(19) **Bundesrepublik Deutschland**
Deutsches Patent- und Markenamt

(10) **DE 600 13 303 T2** 2005.09.22

(12) # Übersetzung der europäischen Patentschrift

(97) **EP 1 200 620 B1**
(21) Deutsches Aktenzeichen: **600 13 303.6**
(86) PCT-Aktenzeichen: **PCT/IB00/00810**
(96) Europäisches Aktenzeichen: **00 935 427.5**
(87) PCT-Veröffentlichungs-Nr.: **WO 00/78991**
(86) PCT-Anmeldetag: **19.06.2000**
(87) Veröffentlichungstag der PCT-Anmeldung: **28.12.2000**
(97) Erstveröffentlichung durch das EPA: **02.05.2002**
(97) Veröffentlichungstag der Patenterteilung beim EPA: **25.08.2004**
(47) Veröffentlichungstag im Patentblatt: **22.09.2005**

(51) Int Cl.[7]: **C12Q 1/68**
G01N 33/00

(30) Unionspriorität:

139639 P	**17.06.1999**	**US**
155173 P	**21.09.1999**	**US**

(84) Benannte Vertragsstaaten:
AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LI, LU, MC, NL, PT, SE

(73) Patentinhaber:
Amersham Biosciences Niagara Inc., St. Catharines, Ontario, CA

(72) Erfinder:
RAMM, Peter, St. Catharines, CA; NADON, Robert, St. Catharines, CA; SHI, Peide, St. Catharines, CA

(74) Vertreter:
Grünecker, Kinkeldey, Stockmair & Schwanhäusser, 80538 München

(54) Bezeichnung: **VERFAHREN ZUM ENTFERNEN SYSTEMATISCHER FEHLER UND ABWEICHUNGEN UND ZUM ABSCHÄTZEN ZUFÄLLIGER FEHLER IN CHEMISCHEN UND BIOLOGISCHEN TESTVERFAHREN**

Beschreibung

1. Gebiet der Erfindung

[0001] Die vorliegende Erfindung betrifft ein Verfahren zum Aufstellen von Bewertungen, welche Analysen von Daten erhalten aus Hybridisierungs- Arrays objektivieren. Die vorliegende Erfindung stellt in einem Aspekt einen Prozess zum Entfernen systematischer Fehler dar, welche in genomischen Wiederholungsproben (replizierten genomischen Proben) vorliegen. Ein zweiter Aspekt ist ein Verfahren zum Detektieren und Entfernen von Extremwertdaten (Ausreißer). Ein dritter Aspekt ist ein Optimierungsverfahren zum Detektieren oder Entfernen von Extremwertdaten (Ausreißer). Ein vierter Aspekt ist ein Prozess zum Abschätzen des Ausmaßes der zufälligen Fehler, die in genomischen Wiederholungsproben bestehend aus einer kleinen Anzahl von Datenpunkten vorliegen.

2. Hintergrund der Erfindung

[0002] Array- gestützte genetische Analysen beginnen mit einer großen Bibliothek von cDNAs oder Oligonucleotiden (Sonden), immobilisiert auf einem Substrat. Die Sonden werden mit einer einfach gelabelten Sequenz hybridisiert oder einer gelabelten komplexen Mischung abgeleitet von einer Messenger- RNA eines Gewebes oder einer Zellinie (Target). Wie hier verwendet, wird sich die Bezeichnung "Sonde" daher so verstehen, dass sie sich auf ein Material bezieht, das an das Array gebunden ist und die Bezeichnung "Target" wird sich auf ein Material beziehen, das auf die Sonden auf dem Array angewandt wird, so dass Hybridisierung auftreten kann.

[0003] Die Bezeichnung "Element" wird sich auf einen Punkt auf einem Array beziehen. Array-Elemente reflektieren die Sonden/ Target- Interaktion. Die Bezeichnung "Hintergrund" wird sich auf eine Fläche auf dem Substrat außerhalb der Elemente beziehen. Die Bezeichnung "Replikat" wird sich auf zwei oder mehrere gemessene Werte derselben Sonden/ Target Interaktion beziehen. Replikate können unabhängig voneinander sein (die gemessenen Werte sind unabhängig) oder abhängig (die gemessenen Werte sind verwandt, statistisch korreliert oder reaktionsgepaart). Replikate (Wiederholungen) können innerhalb von Arrays sein, über Arrays hinweg, innerhalb Experimenten, über Experimente hinweg oder irgendeine Kombination davon.

[0004] Gemessene Werte der Sonden/ Target – Interaktionen sind eine Funktion ihrer tatsächlichen Werte und von Messfehlern. Die Bezeichnung "Ausreißer" wird sich auf einen Extremwert in einer Verteilung von Werten beziehen. Ausreißerarten resultieren häufig aus nicht korrigierbaren Messfehlern und werden typischerweise aus weiteren statistischen Analysen gestrichen.

[0005] Es gibt zwei Arten von Fehlern, zufällige und systematische, welche das Ausmaß, in welchen beobachtete (gemessene) Werte von ihren wirklichen Werten abweichen, beeinflussen.

[0006] Zufällige Fehler erzeugen Fluktationen in den beobachteten Werten desselben Prozesses oder Attributs. Das Ausmaß und die Verteilungsform von zufälligen Fehlern kann detektiert werden durch wiederholte Messungen des gleichen Prozesses oder Attributs. Kleine zufällige Fehler korrespondieren mit hoher Präzision.

[0007] Systematische Fehler erzeugen Verschiebungen (Offsets) in gemessenen Werten. Gemessene Werte mit systematischen Fehlern nennt man "tendenziös". Systematische Fehler können nicht durch wiederholte Messungen des gleichen Prozesses oder Attributs detektiert werden, da die Tendenz in gleicher Weise beeinflusst wird. Geringe systematische Fehler korrespondieren mit hoher Treffergenauigkeit. Die Bezeichnungen "systematischer Fehler", "Tendenz" und "Offset" werden austauschbar untereinander im vorliegenden Dokument verwendet.

