(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 588 908**

(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

(86) Fecha de presentación y número de la solicitud internacional: **21.11.2008 PCT/US2008/084403**

(87) Fecha y número de publicación internacional: **09.07.2009 WO09085473**

(96) Fecha de presentación y número de la solicitud europea: **21.11.2008 E 08867952 (7)**

(97) Fecha y número de publicación de la concesión europea: **03.08.2016 EP 2229587**

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

(30) Prioridad:

21.11.2007 US 989641 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
07.11.2016

(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:

LAZCANO GAINZA, Jesús

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín Europeo de Patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre Concesión de Patentes Europeas).

DESCRIPCIÓN

Sistema de identificación de genoma

Campo de la invención

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

Antecedentes

Hay una amplia variedad de circunstancias que amenazan la vida que sería útil analizar, y la secuencia de una muestra de ADN o ARN, por ejemplo, en respuesta a un acto de bioterrorismo, donde un agente patógeno mortal había sido liberado al ambiente. En el pasado, estos 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 que el personal de salud pública del país rápidamente, identifique con precisión y de forma fiable los agentes patógenos y las enfermedades que provocan. Los agentes patógenos pueden estar contenidos en fuentes de muestras, tales como alimentos, aire, suelo, agua, tejido y presentación clínica de los agentes patógenos. Debido a que los agentes y/o enfermedades potenciales pueden poner en peligro la vida y ser altamente contagiosos, debe hacerse rápidamente este proceso de identificación. Esta es una debilidad significativa en la respuesta actual de seguridad nacional al bioterrorismo.

Puede ser necesario un sistema y método que identifique más de un solo organismo (multiplexación) e indique si está presente una especie, sobre la base de la comparación del genoma de los ácidos nucleicos presentes en una muestra.

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

Tradicionalmente la secuenciación de ADN se llevó a cabo por un método didesoxi, comúnmente conocido como el método de Sanger [Sanger et al, 1977], que utiliza 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 estrategias de secuenciación. Por ejemplo, la llegada de micromatrices de ADN hace posible la construcción de una serie de secuencias e hibridan secuencias complementarias en un proceso comúnmente conocido como secuenciación por hibridación. Otra técnica considerado estado actual de la técnica emplea la extensión del cebador seguido de la adición cíclica de un solo nucleótido con cada ciclo seguido de la detección del evento de incorporación. La técnica, comúnmente conocida como la secuenciación por síntesis o pirosecuenciación, incluyendo secuenciación fluorescente in situ (FISSEQ), es reiterativa en la práctica e implica un proceso en serie de ciclos repetidos de extensión de cebador mientras que la secuencia de nucleótidos objetivo se secuencia.

El documento US2002/120408 A1 se dirige hacia el control de infecciones en tiempo real a través de una red informática. El método comprende obtener una muestra de un microorganismo en un centro de atención de salud, secuenciar una primera región de un ácido nucleico de la muestra de microorganismo, comparar la primera región secuenciada con la secuencia de datos históricos almacenados en una base de datos, determinar una medida de la relación filogenética entre la muestra de microorganismos y muestras históricas almacenadas en la base de datos, y proporcionar información de control de infección basada en la determinación de las relaciones filogenéticas a las instalaciones de atención de salud para utilizar, controlar o prevenir la propagación de una infección.

Existe una necesidad de métodos y sistemas de identificación de genoma rápidos, que incluye comunicaciones electrónicas multidireccionales de datos de secuencias de ácidos nucleicos, datos clínicos, intervención terapéutica, y entrega a la medida de terapéuticos a la población adecuada para optimizar respuestas, conservar suministros médicos valiosos, y contener el bioterrorismo, la liberación inadvertida, y epidemias emergentes de patógenos.

El sistema presente está diseñado para analizar cualquier muestra que contenga material biológico para determinar la presencia de especies o genomas en la muestra. Esto se consigue mediante la obtención de información de la secuencia de material biológico y comparación de la información de secuencia en contra de una base de datos. La información de secuencia que coincide indicará la presencia de un genoma o especie. El emparejamiento probabilístico calculará la probabilidad de que se presenten especímenes. Los métodos se pueden aplicar en sistemas de secuenciación masivamente paralelos.

Resumen de la invención

La presente invención proporciona un método ex vivo de identificación de un material biológico en una muestra, que comprende:

(I) obtener una muestra que comprende dicho material biológico;

(II) extraer una o más molécula (s) de ácido nucleico a partir de dicha muestra;

(III) generar información de secuencia, que comprende una secuencia de un fragmento de nucleótidos de dicha una o más moléculas de ácido nucleico;

(IV) comparar dicha secuencia de un fragmento de nucleótidos con secuencias de ácidos nucleicos en una base de datos mediante comparación probabilística; y si dicha comparación de dicha secuencia de un fragmento de nucleótidos no resulta en una identificación de coincidencia del material biológico en dicha muestra, en virtud de la probabilidad de coincidencia del fragmento de nucleótidos de ser inferior a un umbral de una coincidencia de destino, entonces el método comprende adicionalmente:

(V) generar información de secuencia adicional de dicha una o más molécula (s) de ácido nucleico, en la que dicha información de secuencia adicional comprende una secuencia de un fragmento de nucleótido con uno o más nucleótidos adicionales;

(VI) comparar dicha información de secuencia adicional con secuencias de ácidos nucleicos en una base de datos inmediatamente después de la generación de dicha información de la secuencia adicional utilizando dicho emparejamiento probabilístico; y

(VII) repetir las etapas (v)-(vi) hasta que dé como resultado una coincidencia en la identificación del material biológico en dicha muestra.

