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

(10) **DE 698 27 154 T2** 2006.03.09

(12) # Übersetzung der europäischen Patentschrift

(97) **EP 1 019 536 B1**
(21) Deutsches Aktenzeichen: **698 27 154.8**
(86) PCT-Aktenzeichen: **PCT/US98/16971**
(96) Europäisches Aktenzeichen: **98 944 444.3**
(87) PCT-Veröffentlichungs-Nr.: **WO 99/009218**
(86) PCT-Anmeldetag: **14.08.1998**
(87) Veröffentlichungstag der PCT-Anmeldung: **25.02.1999**
(97) Erstveröffentlichung durch das EPA: **19.07.2000**
(97) Veröffentlichungstag der Patenterteilung beim EPA: **20.10.2004**
(47) Veröffentlichungstag im Patentblatt: **09.03.2006**

(51) Int Cl.[8]: ***C12Q 1/68*** (2006.01)
C12P 19/34 (2006.01)
G06K 9/00 (2006.01)
G05B 15/00 (2006.01)
G06F 19/00 (2000.01)

(30) Unionspriorität:
55939 P **15.08.1997** **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:
Affymetrix, Inc. (n.d.Ges.d.Staates Delaware), Santa Clara, Calif., US

(72) Erfinder:
BERNO, Anthony, San Jose, US

(74) Vertreter:
Haseltine Lake Partners GbR, 81669 München

(54) Bezeichnung: **POLYMORPHISMUSERKENNUNG MIT HILFE CLUSTER-ANALYSE**

Beschreibung

HINTERGRUND DER ERFINDUNG

[0001] Die Erfindung betrifft das Erkennen von Unterschieden in Polymeren. Sie betrifft insbesondere das Erkennen von Polymorphismen in Testnukleinsäuresequenzen durch Clustern der Hybridisierungsaffinitätsdatensätze.

[0002] Es gibt Vorrichtungen und Computersysteme zur Bildung und Verwendung von Material-Gruppierungen auf einem Chip oder Substrat. Die PCT-Anmeldungen WO92/10588 und 95/11995 beschreiben beispielsweise Techniken zur Sequenzierung oder Sequenzüberprüfung von Nukleinsäuren und anderen Materialien. Gruppierungen zur Durchführung dieser Arbeitsschritte können beispielsweise gemäß den Verfahren der Pioniertechniken erzeugt werden, die in den US-Patenten 5 445 934, 5 384 261 und US 5 571 639 offenbart sind.

[0003] Gemäß einem Aspekt der hier beschriebenen Techniken werden Nukleinsäuresonden an bekannten Stelen auf einem Chip gruppiert. Eine markierte Nukleinsäure wird dann mit dem Chip zusammengebracht, und ein Scanner erzeugt eine Bilddatei, die diejenigen Stellen anzeigt, an denen die markierten Nukleinsäuren an dem Chip gebunden sind. Mit Hilfe der Bilddatei und der Identitäten der Sonden an bestimmten Stellen lässt sich Information, wie die Nukleotid- oder Monomersequenz von DNA oder RNA, erhalten. Mit diesen Systemen werden beispielsweise DNA-Gruppierungen erzeugt, die sich zur Untersuchung und zur Erkennung von Mutationen verwenden lassen, die für genetische Erkrankungen, Krebserkrankungen, Infektionserkrankungen, HIV und andere genetische Eigenschaften wichtig sind.

[0004] Die VLSIPS™ Technologie bietet Verfahren zur Herstellung sehr großer Gruppierungen von Oligonukleotidsonden auf sehr kleinen Chips. Siehe US-Patent 5 143 854 und PCT-Patentveröffentlichungen Nr. WO 90/15070 und 92/10092. Die Oligonukleotidsonden auf der DNA-Sondengruppierung werden zum Erkennen komplementärer Nukleinsäuresequenzen in einer interessierenden Testnukleinsäureprobe (der "Ziel"-Nukleinsäure) verwendet.

[0005] Für Sequenz-Überprüfungsanwendungen kann der Chip für eine spezifische Zielnukleinsäuresequenz kachelförmig unterteilt sein. Der Chip kann beispielsweise Sonden enthalten, die zur Zielsequenz vollständig komplementär sind, und Sonden, die sich von der Zielsequenz um eine einzige Basen-Fehlpaarung unterscheiden. Für De-novo-Sequenzier-Anwendungen kann der Chip sämtliche möglichen Sonden einer spezifischen Länge enthalten. Die Sonden sind auf einem Chip in Reihen und Spalten von Zellen angeordnet, wobei jede Zelle mehrere Kopien einer bestimmten Sonde enthält. Zusätzlich können "leere" Zellen auf dem Chip vorhanden sein, die keinerlei Sonden aufweisen. Da die Leerzellen keine Sonden aufweisen, sollten markierte Ziele nicht spezifisch an den Chip in diesem Bereich binden. Somit liefert die Leerzelle ein Maß für die Hintergrundintensität.

[0006] Die Interpretation für die Hybridisierungsdaten von hybridisierten Chips kann mehrere Schwierigkeiten aufwerfen. Statistische Fehler, wie physikalische Fehler auf dem Chip, können dazu führen, dass einzelne Sonden oder räumlich verwandte Gruppen von Sonden anormal hybridisieren (beispielsweise eine anormale Fluoreszenz aufweisen). Systematische Fehler, wie die Bildung von Sekundärstrukturen in den Sonden oder dem Ziel, können auch reproduzierbare, aber dennoch irreführende Hybridisierungsdaten ergeben.

[0007] Für viele Anwendungen möchte man bestimmen, ob es Unterschiede zwischen und unter Testnukleinsäuresequenzen, wie Polymorphismen an einer Basenposition, gibt. Man möchte zum Erkennen dieser Unterschiede gerne solche Systeme und Verfahren besitzen, die nicht übermäßig von statistischen und systematischen Fehlern beeinträchtigt sind.

[0008] WO 97/29212 offenbart Oligonukleotidgruppierungen und Verfahren zur artenbildenden Phänotypbestimmung von Organismen. Die Gruppen oder Arten, denen ein Organismus angehört, können bestimmt werden, indem man die Hybridisierungsmuster einer Zielnukleinsäure aus den Organismen mit Hybridisierungsmustern in einer Datenbank vergleicht.

ZUSAMMENFASSUNG DER ERFINDUNG

[0009] Die Erfindung stellt innovative Systeme und Verfahren zur Erkennung von Unterschieden in Testpolymeren, wie Nukleinsäuresequenzen, bereit. Die Hybridisierungsaffinitätsdaten für die Testpolymere werden derart geclustert, dass möglicherweise vorhandene Unterschiede zwischen oder unter Testpolymeren leicht identifiziert werden können. Durch Clustern der Hybridisierungsaffinitätsdaten der Probenpolymere können Unterschiede zwischen den Testpolymeren sogar bei Vorhandensein von statistischen und systematischen Fehlern genau erfasst werden. Außerdem können Polymorphismen in den Testnukleinsäuren erkannt werden, unabhängig davon, was das Basecalling ergeben hat.

[0010] Bei einer Ausführungsform stellt die Erfindung ein Verfahren zum Erkennen von Polymorphismen in Testnukleinsäuresequenzen bereit. Mehrfache Hybridisierungsaffinitätsdatensätze werden eingegeben, wobei jeder Hybridisierungsdatensatz Hyb-

ridisierungsaffinitäten zwischen einer Testnukleinsäuresequenz und Nukleinsäuresonden beinhaltet. Die mehrfachen Hybridisierungsdatensätze sind hierarchisch in eine Anzahl von Clustern gegliedert, so dass die Hybridisierungsaffinitätsdatensätze in den Clustern alle untereinander ähnlicher sind als zu den Hybridisierungsaffinitätsdatensätzen anderer Cluster. Die Mehrfach-Cluster können dann zur Bestimmung einer Anzahl enger Cluster analysiert werden, wobei ein enger Cluster ein Cluster ist, bei dem die durchschnittliche Distanz zwischen dem Clustermittel und den Mitteln seiner Subcluster geringer ist als die Distanz zum nächsten Schwestercluster, und zwar über einen Ähnlichkeitsfaktor, der derart ist, dass die Anzahl enger Cluster anzeigt, ob ein Polymorphismus in den Test-Nukleinsäuresequenzen vorliegt. Die Polymorphismen können Mutationen, Insertionen und Deletionen beinhalten.

[0011] Andere Aufgaben und Vorteile der Erfindung werden dem Fachmann beim Durchlesen der folgenden eingehenden Beschreibung zusammen mit den beigefügten Zeichnungen ersichtlich.

KURZE BESCHREIBUNG DER ZEICHUNGEN

[0012] Es zeigt:

[0013] Fig. 1 ein Beispiel für ein Computersystem, das zur Ausführung der Software einer Ausführungsform der Erfindung verwendet werden kann;

[0014] Fig. 2. ein System-Blockdiagramm des Computersystems von Fig. 1;

[0015] Fig. 3 ein Gesamtsystem zur Erzeugung und Analyse von Gruppierungen biologischer Materialien, wie DNA oder RNA.