[0008] Eine Erfindung zum Abschätzen zufälliger Fehler, die in genomischen Wiederholungsproben vorliegen, bestehend aus einer kleinen Anzahl von Daten, wurde von Ramm und Nadon in "Process for Evaluating Chemical and Biological Assays" WO 99/ 54724 beschrieben.

[0009] In einer bevorzugten Ausführungsform nahm der Prozess, der darin beschrieben ist, an, dass vor dem Durchführen statistischer Test systematische Fehler in der Messung entfernt worden waren und Ausreißer beseitigt worden waren.

[0010] In Übereinstimmung mit einem Aspekt stellt die vorliegende Erfindung einen Prozess dar, welcher ei-

nen systematischen Fehler aus gemessenen Werten abschätzt und entfernt. In einem weiteren Aspekt stellt sie einen Prozess zum Optimieren der Detektion und Deletion von Ausreißern dar. Ein zweiter Aspekt ist ein Prozess zum Detektieren und Löschen von Ausreißern. Ein dritter Aspekt ist ein Prozess zum Optimieren der Detektion von Ausreißern und ihrer automatischen Beseitigung. Ein vierter Aspekt ist ein Prozess zum Abschätzen des Ausmaßes des zufälligen Fehlers, der in genomischen Wiederholungsproben vorliegt, bestehend aus einer kleinen Anzahl von Daten.

[0011] Es gibt zwei Typen von systematischen Fehlern, die potentiell im Hybridisierungsarrays auftreten.

[0012] Arrayelemente können mit Offsets innerhalb Arrays vorliegen. Typischerweise ist dieser Offset additiv. Er kann von verschiedenen Ursachen herrühren, einschließend Distorsionen im Nylon- Membran- Substrat (Duggan, Bittner, Chen, Meltzer & Trent "Expression profiling using cDNA microarrays", Nature Genetics, 21, 10- 14 (1999).

[0013] Falls vorliegend, wird der Offset korrigiert durch eine Prozedur, die man "Hintergrundkorrektur" nennt, welche das Subtrahieren der Intensität einer Hintergrundfläche außerhalb des Arrayelements von besagtem Arrayelement einschließt.

[0014] Flächen, verwendet für die Kalkulation des Hintergrundes, können nahe an dem Arrayelement (wie beispielsweise ein Kreis, der um das Element herumliegt), oder entfernt (ein Rechtwinkel, um das gesamte Array) liegen. Da der Offset innerhalb eines Arrays dazu tendiert, spezifisch für individuelle Arrayelemente zu sein (selbst bei relativ einheitlichem Hintergrund) werden Flächen in der Nähe des Elementes im allgemeinen zur Hintergrundkorrektur bevorzugt.

[0015] Alternativ können Hintergrundabschätzungen aus "Blindproben"- Elementen (d.h. Elementen ohne Probenmaterial) erhalten werden. In dieser Prozedur wird der "Hintergrund" unterschiedlich von dem typischeren Verfahren beschrieben im vorangegangenen Abschnitt definiert. Theoretisch werden die Blindproben- Element- Intensitäten durch die gleichen Fehlerfaktoren beeinflusst, welche die nicht- Element- Hintergrundflächen (beispielsweise Waschprozeduren) beeinflussen und auch durch Fehlerfaktoren, welche die Elementquantifizierung beeinflussen, welche jedoch ohne Beziehung zum biologischen Signal von Interesse ist (beispielsweise Verteilungsfehler).

[0016] Die vorliegende Erfindung adressiert nicht die Frage der Hintergrundkorrektur. In einer bevorzugten Ausführungsform wurde die Hintergrundkorrektur wenn nötig vor der Abschätzung des systematischen Fehlers und der Ausreißerdetektion vorgenommen. In einer nicht bevorzugten Ausführungsform kann der Prozess noch angewandt werden auf Arrays, die nicht in punkto Hintergrund- Offset korrigiert wurden.

[0017] In einem Aspekt stellt die vorliegende Erfindung einen Prozess dar zum Abschätzen und Entfernen systematischer Fehler über Arrays. Im Gegensatz zum Hintergrundbeitrag tendiert der Beitrag über Arrays hinweg dazu, proportional zu sein.

[0018] Beiträge über Arrays können von verschiedenen Ursachen herrühren. Für Microarray-Untersuchungen, welche Fluoreszenz- Labeling verwenden, schließen Faktoren solcher Beiträge die Targetmenge, das Ausmaß des Target- Labelings, die Fluoreszenzanregungs- und Emissionseffizienten und die Detektoreffizienz ein. Diese Faktoren können alle Elemente gleichermaßen beeinflussen oder können teilweise spezifisch für Elementuntereinheiten des Arrays sein. Beispielsweise kann die Menge des Targetmaterials für verschiedene fleckenbildende Kontaktstellen eines Roboter- Arrays unterschiedliche Beiträge aufweisen (siehe Bowtell "Options available – from start to finish – for obtaining expression data by microarray" Nature Genetics, 21, 25- 32, Seite 31 (1999).

[0019] Für Radio- gelabelte Macroarray- Untersuchungen schließen proportionale Beitragsfaktoren die Target- Menge und die Target- Zugänglichkeit ein (Perret, Ferrán, Marinx, Liauzun, et al. in "Improved differential screening approach to analyse transcriptional variations in organized cDNA libraries" Gene, 208, 103- 115 (1998).

[0020] Die Tageszeit, zu welcher die Arrays betrieben werden (Lander "Array of hope" Nature Genetics, 21, 3- 4 (1999)) und die Variation in chemischen Waschprozeduren über die Experimente (Shalon, Smith & Brown" A DNA microarray system for analyzing complex DNA samples using two- color fluorescent probe hybridization" Genome Research 6, 639-645 (1996)) sind auch als Faktoren des Offsets zitiert worden.