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

En una segunda realización, dicha información de secuencia adicional comprende x nucleótidos adicionales, en donde x es menor que 50.

En una tercera forma de realización de la invención, dicha información de la secuencia adicional comprende x nucleótidos adicionales, en el que x es mayor que 50.

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

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

En una sexta realización, la invención comprende además la amplificación de dicho una o más moléculas de ácido nucleico para producir una pluralidad “i” de moléculas de ácido nucleico, antes de generar dicha información de secuencia.

La presente invención también está dirigida a un sistema para la detección de material biológico, que comprende:

(I) una unidad de recepción de muestras configurado para recibir una muestra que comprende material biológico;

(II) una unidad de extracción en comunicación con dicha unidad de recepción de muestra, dicha unidad de extracción está configurada para extraer por lo menos una molécula de ácido nucleico de dicha muestra;

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

(IV) una base de datos que comprende secuencias de ácidos nucleicos de referencia; y

(V) una unidad de procesamiento en comunicación con dicho casete de secuenciación y dicha base de datos, caracterizado porque dicha unidad de procesamiento está configurada para realizar las etapas (iv) a (vii) del método de la presente invención.

Breve descripción de los dibujos

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

La figura 1 es una ilustración esquemática de un sistema descrito.

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

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

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

La figura 5 es un diagrama de flujo que ilustra un proceso de funcionamiento del sistema de la figura 1.

La figura 6 es una ilustración esquemática de la interacción del sistema de la fig. 1 con diversas entidades potencialmente implicadas en el sistema.

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

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

La figura 9 muestra el porcentaje de secuencias únicas como una función de longitud de lectura.

La figura 10 es un resumen de las principales etapas de secuenciación.

Descripción detallada de la invención

El método y el sistema descrito en la presente invención utilizan la información de secuencia única más corta, que en una mezcla de ácidos nucleicos en una muestra no caracterizada tiene la longitud única mínima (n) con respecto a la información de secuencia completa generada o recogida. Además de las secuencias de longitud únicas, también se comparan las no únicas. La probabilidad de la identificación de un genoma aumenta con múltiples coincidencias. Algunos genomas tendrán secuencias únicas mínimas más largas que otros genomas. El método de emparejamiento de secuencias de longitud corta (n) continúa en paralelo con la recolección o generación de información de secuencia. Las comparaciones ocurren tan rápido como se generan o recogen secuencias (en tiempo real) posteriores más largas. Esto se traduce en una considerable reducción del espacio de decisión, porque los cálculos se realizan temprano en términos de generación/recolección de información de secuencia. El emparejamiento probabilístico puede incluir, pero no limitarse a, emparejamiento perfecto, singularidad posterior, emparejamiento patrón, emparejamiento de sub-secuencia múltiple dentro de longitud n, emparejamiento inexacto, mediciones de distancia, de semillas y extendida, y mapeo de árbol filogenético. Esto proporciona conductos automatizados para que coincidan con la información de secuencia tan rápido como se genera o en tiempo real. El instrumento de secuenciación puede continuar recolectando más cadenas y cadenas más largas de información secuencial en paralelo con la comparación. También se puede comparar información de la secuencia posterior y puede aumentar la confianza de un genoma o identificación de una especie en la muestra. No se necesita que el método espere para el montaje de información de secuencia de lecturas cortas en cóntigos más grandes.

El sistema y el método descrito en este documento proporcionan la ingesta de ácido nucleico, el aislamiento y la separación, la secuenciación de ADN, la creación de redes de bases de datos, procesamiento de la información, almacenamiento de datos, visualización de datos y la comunicación electrónica para acelerar la entrega de los datos relevantes que permitan un diagnóstico o identificación de organismos con aplicaciones de brotes de patógenos y las respuestas apropiadas. El sistema incluye un dispositivo de secuenciación portátil que transmite electrónicamente datos a una base de datos para la identificación de organismos relacionados con la determinación de la secuencia de ácidos nucleicos y otras moléculas de tipo polimérico o de cadena y el emparejamiento de datos probabilísticos.

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

El dispositivo está configurado para recibir una muestra de sujeto (SS) y una muestra de medio ambiente (ES), respectivamente. La muestra de sujeto (tal como sangre, saliva, etc.), puede incluir ADN del sujeto, así como el ADN de cualquier organismo (patógenos o no) en el sujeto. La muestra de medio ambiente (ES) puede incluir, pero no limitado a, los organismos en su estado natural en el medio ambiente (en especial la alimentación, el aire, el agua, el suelo, el tejido). Ambas muestras (SS, ES) pueden verse afectados por un acto de bioterrorismo o por una epidemia emergente. Ambas muestras (SS, ES) se recogen de forma simultánea a través de un tubo o un hisopo y se reciben en una solución o sólido (como una perla) en una membrana o un portaobjetos, placa, capilar, o canal. Las muestras (SS, ES) se secuenciaron entonces simultáneamente. Situaciones específicas circunstanciales pueden requerir el análisis de una muestra compuesta de una mezcla de las muestras (SS, ES). Un primer nivel de respuesta puede hacer contacto una vez que se ha identificado un emparejamiento probabilístico y/o durante la recolección en tiempo real de datos e interpretación de los mismos. A medida que pasa el tiempo un porcentaje creciente de la secuencia puede ser identificado.

