(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 694 573**

(51) Int. CI.:

G01N 33/48 (2006.01)
G01N 33/50 (2006.01)
G06F 7/00 (2006.01)
G06F 17/30 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **21.11.2008 E 16176264 (6)**

(97) Fecha y número de publicación de la concesión europea: **22.08.2018 EP 3144672**

(54) Título: **Sistema de identificación de genomas**

(30) Prioridad:

21.11.2007 US 989641 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
21.12.2018

(73) Titular/es:

COSMOSID INC. (100.0%)
5010 River Hill Road
Bethesda, MD 20816, US

(72) Inventor/es:

COLWELL, RITA, R.;
JAKUPCIAK, JOHN, P. y
CHUN, JONGSIK

(74) Agente/Representante:

SÁEZ MAESO, Ana

DESCRIPCIÓN

Sistema de identificación de genomas

Campo técnico de la invención

Esta invención se refiere a un sistema y a un método para la identificación de organismos y, más particularmente, a la determinación de la secuencia de ácidos nucleicos y otras moléculas de tipo polimérico o de cadena mediante el emparejamiento de datos probabilísticos en un dispositivo electrónico portátil o más grande.

Antecedentes

Hay una gran variedad de circunstancias que ponen en peligro la vida en las que sería útil analizar y secuenciar una muestra de ADN o ARN, por ejemplo, en respuesta a un acto de bioterrorismo en el que se habría liberado un agente patógeno mortal en el ambiente. En el pasado, tales resultados han requerido la participación de muchas personas, lo que exige demasiado tiempo. Como resultado, la rapidez y la precisión pueden sufrir.

En el caso de un ataque bioterrorista o de una epidemia emergente, es importante que los primeros en responder, es decir, los médicos en la sala de emergencias (sus opciones o tratamientos de cabecera), así como para los fabricantes de alimentos, distribuidores, minoristas y para el personal de salud pública en todo el país para identificar de manera rápida, precisa y confiable los agentes patógenos y las enfermedades que causan. Los agentes patógenos pueden estar contenidos en fuentes de muestras tales como alimentos, aire, suelo, agua, tejido y presentación clínica de agentes patógenos. Debido a que los agentes y/o posibles enfermedades pueden poner en peligro la vida y ser altamente contagiosos, este proceso de identificación se debe hacer rápidamente. Esta es una debilidad significativa en la respuesta actual de seguridad nacional al bioterrorismo.

Se necesitan un sistema y un método que puedan identificar más de un solo organismo (multiplexación) e indicar si una especie está presente, con base en la comparación del genoma de los ácidos nucleicos presentes en una muestra.

Los rápidos avances en ingeniería biológica han impactado dramáticamente el diseño y las capacidades de las herramientas de secuenciación de ADN, es decir, la secuenciación de alto rendimiento, que es un método para determinar el orden de las bases en el ADN, produciendo un mapa de variación genética que puede dar pistas al apuntalamiento genético de las enfermedades humanas. Este método es muy útil para secuenciar muchas plantillas diferentes de ADN con cualquier número de cebadores. A pesar de estos importantes avances en ingeniería biológica, se ha avanzado poco en la creación de dispositivos para identificar rápidamente la secuencia [información] y transferir datos de manera más eficiente y efectiva.

Tradicionalmente, la secuenciación del ADN se realizó mediante un método didesoxi, comúnmente denominado método de Sanger [Sanger et al, 1977], que usaba inhibidores de terminación de cadena para detener la extensión de la cadena de ADN mediante la síntesis de ADN.

Se siguen desarrollando métodos novedosos para las estrategias de secuenciación. Por ejemplo, el advenimiento de los chips de ADN hace posible construir un arreglo de secuencias e hibridar secuencias complementarias en un proceso comúnmente denominado secuenciación por hibridación. Otra técnica considerada como el estado actual de la técnica emplea la extensión del cebador seguida de la adición cíclica de un solo nucleótido con cada ciclo seguido de la detección del evento de incorporación. La técnica, comúnmente conocida como secuenciación por síntesis o pirosecuenciación, incluida la secuenciación fluorescente in situ (FISSEQ), es reiterativa en la práctica e implica un proceso en serie de ciclos repetidos de extensión del cebador, mientras que la secuencia de nucleótidos objetivo se secuencia.

El documento US2002/120408 A1 está dirigido hacia el control de infecciones en tiempo real en una red informática. El método comprende obtener una muestra de un microorganismo en una instalación de atención médica, secuenciar una primera región de un ácido nucleico de la muestra de microorganismos, comparar la primera región secuenciada con datos históricos almacenados de secuencias en una base de datos, determinar una medida de la relación filogenética entre la muestra de microorganismos y las muestras históricas almacenadas en la base de datos y proporcionar información de control de infecciones basada en la determinación de la relación filogenética a la instalación de atención médica para uso en el control o prevención de la propagación de una infección.

Existe una necesidad de métodos y sistemas de identificación rápida del genoma, que incluyen comunicaciones electrónicas multidireccionales de datos de secuencias de ácidos nucleicos, datos clínicos, intervención terapéutica y administración personalizada de terapias a la población adecuada para agilizar las respuestas, conservar valiosos suministros médicos y contener el bioterrorismo, liberación involuntaria y nuevas epidemias patógenas.

El sistema actual está diseñado para analizar cualquier muestra que contenga material biológico para determinar la presencia de especies o genomas en la muestra. Esto se logra obteniendo la información de secuencia del material biológico y comparando la información de secuencia con una base o bases de datos. La información de secuencia que coincida indicará la presencia de un genoma o especie. El emparejamiento probabilístico calculará la

probabilidad de que las especies estén presentes. Los métodos pueden aplicarse en sistemas de secuenciación masivamente paralelos.

Resumen de la invención

La presente invención proporciona un método *ex vivo* para identificar un genoma en una muestra que comprende una pluralidad de genomas, como se reivindica en la reivindicación 1 más adelante, que comprende:

(i) obtener una muestra que comprenda la pluralidad de genomas;

(ii) extraer una o más moléculas de ácido nucleico de la muestra;

(iii) generar información de secuencia, la información de secuencia que comprende una secuencia de un fragmento de nucleótido de una o más moléculas de ácido nucleico;

(iv) comparar dicha secuencia de un fragmento de nucleótido con secuencias de ácido nucleico en una base de datos utilizando un emparejamiento probabilístico; y si la comparación de la secuencia de un fragmento de nucleótido no resulta en una coincidencia que identifique un genoma en la muestra en virtud de que la probabilidad de coincidencia del fragmento de nucleótido sea menor que un umbral de una coincidencia objetivo, entonces el método comprende además:

(v) generar información adicional de secuencia a partir de una o más moléculas de ácido nucleico;

(vi) comparar dicha información adicional de secuencia con secuencias de ácido nucleico en una base de datos inmediatamente después de la generación de la información adicional de secuencia utilizando el emparejamiento probabilístico; y

(vii) repetir las etapas (v) - (vi) hasta que resulte una coincidencia en la identificación de un genoma en la muestra.

En una primera realización de la invención, la información adicional de secuencia comprende la secuencia del fragmento de nucleótido que comprende un nucleótido adicional.

En una segunda realización, la secuencia del fragmento de nucleótido de (iv) es una secuencia de nucleótidos de longitud "n" y la información adicional de secuencia comprende una secuencia de nucleótidos de longitud "n + 1", "n + 2" hasta "n + x", donde x es menor que 50.

En una realización, generar la información de secuencia comprende pirosecuenciación.

En una realización, la generación de la información de secuencia comprende la secuenciación por hibridación.

En otra realización de la invención, la secuencia del fragmento de nucleótido de (iv) es una secuencia de nucleótidos de longitud "n"; y la información adicional de secuencia comprende una secuencia de nucleótidos de longitud "n + 1", "n + 2" a "n + x", donde x es mayor que 50.

En otra realización más, la invención comprende además la amplificación de una o más moléculas de ácido nucleico para producir una pluralidad "i" de moléculas de ácido nucleico, antes de generar la información de secuencia. La información de secuencia generada después de la amplificación puede comprender fragmentos de nucleótidos de longitud "n", de manera que una pluralidad de un número "i(n)" de fragmentos se compara con las secuencias de ácido nucleico en una base de datos.

Si la probabilidad de coincidencia de la pluralidad "i(n)" de información de secuencia es menor que un umbral de coincidencia objetivo, entonces se genera una pluralidad de información de secuencia "i(n+1)", "i(n+2)" ... "i(n+x)".

En una realización de la invención, el fragmento de nucleótido se compara con las secuencias de ácido nucleico en una base de datos mediante emparejamiento probabilístico, que incluye, entre otros, el enfoque bayesiano, el enfoque bayesiano recursivo o el enfoque bayesiano sin modificación.

Los enfoques probabilísticos pueden usar probabilidades bayesianas para considerar dos factores importantes para llegar a una conclusión precisa: (i) *P(I;/R)* es la probabilidad de que un organismo que exhibe un patrón de prueba R pertenezca a taxón *Ii*, y (ii) *P(RI/i)* es la probabilidad de que los miembros de taxón *ti* exhiban el patrón de prueba R. El patrón mínimo dentro de una ventana deslizante integrada en las herramientas ayudará a los investigadores a determinar "si" y "cómo" han sido modificados genéticamente los organismos.

En una realización de la invención, el emparejamiento probabilístico proporciona un marco estadístico jerárquico para identificar un genoma en la muestra.

En otra realización de la invención, la comparación de la información de secuencia se realiza, en tiempo real, o tan rápido como, o inmediatamente después de que se genere dicha información de secuencia.

La comparación de dicha información de secuencia se realiza, en tiempo real, o tan rápido como se genera la

información de secuencia, mientras que la información adicional de secuencia continúa generándose a partir de dichas una o más moléculas de ácido nucleico, en donde dicha información adicional de secuencia puede comprender nucleótidos de longitudes variables, que incluyen, pero no se limitan a, aumento, disminución o la misma longitud de información de secuencia en comparación con la información de secuencia generada previamente.

La información de secuencia incluye, pero no se limita a, un cromatograma, imagen de fragmentos de ADN o ARN marcados, interrogación física de una molécula de ácido nucleico para determinar el orden de los nucleótidos, análisis de nanoporos y otros métodos conocidos en la técnica que determinan la secuencia de una cadena de ácido nucleico.

En una realización de la invención, "x" se puede seleccionar entre 1-10, 10-20, 20-30, 30-40, 40-50, 50-60, 60-70, 70-80, 80-90 o 90-100 nucleótidos. En otra realización, "x" puede ser 100-200, 200-300, 300-400 o 400-500 nucleótidos.

Si la probabilidad de coincidencia de la información de secuencia del fragmento de nucleótido de longitud "n" es menor que un umbral de una coincidencia objetivo, entonces "n+x" representa una longitud mínima del fragmento de nucleótido para una identificación positiva de la molécula o moléculas de ácido nucleico obtenidas de dicha muestra.

La presente invención también se refiere a un sistema para identificar un genoma en una muestra que comprende una pluralidad de genomas, que comprende:

(i) una unidad receptora de muestras configurada para recibir la muestra;

(ii) una unidad de extracción en comunicación con la unidad receptora de muestras, estando configurada la unidad de extracción para extraer una o más moléculas de ácido nucleico de la muestra;

(iii) un casete de secuenciación en comunicación con la unidad de extracción, estando el casete de secuenciación configurado para recibir una o más moléculas de ácido nucleico de la unidad de extracción y generar la información de la secuencia de una o más moléculas de ácido nucleico;

(iv) una base de datos que comprende secuencias de referencia de ácido nucleico; y

(v) una unidad de procesamiento en comunicación con el casete de secuenciación y la base de datos, caracterizada porque la unidad de procesamiento está configurada para realizar las etapas (iv) - (vii) del método de la presente invención.

Las variaciones adicionales abarcadas dentro del sistema se describen en la descripción detallada de la invención a continuación.

Breve descripción de los dibujos

Se describen diversas realizaciones con referencia a los dibujos adjuntos. En los dibujos, los números de referencia similares indican componentes idénticos o funcionalmente similares.

La FIG. 1 es una ilustración esquemática de un sistema divulgado.

La FIG. 2 es una ilustración esquemática más detallada del sistema de la FIG. 1.

La FIG. 3 es una ilustración esquemática de la interacción funcional entre el casete intercambiable y otros componentes en una realización del sistema de la FIG. 1.

La FIG. 4 es una vista en perspectiva frontal de una realización de un dispositivo de secuenciación electrónico portátil.

La FIG. 5 es un diagrama de flujo que ilustra un proceso de operación del sistema de la FIG. 1.

La FIG. 6 es una ilustración esquemática de la interacción del sistema de la FIG. 1 con varias entidades potencialmente involucradas con el sistema.

La FIG. 7 es una ilustración esquemática de la interacción funcional entre un dispositivo electrónico portátil de secuenciación con el centro de análisis remoto.

La FIG. 8 es una ilustración esquemática de la arquitectura general del módulo de software probabilístico.

La FIG. 9 muestra el porcentaje de secuencias únicas en función de la longitud de lectura.

La FIG. 10 es un resumen de las etapas principales de la secuenciación.