[0016] Fig. 4 das Konzept der Bindung von Sonden auf Chips;

[0017] Fig. 5 ein High-Level-Fließschema eines Verfahrens zur Analyse von Testpolymeren;

[0018] Fig. 6 ein Fließschema eines Verfahrens zum Clustern von Hybridisierungsaffinitätsdaten;

[0019] Fig. 7 ein Fließschema eines Verfahren zur Analyse von Testnukleinsäuresequenzen;

[0020] Fig. 8 graphisch wie die Normalisierung die Hybridisierungsaffinitäten beeinflusst;

[0021] Fig. 9 eine Bildschirmanzeige mit einem Dendrogramm, welches zeigt, dass kein Polymorphismus an der interessierenden Basenposition zu sein scheint;

[0022] Fig. 10 das Dendrogramm von Fig. 9;

[0023] Fig. 11 ein Dendrogramm, welches zeigt, dass wahrscheinlich ein Polymorphismus an der interessierenden Basenposition vorliegt;

[0024] Fig. 12 ein Dendrogramm, welches zeigt, dass wahrcheinlich mehr als ein Polymorphismus an der interessierenden Basenposiotion vorliegt.

EINGEHENDE BESCHREIBUNG DER BEVORZUGTEN AUSFÜHRUNGSFORMEN

[0025] In der folgenden Beschreibung wird die Erfindung anhand bevorzugter Ausführungsformen, die die VLSIPS™-Technologie zur Erzeugung sehr großer Gruppierungen von Oligonukleotidsonden auf Chips verwenden, erläutert. Die Erfindung ist jedoch nicht auf Nukleinsäuren oder auf diese Technologie eingeschränkt und kann vorteilhafterweise bei anderen Polymeren und Herstellungsverfahren angewendet werden. Daher dient die folgende Beschreibung der Ausführungsformen lediglich Veranschaulichungszwecken und nicht der Einschränkung.

[0026] Fig. 1 veranschaulicht ein Beispiel für ein Computersystem, das zur Ausführung der Software einer erfindungsgemäßen Ausführungsform verwendet wird. Fig. 1 zeigt ein Computersystem **1**, das eine Anzeige **3**, einen Bildschirm **5**, ein Gehäuse **7**, eine Tastatur **9** und eine Maus **11** umfasst. Die Maus **11** kann einen oder mehrere Knöpfe aufweisen, die mit der Grafik-Benutzerschnittstelle interagieren. Das Gehäuse **7** birgt ein CD-ROM-Laufwerk **13**, System-Speicher und ein Festplattenlaufwerk (siehe Fig. 2), die zum Speichern und Aufrufen von Softwareprogrammen verwendet werden können, die den Computercode zur Durchführung der Erfindung, Daten zur erfindungsgemäßen Verwendung und dergleichen beinhalten. Es ist zwar ein CD-ROM **15** als beispielhaftes computerlesbares Speichermedium gezeigt, jedoch können andere Speichermedien verwendet werden, wie Floppy-Disk, Band, Flash-Memory, System-Memory und Hard-Drive. Zusätzlich kann ein Datensignal, das in einer Trägerwelle (beispielsweise in einem Netzwerk, wie dem Internet) aufgenommen ist, ein computerlesbares Speichermedium sein.

[0027] Fig. 2 zeigt ein System-Blockdiagramm von Computersystem **1**, das zur Ausführung der Software einer erfindungsgemäßen Ausführungsform verwendet wird. Wie in Fig. 1 beinhaltet das Computersystem **1** den Monitor **3** und die Tastatur **9** sowie Maus **11**. Das Computersystem **1** beinhaltet weiterhin Untersysteme, wie den Zentralprozessor **51**, System-Memory **53**, Festspeicher **55** (beispielsweise Festplattenlaufwerk), Wechselspeichermedium **57** (beispielsweise CD-ROM-Laufwerk), Display-Adapter **59**, Sound-Karte **61**, Lautsprecher **63** und Netzwerk-Schnittstelle **65**. Andere Computersysteme, die sich zur erfindungsgemäßen Verwendung eignen,

können zusätzliche oder wenigere Untersysteme umfassen. Ein anderes Computersystem kann mehr als einen Prozessor **51** (d.h. ein Multiprozessorsystem) oder ein Cache-Speicher umfassen.

[0028] Die Systembus-Architektur von Computersystem **1** wird durch die Pfeile **67** veranschaulicht. Diese Pfeile veranschaulichen aber jedes Zwischenverbindungsschema, das dazu dient, die Untersysteme miteinander zu koppeln. Ein Local Bus kann beispielsweise zur Verbindung von Zentralprozessor mit dem Systemspeicher und Anzeigeadapter verwendet werden. Das in Fig. 2 gezeigte Computersystem **1** ist aber lediglich ein Beispiel für ein Computersystem, das sich zur erfindungsgemäßen Verwendung eignet. Andere Computer-Architekturen mit verschiedenen Konfigurationen der Untersysteme können ebenfalls eingesetzt werden.

[0029] Die Erfindung ist für Veranschaulichungszwecke als Teil eines Computersystems beschrieben, das eine Chip-Maske entwirft, die Sonden auf dem Chip synthetisiert, die Nukleinsäuren markiert und die hybridisierten Nukleinsäuresonden durchmustert. Ein solches System ist vollständig in US-Patent 5 571 639 beschrieben. Die Erfindung kann jedoch gesondert von dem Gesamtsystem zur Analyse der durch solche Systeme erzeugten Daten verwendet werden.

[0030] Fig. 3 veranschaulicht ein computergesteuertes System zum Erzeugen und Analysieren von Gruppierungen biologischer Materialien, wie RNA oder DNA. Ein Computer **100** wird zum Entwerfen von Gruppierungen biologischer Polymere, wie RNA und DNA, verwendet. Der Computer **100** kann beispielsweise eine geeignet programmierte Sun-Workstation oder ein Personal Computer oder eine Workstation sein, wie ein IBM PC-Äquivalent, mit einem geeigneten Speicher und einer CPU, wie in den Fig. 1 und Fig. 2 gezeigt. Das Computersystem **100** erhält Eingaben von einem Anwender in Bezug auf die Eigenschaften eines interessierenden Gens, sowie andere Dateneingaben, die die gewünschten Eigenschaften der Gruppierung betreffen. Das Computersystem kann gegebenenfalls Information erhalten, die eine bestimmte interessierende Gensequenz von einer externen oder internen Datenbank **102**, wie GenBank, betreffen. Die Ausgabe des Computersystems **100** ist ein Satz von Chip-Design-Computerdateien **104**, beispielsweise in der Form einer Switch-Matrix, wie sie in der PCT-Anmeldung WO 92/10092 beschrieben ist, und anderen angegliederten Computerdateien.

[0031] Die Chip-Design-Dateien werden einem System **106** zugeführt, das die lithographischen Masken entwirft, die bei der Herstellung von Gruppierungen von Molekülen, wie DNA, verwendet werden. Das System oder das Verfahren **106** kann eine Hardware enthalten, die zur Herstellung der Masken **110** nötig ist, und auch die nötige Computer-Hardware und Software **108**, die zum Auflegen der Maskenmuster auf der Maske auf effiziente Weise nötig ist. Entsprechend der anderen Eigenschaften in Fig. 3 kann sich eine solche Ausrüstung an der gleichen physikalischen Stelle befinden oder nicht, ist aber zur einfacheren Veranschaulichung in Fig. 3 zusammen gezeigt. Das System **106** erzeugt die Masken **110** oder andere Synthesemuster, wie Chrom-auf-Glas-Masken zur Verwendung bei der Herstellung von Polymergruppierungen.

[0032] Die Masken **110**, sowie die ausgewählte Information in Bezug auf das Design der Chips aus System **100** werden in einem Synthesesystem **112** verwendet. Das Synthesesystem **112** beinhaltet die notwendige Hard- und Software, die zur Herstellung von Polymergruppierungen auf einem Substrat oder einem Chip **114** verwendet werden. Das Synthesegerät **112** beinhaltet eine Lichtquelle **116** und eine chemische Fließzelle **118**, auf der das Substrat oder Chip **114** untergebracht wird. Die Maske **110** wird zwischen der Lichtquelle und dem Substrat bzw. Chip untergebracht, und die beiden werden bei geeigneten Zeiten zum Entfernen des Schutzes ausgewählter Bereiche des Chips in Bezug zueinander verschoben. Ausgewählte chemische Reagenzien werden durch die Fließzelle **118** zum Kuppeln an die Bereiche, bei denen der Schutz entfernt wurde, sowie für Wasch- und andere Vorgänge geleitet. Sämtliche Vorgänge werden vorzugsweise durch einen geeignet programmierten Computer **119** gesteuert, der der gleiche Computer wie der oder die Computer sein kann, wie sie zum Masken-Entwurf und zur Maskenherstellung verwendet werden.

