be purchased commercially from a number of suppliers. The I.M.A.G.E. consortium (coordinated by the Lawrence Livermore National Laboratory, Livermore CA), for example, has made over one million cDNAs available through suppliers. The three authorized suppliers in the United States are the American Type Culture Collection (“ATCC”, Manassas, VA), Genome Systems, Inc. (St. Louis, MO), and Research Genetics, Inc. (Huntsville, AL). The ATCC is also the supplier of thousands of cDNAs from other sources, such as The Institute for Genomic Research.

**H. LABELS**

Either the probes or the nucleic acids of the samples can be labeled to permit detection of hybridization. If the probes are labeled, preferably each probe has a label which is separately detectable, such as a fluorophore with a color different from that of the other fluorophores used. A wide variety of labels and conjugation techniques are known and are reported extensively in both the scientific and patent literature, and are generally applicable to the present invention. Suitable labels include radionucleotides, enzymes, substrates, cofactors, inhibitors, fluorescent moieties, chemiluminescent moieties, magnetic particles, and the like. Labeling agents optionally include *e.g.*, proteins. Detection of labeled nucleic acids or proteins may proceed by any of a number of methods, including immunoblotting, tracking of radioactive or bioluminescent markers, or methods which track a molecule based upon size, charge or affinity. The particular label or detectable moiety used and the particular assay are not critical aspects of the invention.

The detectable moiety can be any material having a detectable physical or chemical property. Such detectable labels have been well developed in the field of gels, columns, and solid substrates, and in general, labels useful in such methods can be applied to the
present invention. Thus, a label is any composition detectable by spectroscopic, photochemical, immunochemical, electrical, optical or chemical means. Useful labels in the present invention include fluorescent dyes (e.g., fluorescein isothiocyanate, Texas red, rhodamine, and the like), radiolabels (e.g., \(^{3}\text{H}\), \(^{125}\text{I}\), \(^{35}\text{S}\), \(^{14}\text{C}\), or \(^{32}\text{P}\)) enzymes (e.g., LacZ, CAT, horse radish peroxidase, alkaline phosphatase and others, commonly used as detectable enzymes, either as marker gene products or in an ELISA), nucleic acid intercalators (e.g., ethidium bromide) and colorimetric labels such as colloidal gold or colored glass or plastic (e.g. polystyrene, poly-propylene, latex,) beads, as well as electronic transponders (e.g., U.S. Patent 5,736,332).

Where multiple fluorophores are used, it should be noted that two fluorophores binding in close proximity can result in combined signals. For example, a yellow fluorophore and a blue fluorophore in close proximity could appear to a viewer as a green color. To increase the ability to interpret results when using fluorophores or other labels which can be combined, the probe set can be chosen so that no more than two probes can combine to produce a given combined signal (for example, that only the colors of two fluorophores used when seen in proximity will appear green), or the probes can be read in a manner which permits the two signals to be distinguished. In the case of fluorophores, for example, this is conveniently done by using filters which permit each color to be seen individually.

It will be recognized that fluorescent labels are not to be limited to single species organic molecules, but include inorganic molecules, multi-molecular mixtures of organic and/or inorganic molecules, crystals, heteropolymers, and the like. Thus, for example, CdSe-CdS core-shell nanocrystals enclosed in a silica shell can be easily derivatized for coupling to a biological molecule (Bruchez et al. Science, 281: 2013-2016 (1998)).

Similarly, highly fluorescent quantum dots (zinc sulfide-capped cadmium selenide) have been covalently coupled to biomolecules for use in ultrasensitive biological detection (Warren and Nie, Science, 281: 2016-2018 (1998)).

In some embodiments, antibodies can be used as labels. For example, a probe can contain a modified base, such as bromodeoxyuridine (“BrdU”), and detected by use of an anti-BrdU antibody.