Descripción detallada de la invención

Los métodos y el sistema descritos en la presente invención usan la información de secuencia única más corta, que en una mezcla de ácidos nucleicos en una muestra no caracterizada tiene la longitud mínima única (n) con respecto a la información de secuencia completa generada o recopilada. Además de las secuencias de longitud únicas, también se comparan las no únicas. La probabilidad de identificación de un genoma aumenta con múltiples coincidencias. Algunos genomas tendrán secuencias únicas mínimas más largas que otros genomas. El método de coincidencia de secuencias de longitud corta (n) continúa en paralelo con la generación o recopilación de información de secuencia. Las comparaciones se producen tan rápido como se generan o recopilan secuencias posteriores más largas (en tiempo real). Esto se traduce en una reducción considerable del espacio de decisión porque los cálculos se realizan de manera temprana en términos de generación/recopilación de información de secuencia. El emparejamiento probabilístico puede incluir, entre otros, emparejamiento perfecto, unicidad de subsecuencia, emparejamiento de patrones, emparejamiento de múltiples secuencias dentro de la longitud n, emparejamiento inexacto, siembra y extensión, mediciones de distancia y mapeo de árboles filogenéticos. Proporciona un canal automatizado para que coincida con la información de secuencia tan rápido como se genera o en tiempo real. El instrumento de secuenciación puede continuar recolectando más cadenas de información de secuencia y más largas en paralelo con la comparación. La información posterior de la secuencia también se puede comparar y puede aumentar la confianza de la identificación de un genoma o especie en la muestra. El método no necesita esperar a que el conjunto de información de secuencia de las lecturas cortas en cóntigos más grandes.

El sistema y los métodos divulgados en este documento proporcionan la entrada, aislamiento y separación de ácido nucleico, la secuenciación de ADN, la creación de redes de bases de datos, el procesamiento de información, el almacenamiento de datos, la visualización de datos y la comunicación electrónica para acelerar el suministro de datos relevantes para permitir el diagnóstico o la identificación de organismos con aplicaciones de brote patógeno y respuestas adecuadas. El sistema incluye un dispositivo de secuenciación portátil que transmite electrónicamente los datos a una base de datos para la identificación de organismos relacionados con la determinación de la secuencia de los ácidos nucleicos y otras moléculas poliméricas o de tipo cadena y la comparación probabilística de datos.

Las Figuras 1 y 2 ilustran una realización de un sistema 100 que incluye un dispositivo 105 manual portátil de secuenciación electrónica. El dispositivo 105 portátil de secuenciación electrónica (denominado en este documento "dispositivo de secuenciación") está configurado para ser fácilmente mantenido y utilizado por un usuario (U), y puede comunicarse a través de una red 110 de comunicaciones con muchas otras entidades potencialmente relevantes.

El dispositivo está configurado para recibir una muestra de un sujeto (SS) y una muestra del entorno (ES), respectivamente. La muestra del sujeto (como sangre, saliva, etc.) puede incluir el ADN del sujeto así como el ADN de cualquier organismo (patógeno o no) en el sujeto. La muestra del entorno (ES) puede incluir, entre otros, organismos en su estado natural en el medio ambiente (incluidos alimentos, aire, agua, suelo, tejido). Ambas muestras (SS, ES) pueden verse afectadas por un acto de bioterrorismo o por una epidemia emergente. Ambas muestras (SS, ES) se recolectan simultáneamente a través de un tubo o hisopo y se reciben en una solución o un sólido (tal como una perla) en una membrana o portaobjetos, placa, capilar o canal. Las muestras (SS, ES) se secuencian simultáneamente. Las circunstancias de situaciones específicas pueden requerir el análisis de una muestra compuesta de una mezcla de las muestras (SS, ES). Se puede contactar a la primera persona en responder una vez que se identifica una coincidencia probabilística y/o durante la recopilación de datos en tiempo real y la interpretación de los datos. A medida que avanza el tiempo se puede identificar un porcentaje creciente de la secuencia.

El dispositivo 105 de secuenciación puede incluir los siguientes componentes funcionales, como se ilustra en la FIG. 3, que permite que el dispositivo 105 analice una muestra del sujeto (SS) y una muestra del entorno (ES), comunica el análisis resultante a una red 110 de comunicaciones.

Los receptores 120 y 122 de muestras se acoplan a un bloque 130 de extracción y aislamiento de ADN, que luego entrega las muestras al bloque 130 a través de un sistema de flujo. El bloque 130 extrae el ADN de las muestras y lo aísla para que pueda procesarse y analizarse adicionalmente. Esto se puede lograr mediante el uso de una plantilla de reactivo (es decir, una cadena de ADN que sirve como un patrón para la síntesis de una cadena complementaria de ácido nucleico), que puede administrarse combinada con las muestras 120, 122 utilizando una tecnología conocida de transporte por fluidos. Los ácidos nucleicos en las muestras 120, 122 están separados por el bloque 130 de extracción y aislamiento, produciendo una corriente de fragmentos de nucleótidos o moléculas individuales sin amplificar. Una realización podría incluir el uso de métodos de amplificación.

Un casete 140 intercambiable puede estar acoplado de manera removible al dispositivo 105 de secuenciación y al bloque 130. El casete 140 puede recibir la corriente de moléculas del bloque 130 y puede secuenciar el ADN y producir datos de secuencia de ADN.

El casete 140 intercambiable se puede acoplar y proporcionar los datos de la secuencia de ADN al procesador 160, donde se lleva a cabo el emparejamiento probabilístico. Una realización podría incluir el rendimiento de 16 GB de datos transferidos a una velocidad de 1 Mb/s. Se prefiere un casete 140 de secuenciación para obtener la información de secuencia. Pueden intercambiarse diferentes casetes que representan diferentes métodos de

secuenciación. La información de la secuencia se compara mediante un emparejamiento probabilístico. Los algoritmos de coincidencia ultrarrápidos y las bases de datos de firmas ponderadas generadas previamente comparan de nuevo los datos de secuencia con los datos de secuencia almacenados.

El procesador 160 puede ser, por ejemplo, un circuito integrado específico de la aplicación diseñado para lograr una o más funciones específicas o habilitar uno o más dispositivos o aplicaciones específicas. El procesador 160 puede controlar todos los otros elementos funcionales del dispositivo 105 de secuenciación. Por ejemplo, el procesador 160 puede enviar/recibir los datos de secuencia de ADN que se almacenarán en un almacén 170 de datos (memoria). El almacén 170 de datos también puede incluir cualquier tipo o formas adecuadas de memoria para almacenar datos en una forma recuperable por el procesador 160.

El dispositivo 105 de secuenciación puede incluir además un componente 180 de comunicación al cual el procesador 160 puede enviar datos recuperados del almacén 170 de datos. El componente 180 de comunicación puede incluir cualquier tecnología adecuada para comunicarse con la red 110 de comunicaciones, tal como por cable, inalámbrica, satelital, etc.

El dispositivo 105 de secuenciación puede incluir un módulo 150 de entrada de usuario, al que el usuario (U) puede proporcionarle entrada al dispositivo 105. Esto puede incluir cualquier tecnología de entrada adecuada, tal como botones, teclado táctil, etc. Finalmente, el dispositivo 105 de secuenciación puede incluir un módulo 152 de salida de usuario que puede incluir una pantalla para la salida visual y/o un dispositivo de salida de audio.

El dispositivo 105 de secuenciación también puede incluir un receptor 102 del sistema de posicionamiento global (GPS), que puede recibir datos de posicionamiento y pasar los datos al procesador 160, y una fuente 104 de alimentación (es decir, batería, adaptador que puede ser enchufado) para suministro de energía eléctrica u otros tipos de energía a una carga de salida o grupo de cargas del dispositivo 105 de secuenciación.

El casete 140 intercambiable se ilustra esquemáticamente con más detalle en la FIG. 3. El casete 140 se puede acoplar de manera extraíble al dispositivo 105 de secuenciación y al bloque 130 e incluye un método de secuenciación del estado de la técnica (es decir, secuenciación de alto rendimiento). El sistema basado en la química húmeda o en estado sólido se puede construir en la plataforma a través de un modo de "conectar y usar" intercambiable de casete. El casete 140 puede recibir la corriente de moléculas del bloque 130 y puede secuenciar el ADN a través del método de secuenciación y puede producir datos de secuencia de ADN. Las realizaciones incluyen métodos basados en, pero no limitados a, secuenciación por síntesis, secuenciación por ligación, secuenciación de moléculas individuales y pirosecuenciación. Otra realización más incluye una fuente 142 de campo eléctrico y aplica el campo 142 eléctrico a la corriente de moléculas para efectuar la electroforesis del ADN dentro de la corriente. El casete incluye una fuente 144 de luz para emitir una luz 144 fluorescente a través de la corriente de ADN. El casete incluye además un sensor 146 biomédico (detector) para detectar la emisión de luz fluorescente y para detectar/determinar la secuencia de ADN de la corriente de muestra. Además de la luz fluorescente, el sensor biomédico es capaz de detectar la luz en todas las longitudes de onda apropiadas para los restos marcados para la secuenciación.

La detección fluorescente comprende la medición de la señal de un resto marcado de al menos uno de los uno o más nucleótidos o análogos de nucleótidos. La secuenciación que utiliza nucleótidos fluorescentes generalmente implica el fotoblanqueo del marcador fluorescente después de detectar un nucleótido agregado. Las realizaciones pueden incluir métodos fluorescentes basados en perlas, FRET, marcadores infrarrojos, pirofosfatasa, métodos de ligasa que incluyen nucleótidos marcados o polimerasa o el uso de terminadores reversibles cíclicos. Las realizaciones pueden incluir métodos directos de nanoporos o guías de ondas ópticas que incluyen moléculas individuales inmovilizadas o en solución. Los métodos de fotoblanqueo incluyen una intensidad de señal reducida, que se acumula con cada adición de un nucleótido marcado de forma fluorescente a la cadena del cebador. Al reducir la intensidad de la señal, las plantillas de ADN más largas se secuencian opcionalmente.

El fotoblanqueo incluye aplicar un pulso de luz al cebador de ácido nucleico en el que se ha incorporado un nucleótido fluorescente. El pulso de luz típicamente comprende una longitud de onda igual a la longitud de onda de la luz absorbida por el nucleótido fluorescente de interés. El pulso se aplica durante aproximadamente 50 segundos o menos, aproximadamente 20 segundos o menos, aproximadamente 10 segundos o menos, aproximadamente 5 segundos o menos, aproximadamente 2 segundos o menos, aproximadamente 1 segundo o menos, o aproximadamente 0. El pulso destruye la fluorescencia de los nucleótidos marcados con fluorescencia y/o el cebador o ácido nucleico marcados con fluorescencia, o lo reduce a un nivel aceptable, por ejemplo, un nivel de fondo, o un nivel lo suficientemente bajo como para evitar la acumulación de señales durante varios ciclos.

El sensor (detector) 146 controla opcionalmente al menos una señal de la plantilla de ácido nucleico. El sensor (detector) 146 opcionalmente incluye o está enlazado operativamente a un ordenador que incluye un software para convertir la información de la señal del detector en información de resultados de secuenciación, por ejemplo, la concentración de un nucleótido, la identidad de un nucleótido, la secuencia del nucleótido de la plantilla, etc. Además, las señales de la muestra se calibran opcionalmente, por ejemplo, calibrando el sistema de microfluidos controlando una señal de una fuente conocida.

Como se muestra en la FIG. 2, el dispositivo 105 de secuenciación puede comunicarse a través de una red 110 de comunicaciones con una variedad de entidades que pueden ser relevantes para notificar en caso de un acto de bioterrorismo o un brote epidémico. Estas entidades pueden incluir un Primer Respondedor (es decir, Red de Respuesta de Laboratorios (es decir, Laboratorios de Referencia, Laboratorios Regionales, Laboratorios Nacionales), GenBank®, Centro para el Control de Enfermedades (CDC), médicos, personal de salud pública, registros médicos, datos del censo, cumplimiento de la ley, fabricantes de alimentos, distribuidores de alimentos y minoristas de alimentos.

Una realización de ejemplo del dispositivo 105 de secuenciación discutido anteriormente se describe ahora con referencia a la FIG. 4 que ilustra una vista anterior del dispositivo. El dispositivo es un dispositivo manual portátil de secuenciación y se ilustra en comparación con el tamaño de las monedas C. El dispositivo 105 tiene aproximadamente 11 pulgadas de largo y es fácilmente transportable. (En la Figura 4, se muestran las monedas para referencia de la escala). Dos puertos 153, 154 están ubicados en un lado del dispositivo y representan receptores 120, 122 de la muestra. El puerto 153 es para recibir una muestra del sujeto (SS) o una muestra del entorno (ES) para ser analizado y secuenciado. El puerto 154 es para control de la secuenciación (SC). Los dos puertos diferentes están diseñados para determinar si una muestra del sujeto (SS) o una muestra del entorno (ES) contiene materiales que dan como resultado una falla de la secuenciación, en caso de que se produzca una falla de la secuenciación, o funcione en una capacidad CLIA. El dispositivo 105 incluye un módulo 150 de entrada de usuario, que el usuario (U) puede proporcionar entrada al dispositivo 105. En esta realización particular, el módulo 150 de entrada de usuario tiene la forma de un panel táctil, sin embargo, se puede usar cualquier tecnología adecuada. El panel táctil incluye botones 150a para visualización, 150b, 150c para grabar datos, 150d para transmisión y recepción de datos en tiempo real, y 150e para control de la energía para activar o desactivar el dispositivo. Alternativamente, el teclado puede incorporarse a la pantalla y todas las funciones pueden controlarse mediante la interfaz de cristal líquido. Las técnicas adecuadas se describen en la publicación de la solicitud de patente de Estados Unidos No. 2007/0263163, cuya divulgación completa se incorpora a la presente memoria por referencia. Esto puede ser mediante el emparejamiento de dispositivos habilitados para Bluetooth o enfoques similares. Las funciones incluyen teclados digitales, marcados con las letras del alfabeto, tal como el lugar común en los teclados del teléfono, como una tecla de borrar, tecla de espacio, tecla de salida, tecla de imprimir, tecla de entrada, desplazamiento hacia arriba/abajo, izquierda/derecha, caracteres adicionales y cualquier otra deseada por el usuario. El dispositivo incluye además un módulo 152 de salida del usuario, en forma de una pantalla visual, para mostrar información para el usuario (U). También se puede proporcionar un dispositivo de salida de audio si se desea, como se ilustra en 157a y 157b. Finalmente, el dispositivo 105 de secuenciación incluye diodos 155 y 156 emisores de luz para indicar la transmisión o recepción de datos. La función de las teclas/botones es controlar todos los aspectos de la secuenciación de muestras, la transmisión de datos y la coincidencia probabilística y los controles de la interfaz, incluidos, entre otros, funciones de encendido/apagado, envío, tecla de navegación, teclas de función, borrado y pantalla LCD, y herramientas de visualización con rango de genoma calculado por algoritmos para enumerar el nivel de confianza de las coincidencias. Una realización incluye un sistema basado en Internet en el que múltiples usuarios pueden transmitir/recibir datos de forma simultánea a/desde un motor de búsqueda de red jerárquico.