[0033] Die durch das Synthesesystem **112** hergestellten Substrate werden gegebenenfalls zu kleineren Chips geschnitten und Markerzielen ausgesetzt. Die Ziele können zu einem oder mehreren Molekülen auf dem Substrat komplementär sein oder nicht. Die Ziele werden mit einer Markierung, wie einer Fluorescein-Markierung (angegeben mit einem Stern in Fig. 3), markiert und in einem Scanningsystem **120** untergebracht. Die bevorzugten Ausführungsformen verwenden zwar Fluoreszenzmarker, jedoch können andere Marker eingesetzt werden, die Unterschiede in der Radioaktivitätsintensität, Lichtstreuung, Brechungsindex, Leitfähigkeit, Elektrolumineszenz oder anderen Erkennungsdaten großer Moleküle bereitstellen. Daher ist die Erfindung nicht auf die Analyse von Hybridisierungs-Fluoreszenzmessungen eingeschränkt, sondern sie kann leicht zur Analyse anderer Hybridisierungsmessungen verwendet werden.

[0034] Das Scanningsystem **120** wirkt wiederum unter der Steuerung eines geeignet programmierten Digitalcomputers **122**, der ebenfalls der gleiche Computer sein kann wie die Computer, die bei der Synthe-

se, Masken-Herstellung und Masken-Design verwendet werden, oder nicht. Der Scanner **120** umfasst eine Erkennungsvorrichtung **124**, wie ein Konfokal-Mikroskop oder eine CCD (ladungsgekoppelte Vorrichtung), die zur Erkennung der Stelle verwendet wird, an der das markierte Ziel (*) an das Substrat gebunden ist. Der Ausgang von Scanner **120** ist ein bzw. mehrere Bilddateien, die im Fall eines fluoresceinmarkierten Ziels die Fluoreszenzintensität (Photonenzahlen oder eine andere ähnliche Messung, wie Spannung) als Funktion der Position auf dem Substrat anzeigt. Da höhere Photonenzahlen beobachtet werden wo die markierten Ziele stärker an die Gruppierung der Polymere gebunden sind (beispielsweise DNA-Sonden an das Substrat), und da die Monomersequenz der Polymere auf dem Substrat als Funktion der Position bekannt ist, lässt bzw. lassen sich die Sequenzen) von dem oder den Polymeren auf dem Substrat bestimmen, die komplementär zum Ziel sind.

[0035] Die Bilddatei **124** wird einem Analysesystem **126** eingegeben, das die erfindungsgemäßen Syntheseintegritätsbewertungstechniken enthält. Wiederum kann das Analysesystem aus einer großen Anzahl von einem oder mehreren Computersystemen ausgewählt werden, jedoch beruht das Analysesystem bei einer bevorzugten Ausführungsform auf der WINDOWS NT Workstation oder einem Äquivalent. Das Analysesystem kann die Bilddateien) analysieren und den geeigneten Ausgang **128**, wie die Identität spezifischer Mutationen in einem Ziel wie DNA oder RNA, erzeugen.

[0036] Fig. 4 veranschaulicht die Bindung einer bestimmten Ziel-DNA an eine Gruppierung von DNA-Sonden **114**. Wie in diesem einfachen Beispiel gezeigt werden die folgenden Sonden in der Gruppierung gebildet:

3'-AGAACGT

AGACCGT

AGAGCGT

AGATCGT

•

•

•

[0037] Wird demzufolge das fluoresceinmarkierte (oder sonst wie markierte) Ziel 5'-TCTTGCA der Gruppierung ausgesetzt, ist es nur komplementär zur Sonde 3'-AGAACGT, und Fluorescein findet sich vorwiegend auf der Oberfläche desjenigen Chips, auf dem sich 3'-AGAACGT befindet. Der Chip enthält Zellen, die mehrere Kopien einer bestimmten Sonde enthalten, und die Zellen können quadratische Bereiche auf dem Chip sein.

[0038] Fig. 5 ist ein High-Level-Fließschema eines Verfahrens zur Analyse von Testpolymeren, wie Nukleinsäuresequenzen. Bei Schritt **201** werden Hybridisierungsaffinitätsdatensätze in ein Computersystem eingegeben. Die Hybridisierungsaffinitätsdaten können in einer beliebigen Zahl von Formen vorliegen, einschließlich Fluoreszenz-, Radioaktivitäts- oder anderer Daten. Die Hybridisierungsaffinitätsdaten können ohne Modifikation als Eingabe für die Clusterinformation verwendet werden. Die Variationen der Daten können reduziert werden, indem man sie normalisiert.

[0039] Die Hybridisierungsaffinitätsdaten von jedem Satz werden bei Schritt **203** normalisiert. Mit Hilfe der Normalisierung lassen sich übereinstimmendere Daten zwischen und innerhalb der Experimente schaffen. Die Normalisierung kann beispielsweise das Dividieren jedes Hybridisierungsaffinitätswertes durch die Summe sämtlicher anderer Hybridierungsaffinitätswerte beinhalten, so dass jeder Hybridisierungsaffinitätswert auf einen Wert zwischen 0 und 1 reduziert wird. Die Normalisierung kann zwar bei einigen Anwendungen vorteilhaft sein, ist aber nicht erforderlich. Daher veranschaulichen die in den Fließdiagrammen gezeigten Schritte spezifische Ausführungsformen, und Schritte können im Geist und Schutzbereich der Erfindung wegegelassen, hinzugefügt, kombiniert und modifiziert werden.

[0040] Bei Schritt **205** werden die Hybridisierungsaffinitätsdatensätze geclustert. Clusteranalyseverfahren akzeptieren als Eingabe gewöhnlich Mehrfachdatenmuster (beispielsweise dargestellt durch Vektoren von Fließkommazahlen) und ordnen die Muster zu Clustern mit ähnlichen Mustern um. Bevorzugte Ausführungsformen gruppieren Datenmuster zu hierarchischen Clustern, wobei jeder Cluster Cluster beinhaltet, die einander ähnlicher sind als zu anderen Clustern.

[0041] Sobald die Cluster gebildet wurden, können sie einem Anwender auf dem Bildschirm zur Analyse bei Schritt **207** angezeigt werden. Zusätzlich zur Anzeige der Cluster kann das Computersystem ebenfalls die Cluster interpretieren und dem Anwender die Anzahl der gefundenen unterschiedlichen Cluster ausgeben. Fig. 5 wurde bei High Level beschrieben, damit der Leser ein anfängliches Verständnis der Erfindung erhält, und die folgende Beschreibung erläutert die Erfindung eingehender.

[0042] Fig. 6 zeigt ein Fließschema eines Verfahrens, das Hybridisierungsaffinitätsdaten clustert. Bei Schritt **301** wird überprüft, ob die Hybridisierungsaffinitätsdatensätze zu einem einzelnen Root-Cluster geclustert wurden. Ein Cluster kann ein oder mehrere Untercluster enthalten, und ein Root-Cluster ist ein Cluster, der nicht in einem anderen Cluster enthalten ist. In der folgenden Beschreibung kann ein Cluster

(oder Untercluster) ein einzelner Hybridisierungsaffinitätsdatensatz sein, oder er kann mehrere Sätze aufweisen.

[0043] Zu Beginn wird jeder Hybridisierungsaffinitätsdatensatz als Einzelcluster angesehen. Mit fortschreitendem Clustern werden Cluster, die für ähnlich genug befunden werden, zu einem neuen Cluster zusammengefasst. Wird bestimmt, dass sämtliche Hybridisierungsaffinitätsdatensätze zu einem einzelnen Root-cluster in Schritt **303** geclustert werden, ist das Clustern beendet.

[0044] Ansonsten werden die zwei engsten Cluster bei Schritt **305** gefunden. Der Begriff "am engsten sein" bedeutet, dass ein metrisches Maß angibt, dass zwei der Cluster Daten enthalten, die einander ähnlicher sind als die anderen Cluster untereinander. Eine Anzahl verschiedener metrischer Abstände kann verwendet werden, einschließlich der euklidischen Strecke, die eingehender anhand der Fig. 7 beschrieben wird. Am stärksten bevorzugt erfüllt der metrische Abstand die Dreicksungleichung insofern als f(a,c)<= f(a,b) + f(b,c) für jeden Datenmustersatz {a,b,c} gilt.

[0045] In den hier beschriebenen Ausführungsformen beinhaltet ein Cluster bis zu zwei Hybridisierungsaffinitätsdatensätze. Es ist jedoch nicht erforderlich, dass die Cluster auf diese Weise eingeschränkt werden. Die Erfindung kann beispielsweise vorteilhaft bei Clustern angewendet werden, die durch eine Erweiterung der hier beschriebenen Prinzipien bis zu drei oder mehr Hybridisierungsaffinitätsdatensätze enthalten können.

[0046] Bei Schritt **307** wird ein neuer Cluster erzeugt, der die zwei engsten Cluster enthält. Zum Vergleichen des neuen Clusters mit anderen Clustern sollte ein Wert berechnet werden, der die Daten in dem neuen Cluster darstellt. Bei einer Ausführungsform wird der Durchschnitt der beiden engsten Cluster für den neuen Cluster bei Schritt **309** mit dem Computer berechnet. Nach dem Erzeugen des neuen Clusters läuft der Fluss zu Schritt **301**, und es wird überprüft, ob nur ein Root-Cluster übrig bleibt.