The label is coupled directly or indirectly to the probe or desired nucleic acid according to methods well known in the art. As indicated above, a wide variety of labels
may be used, with the choice of label depending on the sensitivity required, ease of conjugation of the compound, stability requirements, available instrumentation, and disposal provisions. Non-radioactive labels are often attached by indirect means. Generally a ligand molecule (e.g., biotin) is covalently bound to a polymer. The ligand then binds to an anti-ligand (e.g., streptavidin) molecule which is either inherently detectable or covalently bound to a signal system, such as a detectable enzyme, a fluorescent compound, or a chemiluminescent compound. A number of ligands and anti-ligands can be used. Where a ligand has a natural anti-ligand, for example, biotin, thyroxine, and cortisol, it can be used in conjunction with labeled anti-ligands. Alternatively, any haptenic or antigenic compound can be used in combination with an antibody.

Labels can also be conjugated directly to signal generating compounds, e.g., by conjugation with an enzyme or fluorophore. Enzymes of interest as labels will primarily be hydrolases, particularly phosphatases, esterases and glycosidases, or oxidoreductases, particularly peroxidases. Fluorescent compounds include fluorescein and its derivatives, rhodamine and its derivatives, dansyl, umbelliferone, fluorescent green protein, and the like. Glass is a preferred substrate when fluorescent labels are used. Chemiluminescent compounds include luciferin, and 2,3-dihydrophthalazinediones, e.g., luminol.

Means of detecting labels are well known to those of skill in the art. Thus, for example, where the label is a radioactive label, means for detection include a scintillation counter, proximity counter (microtiter plates with scintillation fluid built in), or photographic film as in autoradiography. Where the label is a fluorescent label, it may be detected by exciting the fluorochrome with the appropriate wavelength of light and detecting the resulting fluorescence, e.g., by microscopy, visual inspection, via photographic film, by the use of electronic detectors such as charge coupled devices (CCDS) or photomultipliers and the like. Commercial analyzers, such as the Storm® and FluorImager® systems (Molecular Dynamics Inc., Sunnyvale, CA) for gel and blot analysis of direct and chemiluminescence, can also be used, and the Molecular Dynamics’ PhosphorImager system can be used for radiographic analysis. Similarly, enzymatic labels may be detected by providing appropriate substrates for the enzyme and detecting the resulting reaction product. Finally, simple colorimetric labels are often detected simply by
observing the color associated with the label. Thus, in various assays, conjugated gold often appears pink, while various conjugated beads appear the color of the bead.

I. COMPARING HYBRIDIZATION PROFILES

1. Compiling hybridization profiles

As noted in the preceding sections, alterations in gene expression can be detected by comparing patterns of hybridization for nucleic acids. If, for example, the question to be answered is whether a chemical composition is potentially toxic and, if so, what kind of toxicity it might possess, duplicate cultures of cells, tissues, embryonic bodies, or the like, can be set up in which one culture (the “test” culture) is contacted with the chemical composition and one (the “control” culture) is not. cDNA libraries can then be made from the test and the control cultures and the hybridization patterns of the two compared. An identical hybridization pattern indicates that the chemical composition has not changed the gene expression of the cell, tissue, or embryoid body. A changed hybridization pattern means that the expression of genes has been changed by the contact with the chemical composition. It should be noted that these tests can be conducted gene by gene or by looking at mixtures of genes to determine if there is a difference in the hybridization patterns overall, indicating a change in the expression of the genes between the test and control cultures. If a difference is noted, the two cultures can be cloned out to determine the differences in gene expression. Additionally, libraries of hybridization patterns can also be compiled by contacting cultures with additional chemical compositions, forming cDNA libraries, and recording the resulting hybridization patterns.

While contacting cells, tissues, or embryoid bodies with chemical agents is an exemplar method of practicing the invention, many others are possible. For example, cDNA libraries can be made of diseased and normal tissue and hybridization patterns compared to determine which genes are expressed differently in the diseased tissue. Similarly, cDNA libraries can be made of species within a genus and the hybridization patterns examined to determine the differences in expression underlying and identifying the differentiation into species.

2. Comparing Hybridization Patterns

Once a group of hybridization patterns has been collected (for example, by contacting cells, tissues, or embryoid bodies with chemical compositions and hybridizing the cDNA to determine changes in hybridization patterns induced by the chemical
compositions), they can be compared. In one embodiment, the patterns can be comparing
by assigning numeric values to the intensity of hybridization of each of the probes
hybridized to the sample. Thus, the hybridization pattern for the sample can be digitized
and represented by a series of numbers, which can then be compared to the series of
numbers resulting from the hybridization of other samples. The series of numbers
representing the hybridization of the probes to the sample can also be stored for later use,
such as comparison with series of numbers representing hybridizations conducted at a later
time with other samples. As noted earlier, it is desirable (both during initial testing of the
probes and during use of the final probe set) to include for each probe the cDNA from
which it was derived to provide a control for 100% hybridization and to facilitate
quantitation of the signal from hybridizations to the test nucleic acids.