La FIG. 5 es un diagrama de flujo que ilustra un proceso de operación del sistema 100 de una realización del sistema 100 como se describió anteriormente. Como se muestra en la FIG. 5, un proceso de operación del dispositivo incluye en 200 las muestras de sujetos (SS) y muestras ambientales (ES) recogidas en los receptores de muestra 120, 122. En 202, las muestras proceden al bloque 130 de extracción y aislamiento de ADN donde se analiza la muestra y se extrae el ADN de las muestras y se aísla. En 203, el casete 140 intercambiable recibe el ADN aislado del bloque 130 y secuencia el ADN. Dependiendo del casete y si es necesario, con la aplicación de un campo 142 eléctrico y de una luz 144 fluorescente, un sensor 146 biomédico dentro del casete 140 detecta/determina la secuencia de ADN de la corriente de muestra. En 204, los datos secuenciados se procesan y almacenan en un dispositivo 170 de almacenamiento de datos. En 205, los datos secuenciados se comparan mediante un emparejamiento probabilístico y se logra la identificación del genoma. El proceso es de naturaleza reiterativa. La información resultante puede transmitirse a través de una red 110 de comunicaciones. Los datos de GPS (sistema de posicionamiento global) también pueden transmitirse opcionalmente en la etapa 205. En 206, el dispositivo recibe electrónicamente datos de las coincidencias. En 207, el dispositivo despliega visualmente los datos recibidos electrónicamente del emparejamiento a través de un módulo 152 de salida de usuario. Si se requiere un análisis adicional, en 208, los datos secuenciados se transmiten electrónicamente a entidades de interpretación de datos (es decir, personal de salud pública, registros médicos, etc.) a través de la red de comunicación.

Un enfoque de investigación de métodos múltiples puede mejorar la respuesta rápida a un incidente e integrar la atención primaria con la detección de organismos. Se puede utilizar una respuesta triangulada, que involucra datos cuantitativos de instrumentos de la secuenciación del ADN para que converja con la atención crítica cualitativa. Puede usarse una infraestructura de listados de verificación de las observaciones y auditorías de datos de secuenciación de ADN recopilados en el campo a través de múltiples ubicaciones para comparar la apariencia de un organismo, por ejemplo, una amenaza biológica entre las ubicaciones. El análisis estadístico inferencial de los datos genómicos puede combinarse con observaciones médicas para desarrollar categorías de prioridades. La información recopilada y compartida entre las bases de datos de centros médicos y centros genómicos puede permitir la triangulación de un incidente, la magnitud del incidente y la entrega de la intervención correcta a las personas

afectadas en el momento adecuado.

La FIG. 6 ilustra la interacción entre el sistema 100 y varias entidades de recursos potenciales. El dispositivo 105 está configurado para interactuar con estas entidades de recursos a través de una red de comunicación inalámbrica o por cable. El dispositivo 105 puede transmitir información de datos de secuenciación triangulados (310) que ilustra los "Datos de la Muestra", los "Datos del Paciente" y la "Intervención de Tratamiento". El dispositivo 105 puede transmitir y recibir datos de secuencia de ADN hacia y desde los recursos 320 de coincidencia de secuencia, que incluyen al GenBank® y una red de respuesta de laboratorios que incluye Laboratorios de Observación, Laboratorios de Referencia y Laboratorios Nacionales.

Cada uno de los laboratorios tiene funciones específicas. Los laboratorios de observación (hospitales y otros laboratorios clínicos comunitarios) son responsables de descartar o remitir a los agentes críticos que encuentren a los laboratorios de referencia LRN más cercanos. Los laboratorios de referencia (laboratorios estatales y locales de salud pública donde se observan las prácticas del Nivel de Seguridad Biológica 3 (BSL-3)) realizan pruebas de confirmación (confirman). Los laboratorios nacionales (BSL-4) mantienen una estructura capaz de manejar agentes virales tales como la Ébola y la Variola mayor y realizan una caracterización definitiva.

El sistema 100 puede además transmitir y recibir datos desde y hacia los Recursos de Interpretación de Datos 330, incluidas las entidades encargadas de hacer cumplir la ley, el personal de salud pública, los registros médicos y los datos del censo. Finalmente, el dispositivo 105 puede transmitir y recibir datos hacia y desde un primer nivel 320 de respuesta, que incluyen los doctores o los médicos en una sala de emergencias. El sistema 100 total está configurado para comunicarse con el Centro de Control de Enfermedades (CDC) 340 para proporcionar la información pertinente al personal apropiado.

La FIG. 7 es una ilustración esquemática de la interacción funcional entre un dispositivo electrónico portátil de secuenciación con el centro de análisis remoto. El dispositivo 105 puede incluir una unidad 103 de llamada base para procesar la secuenciación recibida por el casete 140 intercambiable. Dichas secuencias y sitios de SNP se ponderan individualmente de acuerdo con su probabilidad encontrada en cada especie. Estas ponderaciones pueden calcularse teóricamente (por simulación) o experimentalmente. El dispositivo también incluye un procesador 109 de coincidencia probabilística acoplado a la unidad 103 de llamada base. La coincidencia probabilística se realiza en tiempo real o tan rápido como la secuencia de llamada base o la recopilación de datos de la secuencia. El procesador 109 de coincidencia probabilística, que utiliza un enfoque bayesiano, puede recibir la secuencia resultante y los datos de calidad, y puede calcular las probabilidades para cada lectura de secuenciación mientras considera las puntuaciones de calidad de la secuenciación generadas por la unidad 103 de llamada base. El procesador 109 de coincidencia probabilística puede usar una base de datos generada y optimizada antes de su uso para la identificación de patógenos. Un sistema 107 de alerta está acoplado al procesador 109 de coincidencia probabilística y puede recopilar información del procesador 109 de coincidencia probabilística (en el sitio) y mostrar el organismo u organismos mejor emparejados en tiempo real.

El sistema 107 de alerta está configurado para acceder a los datos del paciente, es decir, el diagnóstico médico o la evaluación de riesgos para un paciente, en particular los datos de las pruebas o ensayos de diagnóstico en el punto de atención, incluidos los inmunoensayos, electrocardiogramas, rayos X y otras pruebas similares, y proporciona una indicación de una condición médica o riesgo o ausencia de los mismos. El sistema de alerta puede incluir software y tecnologías para la lectura o la evaluación de los datos de prueba y para convertir los datos en información de diagnóstico o evaluación de riesgos. Dependiendo de la identidad del genoma del agente biológico y los datos médicos sobre el paciente, se puede administrar una "intervención de tratamiento" eficaz. El tratamiento puede basarse en la mitigación o neutralización efectiva del agente biológico y/o sus efectos secundarios y en la historia del paciente si existen contraindicaciones. El sistema de alerta puede basarse en el grado y número de ocurrencias. El número de ocurrencias puede basarse en la identificación genómica del agente biológico. Se puede pronunciar un valor cuando el resultado está dentro o sobrepasa un umbral según lo determinan las agencias gubernamentales, como el CDC o DoD o la Seguridad Nacional. El sistema de alerta está configurado para permitir a los médicos utilizar la funcionalidad de los datos de identificación genómica con los datos del paciente. La comunicación permite un flujo rápido de información y una toma de decisiones precisa para las acciones de los primeros respondedores u otros sistemas clínicos.

El dispositivo 105 incluye además un compresor 106 de datos acoplado a la unidad 103 de llamada base, configurado para recibir la secuencia resultante y los datos de calidad para la compresión. El dispositivo de almacenamiento 170 de datos está acoplado al compresor 106 y puede recibir y almacenar la secuencia y los datos de calidad.

El dispositivo 105 de secuenciación interactúa con un centro 400 de análisis remoto, que puede recibir datos transferidos electrónicamente desde el componente 180 de comunicación del dispositivo 105 de secuenciación a través de un método de comunicación cableado y/o inalámbrico. El centro 400 de análisis remoto contiene una gran base de datos de secuencias que incluye todas las secuencias de nucleótidos y aminoácidos y los datos de SNP disponibles hasta la fecha. Esta base de datos también contiene información epidemiológica y terapéutica asociada (por ejemplo, resistencia a los antibióticos). El centro 400 de análisis remoto incluye además un dispositivo 401 de almacenamiento de datos. El dispositivo 401 de almacenamiento de datos puede recibir información de datos de

secuencia descomprimida a través de la transmisión electrónica desde el componente 180 de comunicación del dispositivo 105 de secuenciación. Un montaje 402 del genoma está acoplado al dispositivo 401 de almacenamiento de datos y puede ensamblar los datos de la secuencia descomprimida. El ADN contaminante obvio, tal como el ADN humano, puede filtrarse antes de un análisis adicional.

El centro 400 de análisis remoto incluye además un procesador 403 equipado con tecnología de emparejamiento probabilístico y algoritmos de búsqueda de homología, que pueden emplearse para analizar datos de secuencia ensamblados para obtener las probabilidades de la presencia de patógenos 403a objetivo, estructura 403b comunitaria, epidemiológica e información 403c terapéutica. Los datos de la secuencia del genoma de los patógenos objetivo se comparan con los de los genomas de los no patógenos, incluidos los humanos y el metagenoma, para identificar secuencias de nucleótidos y sitios polimorfos de un solo nucleótido (SNP), que solo aparecen en los organismos objetivo. El análisis en el centro 400 de análisis remoto se realiza sobre la marcha durante la transferencia de datos desde el dispositivo 105 de secuenciación. El centro 400 de análisis remoto puede incluir además una unidad 404 de comunicación desde la cual los resultados del análisis se transfieren electrónicamente al sistema 107 de alerta dentro de el dispositivo 105 de secuenciación, así como otras autoridades (por ejemplo, DHS, CDC, etc.).

Clasificación probabilística: La presente invención proporciona motores de base de datos, diseño de bases de datos, técnicas de filtrado y el uso de la teoría de la probabilidad como lógica extendida. Los presentes métodos y sistemas utilizan los principios de la teoría de probabilidades para hacer razonamientos plausibles (decisiones) sobre los datos producidos por la secuenciación de ácidos nucleicos. Usando el enfoque de la teoría de la probabilidad, el sistema descrito aquí analiza los datos tan pronto como alcanza un número mínimo de nucleótidos en longitud (n), y calcula la probabilidad de la mer n, además cada aumento posterior en la longitud (n + par o pares de bases) se utiliza para calcular la probabilidad de una coincidencia de secuencia. El cálculo de cada mer n y los mer n posteriores más largos se procesa adicionalmente para recalcular las probabilidades de todas las longitudes crecientes para identificar la presencia de un genoma o genomas. A medida que aumenta la longitud de la unidad, se comparan múltiples subunidades dentro del mer n para el reconocimiento de patrones, lo que aumenta aún más la probabilidad de una coincidencia. Dicho método, incluidos otros métodos bayesianos, permite eliminar coincidencias e identificar un número significativo de muestras biológicas que comprenden un fragmento de nucleótido muy corto o leer sin tener que completar la secuenciación completa del genoma o ensamblar el genoma. Como tal, la asignación de la probabilidad de coincidencia con los organismos existentes y se mueve a la siguiente secuencia lectura de secuencia de ácido nucleico para mejorar aún más la probabilidad de coincidencia. El sistema descrito aquí aumenta la velocidad, reduce el consumo de reactivos, permite la miniaturización y reduce significativamente la cantidad de tiempo requerido para identificar el organismo.

Con el fin de construir clasificadores probabilísticos para tomar una decisión sobre secuencias cortas de ácido nucleico, se pueden utilizar una variedad de enfoques para filtrar primero y clasificar posteriormente los datos de secuenciación entrantes. En el presente caso, se utiliza el formalismo de las redes bayesianas. Una red bayesiana es un gráfico acíclico dirigido que representa de manera compacta una distribución de probabilidad. En dicho gráfico, cada variable aleatoria se denota por un nodo (por ejemplo, en un árbol filogenético de un organismo). Un borde dirigido entre dos nodos indica una dependencia probabilística de la variable denotada por el nodo principal con la del secundario. En consecuencia, la estructura de la red denota la suposición de que cada nodo en la red es condicionalmente independiente de sus no descendientes dados sus padres. Para describir una distribución de probabilidad que satisfaga estos supuestos, cada nodo en la red está asociado con una tabla de probabilidad condicional, que especifica la distribución sobre cualquier posible asignación de valores a sus padres. En este caso, un clasificador bayesiano es una red bayesiana aplicada a una tarea de clasificación para calcular la probabilidad de cada nucleótido proporcionado por cualquier sistema de secuenciación. En cada punto de decisión, el clasificador bayesiano se puede combinar con una versión del algoritmo de gráfico de ruta más corta, tal como el de Dijkstra o el de Floyd.

El sistema actual puede implementar un sistema de clasificadores bayesianos (por ejemplo, clasificador bayesiano no modificado, clasificador bayesiano y clasificador de estimación bayesiana recursiva) y fusionar los datos resultantes en la base de datos de decisiones. Una vez fusionados los datos, cada clasificador puede alimentar un nuevo conjunto de resultados con probabilidades actualizadas.

La FIG. 8 muestra una ilustración esquemática de la arquitectura general del módulo de software probabilístico.