El dispositivo 105 de secuenciación puede incluir los siguientes componentes funcionales, como se ilustra en la figura 3, que permiten que el dispositivo 105 analice una muestra del sujeto (SS) y una muestra del medio ambiente (ES), y comunique el análisis resultante a una red 110 de comunicación.

Receptores de muestras 120 y 122 están acoplados un Bloque 130 de Aislamiento y Extracción de ADN, que luego entrega las muestras al Bloque 130 por medio de un sistema de flujo. El Bloque 130 extrae ADN de las muestras y lo aísla de modo que se puede procesar y analizar más. 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 suministrarse en combinación con las muestras 120, 122 utilizando tecnología conocida de transporte de fluidos. Los ácidos nucleicos en las muestras 120, 122 están separados por el Bloque 130 de Aislamiento y Extracción, produciendo una corriente de fragmentos de nucleótidos o moléculas individuales no amplificadas. Una realización podría incluir el uso de métodos de amplificación.

Un casete 140 intercambiable se puede acoplar de manera desmontable al dispositivo 105 de secuenciación y el bloque 130. El casete 140 puede recibir la corriente de moléculas desde el bloque 130 y se puede secuenciar el ADN y producir datos de secuencias de ADN.

El casete 140 intercambiable se puede acoplar a, y proporcionar los datos de la secuencia de ADN al procesador 160, donde se realiza el emparejamiento probabilístico. Una realización podría incluir el rendimiento de 16 GB de datos transferidos a una velocidad de 1 Mb/seg. Se prefiere que un casete 140 de secuenciación obtenga la información de secuencia. Se pueden intercambiar diferentes casetes que representan diferentes métodos de secuenciación. La información de la secuencia se compara a través de emparejamiento probabilístico. Algoritmos de emparejamiento ultra-rápido y bases de datos de firmas ponderadas generados previamente comparan los datos de secuencia de novo con los datos de secuencias 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 la secuencia de ADN para ser almacenados en un almacén de datos (memoria) 170. El almacén de datos 170 también puede incluir cualesquier tipos adecuados o formas 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 a la que el procesador 160 puede enviar datos recuperados del almacén 170 de datos. El componente 180 de comunicación puede incluir cualquier tecnología adecuada para la comunicación con la red 110 de comunicación, tal como cableada, inalámbrica, satelital, etc.

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

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

El casete 140 intercambiable se ilustra esquemáticamente en más detalle en la figura 3. El casete 140 puede estar acoplado de manera desmontable a un dispositivo 105 de secuenciación y el bloque 130 e incluye un método de secuenciación del estado de la técnica (es decir, secuenciación de alto rendimiento). Se pueden construir sistemas basados en estado sólido o química húmeda en la platina a través de una forma de casete intercambiable del tipo "conexión y arranque". El casete 140 puede recibir el flujo de moléculas desde el bloque 130 y puede secuenciar el ADN mediante el método de secuenciación y puede producir datos de secuencias de ADN. Las realizaciones incluyen métodos basados en, pero no limitado a, la secuenciación por síntesis, secuenciación por ligación, secuenciación de una sola molécula y pirosecuenciación. Otra realización incluye una fuente de campo 142 eléctrico y se 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 fluorescente 144 a través de la corriente de ADN. El casete incluye además un sensor biomédico (detector) 146 para detectar la emisión de luz fluorescente y para la 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 adecuadas para grupos funcionales etiquetados para su secuenciación.

La detección fluorescente comprende la medición de la señal de un grupo funcional etiquetado de por lo menos uno de uno o más nucleótidos o análogos de nucleótidos. La secuenciación que utiliza nucleótidos fluorescentes típicamente implica fotoblanqueo la etiqueta fluorescente después de detectar una adición de nucleótido. Las realizaciones pueden incluir métodos fluorescentes basados en perlas, FRET, etiquetas infrarrojas, pirofosfatasa, ligasa que incluye nucleótidos etiquetados 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 construye con cada adición de un nucleótido marcado con fluorescencia a la cadena del cebador. Al reducir la intensidad de la señal, se secuencian opcionalmente las plantillas de ADN más largas.

El fotoblanqueo incluye la aplicación de un pulso de luz al cebador de ácido nucleico en el que se ha incorporado un nucleótido fluorescente. El pulso de luz comprende típicamente 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 unos 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 etiquetados con fluorescencia y/o el ácido nucleico o cebador etiquetado 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 la señal en varios ciclos.

El sensor (detector) 146 monitoriza opcionalmente por lo menos una señal de la plantilla de ácido nucleico. El sensor (detector) 146 incluye opcionalmente o se conecta funcionalmente con un ordenador que incluye software para la conversión de información de la señal del detector en información de resultado de secuenciación, por ejemplo, concentración de un nucleótido, identidad de un nucleótido, secuencia de nucleótidos de la plantilla, etc. Además, opcionalmente se calibran señales de muestra, por ejemplo, al calibrar el sistema microfluídico mediante el control de una señal de una fuente conocida.