[0021] Duggan et al. (Nature Genetics, 21, 10- 14 (1999)) beschreiben die Expressions-Formgebung unter Verwendung von cDNA Microarrays. DNA Targets in der Form von Expressed Sequenz Tags werden auf Glas angeordnet und mit Fluoreszenz- oder radioaktiv- gelabelter cDNA nachgewiesen. Eine Prozedur der Normalisierung ist, eine gekennzeichnete Untereinheit von Genen mit einer konstanten Expressions- Formgebung zu betrachten. Die Varianz des Normalisierungs- Satzes kann verwendet werden, um Abschätzungen der erwarteten Varianz zu erzeugen, was zur Vorhersage von Vertrauensbereichen führt, die geeignet sind, die Signifikanz der beobachteten Veränderungen im kompletten Datensatz zu bewerten.

[0022] Verfahren aus dem Stand der Technik zum Entfernen des systematischen Fehlers nennt man "Normalisierungs"- Prozeduren. Diese Prozeduren schließen das Dividieren der Werte der Array- Elemente durch einen Referenzwert ein. Diese Referenz kann basiert sein auf allen Sonden oder einer Untereinheit (Teilsatz) ("haushaltende Gene", deren theoretische Expressionsgerade sich nicht mit den Bedingungen ändern). Einmal erhalten kann die Referenz jedoch abgeschätzt werden, durch einen oder verschiedene summative Werte (beispielsweise Mittelwert oder ein spezifiziertes Perzentil).

[0023] Sobald ein systematischer Fehler entfernt worden ist, sind alle zurückbleibenden Mess-Fehler theoretisch zufällige. Zufällige Fehler reflektieren die erwartete statistische Variation eines gemessenen Wertes. Ein gemessener Wert kann beispielsweise aus einem einzelnen Wert bestehen, einer Summe von Werten (Mittelwert, Median), einer Differenz zwischen einzelnen oder Summenwerten oder einer Differenz zwischen Differenzen. Damit zwei Werte als signifikant unterschiedlich voneinander gelten, muss ihre Differenz einen Schwellenwert überschreiten, der gemeinsam definiert wird durch den Messfehler assoziiert mit der Differenz und einer spezifizierten Wahrscheinlichkeit für fehlerhafte Schlüsse, dass die beiden Werte unterschiedlich sind (Typ 1 Fehlerrate). Statistische Tests werden durchgeführt, um zu bestimmen, ob Werte sich signifikant voneinander unterscheiden.

[0024] Alle die Normalisierungsprozeduren des Standes der Technik schätzen systematische Fehler als außerhalb des Kontexts eines statistisch liegenden Modells ab. Da diese informellen Prozeduren implizit (und oft unkorrekter Weise) Annahmen über die Struktur der Daten machen (beispielsweise über Form und Ausmaß sowohl von systematischen als auch zufälligen Fehlern) scheitern sie oft bei der adäquaten Elimination von systematischen Messabweichungen und können zusätzliche Messabweichungen aufgrund der Normalisierungsprozedur selbst einbringen. In einem anderen wissenschaftlichen Zusammenhang beschrieben Freedman und Navidi, in "Regression models for adjusting the 1980 census", Statistical Science, 1, 3- 11 (1986) die Probleme die inherent in Ermangelung des korrekten Modellierens von Daten, welche Messfehler ("Unsicherheit" in ihrer Terminologie) enthalten, sind.

[0025] Modelle werden oft verwendet, um Probleme in Situationen zu entscheiden, die durch Unsicherheit gekennzeichnet sind. Jedoch hängen statistische Schlussfolgerungen von Daten von Annahmen über die Prozesse ab, welche diese Daten generierten. Falls die Annahmen nicht standhalten, können die Schlussfolgerungen auch nicht verlässlich sein. Diese Begrenzung wird oft durch die Anwender ignoriert, die daran scheitern, die entscheidenden Annahmen zu identifizieren oder sie jeglicher Art von empirischen Tests zu unterziehen. Unter solchen Umständen kann die Verwendung statistische Prozeduren nur die Unsicherheit vergrößern (S. 3).

[0026] Zusätzlich zur korrekten Entfernung des systematischen Fehlers verlangen viele statistische Tests die Annahme, dass Reste (=Residualwerte) normal verteilt sind. Reste reflektieren die Differenz zwischen den abgeschätzten wahren Größen der Treffer und ihren beobachteten (gemessenen) Größen. Falls die Größe eines Rests extrem ist (relativ zu anderen Größen in der Verteilung) nennt man diesen einen Ausreißer. Der Ausreißer wird typischerweise von der weiteren statistischen Analyse entfernt, da er im allgemeinen darauf hindeutet, dass der gemessene Wert einen exzessiven Messfehler enthält, der nicht korrigiert werden kann. Um normal verteilte Reste zu erhalten, ist oft eine Datentransformation von Nöten (z.B. eine logarithmische Transformation).

[0027] In einem Aspekt stellt die vorliegende Erfindung einen Prozess zum Detektieren und Entfernen von Ausreißern durch Untersuchen der Verteilung von Resten dar. In einem anderen Aspekt stellt sie einen Prozess zum Detektieren und Entfernen von Ausreißern in automatischer Art und Weise durch einen iterativen Prozess dar, welcher die Charakteristik hat der Verteilung der Reste untersucht (z.B. Schiefe, Kurtosis).

[0028] Wie bei der Korrektur von Offsets über Arrays (Normalisierung) verlässt sich der Stand der Technik bei der Detektion von Ausreißern auf informelle und zufällige Prozeduren außerhalb eines Kontexts von statistischen Modellen. Beispielsweise verglichen Perret, Ferrán, Marinx Liauzun, et al., Improved differential scree-

ning approach to analyse transcriptional variations in organized cDNA libraries" Gene, 208, 103- 115 (1998), die Intensitäten von Sätzen von zwei Wiederholungsarrayelementen nach Normalisierung. Jeder Wiederholungssatz, der eine größere als zweifache Differenz zeigte (oder äquivalent weniger als eine halbfache Differenz) wurde als Ausreißer betrachtet.