In practice, due to normal variations in the experiment and margins of error in
reading the labels, it will not be possible to read intensity levels to a two digit level of
accuracy (that is, it is unlikely that a reading of 53% will represent a statistically
meaningful difference from 56%). Thus, the practitioner will normally follow the common
statistical practice of assigning various intensities to a “band” or a “bin.” For example,
values 0-10 might be assigned to band 1, values 11-20 to band 2, and so on, so that a
reading of 15 will fall in band 2 while a reading of 22 will fall in band 3. The practitioner
will generally set up the number of bands into which the numbers will be placed according
to the accuracy of the system and labeling used. For systems with a substantial margin of
error, for example, the practitioner might set up 5 bands of intensity readings. For systems
with a smaller margin of error, the practitioner might set up 10 bands, or even 20, if the
accuracy of the system permits.

The patterns can be graphically represented by bars whose height or width
corresponds to the band assigned to the intensity of hybridization of each probe, to form a
“bar code” representing the hybridization of the different probes to a sample of interest.
Figure 1 demonstrates how a bar code can be created from the determination of
hybridization patterns. cDNA clones representing eight individual genes, designated as
numbers 1-8, are spotted onto filters labeled A-G. Each filter is hybridized with a
particular probe. In this system, it has been assumed that the accuracy of reading permits
the grouping of readings of 5 bands. The intensity of the hybridization of the probe to each
gene on the filter (represented by the degree of darkness of the spot) has therefore been
assigned a corresponding quantitative value on a scale of 1 to 5, which for convenience has been printed just to the lower right of each spot. The graphs to the right of the figure depict the intensity with which each probe hybridized to a particular by vertical bars, whose height is proportional to the intensity of hybridization of the corresponding probe. The resulting graph presents a bar code representing the hybridization of that gene to the probe set used for the hybridizations.

The hybridization patterns can be read at the time of hybridization, or stored for later analysis. If the patterns are determined by autoradiographs (in the case of radioactive labels) or fluorescence (in the case of fluorescent labels), for example, they can be photographed. The photographs can be stored, for example, in file folders or the like, and examined visually to discern common patterns of expression compared to the control, as well as differences. Conveniently, however, the data can be stored on and compared by a computer. Preferably, the results are placed into a computer database, with information pertaining to the sample recorded in searchable data fields.

Entries of data from other forms of detecting alterations in gene expression can also be reviewed and recorded manually or in a computer database. For example, the values from an ELISA, or the proteins identified on a Western blot can be recorded to identify the types and amounts of proteins expressed in control and test samples. Similarly, Northern blots or PCR can be run and recorded to confirm the identity of the genes expressed in control and test samples. The information from these other sources can be correlated to that acquired through use of the invention. The information can be kept manually, but preferably is compiled and maintained in a computer searchable form. Standard database programs, such as Enterprise Data Management (Sybase, Inc., Emeryville, CA) or Oracle8 (Oracle Corp., Redwood Shores, CA) can be used to store and compare information.

Companies such as Incyte Pharmaceuticals, Inc. (Palo Alto, CA), which provide oligonucleotide hybridization services maintain proprietary image recognition algorithms to record and analyze the scanned images, of hybridization arrays. Other image analysis software, such as ImageQuant™ from Molecular Dynamics Inc. (Sunnyvale, CA), is also commercially available to quantify and analyze scanned images.

In a preferred embodiment, the data can be recorded and analyzed by neural network technology. Neural networks are complex non-linear modeling equations which are specifically designed for pattern recognition in data sets. One such program is the
NeuroShell Classifier™ classification algorithm from Ward Systems Group, Inc. (Frederick, MD). Other neural network programs are available from, e.g., Partek, Inc., BioComp Systems, Inc. (Redmond WA) and Z Solutions, LLC (Atlanta, GA).