Fragmento de secuenciación de ADN: se puede usar cualquier método de secuenciación para generar la información del fragmento de secuencia. El módulo 160 en la FIG. 2 o 109 en la FIG. 7 es responsable de procesar los datos que ingresan desde el módulo de secuenciación en el casete intercambiable. Los datos se encapsulan con los datos de secuenciación, así como con la información anterior sobre el inicio y el final de la secuencia, ID de la secuencia, ID de la cadena de ADN. El módulo formatea los datos y los pasa al módulo de filtro de la taxonomía. El formato incluye la adición de los datos del sistema y la alineación en trozos.

El módulo de secuenciación del ADN tiene 2 interfaces. Está conectado al módulo de preparación de ADN y al filtro de taxonomía.

I. Interfaz de preparación del ADN: varios métodos disponibles comercialmente para lograr la preparación de muestras se pueden integrar mediante técnicas de microfluidos. La preparación típica de la muestra se basa en la solución e incluye la lisis celular y la eliminación de inhibidores. Los ácidos nucleicos se recuperan o extraen y se concentran. Las realizaciones de la lisis incluyen detergente/enzimas, métodos mecánicos, microondas, presión y/o ultrasónicos. Las realizaciones de extracción incluyen afinidad de fase sólida y/o exclusión de tamaño.

II. Filtro de taxonomía: el filtro de taxonomía tiene dos tareas principales: (i) filtrar tantos organismos como sea posible para limitar el módulo clasificador a un espacio de decisión más pequeño, y (ii) ayudar a determinar la estructura de la red bayesiana, que implica el uso de técnicas de aprendizaje de máquina.

Filtro del árbol filogenético: este submódulo de filtro de taxonomía interactúa con las "bases de datos de decisiones" para conocer los resultados de la ronda previa del análisis. Si no se encuentran resultados, el módulo pasa los nuevos datos al módulo de clasificación. Si se encuentran los resultados, el filtro de taxonomía ajusta los datos del clasificador para limitar el posible espacio de decisión. Por ejemplo, si los datos anteriores indican que esta es una secuencia de ADN del virus que se está analizando, el espacio de decisión para el clasificador se reducirá a los datos virales únicamente. Esto se puede hacer modificando los datos de los clasificadores bayesianos recopilados durante el funcionamiento.

Aprendizaje de máquina: los algoritmos de aprendizaje de máquina se organizan en una taxonomía, en función del resultado deseado del algoritmo. (i) Aprendizaje supervisado, en el que el algoritmo genera una función que asigna entradas a las salidas deseadas. Una formulación estándar de la tarea de aprendizaje supervisado es el problema de clasificación: se requiere que el alumno aprenda (para aproximar) el comportamiento de una función que asigna un vector $[X_1, X_2, ... X_N]$ en una de varias clases observando varios ejemplos de entrada-salida de la función. (ii) Aprendizaje semi-supervisado, que combina ejemplos etiquetados y no etiquetados para generar una función o clasificador apropiado. (iii) Aprendizaje de refuerzo, en el que el algoritmo aprende una política de cómo actuar dada una observación del mundo. Cada acción tiene algún impacto en el entorno, y el entorno proporciona retroalimentación que guía el algoritmo de aprendizaje. (iv) Transducción, predice nuevos resultados en función de las entradas de entrenamiento, las salidas de entrenamiento y las entradas de prueba que están disponibles durante el entrenamiento. (v) Aprendizaje para aprender, en el que el algoritmo aprende su propio sesgo inductivo basado en la experiencia anterior.

Módulo caché de taxonomía: El módulo caché almacena la información de taxonomía producida por el filtro de taxonomía. Puede actuar como una interfaz entre el filtro de taxonomía y la base de datos de taxonomía que contiene toda la información en la base de datos SQL. El caché de taxonomía se implementa como una base de datos en memoria con tiempo de respuesta de microsegundos. Las consultas a la base de datos SQL se manejan en un subproceso separado del resto del submódulo. La información de caché incluye el gráfico de red creado por el módulo de filtro de taxonomía. El gráfico contiene toda la taxonomía a medida que el sistema comienza el análisis. El análisis de secuencia de ADN reduce el gráfico de taxonomía con el caché de taxonomía implementando las reducciones en el tamaño de los datos y la eliminación de los conjuntos de datos apropiados.

Selector clasificador: el presente sistema puede utilizar múltiples técnicas de clasificación que se ejecutan en paralelo. El selector clasificador puede actuar como árbitro de datos entre diferentes algoritmos de clasificación. El selector clasificador puede leer información de la base de datos de decisiones y enviar dicha información a los módulos de clasificación con cada unidad de secuenciación de ADN recibida para su análisis desde el Módulo de secuenciación de ADN. El filtro de taxonomía actúa como paso de datos para los datos de secuenciación de ADN.

Clasificador bayesiano recursivo: El clasificador bayesiano recursivo es un enfoque probabilístico para estimar una función de densidad de probabilidad desconocida recursivamente a lo largo del tiempo utilizando mediciones entrantes y un modelo de proceso matemático. El módulo recibe datos del selector clasificador y de la base de datos de decisiones donde se almacenan las decisiones anteriores. El conjunto de datos se recupera de las bases de datos y la identificación de la decisión anterior se coloca en la memoria local del módulo donde se produce el filtrado. El clasificador toma la secuencia de ADN e intenta asociarla con o sin las firmas, códigos de barras, existentes, de las base de datos de taxonomía, filtrando rápidamente las familias de organismos que no coinciden. El algoritmo funciona calculando las probabilidades de creencias múltiples y ajustando creencias basadas en los datos entrantes. Los algoritmos utilizados en este módulo pueden incluir métodos secuenciales Monte Carlo y muestreos de importancia de nuevo muestreo. El modelo oculto de Markov, el filtro de ensamble Kalman y otros filtros de partículas también pueden usarse junto con la técnica de actualización bayesiana.

Clasificador bayesiano sin modificar: clasificador probabilístico simple basado en la aplicación del teorema de Bayes. El clasificador toma todas las decisiones basadas en el conjunto de reglas predeterminado que se proporciona como entrada del usuario en el inicio. El módulo se puede reinicializar con un nuevo conjunto de reglas mientras se está ejecutando el análisis. Las nuevas reglas pueden provenir del usuario o pueden ser un producto de la fusión de reglas del módulo de fusiones de resultados.

Clasificador de red bayesiana: el clasificador de red bayesiana implementa una red bayesiana (o una red de creencias) como un modelo gráfico probabilístico que representa un conjunto de variables y sus independencias probabilísticas.

Base de datos de decisiones: La base de datos de decisiones es un caché de trabajo para la mayoría de los módulos en el sistema. La mayoría de los módulos tienen acceso directo a este recurso y pueden modificar sus regiones individuales. Sin embargo, solo el módulo de fusión de resultados puede acceder a todos los datos y modificar los conjuntos de reglas bayesianas en consecuencia.

Datos de reglas bayesianas: el módulo recopila todas las reglas bayesianas en forma binaria compilada previamente. Las reglas son de lectura y escritura para todos los clasificadores bayesianos, así como para los módulos de filtro de taxonomía y fusiones de resultados. Las reglas se compilan nuevamente en forma dinámica a medida que se realizan los cambios.

Fusión de resultados: el módulo fusiona la fecha de múltiples clasificadores bayesianos así como otros clasificadores estadísticos que se utilizan. El módulo de fusion de resultados analiza la varianza media entre las respuestas generadas para cada clasificador y fusiona los datos si es necesario.

Interfaz de base de datos: Interfaz con la base de datos SQL. La interfaz se implementa mediante programación con funciones de lectura y escritura separadas en diferentes subprocesos. MySQL es la base de datos de elección, sin embargo, puede utilizarse sqLite para acelerar la velocidad de la base de datos.

Base de datos de taxonomía: La base de datos tendrá varias bases de datos internas: árbol de taxonomía, árbol indexado previamente procesado, entrada de usuario y reglas.

Reglas almacenadas en caché: caché en memoria de las reglas procesadas posteriormente proporcionadas por el usuario.

Gestión de reglas: interfaz de gestión gráfica para el módulo

Entrada de usuario: reglas de inferencia creadas por el usuario. Los clasificadores bayesianos utilizan las reglas para tomar decisiones.

El sistema y el método de la invención se describen en el presente documento como incorporados en programas informáticos que tienen un código para realizar una variedad de funciones diferentes. Las mejores tecnologías de clase (actuales o emergentes) pueden ser componentes con licencia. Los métodos existentes para la extracción de ADN incluyen el uso de fenol/cloroformo, separación por saturación salina, el uso de sales caotrópicas y resinas de sílice, el uso de resinas de afinidad, cromatografía de intercambio iónico y el uso de perlas magnéticas. Los métodos se describen en las patentes de Estados Unidos Nos. 5.057.426, 4.923.978, las patentes EP 0512767 A1 y EP 0515484B y los documentos WO 95/13368, WO 97/1033I y WO 96/1873I. Debe entenderse, sin embargo, que los sistemas y métodos no se limitan a un medio electrónico, y varias funciones pueden practicarse alternativamente en un ajuste manual. Los datos asociados con el proceso pueden transmitirse electrónicamente a través de una conexión de red a través de Internet. Los sistemas y técnicas descritos anteriormente pueden ser útiles en muchos otros contextos, incluidos los descritos a continuación.

Estudios de asociación de enfermedades: muchas enfermedades y afecciones comunes implican factores genéticos complejos que interactúan para producir las características visibles de esa enfermedad, también llamada fenotipo. Múltiples genes y regiones reguladoras a menudo se asocian con una enfermedad o síntoma particular. Al secuenciar los genomas o los genes seleccionados de muchos individuos con una condición dada, puede ser posible identificar las mutaciones causales subyacentes a la enfermedad. Esta investigación puede conducir a avances en la detección, prevención y tratamiento de enfermedades.

Investigación del cáncer: la genética del cáncer implica comprender los efectos de mutaciones heredadas y adquiridas y otras alteraciones genéticas. El desafío de diagnosticar y tratar el cáncer se ve agravado por la variabilidad individual del paciente y las respuestas difíciles de predecir al tratamiento farmacológico. La disponibilidad de la secuenciación del genoma a bajo costo para caracterizar los cambios adquiridos del genoma que contribuyen al cáncer a partir de pequeñas muestras o biopsias de células tumorales, puede permitir un mejor diagnóstico y tratamiento del cáncer.

Investigación y desarrollo farmacéutico: una promesa de la genómica ha sido acelerar el descubrimiento y el desarrollo de nuevos fármacos más efectivos. El impacto de la genómica en esta área ha emergido lentamente debido a la complejidad de las vías biológicas, los mecanismos de enfermedad y los múltiples objetivos farmacológicos. La secuenciación de una sola molécula podría permitir la detección de alto rendimiento de una manera rentable utilizando el análisis de expresión génica a gran escala para identificar mejor los prospectos prometedores de los medicamentos. En el desarrollo clínico, la tecnología descrita podría potencialmente usarse para generar perfiles de genes individuales que pueden proporcionar información valiosa sobre la posible respuesta a la terapia, toxicología o riesgo de eventos adversos, y posiblemente para facilitar la selección del paciente y la individualización de la terapia.

Enfermedad infecciosa: todos los virus, bacterias y hongos contienen ADN o ARN. La detección y secuenciación de ADN o ARN de patógenos a nivel de una sola molécula podría proporcionar información útil desde el punto de vista médico y ambiental para el diagnóstico, tratamiento y control de infecciones y para predecir la posible resistencia a

los medicamentos.

Afecciones autoinmunes: se cree que varias afecciones autoinmunes, que van desde la esclerosis múltiple y el lupus hasta el riesgo de rechazo de un trasplante, tienen un componente genético. El monitoreo de los cambios genéticos asociados con estas enfermedades puede permitir un mejor manejo del paciente.

Diagnóstico clínico: los pacientes que presentan los mismos síntomas de la enfermedad a menudo tienen diferentes pronósticos y respuestas a los fármacos en función de sus diferencias genéticas subyacentes. La entrega de información genética específica del paciente abarca diagnósticos moleculares que incluyen kits y servicios de diagnóstico basados en genes o expresiones, productos de diagnóstico complementarios para seleccionar y monitorear terapias particulares, así como la detección de pacientes para detección temprana de enfermedades y monitoreo de enfermedades. La creación de diagnósticos moleculares y pruebas de detección más efectivos y específicos requiere una mejor comprensión de los genes, factores reguladores y otros factores relacionados con enfermedades o fármacos, que la tecnología de secuenciación de una sola molécula descrita tiene el potencial de permitir.

Agricultura: la investigación agrícola ha recurrido cada vez más a la genómica para el descubrimiento, desarrollo y diseño de animales y cultivos genéticamente superiores. La industria de los negocios agrícolas ha sido una gran consumidora de tecnologías genéticas, particularmente de chips, para identificar variaciones genéticas relevantes entre variedades o poblaciones. La tecnología de secuenciación descrita puede proporcionar un enfoque más poderoso, directo y rentable para el análisis de la expresión génica y los estudios de población para esta industria

Una oportunidad adicional estará en el campo de las aplicaciones de repetición de secuencias en las que los métodos se aplican a la detección de variaciones genéticas sutiles. El análisis genómico comparativo ampliado a través de las especies puede proporcionar una gran comprensión de la estructura y función del genoma humano y, en consecuencia, la genética de la salud y la enfermedad humanas. Los estudios sobre la variación genética humana y su relación con la salud y la enfermedad se están expandiendo. La mayoría de estos estudios utilizan tecnologías que se basan en patrones de variación conocidos y relativamente comunes. Estos poderosos métodos proporcionarán información nueva e importante, pero son menos informativos que para determinar la secuencia completa y contigua de los genomas humanos individuales. Por ejemplo, es probable que los métodos actuales de genotipificación no detecten diferencias raras entre personas en cualquier ubicación genómica en particular y tengan una capacidad limitada para determinar reordenamientos de largo alcance. La caracterización de los cambios somáticos del genoma que contribuyen al cáncer actualmente emplea combinaciones de tecnologías para obtener datos de secuencia (en muy pocos genes) más información limitada sobre cambios en el número de copias, reordenamientos o pérdida de heterocigosidad. Dichos estudios sufren de mala resolución y/o cobertura incompleta del genoma. La heterogeneidad celular de las muestras tumorales presenta desafíos adicionales. La secuenciación completa del genoma a bajo costo a partir de muestras extremadamente pequeñas, tal vez incluso células individuales, alteraría la batalla contra el cáncer en todos los aspectos, desde el laboratorio de investigación hasta la clínica. El recientemente lanzado proyecto piloto Atlas del Genoma del Cáncer (TCGA) se mueve en la dirección deseada, pero sigue siendo dramáticamente limitado por los costos de secuenciación. Se necesitan secuencias genómicas adicionales de animales y plantas de importancia agrícola para estudiar la variación individual, diferentes razas domesticadas y varias variantes silvestres de cada especie. El análisis de secuencia de las comunidades microbianas, muchos de los cuales no pueden ser cultivadas, proporcionará una fuente rica de información médica y ambientalmente útil. Y la secuenciación precisa y rápida puede ser el mejor enfoque para el monitoreo microbiano de los alimentos y el medio ambiente, incluida la detección rápida y la mitigación de amenazas de bioterrorismo.