[0047] Die Fig. 7 zeigt ein Fließdiagramm eines Verfahrens zur Analyse von Testnukleinsäuresequenzen. Für diese Ausführungsform werden Hybridisierungsdaten von einem Chip mit Sense- und Antisense-Sonden verwendet. Fragmente der Sense- und Antisense-Stränge eines Ziels werden markiert und dem Chip ausgesetzt, was zu vier Hybridisierungsaffinitätsmessungen für den Sense-Strang und vier Hybridisierungsaffinitätsmessungen für den Antisense-Strang bei jeder Abfrageposition führt.

[0048] Ist beispielsweise der Sense-Strang einer Zielsequenz (oder eines Abschnitts davon) 5'-GTAACGTTG, dann würden die folgenden Sense-Sonden die unterstrichene Basenposition abfragen:

3'-TTACA

3'-TTCCA

3'-TTGCA

3'-TTTCA

[0049] Der Antisense-Strang der Zielsequenz (oder eines Abschnitts davon) ist 3'-CATTGCAAC, und die folgenden Sense-Sonden fragen die unterstrichene Basenposition für den Antisense-Strang ab:

5'-AAAGT

5'-AACGT

5'-AAGGT

5'-AATGT

[0050] Folglich gibt es in dieser Ausführungsform acht Hybridisierungsaffinitäten, und zwar eine für jede Sonde bei jeder Abfrageposition.

[0051] Bei Schritt **401** werden die Hybridisierungsaffinitätsdatensätze in ein Computersystem eingegeben. Dies kann das Lesen einer Datei beinhalten, die die Hybridisierungsaffinitätsdaten für jede Basenposition enthält, die in dem Ziel abgefragt wird. Wie vorstehend erörtert können die Hybridisierungsaffinitätsdaten für eine Basenposition 8 gemessene Hybridisierungsaffinitäten beinhalten. Die 8 gemessenen Hybridisierungsaffinitäten können als Satz oder Muster von 8 Werten (beispielsweise Photonenzahl), wie $\{A_1, A_2, ..., A_8\}$ gespeichert werden.

[0052] Die Hybridisierungsaffinitätsdaten jedes Satzes werden bei Schritt **403** normalisiert. Die Normalisierung der Hybridisierungsaffinitätsdaten kann die Bedeutung der Unterschiede, die sich nicht direkt auf die Zielsequenz-Zusammensetzung beziehen, mindern. Eine effiziente Strategie zur Normalisierung der Hybridisierungsaffinitäten eines Satzes ist es, erst den Durchschnitt der Hybridisierungsaffinitäten für einen Satz zu berechnen und diesen Durchschnitt von jeder Hybridisierungsaffinität in dem Satz zu subtrahieren. Dann wird jede Hybridisierungsaffinität, von der der Durchschnitt subtrahiert wurde, durch die Quadratwurzel der Summe der quadrierten Hybridisierungsaffinitäten des Satzes abzüglich der durchschnittlichen Hybridisierungsaffinität dividiert. Mit anderen Worten wird die folgende Formel zur Normalisierung jeder Hybridisierungsaffinität eines Satzes verwendet.

$A_l = (A_l - \bar{A})/\text{Quadratwurzel}\ ((A_1 - \bar{A})^2 + (A_2 - \bar{A})^2 + ... + (A_8 - \bar{A})^2)$,

wobei 1 von 1 bis 8 reicht, und $\bar{A}$ der Durchschnitt von $A_1, A_2, ..., A_8$ ist.

[0053] Fig. 8 zeigt graphisch, wie die Normalisierung die Hybridisierungsaffinitäten beeinflussen kann. Die Hybridisierungsaffinitäten **451** sind die von dem Chip gemessenen Rohdaten, und die Höhe der Balken zeigt die relative gemessene Hybridisierungsaffinität.

[0054] Die Hybridisierungsaffinitäten **453**, von denen der Durchschnitt subtrahiert wurde, zeigen, dass die Hybridisierungsaffinitäten nun Vektoren in zwei möglichen Richtungen sind. Die Hybridisierungen, von denen der Durchschnitt subtrahiert wurde, werden in einem Zwischenvektormuster **455** vereinigt. Die Normalisierung des Vektormusters **455** wird beendet, indem die Vektoren jeweils durch den vorstehenden Nenner geteilt werden, so dass ein endgültiges normalisiertes Vektormuster **457** erhalten wird.

[0055] Die Normalisierung kann verschiedene Hintergründe und Gesamthybridisierungsaffinitätswerte korrigieren, während die Stellung jeder Hybridisierungsaffinität in dem Satz sowie der Unterschied der Gesamthybridisierungsaffinitätzwischen den Sense- und Antisense-Sonden beibehalten werden. Zudem wird durch Normalisieren des Satzes von 8 Werten in der beschriebenen Weise die Distanz zwischen zwei Mustern auf (0,2) begrenzt, so dass eine konsistente Skala erhalten wird, auf der die Musterunterschiede bewertet werden können.

[0056] Ebenfalls in Fig. 7 werden die Hybridisierungsaffinitätsdatensätze bei Schritt **405** hierarchisch geclustert. Es kann jede Zahl von Clusteralgorithmen verwendet werden. Bei bevorzugten Ausführungsformen wird eine Modifikation des mittleren Linkage-Clusteralgorithmus verwendet. Der Wert für einen Cluster, der nur einen einzigen Satz von Hybridisierungsaffinitäten enthält, ist das Muster von 8 Hybridisierungsaffinitäten. Der Wert von einem Cluster C, der zwei Cluster A und B enthält, ist wie folgt: C_l = Durchschnitt (A_l, B_l), wobei l von 1 bis 8 reicht. Somit wird jedes Cluster durch ein 8-Werte-Muster dargestellt. Andere Linkage-Berechnungen lassen sich verwenden, einschließlich des herkömmlichen mittleren Linkage, wobei das Mittel der Distanzen zwischen den einzelnen Mitglieder eines Musters verwendet wird. Zusätzlich kann die größte (oder kleinste) Distanz zwischen den beiden Mitgliedern von mindestens 2 Clustern als Linkage-Formel verwendet werden.

[0057] Die Distanz zwischen den beiden Clustern wird gewöhnlich durch einen metrischen Abstand bestimmt. Viele verschiedene metrische Abstände können verwendet werden, einschließlich der euklidischen Distanz, City-Block-Distanz Korrelationsdistanz, Winkeldistanz und dergleichen. Am stärksten bevorzugt wird die euklidische Strecke verwendet, die wie folgt berechnet wird:

$$D_{AB} = \text{Quadratwurzel}\left((A_1 - B_1)^2 + (A_2 - B_2)^2 + \ldots + (A_8 - B_8)^2\right)$$

wobei lvon 1 bis 8 reicht. Die Cityblock-Distanz kann wie folgt errechnet werden:

$$D_{AB} = |(A_1 - B_1)| + |(A_2 - B_2)| + \ldots + |(A_8 - B_8)|$$

wobei lvon 1 bis 8 reicht, und |(X)| der absolute Wert von X ist.

[0058] Bei Schritt **407** wird die Anzahl "enger" Cluster gezählt. Ein "enger" Cluster ist definiert als ein beliebiger Cluster, bei dem die mittlere Distanz zwischen dem Clustermittel und den Mitteln seiner Untercluster geringer ist als die Distanz zum nächsten Schwestercluster, und zwar über einen Ähnlichkeitsfaktor (beispielsweise einen Faktor von 3). Für den Anwender ist es ziemlich einfach, optisch Cluster zu identifizieren, aber die Anzahl enger Cluster kann als berechnete Bestimmung der Anzahl von Clustern verwendet werden. Gibt es zwei oder mehrere enge Cluster, ist die Abfrageposition wahrscheinlich polymorph. Man beachte, dass die Steigerung der Zahl der Dimensionen in einem Eingabemuster die Möglichkeit stark senkt, dass zwei Muster zufällig ähnlich sind, und der Wert des Ähnlichkeitsfaktors kann entsprechend eingestellt werden.

[0059] Die Cluster werden bei Schritt **409** angezeigt. Die Cluster können auf beliebige Weise angezeigt werden, aber in bevorzugten Ausführungsformen werden sie als Dendrogramme angezeigt. Dendrogramme sind Diagramme, die Cluster darstellen. Die Distanz zwischen den Clustern kann auf dem Dendrogramm dargestellt werden, so dass der Anwender die Cluster leichter identifizieren kann, die einen Polymorphismus anzeigen, wie eine Mutation, Insertion oder Deletion. Mit anderen Worten variiert die Distanz zwischen den Clustern mit der Ähnlichkeit der Cluster.

[0060] Fig. 9 veranschaulicht beispielsweise eine Bildschirmanzeige mit einem Dendrogramm, das anzeigt, dass kein Polymorphismus an der interessierenden Position zu sein scheint. Eine Bildschirmanzeige **501** beinhaltet ein Dendrogramm **503**. Das Dendrogramm wird anhand der Fig. 10 eingehender beschrieben.