Como se muestra en la figura 2, el dispositivo 105 de secuenciación puede comunicarse a través de una red 110 de comunicación 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 una Primera Respuesta (es decir, la Red de Laboratorios de Respuesta (es decir, Laboratorios de Referencia, Laboratorios Seminales, Laboratorios Nacionales), GenBank ®, Centro para el Control de Enfermedades (CDC), médicos, personal de salud pública, registros médicos, datos del censo, aplicación 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 figura 4 que ilustra una vista anterior del dispositivo. El dispositivo es un dispositivo portátil de secuenciación y se ilustra en comparación con el tamaño de monedas C. El dispositivo 105 es de aproximadamente 11 pulgadas de largo y fácilmente transportable. (En la figura 4, se muestran monedas a escala.) Dos puertos 153, 154 están situados a un lado del dispositivo y representa a los receptores 120, 122 de muestras. El puerto 153 es para recibir una muestra de sujetos (SS) o una muestra de medio ambiente (ES) que se va a analizar y a secuenciar. El puerto 154 es para el control de la secuencia (SC). Los dos puertos diferentes están diseñados para determinar si una muestra de sujeto (SS) o muestra de medio ambiente (ES) contiene materiales que resultan en falta de secuenciación, si ocurre fallo en secuenciación, o función 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 está en la forma de una almohadilla de contacto, sin embargo, se puede utilizar cualquier tecnología adecuada. La almohadilla de contacto incluye botones 150a para la presentación visual, 150b, 150c para los datos de grabación, 150d para la transmisión de datos en tiempo real y recepción, y 150e de control de potencia para activar o desactivar el dispositivo. Alternativamente, el teclado puede ser incorporado en la pantalla de visualización y se pueden controlar todas las funciones mediante la interfaz de cristal líquido. Las técnicas adecuadas se describen en el documento Publicación de Patente de Los Estados Unidos No. 2007/0263163, cuya descripción se incorpora aquí mediante referencia. Esto puede ser por métodos de

emparejamiento de dispositivos habilitados para Bluetooth o similares. Las funciones incluyen teclas numéricas, marcados con las letras del alfabeto, tal como en el lugar común en los teclados de teléfono, tal como una tecla de borrar, barra espaciadora, tecla de escape, tecla de impresión, tecla Intro, caracteres adicionales, arriba/abajo, izquierda/derecha, y cualquier otro deseado por el usuario. El dispositivo incluye además un módulo 152 de salida de usuario, en la forma de una pantalla de visualización, para visualizar información por 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 los botones/teclas es controlar todos los aspectos de la secuenciación de muestras, transmisión de datos y emparejamiento probabilístico y controles de interfaz, que incluyen pero no limitan a encendido/apagado, enviar, tecla de navegación, teclas de función, funciones de pantalla LCD y borrado y herramientas de visualización con rango de genoma calculados por algoritmos para la lista de confianza de emparejamientos. Una realización incluye un sistema basado en Internet, donde varios usuarios pueden transmitir/recibir datos simultáneamente hacia/desde un motor de búsqueda de red jerárquica.

La figura 5 es un diagrama de flujo que ilustra un proceso de funcionamiento del sistema 100 de una realización del sistema 100 como se describió anteriormente. Como se muestra en la figura 5, un proceso de la operación del dispositivo incluye en 200 recibir las muestras de sujetos (SS) recogidas y la muestra de medio ambiente (ES) en los receptores 120, 122 de la muestra. En 202, las muestras proceden a extracción de ADN y aislamiento del Bloque 130 de donde se analiza la muestra y el ADN se extrae de las muestras y se aísla. En 203, el casete 140 intercambiable recibe el ADN aislado a partir del bloque 130 y las secuencias de ADN. Dependiendo del casete y, si es necesario, con la aplicación de un campo 142 eléctrico y de una luz 144 fluorescente de, un sensor 146 biomédico dentro del casete 140 detecta/determina la secuencia de ADN de la corriente de muestra. En 204, los datos de secuenciado se procesan y se almacenan en un almacén 170 de datos. En 205, los datos de secuenciado se comparan a través de emparejamiento probabilístico y se lleva a cabo la identificación del genoma. El proceso tiene naturaleza reiterativa. La información resultante se puede transmitir a través de una red 110 de comunicación. Los datos GPS (sistema de posicionamiento global) pueden estar opcionalmente transmitidos también en la etapa 205. En 206, el dispositivo recibe electrónicamente los datos del emparejamiento. En 207, el dispositivo muestra visualmente los datos recibidos electrónicamente de emparejamiento a través de un módulo 152 de salida de usuario. Si se requiere más análisis, en 208, los datos de secuencia se transmiten electrónicamente a las entidades de interpretación de datos (es decir, Personal de Salud Pública, Informes Médicos, etc.) a través de la red de comunicación.

Un enfoque de investigación de múltiples métodos puede mejorar la rápida respuesta a un incidente e integrar la atención primaria con la detección de organismos. Se puede utilizar una respuesta triangular, que consiste en datos de instrumentos cuantitativos de la secuenciación del ADN para converger con cuidados críticos cualitativos. Se puede utilizar una infraestructura de listas de verificación y auditorías de los datos de secuenciación de ADN recogidas en el campo a través de múltiples sitios para comparar la aparición de un organismo, por ejemplo, una amenaza biológica entre ubicaciones. El análisis estadístico inferencial de los datos genómicos se puede combinar con observaciones médicas para desarrollar categorías de prioridades. La información recopilada y compartida entre las bases de datos de los centros médicos y centros de genómica 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 apropiado.

La figura 6 ilustra la interacción entre el sistema 100 y diversas 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 triangulados secuenciados (310) que ilustran los “Datos de muestra”, los “Datos del Paciente”, e “intervención de tratamiento”. El dispositivo 105 puede transmitir y recibir datos de secuencias de ADN hacia y desde recursos 320 de emparejamiento de secuencias que incluyen GenBank ® y una red de Laboratorios de Respuesta, que incluyen Laboratorios de Vigilancia, Laboratorios de Referencia, y Laboratorios Nacionales.