3. **Generating libraries of profiles**

As noted, one use for the invention involves contemplates multiple iterations of contacting cells, tissues, embryoid bodies or organisms with an ever-widening group of chemical compositions. The toxicities and biological effects of many chemical compositions are already known through previous animal or clinical testing. Any such information on any chemical composition of interest can be carefully noted and the hybridization pattern of RNA or DNA from cells, tissues, embryoid bodies or organisms contacted with the chemical composition determined. As the data from tests on a number of chemical compositions, or agents, is gathered, it is assembled to form a library. Separate libraries can be maintained for each type of toxicity; preferably, a single database can be maintained recording the results of all the tests conducted and any available toxicity information on the agents to which the cells, tissues, embryoid bodies, or organisms were exposed. In a preferred embodiment, the tests are conducted using embryoid bodies. Preferably, biological effects are also noted. Past experience has indicated that biological effects often become associated with, or markers for, particular toxicities as the biology of the toxicity becomes better understood.

The invention contemplates that each iteration of contacting cells, tissues, embryoid bodies or organisms with a chemical composition and then determining the hybridization patterns for its cDNA will result in a pattern of gene expression that is characteristic of the response of the cell, tissue, embryoid body or organism to that chemical composition. The determination of the alterations in gene expression caused by a reasonably large number of chemical compounds of similar toxicity is desirable so that patterns of gene expression associated with that toxicity can be determined. The expression of a single gene, by itself, might not be significant as a marker of any particular toxicity. A change in the combination of genes expressed, however, would be highly predictive that a chemical composition has a type of toxicity similar to other agents which induce the same combination of expression. The correlation of these changes in gene expression and toxicities of the chemical compositions tested provides the power to predict the toxicity of previously untested compounds. (The use of alterations in gene or protein expression in
embryoid bodies to predict toxicities of chemical compositions is the subject of a
pending patent application.)

The correlation of hybridization patterns with toxicities can be performed by any
convenient means. For example, visual comparisons of patterns can be performed to
determine patterns associated with different types of toxicities. More conveniently, the
correlation can be done by computer, using one of the database programs discussed in
preceding sections. Preferably, the correlation is performed by a computer using a neural
network program, since neural network programs are specifically designed for pattern
recognition. Once a correlation has been made of hybridization patterns and patterns of
gene expression which are biomarkers for a particular toxicity, a comparison can be made,
again conveniently by computer, of known hybridization patterns induced by a new or
unknown chemical composition to provide the closest matches. The patterns can then be
reviewed to predict the likely toxicity of the new or unknown chemical.

J. ADAPTING ARRAY READERS

In one embodiment, the invention relates to the formation of arrays of hybridized
oligonucleotides to detect changes in gene expression. Such arrays can be scanned or read
by array readers.

Typically, the array reader will have an optical scanner adapted to read the pattern
of labels on an array, such as of hybridized oligonucleotides, operably linked to a computer
which has stored on it, or accessible to it (for example, on an external drive or through the
internet) one or more data files having a plurality of gene expression profiles of, for
example, mammalian embryoid bodies contacted with known or unknown toxic chemical
compositions.

The computer can be, for example, a PC, an Apple, a Sun workstation, or a
computer compatible with one of these formats. The operating system can be, for example,
a Microsoft operating system, an Apple operating system, a Unix based system, a Linux
based system, or a Java based system.

The array reader can be adapted with a detection device suitable to “read” labels
that can not be read optically, such as electronic transponders. The detection device can
further be, for example, a fluorescence detector, a radioactivity detector, such as a
scintillation counter or a Geiger counter. Further, it can be a CCD, a photomultiplier, or a
microscope. If the labels are radioactive, the hybridization patterns can be
autoradiographed and read by a device for determining the density of an image.

The array reader, in combination with a computer, a detection device, or both, a
library of hybridization patterns, and an algorithm for comparing the hybridization pattern
of a sample with members of its library of hybridization patterns, constitutes an integrated
system for detecting changes in gene expression or the presence or absence of a
polymorphism or gene of interest in a sample.