[0029] In Übereinstimmung mit einem Aspekt der vorliegenden Erfindung ist die vorliegende Erfindung ein Prozess zum Abschätzen des Ausmaßes des zufälligen Fehlers, der in genomischen Wiederholungs- Proben, bestehend aus einer kleinen Anzahl von Daten vorliegt und zum Durchführen eines statistischen Tests, der die Expressionsgrade über die Bedingungen (beispielsweise erkranktes gegen normales Gewebe) vergleicht. Sie ist eine Alternative zum Verfahren beschrieben von Ramm und Nadon in "Process for Evaluating Chemical and Biological Assays", International Application No. PCT/ IB99/ 00734. Als solche kann sie verwendet werden zusätzlich (oder anstelle von) den Prozeduren beschrieben von Ramm und Nadon (ibid).

[0030] Nachteile aller Prozeduren des Standes der Technik schließen ein:

1. Der Wert der als Normalisierungsreferenz gewählt wird (z.B. 75. Perzentil, etc.) ist zufällig;
2. Geht man davon aus, dass die Wahl der Normalisierungsreferenz zufällig ist, führt das Dividieren des Referenzwertes zur Überkorrektur einiger Elemente und zur Unterkorrektur anderer;
3. Da die Prozeduren des Standes der Technik einen systematischen Fehler nicht innerhalb des Kontexts eines statistischen Models abschätzen, werden Datentransformationen, die notwendig sind, um korrekte Rückschlüsse zu machen, nicht durchgeführt oder können inkorrekt angewandt werden;
4. Da die Prozeduren des Standes der Technik den systematischen Fehler nicht innerhalb des Kontexts eines statistischen Models abschätzen, kann die Normalisierung die wirkliche Struktur der Daten verändern;
5. Da die Prozedur nach dem Stand der Technik die Ausreißer nicht innerhalb des Kontexts eines statistischen Models detektieren, können wirkliche Ausreißer undetektiert bleiben und nicht- Ausreißer können unkorrekterweise als Ausreißer klassifiziert werden;
6. Die Klassifikation von Werten als Ausreißer ist nicht zufällig und subjektiv;
7. Theoretische Annahmen über Datenstrukturen (beispielsweise dass Reste normal verteilt sind) werden nicht empirisch überprüft.
8. Normalisierungsprozeduren können weitere Messfehler kreieren, die nicht in den originalen nicht- normalisierten Messungen vorliegen.

[0031] Die Bezeichnung "Behandlungs- Bedingung" wird sich auf einen Effekt von Interesse beziehen. Solch ein Effekt kann von vornherein existieren (beispielsweise Unterschiede über verschiedene Gewebe oder über die Zeit) oder kann durch eine experimentelle Manipulation induziert werden.

[0032] Hybridisierungsarrays erzeugt unter verschiedenen Behandlungszuständen können statistisch abhängig oder unabhängig sein. Die Mikroarray- Technologie, in welcher zwei verschiedene Target- Behandlungs-Proben mit verschiedenen Fluoreszenzfarbstoffen gelabelt werden und dann auf jedes Element des Arrays cohybridisiert werden, repräsentieren ein Beispiel von statistischer Abhängigkeit. Typischerweise werden die Expressionsverhältnisse der Roh- Signale erzeugt von den beiden Fluoreszenzfarbstoffen hinsichtlich des Nachweises von Unterschieden über die Behandlungsbedingungen untersucht.

[0033] Chen, Dougherty & Bittner "Ratio- based decisions and the quantitative analysis of cDNA microarray immages", Journal of Biomedical Optics, 2, 364- 374 (1997) haben einen analystischen mathematischen Ansatz vorgestellt, welcher die Verteilung von nichtwiederholten differenziellen Verhältnissen unter der Null- Hypothese abschätzt. Dieser Ansatz ist ähnlich zur vorliegenden Erfindung dahingehend, dass er ein Verfahren ableitet zum Erhalten von Vertrauensbereichen und Wahrscheinlichkeitsabschätzungen für Unterschiede in Probenintensität über verschiedene Bedingungen. Er unterscheidet sich von der vorliegenden Erfindung dahingehend, wie er diese Abschätzungen erhält. Anders als in der vorliegenden Erfindung erhält der Chen et al. Ansatz keine Messfehlerabschätzungen von Wiederholungsproben- Werten. Statt dessen wird der Messfehler assoziiert mit Verhältnissen von Probenintensitäten zwischen Bedingungen über mathematische Ableitung der Null- Hypothese Verteilung von Verhältnissen erhalten. D.H. Chen et al. leiten ab, wie die Verteilung der Verhältnisse sein würde, falls keine der Proben Unterschiede in gemessenen Werten über Bedingungen zeigen würde, die größer wären als sie durch "Chance" zu erwarten würden. Basierend auf dieser Ableitung etablieren sie Schranken für statistisch verlässliche Verhältnisse von Probenintensitäten über zwei Bedingungen. Das Verfahren wie abgeleitet, wird nur anwendbar für Unterschiede über zwei Bedingungen. Darüber hinaus schätzt es ab, dass der Messfehler assoziiert mit Probenintensitäten normal verteilt ist. Das Verfahren wie abgeleitet, kann nicht andere Messfehlermodelle anpassen (beispielsweise "lognormal"). Es nimmt auch die gemessene Werte als erwartungstreu und als verlässliche Abschätzungen der "wirklichen" Probenintensität an. D.h. es wird abgeschätzt, dass keine der Proben- Intensitäten "Ausreißer"- Werte darstellen, die von der Ana-

lyse ausgeschlossen werden würden. Tatsächlich ist eine Ausreißer-Detektion mit dem Ansatz beschrieben von Chen et al. nicht möglich.

[0034] Die vorliegende Erfindung wendet die Prozesse beschrieben von Ramm und Nadon in "Process for Evaluating Chemical and Biological Assays", International Application No. PCT/ IB99/ 00734 und von Ramm, Nadon und Shi in "Process for Removing Systematic Error and Outlier Data and for Estimating Error in Chemical and Biological Assays", Provisional Application No. 60/ 139,639 (1999) auf zwei oder mehr statistisch abhängig genomische Proben an.