Cada uno de los laboratorios tiene funciones específicas. Los Laboratorios de Vigilancia (hospital y otros laboratorios centros clínicos comunitarios) son responsables de descartar o referirse a agentes críticos que se enfrentan a laboratorios de referencia de LRN. Los laboratorios de referencia (laboratorios de salud pública estatales y locales en donde se observan prácticas del nivel de seguridad biológica de 3 (BSL-3)) realizan pruebas de confirmación (en regla). Los Laboratorios nacionales (BSL-4) mantienen una capacidad capaz de manejar agentes virales tal como Ébola y Variola major y llevar a cabo la caracterización definitiva.

El sistema 100 puede transmitir más y recibir datos hacia y desde los Recursos 330 de Interpretación de Datos, que incluyen entidades de orden público, personal de salud pública, registros médicos, y datos del censo. Por último, el dispositivo 105 puede transmitir y recibir datos hacia y desde un primer nivel de respuesta 320 que incluyen médicos o doctores en una sala de emergencias. El sistema 100 general está configurado para comunicarse con el Centro de Control de Enfermedades (CDC) 340 para proporcionar la información pertinente para el personal apropiado.

La figura 7 es una ilustración esquemática de la interacción funcional entre un dispositivo portátil electrónico de secuenciación con el centro de análisis remoto. El dispositivo 105 puede incluir una unidad 103 de llamada base

para procesamiento de secuenciación recibida por el casete 140 intercambiable. Tales secuencias y sitios SNP se ponderan de forma individual de acuerdo con su probabilidad que se encuentra en cada especie. Estas ponderaciones se pueden calcular teóricamente (mediante simulación) o experimentalmente. El dispositivo también incluye un procesador 109 de emparejamiento probabilístico acoplado a la unidad 103 base. El emparejamiento probabilístico se realiza en tiempo real o tan rápido como la llamada de base de secuencia o la recolección de datos de secuencia. El procesador 109 de emparejamiento probabilístico, utilizando un método bayesiano, puede recibir datos de calidad y de secuencias resultantes, y puede calcular las probabilidades para cada lectura de secuenciación teniendo en cuenta las puntuaciones de calidad de secuenciación generadas por la unidad 103 de llamada base. El procesador 109 de emparejamiento probabilístico puede utilizar una base de datos generada y optimizada antes de su uso la identificación de patógenos. Un sistema 107 de alerta se acopla al procesador 109 de emparejamiento probabilístico y puede recopilar información desde el procesador 109 de emparejamiento probabilístico (en el sitio) y mostrar el mejor organismo emparejado en tiempo real.

El sistema 107 de alerta está configurado para a los datos del paciente, es decir, la evaluación del riesgo o de diagnóstico médico para datos particulares de un paciente en los desde el punto de pruebas o ensayos de diagnóstico para el cuidado, que incluye inmunoensayos, electrocardiogramas, radiografías y otras pruebas, y proporcionar 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 riesgo. 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 se puede basar en la mitigación o neutralización efectiva del agente biológico y/o sus efectos secundarios y se basa en la historia clínica del paciente si hay contraindicaciones. El sistema de alerta se puede basar en el grado y el número de ocurrencias. El número de ocurrencias se puede basar en la identificación genómica del agente biológico. Un valor puede ser pronunciado cuando el resultado está dentro o supera un umbral como se determina por las agencias gubernamentales, tales como la CDC o DoD o 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 el flujo de información rápido y preciso para tomar acciones por los primeros en responder u otros sistemas clínicos.

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

El dispositivo 105 de secuenciación interactúa con un centro 400 de análisis remoto, que puede recibir electrónicamente datos transferidos desde el componente 180 de comunicación del dispositivo 105 de secuenciación a través de un método de comunicación inalámbrica y/o por cable. El centro 400 de análisis remoto contiene una gran base de datos de secuencia que incluye todas las secuencias de nucleótidos y de aminoácidos y datos SNP disponibles hasta la fecha. Esta base de datos también contiene información epidemiológica y terapéutica asociada (por ejemplo, resistencia a antibióticos). El centro 400 de análisis remoto incluye además un almacén 401 de datos. El almacén 401 de datos puede recibir información de datos de secuencias descomprimidas a través de la transmisión electrónica del componente 180 de comunicación del dispositivo 105. Se acopla un ensamble 402 de genoma al almacén 401 de datos y puede ensamblar los datos de secuencia descomprimidos. ADN contaminante obvio, como el ADN humano, se puede filtrar antes de su posterior análisis.

El centro 400 de análisis remoto incluye además un procesador 403 equipado con tecnología de emparejamiento probabilístico y algoritmos de búsqueda de homología, que se puede emplear para analizar los datos de secuencia reunidos para obtener las probabilidades de presencia de patógenos 403a objetivo, estructura 403b de comunidad, información 403c epidemiológica y terapéutica. Los datos de la secuencia del genoma de patógenos objetivo se comparan con los de los genomas de los no patógenos, que incluye metagenoma humano para identificar secuencias de nucleótidos y sitios polimórficos de nucleótidos únicos (SNP), que sólo se producen en los organismos objetivos. El análisis en el centro 400 de análisis remoto se lleva a cabo 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 de comunicación 404 desde la que los resultados del análisis se transfieren electrónicamente de vuelta al sistema 107 de alerta dentro del dispositivo 105 de secuenciación, así como otras autoridades (por ejemplo, el DHS, CDC, etc.).