In some embodiments, the array reader can track reactions by means of substrates
with distinguishing characteristics, such as differing spectral properties. For example,
molecules of a particular cDNA can be coupled to microspheres with characteristics such
as a color-code readable by an appropriate device, such as a laser, and then hybridized with
one or more probes. If all the molecules of that cDNA are coupled to microspheres of the
same characteristic (such as a color) and the molecules of other cDNAs are coupled to
microspheres of different characteristics (such as different colors), each cDNA species can
be distinguished from the others by simply noting the characteristic of the microsphere to
which they are bound. The intensity of binding of the probe to each cDNA can then be
determined.

In a preferred embodiment, the microspheres are color-coded microspheres
available from Luminex Corp. (Austin, TX). Luminex currently provides microspheres of
some 100 different colors, which can be read individually even when mixed together in, for
example, the wells of a microtiter plate. Thus, some 100 different cDNAs can be coupled
to microspheres of colors which are different for and attributable to each of the respective
cDNAs, and the microspheres placed in the well of a microtiter plate or the reaction
chamber of a microfluidic device (a reaction chamber can be any defined space in which
reactions can be conducted). The cDNAs can then be hybridized with a probe set. The
Luminex 100 benchtop analyzer then uses lasers to determine the color of each bead. If the
probes hybridized to the cDNAs are also color-labeled, the equipment can also capture the
color of the probe, at a rate of 20,000 microspheres a second.

K. USE IN HIGH THROUGHPUT SCREENING

The methods of the invention can be adapted to high throughput screening of large
numbers of cDNAs. High throughput ("HTP") screening is highly desirable in a variety of
contexts. For example, assessing the many interrelated effects of a compound under
consideration for development as a pharmaceutical is a complicated, costly, and multiyear effort. The invention permits a large number of genes to be screened for changes in expression levels between control and test samples. For example, a library of cDNAs can be arrayed on a substrate with the positions of the various cDNAs recorded. A change in the hybridization pattern of a particular cDNA can be detected and quantitated, and the cDNA identified. In this manner, the biological effect of the compound can be more readily determined and important information gained on the suitability of the compound as a drug.

HTP screening can be facilitated by using automated and integrated culture systems, sample preparation (RNA/cDNA), and analysis. These steps can be performed in regular labware using standard robotic arms, or in more recently developed microchip and microfluidic devices, such as those developed by Caliper Technologies Corp. (Palo Alto, CA), as described in U.S. Patent 5,800,690, by Orchid Biocomputer, Inc. (Princeton, NJ), as described in the October 25, 1997 New Scientist, and by other companies, which provide methods of automated analysis using very low volumes of reagents. See, e.g., McCormick, R., et al., Anal. Chem. 69:2626-2630 (1997); Turgeon, M., Med Lab. Management Rept, Dec. 1997, page 1.

The LabMAP system provided by Luminex Corp. (Austin, TX), provides HTP analysis of samples coupled to microspheres using a combination of microfluidics, lasers, and optical readers. Microspheres can be placed in a reaction chamber, such as a well of a microtiter plate, coupled to molecules of a particular cDNA, blocked, and a second set of microspheres added and coupled to molecules of a second cDNA, blocked, and so on. The cDNAs can then be hybridized in the reaction chamber to a probe set and the hybridizations determined. The use of color-coded microspheres, such as those available from Luminex, permits a number of cDNAs to be physically co-located (as in a well of a microtiter plate or in microfluidic chambers), yet remain distinguishable from each other for reading the results of hybridization to the probes. This permits the cDNAs to be tested and analyzed in compact spaces and speeds up the ability to read and quantitate the results. As noted in the previous section, some 20,000 microspheres can be read per second. If desired, the probes, rather than the cDNAs, can be bound to the microspheres.
All publications and patent applications cited in this specification are herein incorporated by reference as if each individual publication or patent application were specifically and individually indicated to be incorporated by reference.