La secuenciación del genoma también podría proporcionar ácidos nucleicos aislados que comprenden regiones intrónicas útiles en la selección de secuencias de la firma clave. Actualmente, las secuencias de firma clave están dirigidas a regiones exónicas.

Una aplicación fundamental de la tecnología de ADN implica varias estrategias de etiquetado para marcar un ADN producido por una ADN polimerasa. Esto es útil en la tecnología de chips: secuenciación de ADN, detección de SNP, clonación, análisis de PCR y muchas otras aplicaciones.

Ejemplo 1

Propósito: El uso de firmas clave y/o códigos de barras para permitir la identificación del genoma con tan solo 8-18 nucleótidos y el análisis de datos de secuencias muy cortas (lecturas) en tiempo real.

Se usaron algoritmos de construcción de matrices de sufijos de tiempo lineal para calcular el análisis de singularidad. El análisis determinó el porcentaje de todas las secuencias que fueron únicas en varios genomas modelo. Se analizaron todas las longitudes de secuencia en un genoma. Se cuentan las secuencias que ocurren solo una vez en un genoma. El algoritmo de matriz de sufijos funciona calculando una gráfica de puntuación de repetición que analiza la frecuencia de subsecuencias específicas dentro de una secuencia para que se produzca en función de una ventana deslizante de dos pares de bases. La información del genoma almacenada en el GenBank se utilizó para el análisis *in silico*. Se analizaron un genoma viral, un fago Lambda, un genoma bacteriano, *E. coli* K12 MG1655 y el genoma humano. El porcentaje de lecturas únicas es una función de la longitud de la secuencia. Se

hizo una suposición con respecto a las secuencias que solo producen coincidencias inequívocas y que producen superposiciones inequívocas para reconstruir el genoma. Las lecturas únicas variaron en tamaño de 7 a 100 nucleótidos. La mayoría de los tamaños únicos fueron más cortos que 9, 13 y 18 nucleótidos, respectivamente.

Resultados: los resultados muestran que las secuencias aleatorias de 12 nt del genoma del fago son 98% únicas para el fago. Esto aumenta lentamente, de modo que las secuencias de 400 nt son un 99% exclusivas del fago. Esto disminuye a 80% para secuencias de fagos de 10 nt. Para las bacterias (*E. coli*), las secuencias de 18 nt del genoma son exclusivas del 97% de *E. coli*. Para los genomas humanos, las secuencias de 25 nt son 80% únicas en humanos y un aumento a 45 nt da como resultado que 90% del genoma sea único. Otros aspectos divulgados incluyen los siguientes:

[1] Un método para identificar un material biológico en una muestra, que comprende:

obtener una muestra que comprende dicho material biológico, extraer una o más moléculas de ácido nucleico de dicha muestra, generar información de secuencia de dicha molécula o moléculas de ácido nucleico con la presente coincidencia probabilística directa para la comparación de dicha información de secuencia con las secuencias de ácido nucleico en una base de datos.

[2] El método de [1], en el que dichas una o más moléculas de ácido nucleico se seleccionan de ADN o ARN.

[3] El método de [1], en el que dicha información de secuencia comprende un fragmento de nucleótido de longitud "n".

[4] El método de [3], en el que dicho fragmento de nucleótido de longitud "n" se compara con las secuencias de ácido nucleico en una base de datos.

[5] El método de [4], en el que dicho fragmento de nucleótido de longitud "n" se compara con las secuencias de ácido nucleico en una base de datos mediante emparejamiento probabilístico.

[6] El método de [4], en el que la comparación de dicho fragmento de nucleótido de longitud "n" se realiza, en tiempo real, o tan rápido como se genera dicho fragmento, o se genera información de secuencia de dicho fragmento.

[7] El método de [4], en el que si la probabilidad de coincidencia de un fragmento de nucleótido de longitud "n" es menor que el umbral de una coincidencia objetivo, entonces un fragmento de ácido nucleico de longitud "n + 1", "n + 2" …"n + x" se genera a partir de dichas una o más moléculas de ácido nucleico y se compara con las secuencias de ácido nucleico en una base de datos, en la que x es menor a 50.

[8] El método de [4], en el que si la probabilidad de coincidencia de un fragmento de nucleótido de longitud "n" es menor que el umbral de una coincidencia objetivo, entonces un fragmento de ácido nucleico de longitud "n + 1", "n + 2" …"n + x" se genera a partir de dichas una o más moléculas de ácido nucleico y se compara con las secuencias de ácido nucleico en una base de datos, en la que "x" es mayor a 50.

[9] El método de [1], que además comprende la amplificación de dichas una o más moléculas de ácido nucleico para producir una pluralidad "i" de moléculas de ácido nucleico, antes de generar información de secuencia.

[10] El método de [8], en el que dicha información de secuencia comprende fragmentos de nucleótidos de longitud "n".

[11] El método de [9], en el que la pluralidad "i" de longitud "n" de fragmentos de nucleótidos se comparan con las secuencias de ácido nucleico en una base de datos.

[12] El método de [11], en el que la pluralidad i(n) de fragmentos de nucleótidos se compara con las secuencias de ácido nucleico en una base de datos mediante emparejamiento probabilístico.

[13] El método de [11], en el que se realiza la comparación de la pluralidad i(n) de fragmentos de nucleótidos, en tiempo real, o tan rápido como se generan dichos fragmentos.

[14] El método de [11], en el que si la probabilidad de coincidencia de la pluralidad i(n) de fragmentos de nucleótidos es menor que un umbral de una coincidencia objetivo, entonces los fragmentos de ácido nucleico de longitud "i(n + 1)", "i(n + 2)" ... "i(n + x)" se genera a partir de dichas una o más moléculas de ácido nucleico y se compara con las secuencias de ácido nucleico en una base de datos, en la que "x" es menos a 50.

[15] El método de [11], en el que si la probabilidad de coincidencia de la pluralidad i(n) de fragmentos de nucleótidos es menor que un umbral de una coincidencia objetivo, entonces los fragmentos de ácido nucleico de longitud "i(n + 1)", "i(n + 2)" ... "i(n + x)" se genera a partir de dichas una o más moléculas de ácido nucleico y se compara con las secuencias de ácido nucleico en una base de datos, en donde "x" es mayor a 50.

[16] El método de acuerdo con [5] o [12], en el que dicho emparejamiento probabilístico se realiza utilizando un enfoque bayesiano.

[17] El método de acuerdo con [5] o [12], en el que dicha comparación probabilística se realiza utilizando un enfoque bayesiano recursivo.

[18] El método de acuerdo con [5] o [12], en el que dicho emparejamiento probabilístico se realiza utilizando un enfoque bayesiano sin modificación.

[19] El método de acuerdo con [5] o [12], en el que dicha comparación probabilística proporciona un marco estadístico jerárquico para identificar las especies de dicha información de secuencia.

[20] El método de [1], en el que la comparación de dicha información de secuencia con las secuencias de ácido nucleico en una base de datos se realiza, en tiempo real, o tan rápido como se genera la información de secuencia, mientras que la información adicional de secuencia continúa siendo generado a partir de dichas una o más moléculas de ácido nucleico.

[21] El método de [20], en el que dicha información adicional de secuencia comprende nucleótidos de longitudes variables.

[22] El método de [1], en el que dicha información de secuencia comprende un fragmento de nucleótido de longitud "n", que se compara, en tiempo real, o tan rápido como el fragmento se genera con las secuencias de ácido nucleico en una base de datos; mientras que los fragmentos de ácido nucleico de longitud "n + 1", "n + 2" ... "n + x" continúan generándose a partir de dichas una o más moléculas de ácido nucleico y se comparan, en tiempo real, o tan rápido como los fragmentos se generan con las secuencias de ácido nucleico en una base de datos.

[23] El método de [1], en el que dichas una o más moléculas de ácido nucleico se amplifican para producir una pluralidad "i" de moléculas de ácido nucleico antes de generar información de secuencia de fragmentos de nucleótidos de longitud "n"; que comprende además comparar la pluralidad i(n) de fragmentos de nucleótidos, en tiempo real, o tan rápido como los fragmentos se generan con las secuencias de ácido nucleico en una base de datos; mientras que una pluralidad "i(n + 1)", "i(n + 2)" ... "i(n + x)" de fragmentos de ácido nucleico continúan generándose a partir de dichas una o más moléculas de ácido nucleico y se compara, en tiempo real, o tan rápido como los fragmentos se generan con las secuencias de ácido nucleico en una base de datos.

[24] Un sistema para detectar material biológico, que comprende:

(i) una unidad receptora de muestras configurada para recibir una muestra que comprende material biológico;

(ii) una unidad de extracción en comunicación con dicha unidad receptora de muestras, estando configurada dicha unidad de extracción para extraer al menos una molécula de ácido nucleico de dicha muestra;

(iii) un casete de secuenciación en comunicación con dicha unidad de extracción, estando configurado dicho casete de secuenciación para recibir dicha al menos una molécula de ácido nucleico de dicha unidad de extracción y generar información de secuencia a partir de dicha al menos una molécula de ácido nucleico;

(iv) una base de datos que comprende secuencias de ácido nucleico de referencia; y una

(v) unidad de procesamiento en comunicación con dicho casete de secuenciación y dicha base de datos, estando configurada dicha unidad de procesamiento para recibir dicha información de secuencia desde dicho casete de secuenciación y comparar dicha información de secuencia con dichas secuencias de ácido nucleico de referencia.

[25] El sistema de [24], que comprende:

un dispositivo de secuenciación portátil que transmite electrónicamente los datos a una base de datos para la identificación de organismos relacionados con la determinación de la secuencia de los ácidos nucleicos.

[26] El sistema de [24], que además comprende una unidad de llamada base configurada para procesar secuencias recibidas por el casete de secuenciación.

[27] El sistema de [26], en el que la unidad de llamada base está acoplada al procesador de coincidencia probabilístico.

[28] El sistema de [27], en el que el procesador de coincidencia probabilística está configurado para utilizar un enfoque bayesiano para recibir la secuencia resultante y calcular las probabilidades para cada lectura de secuenciación mientras considera las puntuaciones de calidad de secuencia generadas por la unidad de llamada base.

[29] El sistema de [27], en el que el procesador de coincidencia probabilística utiliza una base de datos generada y optimizada antes de su uso para la identificación de patógenos.

[30] El sistema de [27], en el que el procesador de coincidencia probabilística utiliza puntuaciones ponderadas que varían de acuerdo con el contenido de la secuencia.

[31] El sistema de [24], que comprende una unidad de almacenamiento en comunicación con dicha unidad de procesamiento, en donde dicha unidad de procesamiento está configurada para transmitir dicha información de secuencia a dicha unidad de almacenamiento de datos y posteriormente recuperar dicha información de secuencia de dicha unidad de almacenamiento de datos para procesamiento.

[32] El sistema de [24], en el que dicha al menos una molécula de ácido nucleico se selecciona del grupo que consiste en ADN y ARN.

[33] El sistema de [24], en el que dicha información de secuencia comprende un fragmento de nucleótido de longitud "n".

[34] El sistema de [33], en el que dicha unidad de extracción está configurada para comparar dicho fragmento de nucleótido de longitud n con dichas secuencias de ácido nucleico de referencia.

[35] El sistema de [34], en el que dicha unidad de extracción está configurada para comparar dicho fragmento de nucleótido de longitud "n" con dichas secuencias de ácido nucleico de referencia mediante emparejamiento probabilístico.

[36] El sistema de [34], en el que dicha unidad de extracción está configurada para comparar dicho fragmento de nucleótido de longitud "n" con dichas secuencias de ácido nucleico de referencia en tiempo real, o tan rápido como se genera dicho fragmento de longitud "n".

[37] El sistema de [34], en el que si la probabilidad de coincidencia de un fragmento de nucleótido de longitud "n" es menor que un umbral de una coincidencia objetivo, entonces dicho casete de secuenciación está configurado para generar información de secuencia de "n + 1 "," n + 2 "..." n + x " fragmentos de nucleótidos de longitud de dichas una o más moléculas de ácido nucleico y dicha unidad de extracción está configurada para comparar dichos fragmentos de nucleótidos de longitud "n + 1", "n + 2"..."n + x" con las secuencias de ácido nucleico en una base de datos.

[38] El sistema de [36], en el que dicho fragmento de nucleótido de longitud "n" se compara con dichas secuencias de ácido nucleico de referencia en tiempo real, o tan rápido como dicho fragmento de longitud "n" se genera, mientras que la unidad de secuenciación continúa generando información de secuencia de "n + 1", "n + 2" ... "n + x" fragmentos de nucleótidos de longitud a partir de dichas una o más moléculas de ácido nucleico, y la unidad de procesamiento compara dicha información de secuencia de "n + 1", "n + 2"..."n + x" fragmentos de nucleótidos en longitud, en tiempo real, o tan rápido como los fragmentos se generan a las secuencias de ácido nucleico en una base de datos.