Clasificación Probabilística: La presente invención proporciona motores de bases de datos, diseños de bases de datos, técnicas de filtrado y el uso de la teoría de la probabilidad como Lógica Extendida. El sistema y métodos actuales utilizan los principios de la teoría de probabilidad para hacer razonamiento convincente (decisiones) sobre los datos producidos por secuenciación de ácidos nucleicos. Utilizando el método de la teoría de probabilidad, el sistema descrito en el presente documento analiza los datos tan pronto como se alcanza un número mínimo de nucleótidos de longitud (n), y calcular la probabilidad del n-mero, cada incremento en longitud posterior (n + pares de bases) se utiliza para calcular la probabilidad de una coincidencia de secuencia. El cálculo de cada n-mero y n-meros posteriores más largos se procesa adicionalmente para volver a calcular las probabilidades de todos los aumentos de longitudes para identificar la presencia de genomas. A medida que aumenta la longitud de la unidad,

múltiples subunidades, en el n-mero se comparan para reconocimiento de patrones, lo que aumenta aún más la probabilidad de una coincidencia. Tal método, incluye otros métodos bayesianos, prevé la eliminación de coincidencias y la identificación de un número significativo de muestras biológicas que comprenden un fragmento de nucleótidos muy corto o lectura sin tener que completar la secuenciación del genoma completo o ensamblar el genoma. Como tal asignar la probabilidad de coincidencia con organismos existentes y pasar a la siguiente lectura de secuencia de ácido nucleico para mejorar aún más la probabilidad de emparejamiento. El sistema descrito en el presente documento aumenta la velocidad, reduce el consumo de reactivos, permite la miniaturización, y reduce significativamente la cantidad de tiempo necesario para identificar el organismo.

Con el fin de construir clasificadores probabilísticos para tomar una decisión sobre secuencias de ácidos nucleicos cortas, se puede utilizar una variedad de métodos para filtrar primero y clasificar después 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 en forma 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). Una arista dirigida entre dos nodos indica una dependencia probabilística de la variable denotada por el nodo padre al hijo. En consecuencia, la estructura de la red indica la suposición de que cada nodo en la red es condicionalmente independiente de sus no descendientes dados sus progenitores. Para describir una distribución de probabilidad que satisface estas suposiciones, cada nodo en la red se asocia con una tabla de probabilidad condicional, que especifica la distribución a través de cualquier posible asignación dada de valores a sus progenitores. En este caso un clasificador bayesiano es una red bayesiana aplicada a una tarea de clasificación de calcular la probabilidad de cada uno de los nucleótidos proporcionada por cualquier sistema de secuenciación. En cada punto de decisión el clasificador bayesiano se puede combinar con una versión de algoritmo de gráfico de ruta más corta tal como de Dijkstra o de Floyd.

El sistema actual puede implementar un sistema de clasificadores bayesianos (por ejemplo, clasificador bayesiano ingenuo, clasificador bayesiano y clasificador de estimación bayesiano recursivo) y fusionar los datos resultantes en la base de datos de decisiones. Después se fusionan los datos, cada clasificador puede alimentar un nuevo conjunto de resultados con probabilidades actualizadas.

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

Fragmento de secuenciación de ADN: Se pueden utilizar métodos de secuenciación para generar la información de fragmento de secuencia. El módulo, 160 en la figura 2 o 109 en la figura 7 son responsables de procesamiento de datos entrantes del módulo de secuenciación en el casete intercambiable. Los datos se encapsulan con datos de secuenciación, así como información sobre inicio y parada de la secuencia, ID de secuencia, ID de cadena de ADN. El módulo da formato a los datos y los pasa al módulo de filtro de taxonomía. El formato incluye la adición de los datos de sistema y alineación en trozos.

El módulo de secuenciación de ADN tiene 2 interfaces. Conexión al módulo de Preparación de ADN y al Filtro de taxonomía.

I. Interfaz de Preparación de ADN: Se puede integrar varios métodos disponibles en el mercado para llevar a cabo la preparación de muestras a través de técnicas de microfluido. La preparación de muestras típicas se basa en solución e incluye lisis celular y eliminación de inhibidor. Los ácidos nucleicos se recuperan o extraen y se concentran. Las realizaciones de lisis incluyen métodos de detergentes/enzimas, mecánicos, microondas, presión y/o ultrasónicos. Las realizaciones de extracción incluyen afinidad en fase sólida y/o exclusión por 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 de árbol filogenético: Este submódulo de interfaces de filtro de taxonomía con “Base de datos de Decisiones” para conocer los resultados de la ronda previa de análisis. Si no se encuentran resultados el módulo pasa a los nuevos datos al módulo de clasificación. Si se encuentran 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 se trata de una secuencia de ADN de virus que se está examinando, el espacio de decisión para el clasificador se redujo a datos virales solamente. Esto se puede hacer al modificar los clasificadores bayesianos de datos recogidos durante el funcionamiento.