Although the foregoing invention has been described in some detail by way of illustration and example for purposes of clarity of understanding, it will be readily apparent to those of ordinary skill in the art in light of the teachings of this invention that certain changes and modifications may be made thereto without departing from the spirit or scope of the appended claims.
CLAIMS

WHAT IS CLAIMED IS:

1. A method of identifying a test nucleic acid as identical to a known nucleic acid, the method comprising:
   i) arraying the test nucleic acid on a solid support;
   ii) hybridizing the test nucleic acid to a set of nucleic acids under low stringency conditions, wherein the set of nucleic acids comprises at least three members which have differing degrees of complementarity to the test nucleic acid and wherein the members of the set of nucleic acids differentially hybridize to the test nucleic acid in relation to their respective degrees of complementarity to the test nucleic acid, thereby providing a first set of hybridization patterns having at least three different intensities of hybridization between the members; and,

   iii) comparing the first set of hybridization patterns to a second set of hybridization patterns, the second set comprising a plurality of hybridization patterns produced by low stringency hybridization of the known nucleic acid to the set of nucleic acids of step (ii), provided that the members of the set of nucleic acids are not degradation products of a single nucleic acid, that the set of nucleic acids are not chosen to be exactly complementary to the known nucleic acid, and that the hybridizations are not performed in situ.

2. The method of claim 1, wherein the hybridization patterns are of the intensity of a label or labels on the test nucleic acid or on the members of the set of nucleic acids.

3. The method of claim 2, wherein the labels are fluorescent labels.

4. The method of claim 3, wherein the fluorescent label of at least one member of the set of nucleic acids is of a different color from the fluorescent label of at least one other member of the set of nucleic acids.

5. The method of claim 1, wherein the solid support is a bead.

6. The method of claim 5, wherein the bead has a detectable color.

7. The method of claim 1, wherein the identity is 95% or greater.

8. The method of claim 1, wherein the first set of hybridization patterns and the second set of hybridization patterns are generated from separate hybridization reactions.

9. The method of claim 1 wherein the set of nucleic acids comprises fifty or fewer members.
10. The method of claim 1, wherein the set of nucleic acids comprises twenty or fewer members.

11. The method of claim 1, wherein the test nucleic acid is from a molecular library.

12. The method of claim 11, wherein the molecular library is selected from the group consisting of a genomic library and a cDNA library.

13. The method of claim 11, wherein the molecular library comprises at least 1,000,000 clones.

14. The method of claim 11, wherein the molecular library comprises about 50,000 to about 1,000,000 clones.

15. The method of claim 11, wherein the molecular library comprises about 1,000 to about 50,000 clones.

16. The method of claim 11, wherein a plurality of members of the molecular library are identified by their hybridization to duplicate sets of the first array of probes.

17. The method of claim 1, wherein the method comprises an additional hybridization step in which the nucleic acid is also hybridized to the set of nucleic acids under stringent hybridization conditions.

18. The method of claim 1, wherein at least 100 new labeling patterns are detected per hour.

19. A method of identifying a test nucleic acid in a sample as identical to a known nucleic acid, the method comprising:

   i) spatially arraying a labeled set of nucleic acids on a solid support in separate regions wherein the set of nucleic acids is comprised of at least three members with differing degrees of complementarity to the test nucleic acid;

   ii) hybridizing the test nucleic acid to the labeled set of nucleic acids under low stringency conditions, wherein the test nucleic acid in the sample differentially hybridizes to a plurality of members of the set of nucleic acids, thereby providing a first set of label patterns having different intensities of label between the members; and,

   iii) comparing the first set of label patterns to a second set of label patterns, the second set comprising a plurality of label patterns produced by low stringency hybridization of the known nucleic acid to the set of nucleic acids of step (ii),
provided that the members of the set of nucleic acids are not degradation products of a single nucleic acid, that the set of nucleic acids are not chosen to be complementary to the known nucleic acid, and that the hybridizations are not performed in situ.

20. The method of claim 19 wherein the hybridization patterns are of the intensity of a label or labels on the test nucleic acid or on the members of the set of nucleic acids.

21. The method of claim 19, wherein the labels are fluorescent labels.

22. The method of claim 21, wherein the fluorescent label of at least one member of the set of nucleic acids is of a different color from the fluorescent label of at least one other member of the set of nucleic acids.