[0061] Die Bildschirmanzeige **501** beinhaltet Rohdaten **505** und die angezeigten Basenaufrufe. Ein Plot **507** der Hybridisierungsaffinitäten gegen die Basenposition ist sowohl für Sense- als auch Antisense-Stränge zur Mustererkennung gezeigt. Eine Tabelle **509** beinhaltet die Information über die Basenpositionen für den Chip. Zusätzlich bietet ein Bild **511** Information für die Abschätzung des Anteils der Mutanten. Das Dendrogramm **503** (und andere) ist das Thema der folgenden Absätze.

[0062] Die Fig. 10 zeigt ein Dendrogramm von Fig. 9, das 8 Hybridisierungsaffinitätsdatensätze (dargestellt durch den Zielnamen) clustert. Eine optische Untersuchung von Dendrogramm **503** ergibt, dass die Distanz zwischen den Clustern (veranschaulicht durch die horizontale Länge des Dendrogramms) relativ konstant ist. Dies zeigt, dass die Muster relativ konstant sind und daher wahrscheinlich kein Polymorphismus an der Abfrageposition vorliegt.

[0063] Fig. 11 veranschaulicht ein Dendrogramm, das anzeigt, dass wahrscheinlich ein Polymorphismus an der interessierenden Position vorliegt. Das Dendrogramm **603** zeigt das Clustern von 8 Hybridisierungsaffinitätsdatensätzen. Eine optische Begutachtung des Dendrogramms ergibt, dass es zwei Cluster **605** und **607** zu geben scheint, wobei die Distanz zwischen den Mitgliedern eines Clusters viel geringer ist als die Distanz zwischen Mitgliedern anderer Cluster. Da die Muster in zwei Cluster fallen, liegt wahrscheinlich ein Polymorphismus an der Abfrageposition vor.

[0064] Als weiteres Beispiel veranschaulicht Fig. 12 eine Bildschirmanzeige mit einem Dendrogramm, das zeigt, dass wahrscheinlich mehr als ein Polymorphismus an der interessierenden Basenposition vorliegt. Ein Dendrogramm **703** zeigt die Clusterung von 8 Hybridisierungsaffinitätsdatensätzen. Eine optische Untersuchung des Dendrogramms ergibt, dass es drei Cluster **705**, **707** und **709** zu geben scheint, bei denen die Distanz zwischen den Mitgliedern eines Clusters viel geringer ist, als die Distanz zwischen den Mitgliedern anderer Cluster. Da die Muster in drei Cluster fallen, gibt es wahrscheinlich 2 Polymorphismen an der Abfrageposition.

[0065] Erfindungsgemäß lassen sich Erscheinungen erkennen, die durch Untersuchung einer einzelnen Hybridisierungsreaktion nicht erkennbar sind. Die Anzahl und die Diversität von Sonden zum Erkennen einer bestimmten Klasse von Erscheinungen kann reduziert werden. Mutationen im BRCA-Gen sind beispielsweise so divers, dass die Konstruktion eines Satzes von Sonden, die jeden möglichen Polymorphismus abdecken würden, unpraktisch ist. Die Erfindung kann zur Erkennung solcher Polymorphismen selbst in Abwesenheit solcher Sonden verwendet werden.

[0066] Das Clustern kann zudem zur Analyse oder Bewertung der Effektivität von Versuchssystemen verwendet werden, z.B. von Genotypbestimmungs-Chips, wobei geeignete Ergebnisse von der Erkennung einer festen Anzahl stark reproduzierbarer Klassen in den resultierenden Daten abhängen. Im Fall der Genotypbestimmung erwartet man drei dicht geclusterte Klassen, die den homozygoten Wildtyp, die homozygote Mutante bzw. die heterozygoten Genotypen veranschaulichen. Metrische Abstände, die mit der Hierarchie von Mustern errechnet werden, und die durch einen Clustering-Algorithmus erzeugt werden, können eine quantitative Bestimmung der Spezifität und der Reproduzierbarkeit des Genotyp-Bestimmungsverfahrens bieten.

[0067] Das Vorstehende ist zwar eine vollständige Beschreibung der bevorzugten Ausführungsformen der Erfindung, jedoch können verschiedene Alternativen, Modifikationen und Äquivalente verwendet werden. Es sollte ersichtlich sein, dass die Erfindung sich gleichermaßen durch geeignetes Modifizieren der vorstehend beschriebenen Ausführungsformen anwenden lässt. Die Erfindung wurde beispielsweise in Bezug auf Nukleinsäuresonden beschrieben, die auf einem Chip synthetisiert werden. Die Erfindung kann jedoch vorteilhafterweise auf andere Monomere (beispielsweise Aminosäuren und Saccharide) sowie andere Hybridisierungstechniken angewendet werden, z.B. solche, bei denen die Sonden nicht an ein Substrat gebunden sind.

Patentansprüche

1. Verfahren zum Erkennen von Polymorphismen in Testnukleinsäuresequenzen, umfassend
Eingeben einer Anzahl Datensätze zu Hybridisierungsaffinitäten, wobei die Hybridisierungsaffinitätsdatensätze die Hybridisierungsaffinitäten zwischen einer Testnukleinsäuresequenz und den Nukleinsäureproben enthalten;
hierarchisches Clustern der Anzahl Hybridisierungsaffinitätsdatensätze in eine Anzahl Cluster, so dass die Hybridisierungsaffinitätsdatensätze in den Clustern alle untereinander ähnlicher sind als zu den Hybridisierungsaffinitätsdatensätzen anderer Cluster; und
Analysieren der Anzahl Cluster zur Bestimmung einer Zahl enger Cluster, wobei ein enger Cluster ein Cluster ist, bei dem die durchschnittliche Distanz zwischen dem Clustermittel und den Mitteln seiner Untercluster geringer ist als die Distanz zum nächsten Schwestercluster und zwar über einen Ähnlichkeitsfaktor, der derart ist, dass die Anzahl enger Cluster anzeigt, ob ein Polymorphismus in den Testnukleinsäuresequenzen vorliegt.

2. Verfahren nach Anspruch 1, wobei die Testnukleinsäurequenz und die Nukleinsäureproben sowohl Sense- als auch Antisense-Stränge enthält.

3. Verfahren nach Anspruch 2, wobei die Hybridisierungsaffinitätsdaten vier Hybridisierungsaffinitäten für die Sensestränge darstellen und vier Hybridisierungsaffinitäten für die Antisense-Stränge.

4. Verfahren nach Anspruch 3, wobei die vier Hybridisierungsaffinitäten für die Antisense-Stränge die Hybridisierungsaffinitäten zwischen den Nukleinsäu-

reproben darstellen, die durch mindestens eine Nukleinsäure an der Abfrageposition verschieden sind.

5. Verfahren nach Anspruch 3, wobei die vier Hybridisierungsaffinitäten für die Antisense-Stränge Hybridisierungsaffinitäten darstellen zwischen Nukleinsäureproben, die durch mindestens eine Nukleinsäure an der Abfrageposition verschieden sind.

6. Verfahren nach Anspruch 1, wobei die Polymorphismen Mutationen, Deletionen und Insertionen an der Abfrageposition aufweisen.

7. Verfahren nach Anspruch 1, zudem umfassend die normalisierten Hybridisierungsaffinitätsdaten für jeden Satz und umfassend das hierarchische Clustern der normalisierten Hybridisierungsaffinitätsdaten.

8. Verfahren nach Anspruch 7, wobei das Normalisieren der Hybridisierungsaffinitätsdaten für jeden Satz beinhaltet das Abziehen einer mittleren Hybridisierungsaffinität von den Hybridisierungsaffinitäten und das Teilen der jeweiligen Hybridisierungsaffinität durch die Quadratwurzel der Summe der quadrierten Abweichungen der Hybridisierungsaffinitäten.

9. Verfahren nach Anspruch 1, wobei der Schritt des Clusterns der Anzahl Hybridisierungsaffinitätsdatensätze beinhaltet das Berechnen einer mittleren Clusterverknüpfung der Cluster.

10. Verfahren nach Anspruch 9, wobei die mittlere Clusterverknüpfung der Proben einen metrischen Abstand für die Unterschiede zwischen den Clustern verwendet.

11. Verfahren nach Anspruch 10, wobei der metrische Abstand eine euklidische Strecke oder eine Cityblock-Distanz ist.

12. Verfahren nach Anspruch 1, zudem umfassend das Darstellen einer Baumstruktur der Anzahl Cluster.

13. Verfahren nach Anspruch 12, wobei die Strecke zwischen den Clustern sich mit der Ähnlichkeit der Cluster ändert.

14. Computerprogramm, umfassend Kodiereinrichtungen zur Durchführung all der Schritte des Verfahrens nach einem der Ansprüche 1 bis 13, sofern auf einem Datenverarbeitungssystem vorgenommen.

15. Computerlesbares Medium mit einer Computercodeeinrichtung darauf, das die Schritte des Verfahrens nach einem der Ansprüche 1 bis 13 vornimmt, wenn auf einem Datenverarbeitungssystem ausgeführt.

16. Computerlesbares Medium nach Anspruch 15, wobei das computerlesbare Medium ausgewählt ist aus der Gruppe Floppy-Disk, Band, Flash-Memory, System-Memory, Hard-Drive und Datensignal auf einer Trägerwelle.