Aprendizaje de Máquina: Se organizan algoritmos de aprendizaje de máquina en una taxonomía, con base en el resultado deseado del algoritmo. (i) aprendizaje supervisado, en el que el algoritmo genera una función que mapea las entradas a las salidas deseadas. Una formulación estándar supervisada de la tarea de aprendizaje es el problema de clasificación: se requiere que el módulo de aprendizaje aprenda (aproximadamente) el comportamiento de una función que mapea un vector $[X_1, X_2, ... X_n]$ en una de varias clases al buscar en varios ejemplos de entrada-salida de la función. (ii) aprendizaje semisupervisado; que combina tanto ejemplos etiquetados y no etiquetados para generar una función o clasificador apropiado. (iii) Refuerzo de aprendizaje, en el que el algoritmo aprende de una política de cómo actuar dada una observación del mundo. Todas las acciones tienen cierto impacto en el medio

ambiente, y el medio ambiente proporciona retroalimentación que guía al algoritmo de aprendizaje. (iv) Transducción, predice nuevos resultados en función de las entradas de datos de capacitación, resultados de capacitación, y entradas de prueba que están disponibles durante la capacitación. (v) Aprender a aprender, en el que el algoritmo aprende de su propio sesgo inductivo basado en la experiencia previa.

Módulo de Reserva de Taxonomía: El módulo reserva 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 mantiene toda la información en la base de datos SQL. La reserva de taxonomía se implementa como base de datos en memoria con tiempos de respuesta de micro segundos. Las consultas a la base de datos SQL se manejan en un subproceso separado del resto del sub-módulo. Información de reserva incluye el gráfico de red creado por el módulo de filtro de taxonomía. El gráfico contiene toda la taxonomía cuando el sistema inicia el análisis. El análisis de secuencia de ADN reduce el gráfico de taxonomía con reserva de taxonomía que implementa las reducciones en el tamaño de datos y la eliminación de los grupos de datos apropiados.

Selector Clasificador: El actual 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 análisis desde el módulo de la secuenciación de ADN. El filtro taxonomía actúa como datos que pasan a través de los datos de secuenciación de ADN.

Clasificador bayesiano recursivo: El clasificador bayesiano recursivo es un método probabilístico para la estimación de una función de densidad de probabilidad desconocida de forma recursiva en el tiempo utilizando mediciones entrantes y un modelo de proceso matemático. El módulo recibe los datos del selector clasificador y de la Base de Datos de Decisiones en donde se almacenan decisiones anteriores. El grupo de datos se recupera de las bases de datos y la identificación de decisión previa colocados en la memoria local del módulo de donde se produce el filtrado. El clasificador toma una secuencia de ADN y trata de hacerla coincidir con o sin firmas, códigos de barras, etc., existentes a partir de la base de datos de taxonomía filtrando rápidamente las familias de los organismos que no coinciden. El algoritmo funciona mediante el cálculo de las probabilidades de múltiples creencias y ajuste de creencias basado en los datos entrantes. Los algoritmos utilizados en este módulo pueden incluir métodos secuenciales de Monte Carlo y remuestreo de importancia de muestreo. También se pueden utilizar filtros Hidden Markov Model, Ensemble Kalman y otros filtros de partículas junto con la técnica de actualización Bayesiana.

Clasificador bayesiano ingenuo: clasificador probabilístico simple basado en la aplicación del teorema de Bayes. El clasificador toma todas las decisiones con base en el conjunto de reglas predeterminado que se proporciona como entrada de usuario al inicio. El módulo se puede reinicializar con un nuevo conjunto de reglas mientras se está ejecutando el análisis. El nuevo conjunto de reglas puede venir por parte del usuario o puede ser un producto de fusión de reglas del módulo de fusiones de resultados.

Clasificador bayesiano de red: el clasificador bayesiano de red 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 una reserva de trabajo para la mayoría de módulos en el sistema. La mayoría de módulos tienen acceso directo a este recurso y pueden modificar sus regiones individuales. Sin embargo sólo el módulo de fusión de resultados puede acceder a todos los datos y modificar la regla bayesiana establecida en consecuencia.

Datos de Reglas Bayesianas: El módulo recoge todas las Reglas Bayesianas en forma binaria, compiladas previamente. Las reglas son de lectura-escritura para todos los clasificadores bayesianos, así como módulos de filtro de Taxonomía y de fusiones de Resultados. Las reglas se vuelven a compilar dinámicamente a medida que se realizan cambios.

Fusión de Resultados: El módulo fusiona la fecha de múltiples clasificadores bayesianos, así como otros clasificadores estadísticos que se utilizan. El módulo de Fusión de Resultados busca en la varianza media entre las respuestas generadas por 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 se puede utilizar sqLite para velocidad de base de datos más rápida.

Base de datos de Taxonomía: La base de datos incluirá múltiples bases de datos internas: árbol de taxonomía, árbol preprocesado indexado, entrada de usuario y reglas.

Reglas de Reserva: Reserva En Memoria de reglas postproceso proporcionadas por el usuario.

Gestión de Reglas: Gestión gráfica que hacer Interfaz con el módulo

Entrada de usuario: reglas de inferencia creadas por el usuario. Las reglas son utilizadas por clasificadores bayesianos para tomar decisiones.

En el presente documento se describen sistemas y métodos de la invención como incorporados en los programas de ordenador que tienen códigos para realizar una variedad de funciones diferentes. Determinadas tecnologías mejores en su clase (presentes o emergentes) pueden tener componentes de licencia. Los métodos existentes para la extracción de ADN incluyen el uso de fenol/cloroformo, precipitación de sales, uso de sales caotrópicas y resinas de sílice, uso de resinas de afinidad, cromatografía de intercambio iónico y uso de perlas magnéticas. Se describen métodos en la patente de EE.UU. Nos. 5.057,426, 4,923,978, Patentes EP 0512767 A1 y EP 0515484B y WO 95/13368, WO 97/10331 y WO 96/18731. Debe entenderse, sin embargo, que los sistemas y métodos no se limitan a un medio electrónico, y alternativamente se pueden practicar diversas funciones en un entorno manual. Los datos asociados con el proceso pueden ser transmitidos electrónicamente a través de una conexión de red utilizando Internet. Los sistemas y técnicas descritas anteriormente pueden ser útiles en muchos otros contextos, incluyendo los descritos a continuación.