[39] Un método para identificar un material biológico en una muestra, que comprende:

(i) obtener una muestra que comprenda dicho material biológico,

(ii) extraer una o más moléculas de ácido nucleico de dicha muestra,

(iii) generar información de secuencia, que comprende una secuencia de un fragmento de nucleótido de dichas una o más moléculas de ácido nucleico,

(iv) comparar dicha secuencia de fragmento de nucleótido con secuencias de ácido nucleico en una base de datos;

y si dicha comparación de dicha secuencia de un fragmento de nucleótido no resulta en una coincidencia que identifique el material biológico en dicha muestra, entonces el método comprende además:

(v) generar información adicional de secuencia a partir de dichas una o más moléculas de ácido nucleico, en la que dicha información adicional de secuencia comprende una secuencia de un fragmento de nucleótido que consiste de un nucleótido adicional,

(vi) comparar dicha información adicional de secuencia con secuencias de ácido nucleico en una base de datos inmediatamente después de la generación de dicha información adicional de secuencia, y repetir las etapas (v) - (vi) hasta que una coincidencia resulte en la identificación del material biológico en dicha muestra.

[40] Un método para identificar un material biológico en una muestra, que comprende:

(i) obtener una muestra que comprenda dicho material biológico,

(ii) extraer una o más moléculas de ácido nucleico de dicha muestra,

(iii) amplificar dichas una o más moléculas de ácido nucleico para producir una pluralidad de una o más moléculas de ácido nucleico,

(iii) generar una pluralidad de información de secuencia, que comprende una pluralidad de secuencias de un fragmento de nucleótido, a partir de dicha pluralidad de una o más moléculas de ácido nucleico,

(iv) comparar dicha pluralidad de secuencias de un fragmento de nucleótido con secuencias de ácido nucleico en una base de datos,

y si dicha comparación de dicha pluralidad de secuencias de un fragmento de nucleótido no resulta en una coincidencia que identifique el material biológico en dicha muestra, entonces el método comprende además:

(v) generar una pluralidad de información adicional de secuencia a partir de dichas una o más moléculas de ácido nucleico, en donde dicha información adicional de secuencia comprende una secuencia de un fragmento de nucleótido que consiste en un nucleótido adicional,

(vi) comparar dicha información adicional de secuencia con secuencias de ácido nucleico en una base de datos inmediatamente después de la generación de dicha información adicional de secuencia, y repetir las etapas (v) - (vi) hasta que resulte una coincidencia en la identificación del material biológico en dicha muestra.

[41] Los métodos de [39] o [40], en los que la comparación con las secuencias de ácido nucleico en una base de datos se realiza a través de un emparejamiento probabilístico tan rápido como se genera la información de la secuencia.

REIVINDICACIONES

1. Un método *ex vivo* de identificación de un genoma en una muestra que comprende una pluralidad de genomas, que comprende:

(i) obtener una muestra que comprenda la pluralidad de genomas;

(ii) extraer una o más moléculas de ácido nucleico de la muestra;

(iii) generar información de secuencia, la información de secuencia que comprende una secuencia de un fragmento de nucleótido de una o más moléculas de ácido nucleico;

(iv) comparar la secuencia de un fragmento de nucleótido con secuencias de ácido nucleico en una base de datos utilizando un emparejamiento probabilístico; y si la comparación de la secuencia de un fragmento de nucleótido no resulta en una coincidencia que identifique un genoma en la muestra en virtud de que la probabilidad de coincidencia del fragmento de nucleótido sea menor que un umbral de una coincidencia objetivo, entonces el método comprende además:

(v) generar información adicional de secuencia a partir de una o más moléculas de ácido nucleico;

(vi) comparar la información adicional de secuencia con las secuencias de ácido nucleico en la base de datos inmediatamente después de la generación de la información adicional de secuencia utilizando el emparejamiento probabilístico; y

(vii) repetir las etapas (v) - (vi) hasta que una coincidencia resulte en la identificación de un genoma en la muestra.

2. El método de la reivindicación 1, en el que la información adicional de secuencia comprende la secuencia del fragmento de nucleótido que comprende un nucleótido adicional.

3. El método de la reivindicación 1, en el que la secuencia del fragmento de nucleótido de (iv) es una secuencia de nucleótidos de longitud "n" y la información adicional de secuencia comprende una secuencia de nucleótidos de longitud "n + 1", "n + 2" hasta "n + x", en la que x es menor que 50.

4. El método de la reivindicación 1, en el que la secuencia del fragmento de nucleótido de (iv) es una secuencia de nucleótidos de longitud "n" y la información adicional de secuencia comprende una secuencia de nucleótidos de longitud "n + 1", "n + 2" hasta "n + x", donde x es mayor a 50.

5. El método de la reivindicación 1, en el que generar la información de secuencia comprende pirosecuenciación.

6. El método de la reivindicación 1, en el que generar la información de secuencia comprende secuenciación por hibridación.

7. El método de la reivindicación 1, que comprende además la amplificación de una o más moléculas de ácido nucleico para producir una pluralidad "i" de moléculas de ácido nucleico, antes de generar la información de la secuencia.

8. El método de la reivindicación 1, en el que la comparación de (vi) se realiza, en tiempo real, o tan rápido como se genera la información adicional de secuencia de (v).

9. El método de la reivindicación 1, en el que la comparación probabilística se realiza utilizando un proceso que comprende un enfoque bayesiano.

10. El método de la reivindicación 9, en el que el enfoque bayesiano comprende un enfoque bayesiano recursivo.

11. El método de la reivindicación 9, en el que el enfoque bayesiano comprende un enfoque bayesiano sin modificar.

12. El método de la reivindicación 1, en el que la comparación probabilística comprende el uso de un marco estadístico jerárquico para identificar un genoma en la muestra.

13. El método de la reivindicación 1, en el que la muestra comprende una muestra del sujeto y una muestra ambiental.

14. Un sistema para identificar un genoma en una muestra que comprende una pluralidad de genomas, que comprende:

(i) una unidad receptora de muestras configurada para recibir la muestra;

(ii) una unidad de extracción en comunicación con la unidad receptora de muestras, estando configurada la unidad de extracción para extraer una o más moléculas de ácido nucleico de la muestra;

(iii) un casete de secuenciación en comunicación con la unidad de extracción, estando el casete de secuenciación configurado para recibir una o más moléculas de ácido nucleico de la unidad de extracción y generar la información de la secuencia de una o más moléculas de ácido nucleico;

(iv) una base de datos que comprende secuencias de ácido nucleico de referencia; y

(v) una unidad de procesamiento en comunicación con el casete de secuenciación y la base de datos, caracterizada porque la unidad de procesamiento está configurada para realizar las etapas (iv) - (vii) de la reivindicación 1.

15. El sistema de la reivindicación 14, en el que el sistema comprende un dispositivo de secuenciación portátil que transmite datos electrónicamente a la base de datos, el dispositivo que comprende la recepción de muestras, la unidad de extracción, el casete de secuenciación y la unidad de procesamiento.

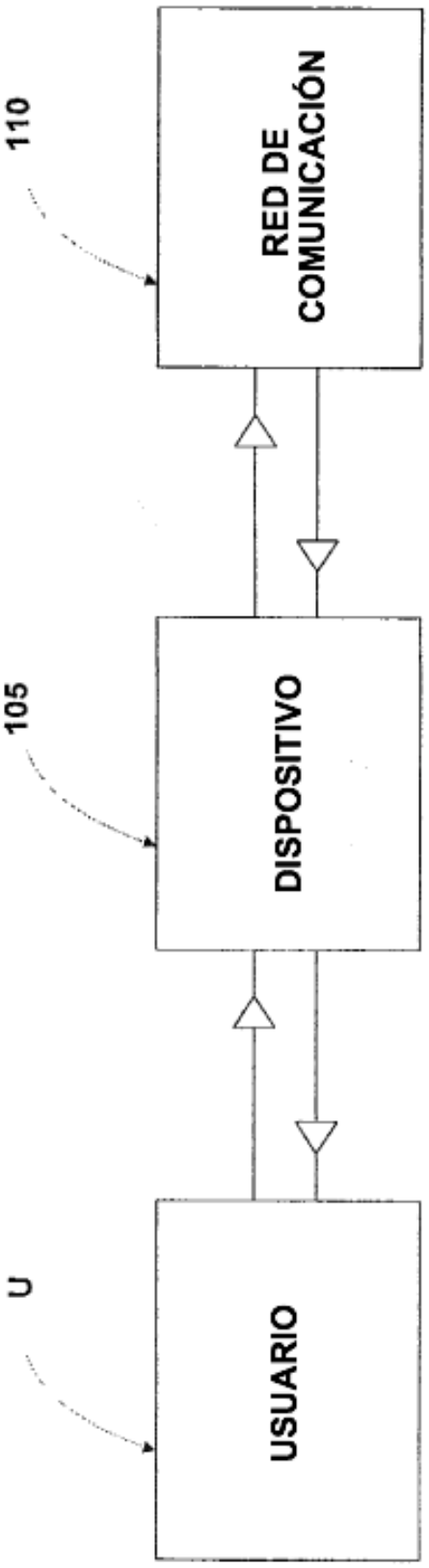

FIG. 1

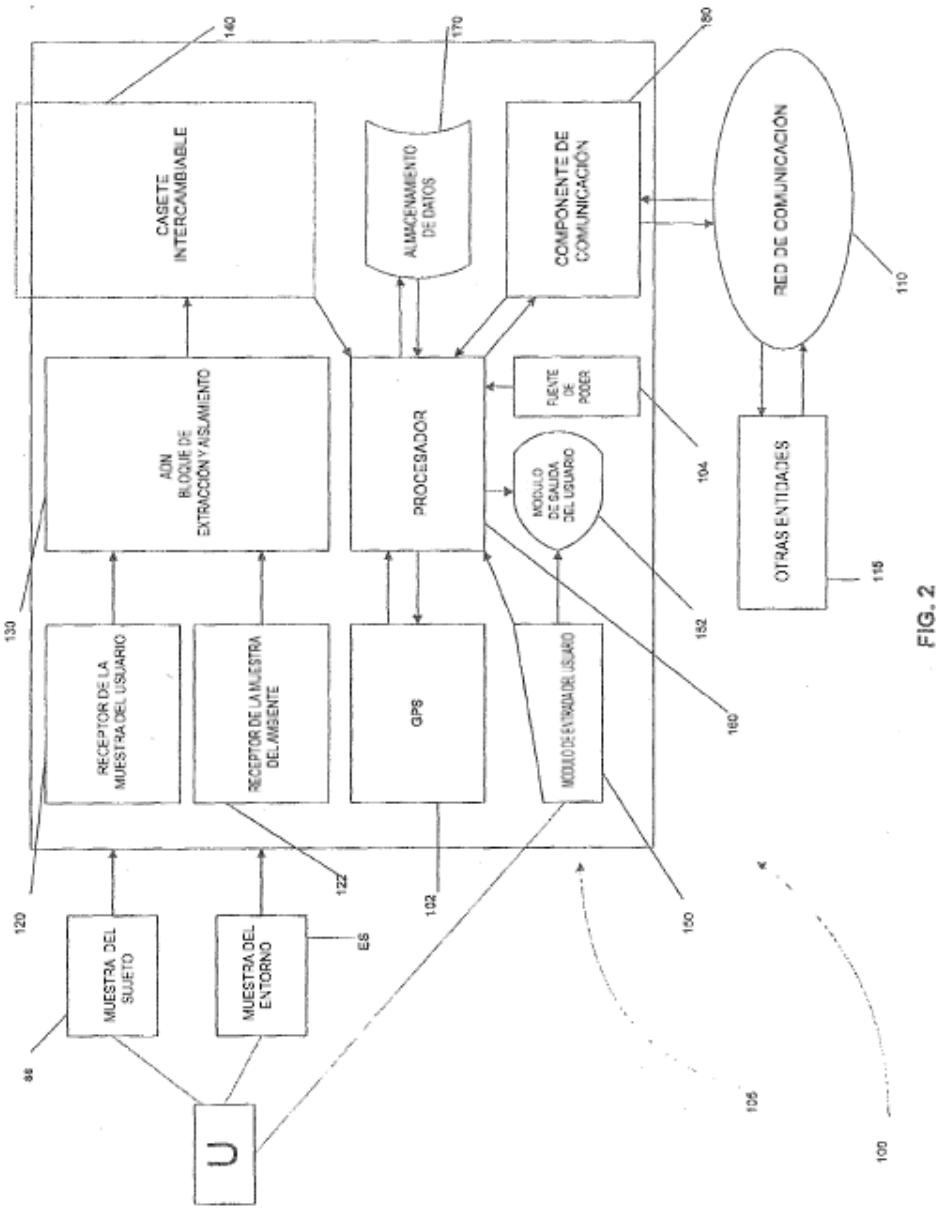
CASETE INTERCAMBIABLE
ALMACENAMIENTO DE DATOS
COMPONENTE DE COMUNICACIÓN
RED DE COMUNICACIÓN
ADN BLOQUE DE EXTRACCIÓN Y AISLAMIENTO
PROCESADOR
FUENTE DE PODER
MÓDULO DE SALIDA DEL USUARIO
OTRAS ENTIDADES
RECEPTOR DE LA MUESTRA DEL USUARIO
RECEPTOR DE LA MUESTRA DEL AMBIENTE
GPS
MÓDULO DE ENTRADA DEL USUARIO
MUESTRA DEL SUJETO
MUESTRA DEL ENTORNO
U
140
170
180
110
130
104
152
115
160
120
122
102
150
88
ES
105
100