[0035] Die vorliegende Erfindung unterscheidet sich vom Stand der Technik dahingehend, dass:

1. sie verschiedene Mess- Fehler- Modelle (z.B. Lognormal) anpassen kann;
2. sie Ausreißer innerhalb des Kontexts eines statistischen Models detektieren kann;
3. sie verwendet werden kann, um theoretische Annahmen über Datenstrukturen zu untersuchen (z.B. dass Reste normal verteilt sind).

Detaillierte Beschreibung der bevorzugten Ausführungsform

[0036] Angenommen sei beispielsweise, dass Expressionsgrade für einen speziellen Datensatz proportionale systematische und proportionale zufällige Fehler über Wiederholungsarrays aufweisen. Dieses Szenario wird symbolisch repräsentiert in Gleichung 1

$$X_{gij} = \mu_{gi} v_{gi} \varepsilon_{gij} \qquad (1)$$

für g = 1,..., G, j = 1,..., m und i = 1,..., n, wobei μ_{gi}; den assoziierten wirklichen Intensitätswert des Arrayelements i repräsentiert (welches unbekannt und fixiert ist), v_{gj} die unbekannten systematischen Verschiebungen oder Beiträge über Wiederholungen repräsentiert und ε_{gij} die beobachteten zufälligen Fehler in einer gegebenen Bedingung g für Spot i und Wiederholung j repräsentiert. Das Interesse liegt im Erhalt einer erwartungstreuen Abschätzung eines "wirklichen" Wertes (μ_{gi}) eines Elements.

[0037] Setzt man die Bedingung g voraus (z.B. normale Zellen oder erkranke Widerparte), das Arrayelement i und die Wiederholung j so wird der assoziierte Intensitätswert als X_{gij} bezeichnet.

[0038] Alternativ würde ein Model mit einem additiven Beitrag und einem additiven zufälligen Fehler symbolisiert werden durch

$$X_{gij} = u_{gi} + V_{gi} + e_{gij} \qquad (2)$$

für g = 1,..., G, J = 1,..., m und i = 1,..., n wobei u_{gi} den assoziierten wirklichen Intensitätswert des Arrayelements i (welches unbekannt und fixiert ist) repräsentiert, V_{gi} die unbekannten systematischen Verschiebungen oder Beiträge über Wiederholungen repräsentiert und ε_{gij} die beobachteten zufälligen Fehler in einer gegebenen Bedingung g für Element i und Wiederholung j repräsentiert. Das Interesse liegt im Erhalt einer erwartungstreuen Abschätzung eines "wahren" Wertes (u_{gi}) eines Elements.

[0039] Das Model dargestellt in Gleichung 1 wird als bevorzugte Ausführungsform präsentiert. Anwendungen eines Prozesses verwendend das Model gezeigt in Gleichung 2, wären jedoch für den Fachmann auf dem Gebiet offensichtlich. Anwendungen unter Verwendung weiterer Modelle (z.B. proportionaler Beiträge und additiver zufälliger Fehler) wären auch offensichtlich für den Fachmann auf dem Gebiet.

[0040] Um die Parameter $v_{gj}(V_{gj})$ identifizierbar in dem Model zu machen, wird die Bedingung an die Gleichung

$$\sum_{j=1}^{m} \log(v_{gj}) = 0 \quad \left(\sum_{j=1}^{m} V_{gj} = 0\right)$$

verlangt.

[0041] Diese Parameter können als fixiert oder zufällig genommen werden. Wenn die Parameter als zufällig angenommen werden, nehmen wir des weiteren an, dass sie unabhängig von den zufälligen Fehlern sind.

[0042] Unter dem Model gezeigt in Gleichung 1 haben wir beispielsweise die maximale Wahrscheinlichkeitsabschätzung (MLE, maximum likelyhood estimate) von μ_{gi} und V_{gj} wie folgt:

$$\hat{\mu}_{gi} = \exp\{\frac{1}{m}\sum_{j=1}^{m}\log(X_{gij})\} \quad (3)$$

und

$$\hat{v}_{gj} = \exp\{\frac{1}{n}\sum_{i=1}^{n}\log(X_{gij}) - \log(\hat{\mu}_{gi}))\} \quad (4)$$

[0043] Das Kombinieren der Gleichung 3 und 4 führt zu Abschätzungen der Reste [$\log(\hat{\varepsilon}_{gij})$] dargestellt in Gleichung 5.

$$\log(\hat{\varepsilon}_{gij}) = \log(X_{gij}) - \log(\hat{\mu}_{gi}) - \log(\hat{v}_{gj}) \quad (5)$$

[0044] Dafür gegebenes g und i

$$\log(X_{gij}) - \log(v_{gj}) = \log(\mu_{gi}) + \log(\varepsilon_{gij}),$$

j = 1,..., m unabhängig und identisch verteilt als Normalverteilung sind, mit Mittelwert $\log(\mu_{gi})$ und Varianz σ^2gi, stellt Gleichung 6 erwartungstreue Abschätzungen von wirklichen Werten von Arrayelementen bereit. D.h. Gleichung 6 liefert die abgeschätzten Werte, wobei die systematischen Fehler entfernt sind.

$$\log(X_{gij}) - \log(\hat{v}_{gj}) \quad (6)$$

[0045] Man nimmt an, dass, falls das Modell korrekt ist, die Reste normal verteilt sein sollten. Diese Annahme kann empirisch durch Untersuchen der Schiefe und der Kurtosis der Verteilung der Reste wie gemäß Gleichung 5 berechnet überprüft werden (Schiefe und Kurtosis Messwerte sind standardisierte statistische Indices; siehe Stuart & Ord "Distribution theory (6th ed.)(Kendall's advanced theory of statistics Vol. 1)", New York, Halsted Press (1994). Schiefe ist ein Maß der Symmetrie einer Verteilung. Kurtosis ist ein Maß der "Überhöhung" einer Verteilung. Unter der Normalitäts- Annahme sollten sowohl die Schiefe als auch die Kurtosis der Reste- Verteilung etwa Null sein.)