Es folgen 10 Blatt Zeichnungen

Anhängende Zeichnungen

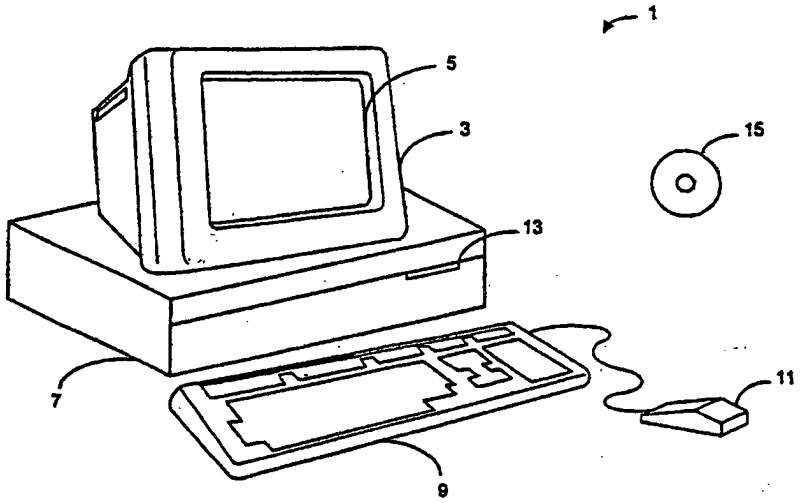

FIG. 1

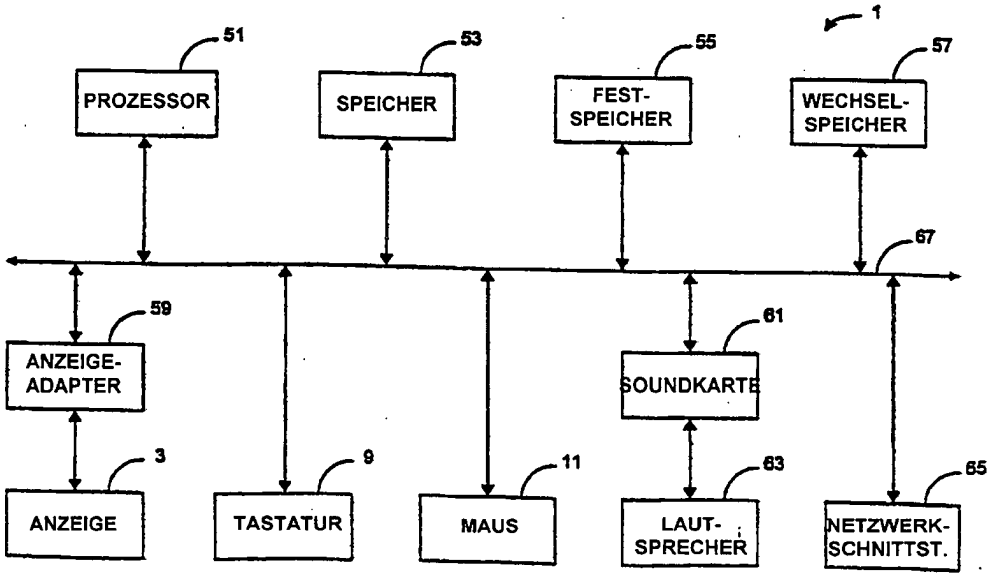

FIG. 2

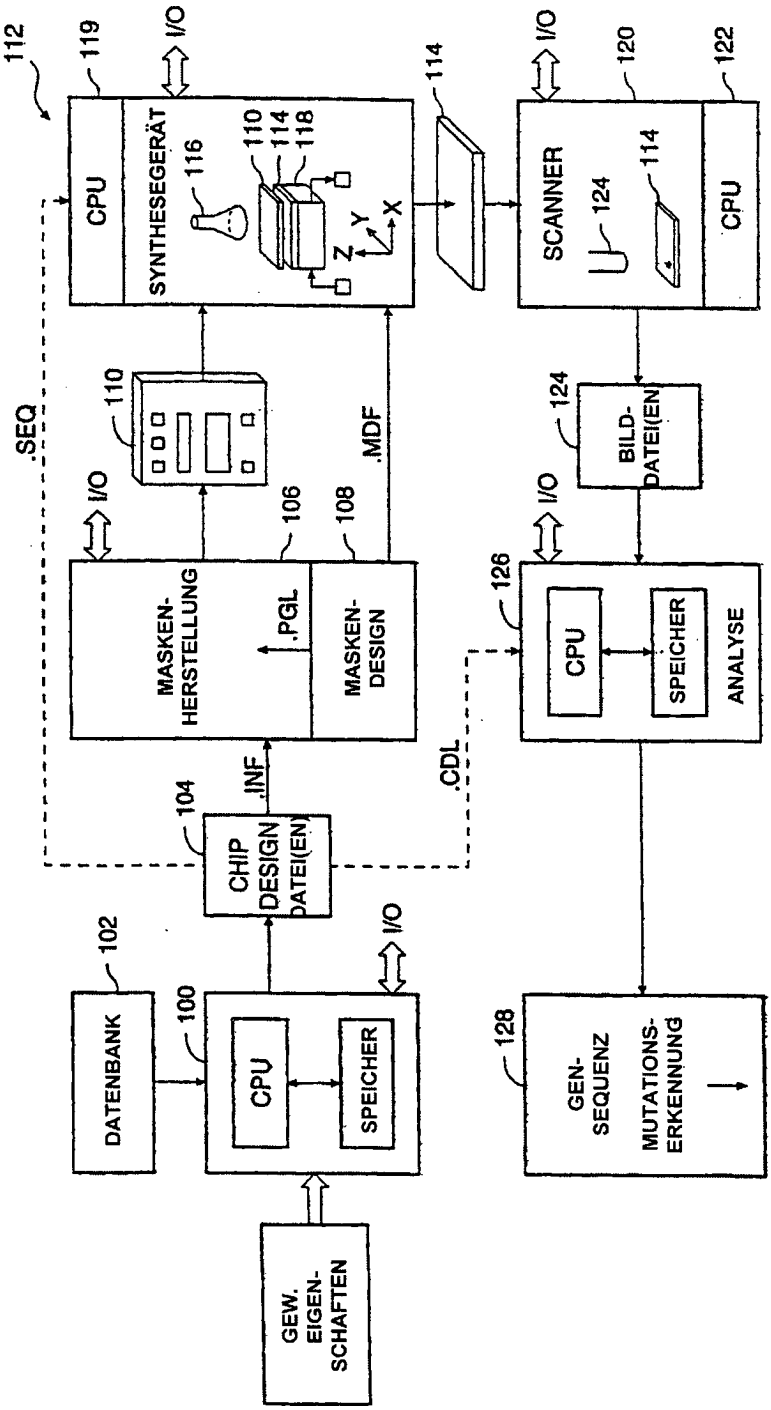
112
119
I/O
CPU
SYNTHESEGERÄT
116
110
114
118
Z
Y
X
114
I/O
120
122
SCANNER
124
114
CPU
.SEQ
110
.MDF
124
BILD-DATEI(EN)
I/O
106
108
I/O
126
MASKEN-HERSTELLUNG
.PGL
MASKEN-DESIGN
CPU
SPEICHER
ANALYSE
.INF
.CDL
104
CHIP DESIGN DATEI(EN)
102
I/O
DATENBANK
100
CPU
SPEICHER
128
GEN-SEQUENZ
MUTATIONS-ERKENNUNG
GEW. EIGEN-SCHAFTEN

FIG. 3

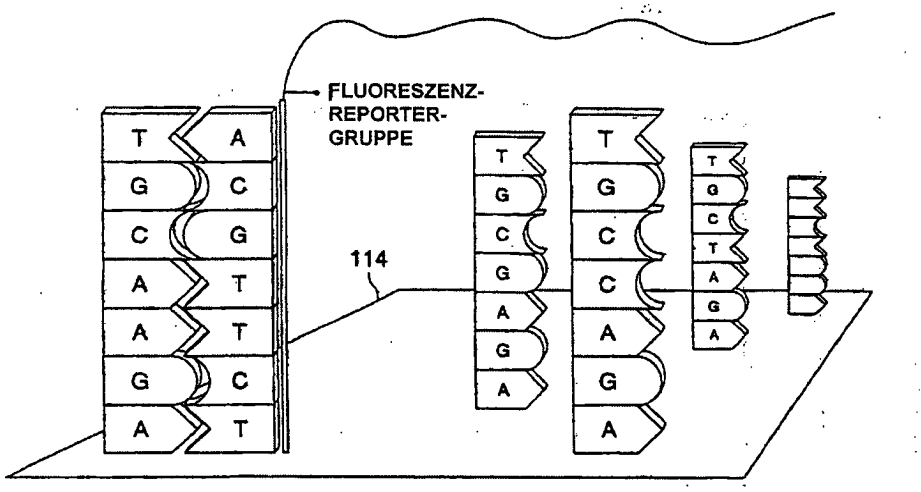
FLUORESZENZ-
REPORTER-
GRUPPE
114
T
A
G
C
C
G
A
T
A
T
G
C
A
T
T
G
C
G
A
G
A
T
G
C
C
A
G
A
T
G
C
T
A
G
A