FIG. 2

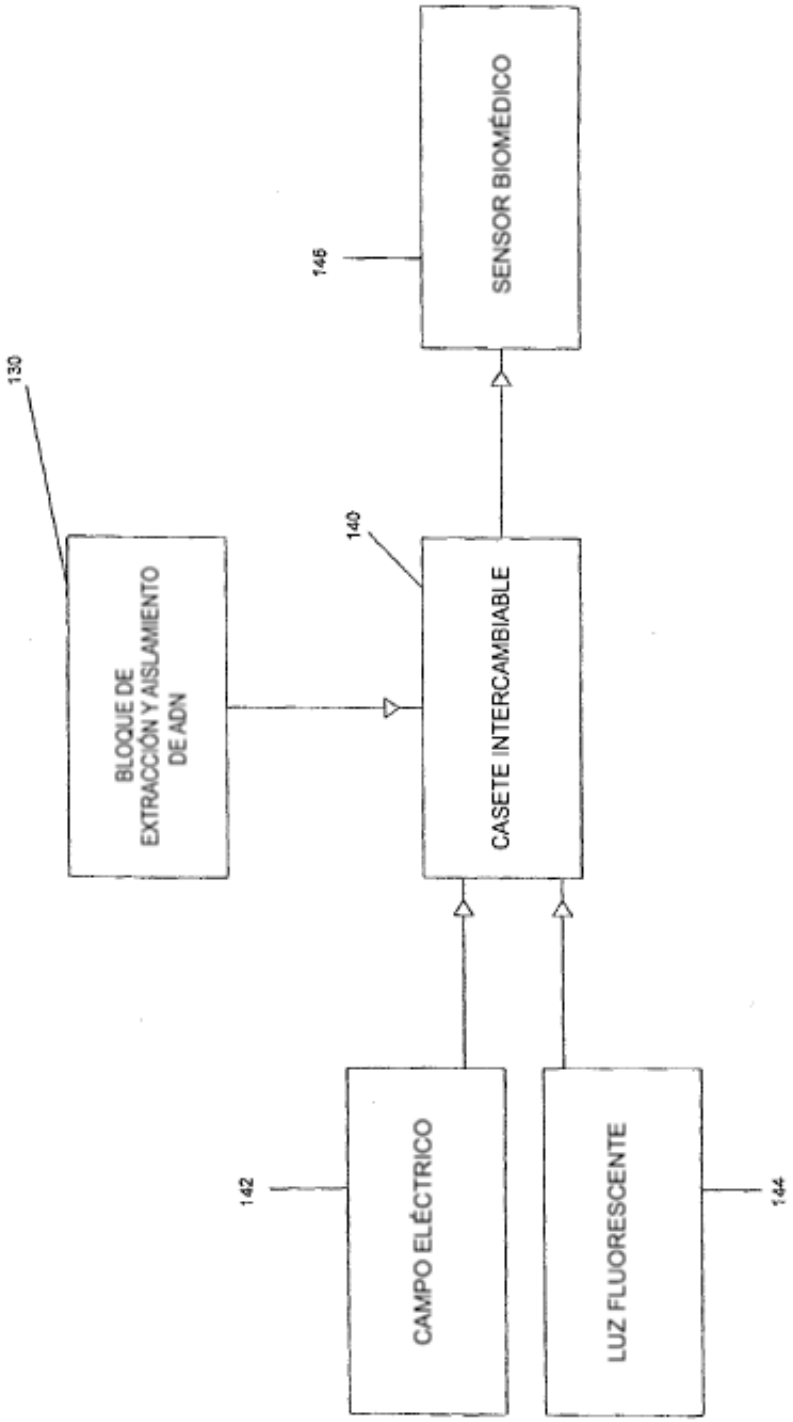
SENSOR BIOMÉDICO
146
130
140
BLOQUE DE EXTRACCIÓN Y AISLAMIENTO DE ADN
CASETE INTERCAMBIABLE
142
CAMPO ELÉCTRICO
LUZ FLUORESCENTE
144

FIG. 3

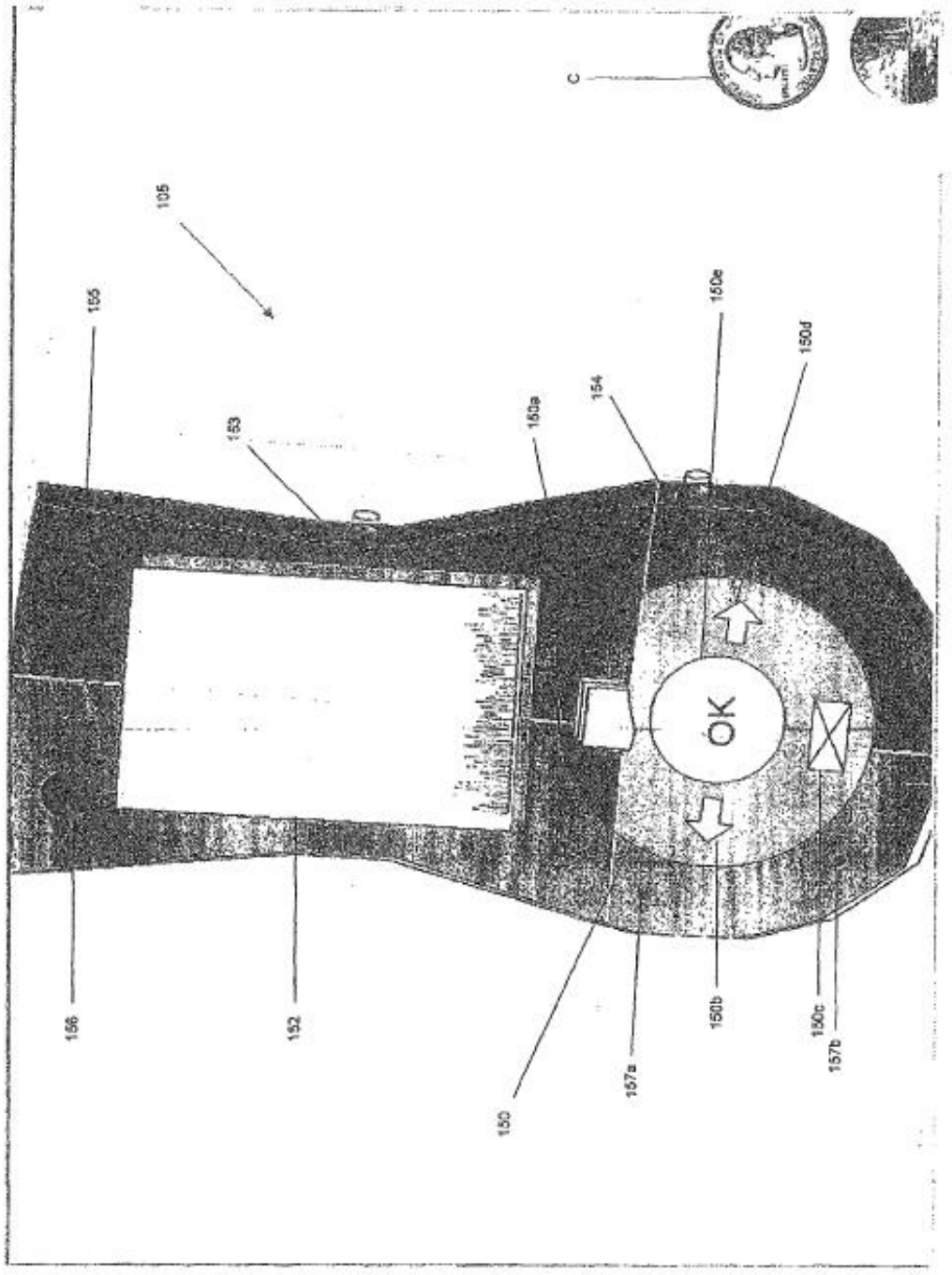
C
105
155
153
150a
154
150e
150d
OK
156
152
150
157a
150b
150c
157b

FIG. 4

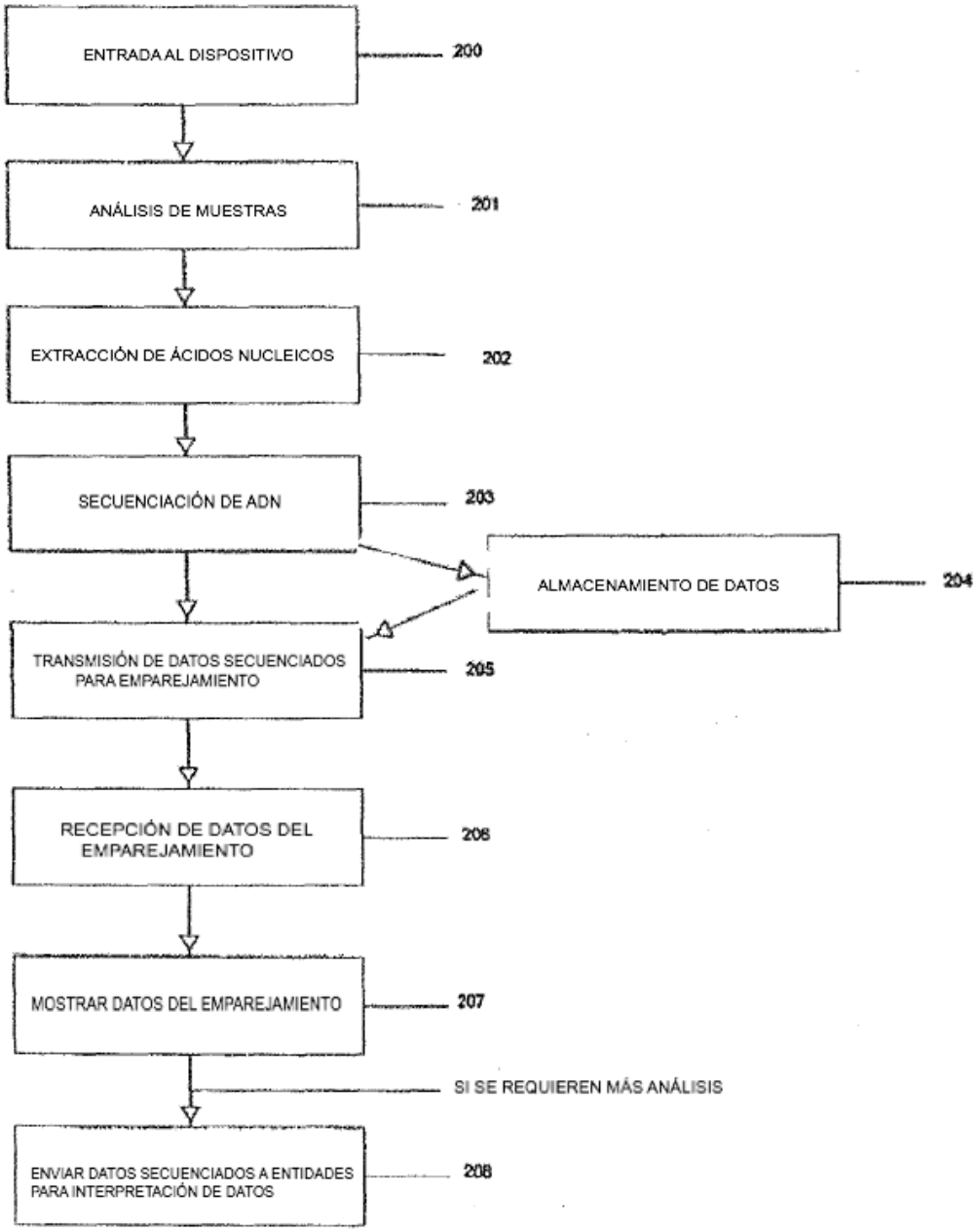
ENTRADA AL DISPOSITIVO
200
ANÁLISIS DE MUESTRAS
201
EXTRACCIÓN DE ÁCIDOS NUCLEICOS
202
SECUENCIACIÓN DE ADN
203
ALMACENAMIENTO DE DATOS
204
TRANSMISIÓN DE DATOS SECUENCIADOS PARA EMPAREJAMIENTO
205
RECEPCIÓN DE DATOS DEL EMPAREJAMIENTO
206
MOSTRAR DATOS DEL EMPAREJAMIENTO
207
SI SE REQUIEREN MÁS ANÁLISIS
ENVIAR DATOS SECUENCIADOS A ENTIDADES PARA INTERPRETACIÓN DE DATOS
208