[0046] Selbst wenn das Modell für die meisten der Daten korrekt ist, können Ausreißer verursachen, dass die Verteilung des gesamten Datensatzes von der Normalität abweicht. Ausreißer können detektiert und entfernt werden, über eine der folgenden Optimierungsprozeduren:

1. Ausreißer können definiert werden über eine Schranke (beispielsweise ± 2 Standardfehler entfernt vom Mittelwert der Reste). In einer bevorzugten Ausführungsform würde jeglicher Rest, dessen absoluter Wert die Schranke überschreitet, von weiteren statistischen Test gelöscht.
2. Ein automatisierter iterativer Prozess, welcher die Schiefe und Kurtosis untersucht, kann auch verwendet werden. In diese Prozedur werden die Schiefe und Kurtosis für eine Mittel- Proportion an Treffern (z.B. die mittleren 80%) berechnet. Die Schiefe und Kurtosis werden wiederholt kalkuliert, wenn die Proportion der Treffer in den nachfolgenden Schritten vergrößert wird. Die Proportion der Treffer, welche optimale Schiefe und Kurtosis- Werte erzeugen (am nächsten bei Null), wird als die optimale Verteilung von Resten gewählt. Treffer, die außerhalb der gewählten mittleren Proportion an Werten fallen, werden als Ausreißer abgeschätzt. In einer bevorzugten Ausführungsform werden diese Treffer von der weiteren Analyse gelöscht.

[0047] Statistische Indices (z.B. Vertrauensbereiche) und statistische Tests (z.B. t- Tests, Analyse der Varianz) wie von Ramm und Nadon in "Process for Evaluating Chemical and Biological Assays", International Application No. PCT/ IB99/ 00734 beschrieben, können auf die Array- Elemente- Daten angewandt werden, deren Reste- Treffer nicht Ausreißer darstellen.

[0048] Zusätzlich dazu oder alternativ können die statistischen Tests, beschrieben in Gleichungen 7 und 8 auf diese Daten angewandt werden.

$$z^* = \sqrt{m}\frac{(\bar{X}_{1i} - \bar{X}_{2i})}{\sqrt{\sigma_1^{2*} + \sigma_2^{2*}}} \quad (7)$$

wobei σ2 für jede Bedingung berechnet wird als:

$$\sigma^2 = [\text{median}\{|x_i - \text{median}(x_i)|\}]^2 \bullet c^2 \qquad (8)$$

wobei x_i = alle Reste für alle wiederholten Array- Elemente innerhalb einer Bedingung und c ein Normalisierungsfaktor zum Abschätzen des Standardfehlers für die Reste ist, wenn sie normal verteilt sind. Vorzugsweise gilt c = 1,0532, jedoch können andere Werte von c eingesetzt werden.

[0049] Der z* Wert von Gleichung 7 wird relativ zu einer Standardnormalverteilung (z- Tabelle) untersucht, um den Grad der statistischen Signifikanz zu bewerten. Die Gleichungen 7 und 8 verallgemeinern sich für drei oder mehr Bedingungen in einer Weise, die für den Fachmann auf dem Gebiet offensichtlich ist.

[0050] Die vorliegende Erfindung schließt nicht die Verwendung von Normalisierungsprozeduren aus dem Stand der Technik aus, die auf die Daten vor der Anwendung des vorliegenden Prozesses angewendet werden. Dies kann notwendig sein, beispielsweise wenn Daten unter verschiedenen Bedingungen und an verschiedenen Tagen erhalten werden müssen. Unter diesen Umständen können Daten innerhalb von Zuständen auf eine Referenz (z.B. haushaltende Gene) normalisiert werden müssen, vor der Anwendung des vorliegenden Prozesses.

Appendix

[0051] Man betrachte einen Fall, in welchem die Expressionsdaten von drei Wiederholungsarrays gesammelt wurden, welche 1280 verschiedene Elemente enthielten. Der systematische Fehler über Wiederholungsarrays wird als proportional angenommen und es sei auch angenommen, dass zufällige Fehler über Wiederholungsarrays proportional sind. Dieses Modell ist in Gleichung 1 gezeigt und im hauptsächlichen Teil des Textes.

Normalisierungsverfahren

[0052] Ein Ansatz ist zu versuchen, die proportionalen systematischen Fehler durch Dividieren eines jeden Elementes innerhalb eines Arrays durch einen Referenzwert (z.B. 75. Perzentil- Wert aller Elemente innerhalb des Arrays) zu entfernen. Falls der systematische Fehler durch die Normalisierungsprozedur entfernt wird, wird die Gleichung 1 zu:

$$x_{gij} = \mu_{gi}\varepsilon_{gij}$$

[0053] Reste werden dann gemäß Gleichung 5 mit der Bezeichnung für den systematischen Fehler entfernt:

$$\log(\hat{\varepsilon}_{gij}) = \log(X_{gij}) - \log(\hat{\mu}_{gi})$$

[0054] Fig. 1 repräsentiert die Verteilung der Reste mit optimierter Schiefe und Kurtosis (d.h. Null am nächsten kommend) und den gelöschten Ausreißern. Von 1280 Resten wurden 40 als Ausreißer detektiert und gelöscht. Die Schiefe und Kurtosis- Werte waren – 0,27, z = 3,88; p < 0,001 und 0,0006, z = 0,04 und p = 0,49. Der Schiefe Wert weicht signifikant von Null ab, was darauf hindeutet, dass die Reste nicht normal verteilt sind. Dieses Ergebnis liegt nahe, dass im Gegensatz zur Annahme des Models die Normalisierung nicht adäquat die Komponente des systematischen Fehlers von den gemessenen Expressionswerten entfernt hat.