23. The method of claim 19, wherein the solid support is a bead.

24. The method of claim 23, wherein the bead has a detectable color.

25. The method of claim 19, wherein the identity is 95% or greater.

26. A method of identifying a test nucleic acid in a sample, said method comprising
   i) hybridizing the test nucleic acid to a set of nucleic acids under low stringency conditions, wherein the set of nucleic acids comprises at least three different nucleic acids, each of which has a differing degree of complementarity to the test nucleic acid, the test nucleic acid binding to the members of the set of nucleic acids in relative proportion to their respective degrees of complementarity,
   ii) measuring the relative proportion of hybridization of the test nucleic acid to the members of the set of nucleic acids to provide a first set of hybridization patterns, and
   iii) comparing the first set of hybridization patterns to a second set of hybridization patterns, the second set comprising a plurality of hybridization patterns produced by low stringency hybridization of the known nucleic acid to the set of nucleic acids of step (ii), provided that the members of the set of nucleic acids are not degradation products of a single nucleic acid, that the set of nucleic acids are not chosen to be complementary to the known nucleic acid, and that the hybridizations are not performed in situ.

27. The method of claim 26, wherein the hybridization patterns are of the intensity of a label or labels on the test nucleic acid or on the members of the set of nucleic acids.

28. The method of claim 27, wherein the labels are fluorescent labels.
29. The method of claim 28, wherein the fluorescent label of at least one member of the set of nucleic acids is of a different color from the fluorescent label of at least one other member of the set of nucleic acids.

30. A method of bar coding components of a molecular library, comprising:
   i) hybridizing arrayed components of a molecular library to a limited set of probes under low-stringency conditions, wherein the set comprises at least three member probes which differ in percentage of complementarity to the test nucleic acid, the library components binding to the members of the set of probes in relative proportion to their respective degrees of complementarity, thereby creating a hybridization profile of the intensity of binding of each probe to each component of the molecular library;
   ii) recording the resulting low-stringency hybridization profile for each library component for each hybridization with a probe; and,
   iii) compiling the low-stringency hybridization profiles for hybridization to each probe to produce a bar code for each library component, which bar code provides a unique identifier for each library component.

31. The method of claim 30, wherein the hybridization profiles are of the intensity of fluorescent labels.

32. The method of claim 31, wherein the colors of the fluorescent labels of at least two probes are different.

33. The method of claim 30, wherein the limited set of probes comprises fifty or fewer probes.

34. The method of claim 30, wherein the limited set of probes comprises twenty or fewer probes.

35. The method of claim 30, wherein the molecular library is selected from the group consisting of a genomic library and a cDNA library.

36. The method of claim 30, wherein the molecular library comprises at least about 1,000,000 clones.

37. The method of claim 30, wherein the molecular library comprises about 50,000 to about one million clones.

38. The method of claim 30, wherein the molecular library comprises about 1,000 to about 50,000 clones.
39. The method of claim 30, wherein a plurality of labeled members of the molecular library are identified by their hybridization to duplicate sets of the first array of probes.

40. The method of claim 30, wherein the differential hybridization of the labeled member comprises a difference in labeling intensity to at least one probe.

41. The method of claim 30, wherein the method further comprises hybridizing the member component of the labeled member to the first array of probes under at least one additional hybridization condition.

42. The method of claim 30, wherein bar codes for essentially all of the molecular library components are recorded.

43. The method of claim 30, wherein a bar code for only a component of interest of the molecular library is recorded.

44. The method of claim 30, wherein the molecular library components recorded represent either known or unknown components.

45. The method of claim 30, wherein the bar code for at least one library member is digitized.

46. The method of claim 30, wherein the member components of the molecular library are affixed to a surface.

47. The method of claim 30, wherein the limited set of probes is affixed to a surface.

48. An integrated system for comparing hybridization patterns, comprising: an array reader adapted to read the hybridization patterns on an array, wherein the hybridization patterns are of the hybridization of a test nucleic acid to at least three probes which differ in percentage of complementarity to the test nucleic acid and binding to the test nucleic acid in relative proportion to their respective degrees of complementarity to the test nucleic acid, operably linked to a digital computer comprising a data file having a set of at least about 500 low-stringency array hybridization patterns in a digital format.

49. The integrated system of claim 48, further comprising a robotic armature for fluid delivery to the array.

50. The integrated system of claim 48, capable of reading the hybridization pattern of 500 or more labels on an array per hour.
51. The integrated system of claim 48, further operably linked to an optical detector for reading the hybridization pattern of labels on an array.
Oligo Probes

Figure 1