FIG. 4

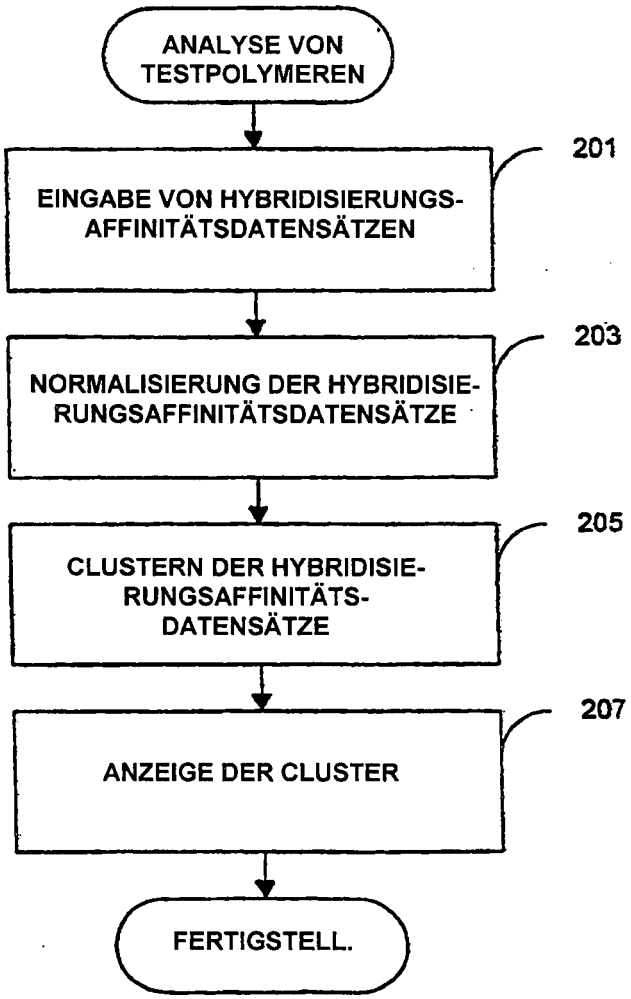
ANALYSE VON TESTPOLYMEREN
201
EINGABE VON HYBRIDISIERUNGS-AFFINITÄTSDATENSÄTZEN
203
NORMALISIERUNG DER HYBRIDISIE-RUNGSAFFINITÄTSDATENSÄTZE
205
CLUSTERN DER HYBRIDISIE-RUNGSAFFINITÄTS-DATENSÄTZE
207
ANZEIGE DER CLUSTER
FERTIGSTELL.

FIG. 5

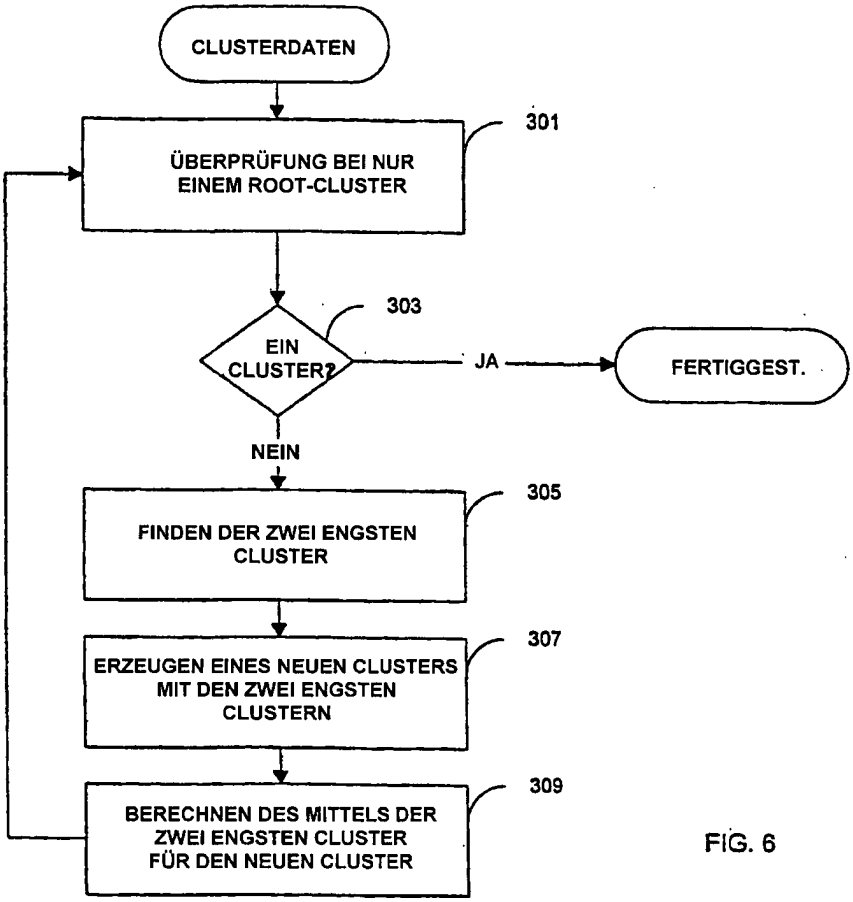
CLUSTERDATEN
301
ÜBERPRÜFUNG BEI NUR EINEM ROOT-CLUSTER
303
EIN CLUSTER?
JA
FERTIGGEST.
NEIN
305
FINDEN DER ZWEI ENGSTEN CLUSTER
307
ERZEUGEN EINES NEUEN CLUSTERS MIT DEN ZWEI ENGSTEN CLUSTERN
309
BERECHNEN DES MITTELS DER ZWEI ENGSTEN CLUSTER FÜR DEN NEUEN CLUSTER

FIG. 6

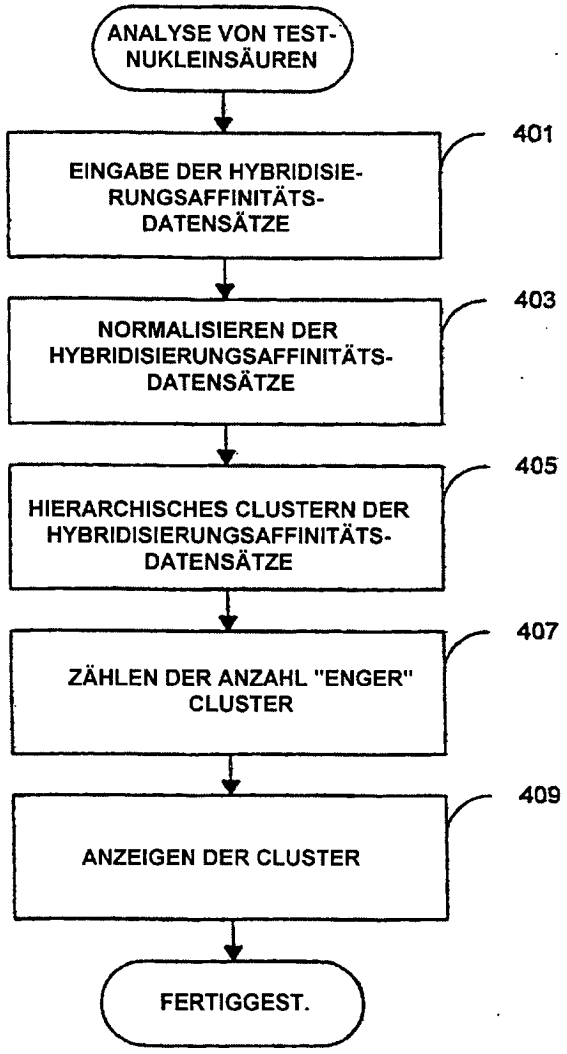
ANALYSE VON TEST-NUKLEINSÄUREN
401
EINGABE DER HYBRIDISIE-RUNGSAFFINITÄTS-DATENSÄTZE
403
NORMALISIEREN DER HYBRIDISIERUNGSAFFINITÄTS-DATENSÄTZE
405
HIERARCHISCHES CLUSTERN DER HYBRIDISIERUNGSAFFINITÄTS-DATENSÄTZE
407
ZÄHLEN DER ANZAHL "ENGER" CLUSTER
409
ANZEIGEN DER CLUSTER
FERTIGGEST.