Verfahren der vorliegenden Erfindung

[0055] In einer bevorzugten Ausführungsform würde die vorliegende Erfindung wie folgt vorgehen:

1. Abschätzen des Messwertmodels, dargestellt in Gleichung 1.
2. Berechnen des Durchschnittes für jede Elementstelle über Wiederholungsarrays (Gleichung 3).
3. Abschätzen des systematischen Fehlers für jedes Array (Gleichung 4).
4. Berechnen der Reste für jede Arrayelementstelle (Gleichung 5).

[0056] Fig. 2 repräsentiert die Verteilung der Reste in optimierter Schiefe und Kurtosis (d.h. Null am nächsten kommend) mit gelöschten Ausreißern. Unter 1280 Resten wurden 65 als Ausreißer detektiert und gelöscht. Schiefe und Kurtosis Werte waren 0,073, z = 1,04; p = 0,15 bzw. 0,039, z = 0,28, p = 0,39. Die Schiefe und Kurtosis Werte waren nicht signifikant verschieden von Null, was darauf hindeutet, dass die Reste annähernd normal verteilt waren. Dieses Ergebnis legt nahe, dass der statistische Modellierungsprozess adäquat die systematische Fehlerkomponente von den gemessenen Expressionswerten entfernt hat.

Schlussfolgerung

[0057] In diesem Beispiel würden die Prozeduren beschrieben von Ramm und Nadon in "Process for Evaluating Chemical and Biological Assays", WO 9 954 724 oder die Prozedur der vorliegenden Erfindung (Gleichung 7 und 8) brauchbare Ergebnisse erzeugen mit dem "Verfahren der vorliegenden Erfindung", jedoch nicht mit dem "Normalisierungsverfahren". Unter anderen Umständen können abhängig vom Messungs- Fehler-Model Normalisierungsprozeduren aus dem Stand der Technik adäquat für diesen Zweck sein (z.B. proportionale systematische Fehler über Arrays mit additiven zufälligen Fehlern). Jedoch ist es wahrscheinlich, dass die Wahl des Referenzwertes für die Normalisierungsprozedur zufällig aus einer statistischen Schlussfolgerungsperspektive erfolgen wird, solange nicht die Prozesse folgen, welche im vorliegenden Dokument beschrieben werden.

Patentansprüche

1. Verfahren zum. Verbessern der Zuverlässigkeit von aus Array-Hybridisierungsstudien erhaltenen physikalischen Messungen, die an einem Array mit einer großen Zahl genomischer Proben durchgeführt wurden, die sich jeweils aus einer kleinen Zahl Replikate zusammensetzen, die nicht ausreicht, um genaue und gültige statistische Schlussfolgerungen zu ziehen, das den Schritt des Abschätzens eines Fehlers in der Messung einer Probe durch Mitteln von beim Messen mindestens einer der großen Zahl von Proben und einer Teilmenge der großen Zahl von Proben erhaltenen Fehlern und die Nutzung des abgeschätzten Probenfehlers als Standard für das Annehmen oder Abweisen der Messung der jeweiligen Probe umfasst.

2. Verfahren nach Anspruch 1, wobei eine physikalische Messgröße basierend auf der Differenz zwischen statistisch abhängigen Größen bestimmt wird.

3. Verfahren nach Anspruch 1, wobei eine aus einer gesamten Array-Population bestimmte physikalische Messgröße benutzt wird, um diskrete Vorkommnisse dieser Größe für die kleine Zahl von Replikatproben innerhalb dieser Population abzuschätzen.

4. Verfahren nach Anspruch 1, wobei die Abschätzungen des Messfehlers benutzt werden, um Array-Hybridisierungsstudien basierend auf
(a) der Wahrscheinlichkeit des Erfassens einer echten Differenz eines vorgegebenen Betrags zwischen physikalischen Messungen einer gegebenen Anzahl Replikate oder
(b) der Anzahl der für die Erfassung einer echten Differenz eines vorgegebenen Betrags erforderlichen Replikate zu planen, zu handhaben und zu steuern.

5. Verfahren zum Verbessern der Zuverlässigkeit und Genauigkeit von aus Array-Hybridisierungsstudien erhaltenen physikalischen Messungen, die an einem Array mit einer großen Zahl genomischer Proben durchgeführt wurden, die sich jeweils aus einer kleinen Zahl Replikate zusammensetzen, die nicht ausreicht, um genaue und gültige statistische Schlussfolgerungen zu ziehen, das den Schritt des Erfassens von Ausreißerwerten in der Messung einer Probe durch Kombinieren von Residualwerten von beim Messen einer der großen Zahl von Proben und einer Teilmenge der großen Zahl von Proben erhaltenen Werten umfasst.

6. Verfahren nach Anspruch 5, wobei Ausreißer basierend auf der Abweichung ihrer Residualwerte vom Mittelwert oder Medianwert oder einer anderen Messung der Residualwerte erfasst werden.

7. Verfahren nach Anspruch 5, wobei Ausreißer manuell, basierend auf Eigenschaften, einschließlich Schiefe und Kurtosis der Verteilung der Residualwerte erfasst werden.

8. Verfahren nach Anspruch 5, wobei Ausreißer basierend auf automatisch und iterativ bezüglich der Eigenschaften, einschließlich Schiefe und Kurtosis der Verteilung der Residualwerte erfasst werden.

9. Verfahren zum Verbessern der Genauigkeit von aus Array-Hybridisierungsstudien erhaltenen physikalischen Messungen, die an einem Array mit einer großen Zahl genomischer Proben durchgeführt wurden, die sich jeweils aus einer kleinen Zahl Replikate zusammensetzen, die. nicht ausreicht, um systematische Fehler über Arrays hinweg abzuschätzen, wobei das Verfahren den Schritt des Mittelns der Differenzen zwischen Einzelproben innerhalb eines Arrays und des Mittelwerts der gewissen Replikate aus anderen Arrays, die dieses eine Array enthalten, umfasst.