FIG. 5

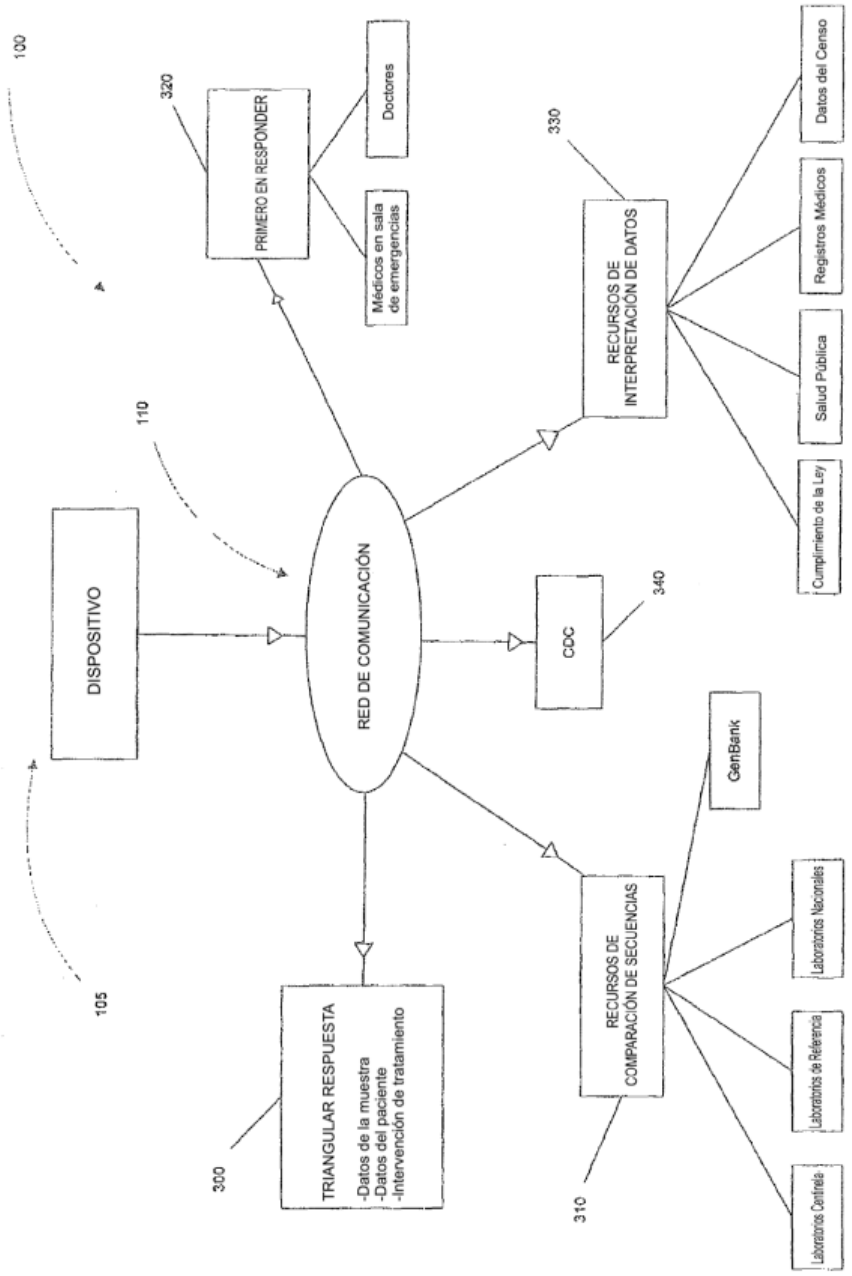
100
105
110
300
310
320
330
340
DISPOSITIVO
RED DE COMUNICACIÓN
TRIANGULAR RESPUESTA
-Datos de la muestra
-Datos del paciente
-Intervención de tratamiento
PRIMERO EN RESPONDER
Doctores
Médicos en sala de emergencias
RECURSOS DE COMPARACIÓN DE SECUENCIAS
GenBank
Laboratorios Nacionales
Laboratorios de Referencia
Laboratorios Centrales
RECURSOS DE INTERPRETACIÓN DE DATOS
Datos del Censo
Registros Médicos
Salud Pública
Cumplimiento de la Ley
CDC

FIG. 6

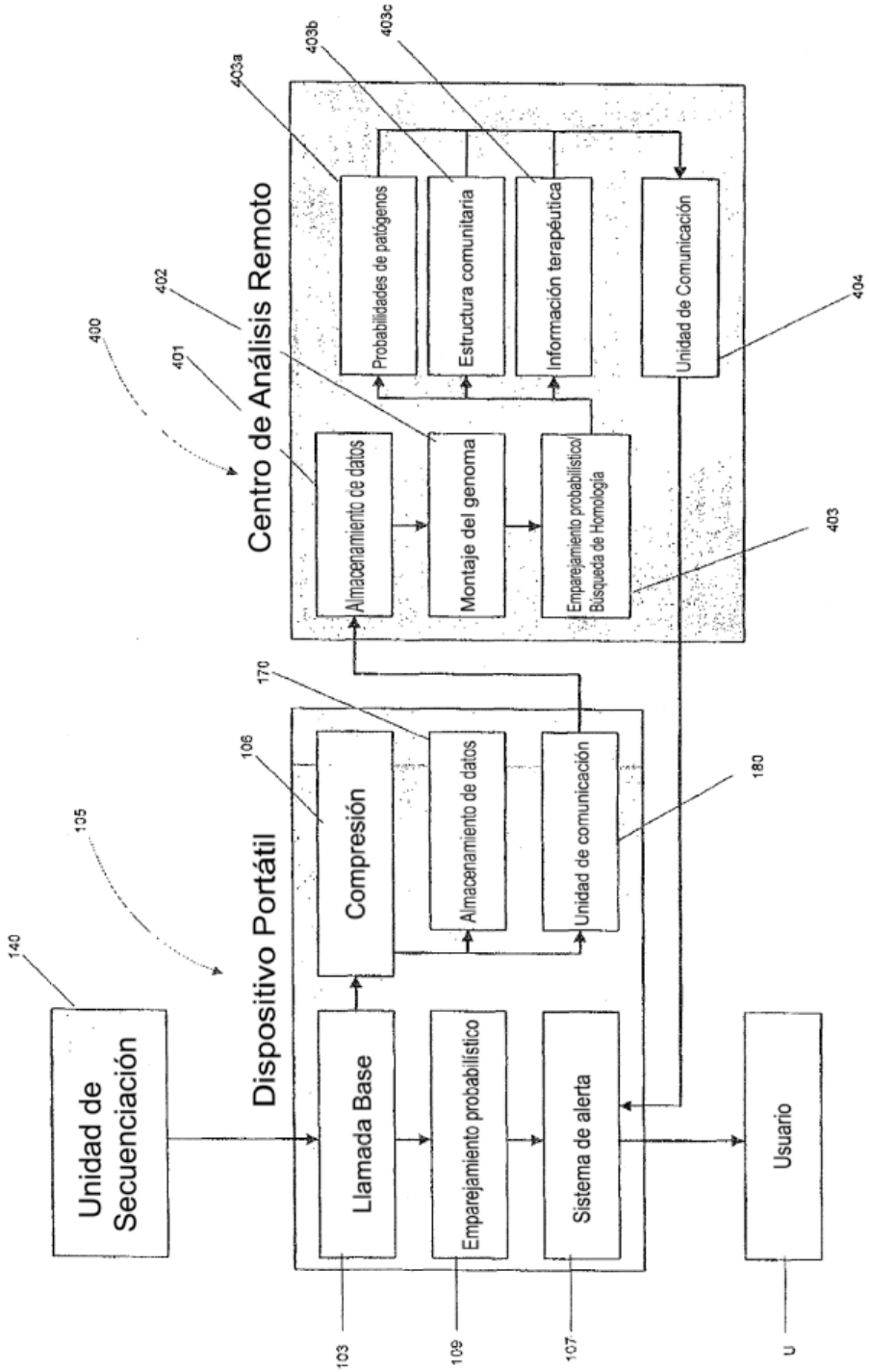
Unidad de Secuenciación
140
Dispositivo Portátil
105
Llamada Base
Compresión
106
Emparejamiento probabilístico
Almacenamiento de datos
170
Sistema de alerta
Unidad de comunicación
180
Usuario
103
109
107
U
Centro de Análisis Remoto
400
401
402
403
403a
403b
403c
404
Almacenamiento de datos
Montaje del genoma
Emparejamiento probabilístico/ Búsqueda de Homología
Probabilidades de patógenos
Estructura comunitaria
Información terapéutica
Unidad de Comunicación

FIG. 7

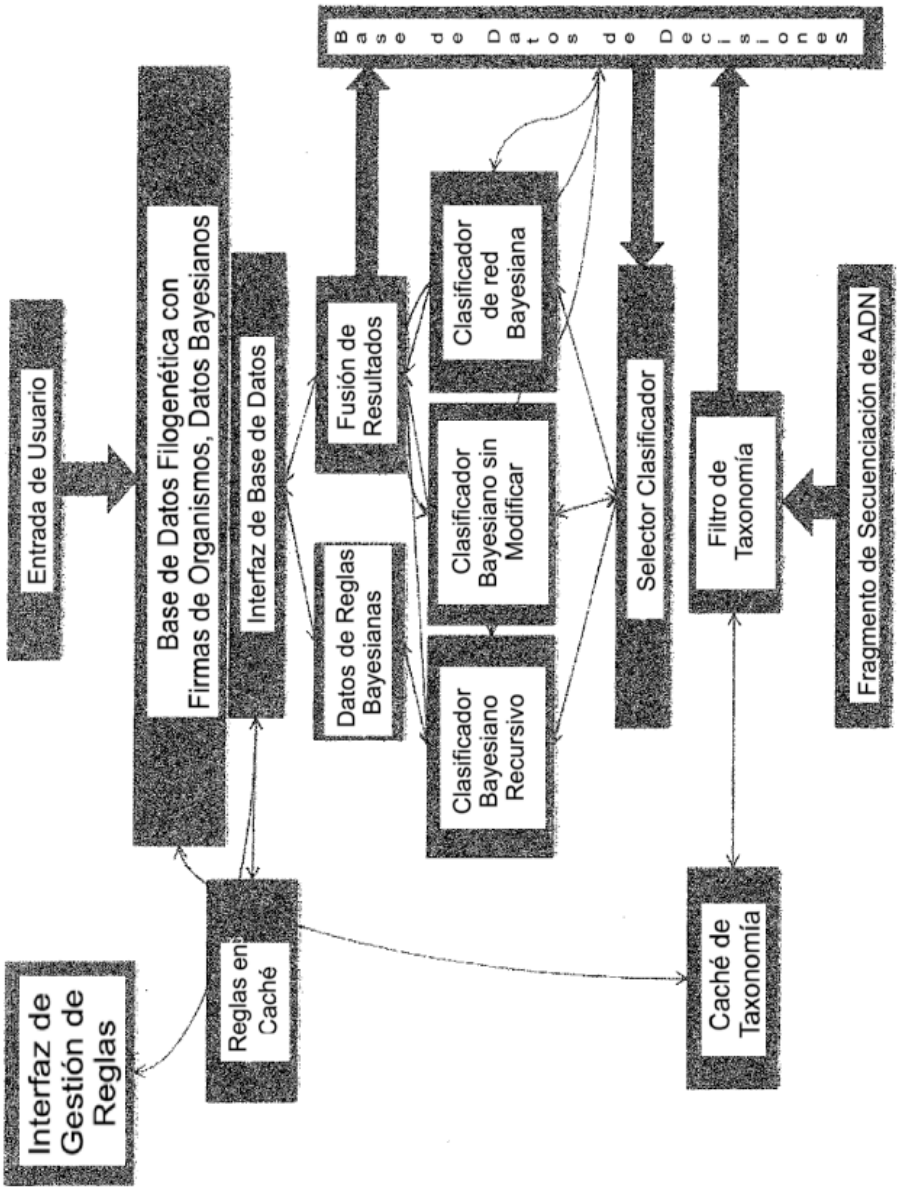
Base de Datos de Decisiones
Entrada de Usuario
Base de Datos Filogenética con Firmas de Organismos, Datos Bayesianos
Interfaz de Base de Datos
Fusión de Resultados
Clasificador de red Bayesiana
Clasificador Bayesiano sin Modificar
Clasificador Bayesiano Recursivo
Datos de Reglas Bayesianas
Selector Clasificador
Filtro de Taxonomía
Fragmento de Secuenciación de ADN
Interfaz de Gestión de Reglas
Reglas en Caché
Caché de Taxonomía

FIG. 8

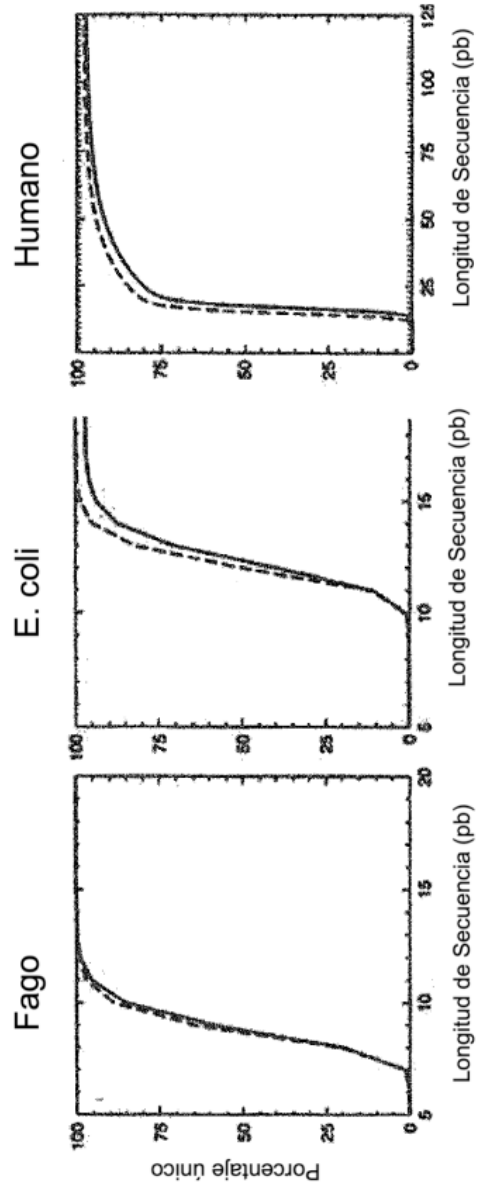
Fago
E. coli
Humano
Porcentaje único
Longitud de Secuencia (pb)
Longitud de Secuencia (pb)
Longitud de Secuencia (pb)

FIG. 9

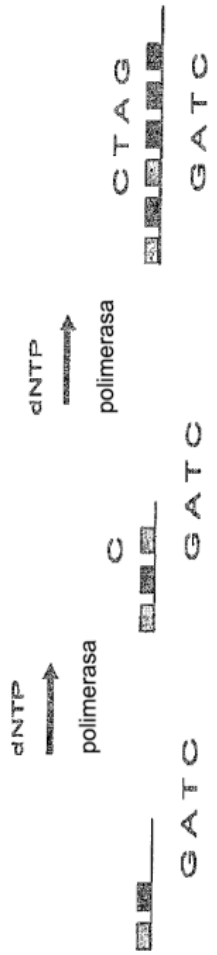

FIG. 10