FIG. 7

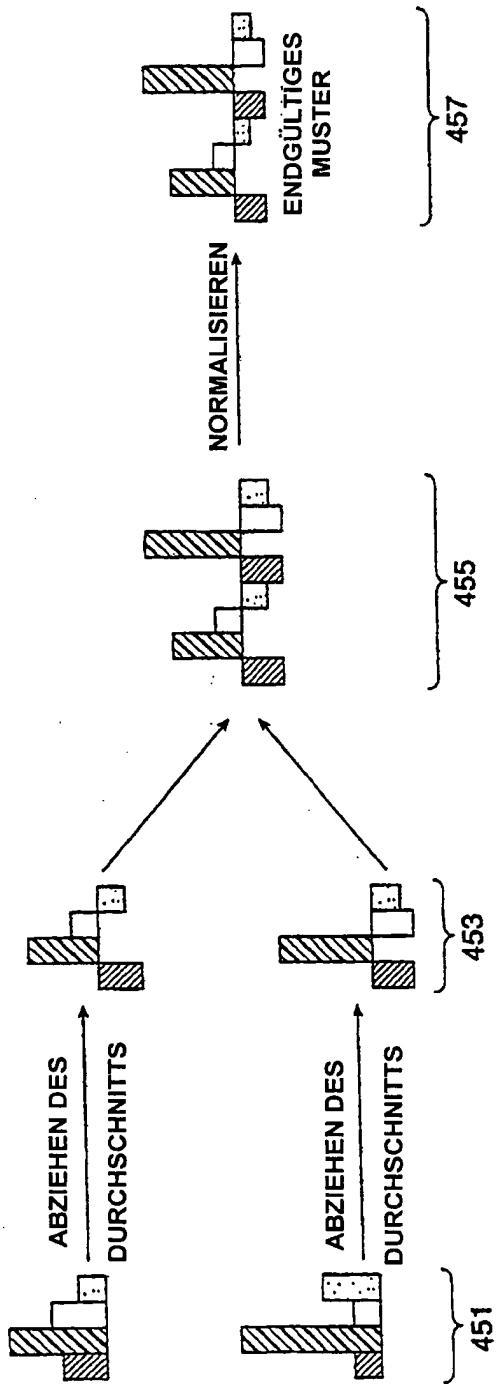
ABZIEHEN DES DURCHSCHNITTS
ABZIEHEN DES DURCHSCHNITTS
NORMALISIEREN
ENDGÜLTIGES MUSTER
451
453
455
457

FIG. 8

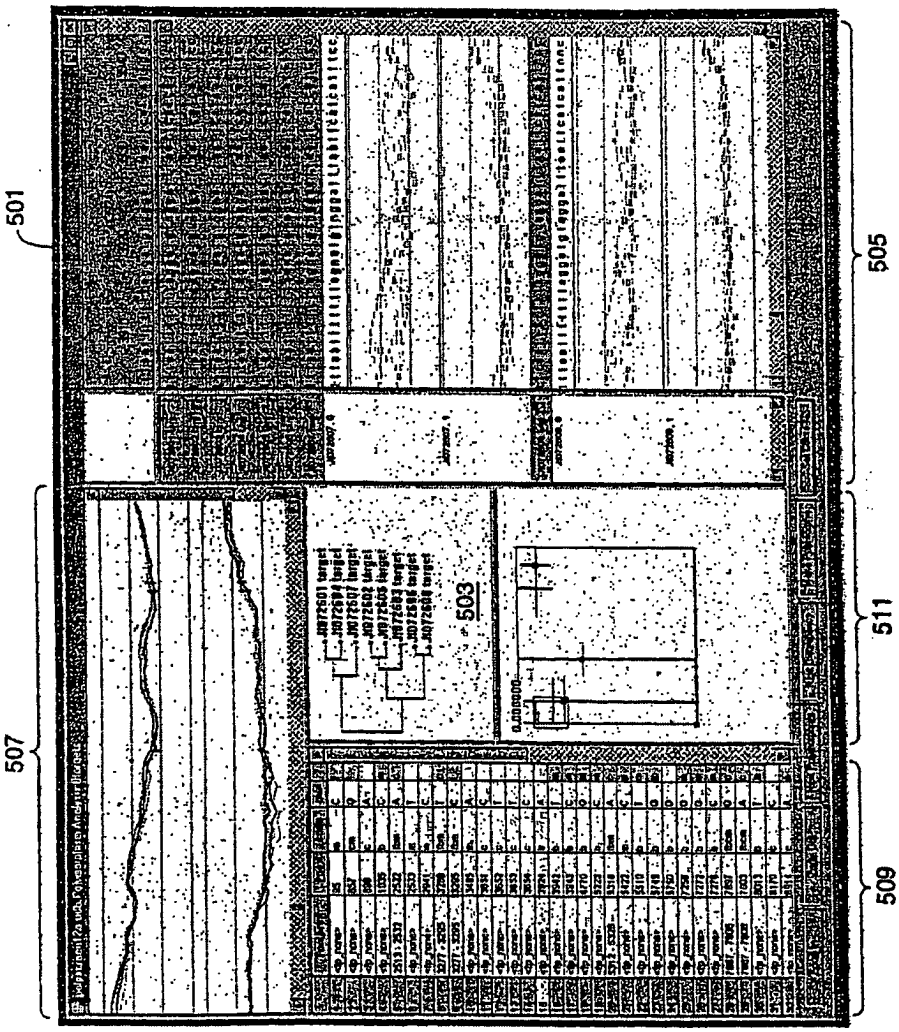

FIG. 9

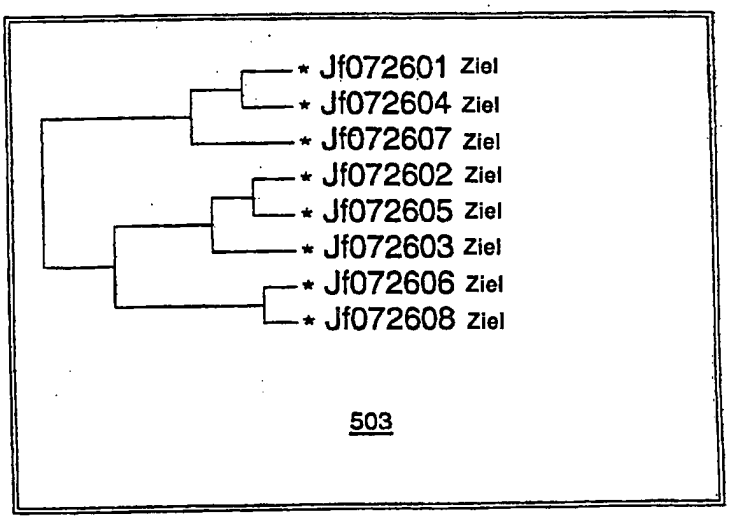
* Jf072601 Ziel
* Jf072604 Ziel
* Jf072607 Ziel
* Jf072602 Ziel
* Jf072605 Ziel
* Jf072603 Ziel
* Jf072606 Ziel
* Jf072608 Ziel
503

FIG. 10

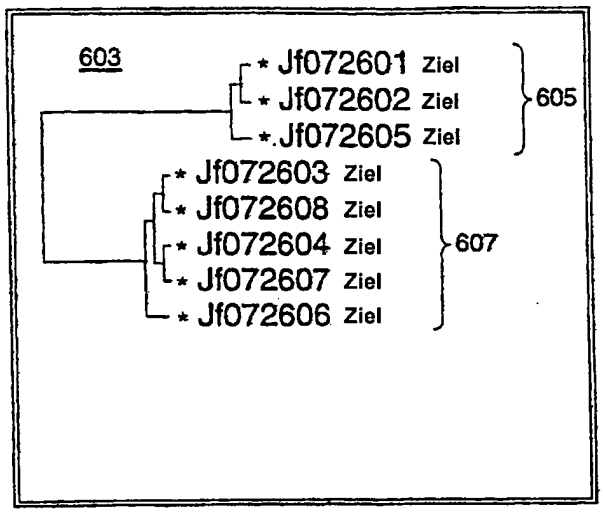
603
* Jf072601 Ziel
* Jf072602 Ziel
*.Jf072605 Ziel
605
* Jf072603 Ziel
* Jf072608 Ziel
* Jf072604 Ziel
* Jf072607 Ziel
* Jf072606 Ziel
607

FIG. 11

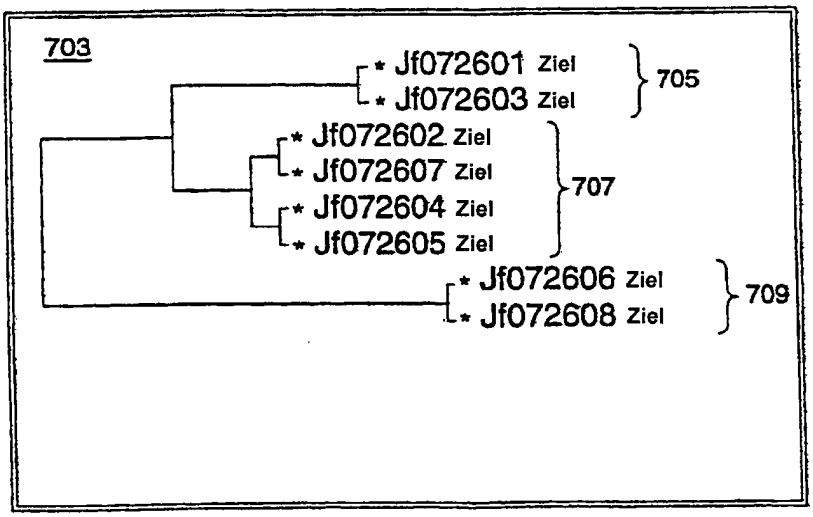
703
* Jf072601 Ziel
* Jf072603 Ziel
705
* Jf072602 Ziel
* Jf072607 Ziel
* Jf072604 Ziel
* Jf072605 Ziel
707
* Jf072606 Ziel
* Jf072608 Ziel
709

FIG. 12