10. Verfahren zum Verbessern der Genauigkeit von aus Array-Hybridisierungsstudien erhaltenen physika-

lischen Messungen, die an einem Array mit einer großen Zahl genomischer Proben über zwei oder mehr Bedingungen durchgeführt wurden, die sich jeweils aus einer kleinen Zahl Replikate zusammensetzen, die nicht ausreicht, um systematische Fehler über Arrays hinweg abzuschätzen, wobei von gewissen Replikaten erhaltene Messungen über Bedingungen korreliert sind und wobei das Verfahren den Schritt des Mittelns der Differenzen zwischen Einzelproben innerhalb eines Arrays und dem Mittelwert der gewissen Replikate aus anderen Arrays, die dieses eine Array enthalten, umfasst.

11. Verfahren nach einem der Ansprüche 5–10, wobei eine physikalische Messgröße basierend auf der Differenz zwischen statistisch abhängigen Größen bestimmt wird.

12. Verfahren nach einem der Ansprüche 5–10, wobei eine physikalische Messgröße, die aus einer gesamten Array-Population bestimmt wurde, benutzt wird, um diskrete Vorkommnisse dieser Größe für die kleine Zahl der Replikatproben innerhalb dieser Population abzuschätzen.

13. Verfahren nach einem der Ansprüche 5–10; wobei die Abschätzungen des Messfehlers benutzt werden, um Array-Hybridisierungsstudien basierend auf
(a) der Wahrscheinlichkeit des Erfassens einer echten Differenz eines vorgegebenen Betrags zwischen physikalischen Messungen einer gegebenen Anzahl Replikate oder
(b) der Anzahl der für die Erfassung einer echten Differenz eines vorgegebenen Betrags erforderlichen Replikate
zu planen, zu handhaben und zu steuern.

14. Verfahren nach einem der Ansprüche 1–10, angewandt, um physikalische Messungen auszuwerten, die aus in Substraten oder in Vertiefungen enthaltenden Substraten oder in Reagenzgläsern durchgeführten biologischen und chemischen Untersuchungen erhalten wurden.

15. Verfahren nach Anspruch 11, angewandt, um physikalische Messungen auszuwerten, die aus in Substraten oder in Substraten enthaltenden Vertiefungen oder in Reagenzgläsern durchgeführten biologischen und chemischen Untersuchungen erhalten wurden.

16. Verfahren nach Anspruch 12, angewandt, um physikalische Messungen auszuwerten, die aus in Substraten oder in Substraten enthaltenden Vertiefungen oder in Reagenzgläsern durchgeführten biologischen und chemischen Untersuchungen erhalten wurden.

17. Verfahren nach Anspruch 13, angewandt, um physikalische Messungen auszuwerten, die aus in Substraten oder in Substraten enthaltenden Vertiefungen oder in Reagenzgläsern durchgeführten biologischen und chemischen Untersuchungen erhalten wurden.

Es folgen 2 Blatt Zeichnungen

Anhängende Zeichnungen

Fig. 1 Normalisierungsverfahren

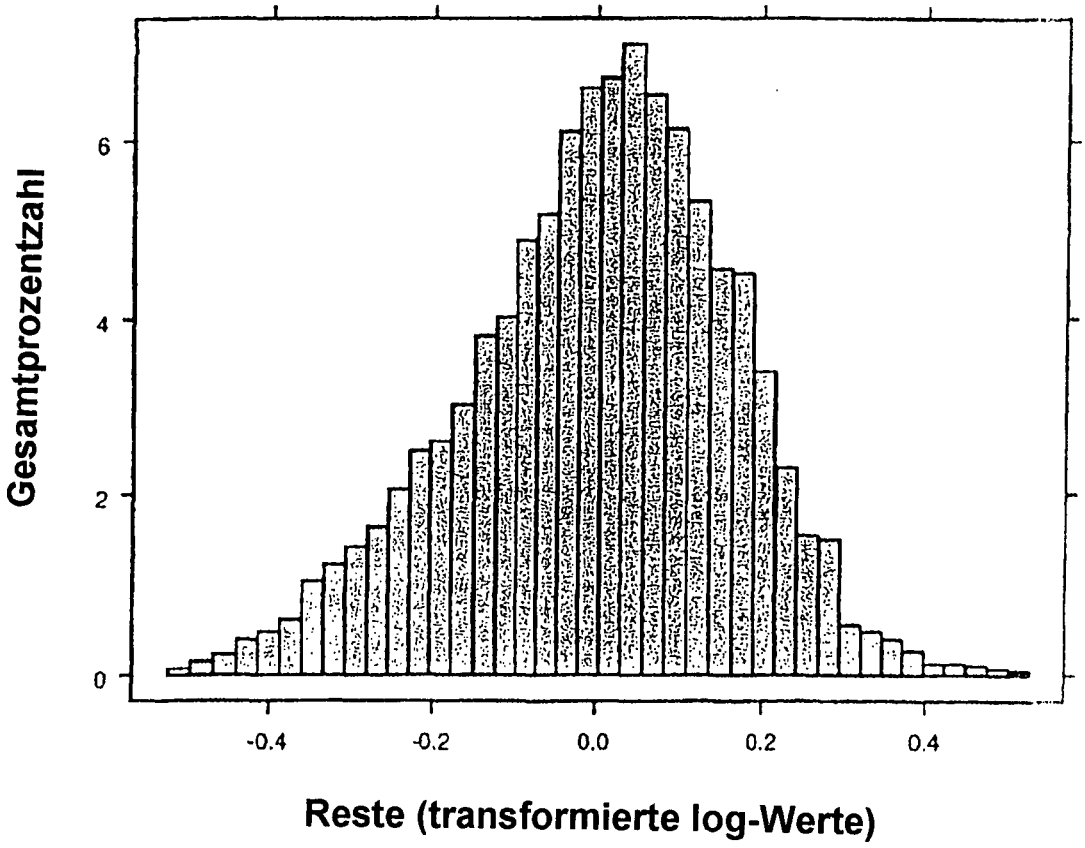

Fig. 2 statistisches Modellierungsverfahren