Estudios de asociación de enfermedades: Muchas enfermedades y condiciones comunes implican factores genéticos complejos que interactúan para producir las características visibles de esa enfermedad, también denominado fenotipo. Múltiples genes y regiones reguladores se asocian a menudo con una enfermedad o síntoma particular. Mediante secuenciación de los genomas o genes seleccionados de muchos individuos con una condición determinada, puede ser posible identificar las mutaciones causales que subyacen a la enfermedad. Esta investigación puede conducir a avances en detección, prevención y tratamiento de la enfermedad.

Investigación del cáncer: La genética del cáncer implica la comprensión de los efectos de las mutaciones heredadas y adquiridas y otras alteraciones genéticas. El reto de diagnosticar y tratar el cáncer se complica aún más por la variabilidad del paciente individual y dificultad de predecir resultados a la terapia con medicamentos. La disponibilidad de secuenciación del genoma de bajo costo para caracterizar cambios adquiridos por el genoma que contribuyen al cáncer con base en pequeñas muestras o biopsias de células tumorales, pueden permitir un mejor diagnóstico y tratamiento del cáncer.

Investigación y desarrollo farmacéutico: Una promesa de la genómica ha sido la de acelerar el descubrimiento y desarrollo de nuevos fármacos más eficaces. El impacto de la genómica en esta área ha surgido lentamente debido a la complejidad de las rutas biológicas, mecanismos de la enfermedad y múltiples objetivos de los medicamentos. La secuenciación de una sola molécula podría permitir cribado de alto rendimiento de una manera rentable mediante análisis de expresión génica a gran escala para identificar mejor fármacos prometedores. En el desarrollo clínico, la tecnología descrita podría utilizarse para generar perfiles de genes individuales que pueden proporcionar información valiosa sobre la probable respuesta a la terapia, toxicología o riesgo de efectos adversos, y, posiblemente, facilitar la selección de pacientes e individualización de la terapia.

Enfermedades infecciosas: 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 médica y ambientalmente útil para el diagnóstico, tratamiento y seguimiento de infecciones y para predecir el potencial de resistencia a los medicamentos.

Condiciones autoinmunitarias: Se considera que varias enfermedades autoinmunitarias, que van desde esclerosis múltiple y lupus hasta riesgo de rechazo de trasplante, tienen un componente genético. La monitorización 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 diferente pronóstico y respuesta a los fármacos en función de sus diferencias genéticas subyacentes. La entrega de información genética específica del paciente abarca el diagnóstico molecular, que incluye kits y servicios de diagnóstico basados en expresión o genes, productos de diagnóstico de compañía para la selección y el seguimiento de terapias particulares, así como selección de pacientes para la detección temprana de la enfermedad y el seguimiento de la enfermedad. Crear pruebas de detección y diagnóstico molecular dirigidas y más eficaces requiere una mejor comprensión de los genes, factores reguladores y otros factores relacionados con enfermedades o con 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 genética para el descubrimiento, desarrollo y diseño de animales y cultivos genéticamente superiores. La industria de la agricultura ha sido un gran consumidor de tecnologías genéticas, en particular micromatrices genéticas, para identificar variaciones genéticas relevantes de variedades o poblaciones. La tecnología de secuenciación descrita puede proporcionar un método más poderoso, directo y rentable para análisis de expresión génica y estudios de población para esta industria.

Oportunidades adicionales estarán en el campo de las aplicaciones de repetición de secuencia en la que se aplican los métodos para detección de variaciones genéticas sutiles. El análisis genómico comparativo ampliado a través de las especies puede generar comprensión sobre la estructura y función del genoma humano y, en consecuencia, la genética de la salud y las enfermedades humanas. Se están expandiendo estudios de la variación genética humana y su relación con la salud y la enfermedad. La mayoría de estos estudios utilizan tecnologías que se basan en patrones de variación relativamente comunes conocidos. Estos métodos poderosos proporcionarán nueva información importante, pero son menos informativos que la determinación de la secuencia contigua completa, de genomas humanos individuales. Por ejemplo, es probable que los métodos de genotipificación actuales pierdan diferencias inusuales entre personas en cualquier ubicación genómica particular y tienen una capacidad limitada para determinar reordenamientos de largo alcance. La caracterización de los cambios somáticos del genoma que contribuyen al cáncer emplea actualmente combinaciones de tecnologías para la obtención de datos de secuencias (en muy pocos genes), además de información limitada sobre cambios de número de copias, reordenamientos, o la pérdida de heterocigocidad. Tales estudios sufren de mala resolución y/o cobertura incompleta del genoma. La heterogeneidad celular de las muestras de tumores presenta retos adicionales. La secuenciación completa del genoma a bajo coste a partir de muestras muy 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 proyecto piloto Cancer Genome Atlas (TCGA) puesto en marcha recientemente se mueve en la dirección deseada, pero sigue siendo limitado drásticamente por los costes de secuenciación. Se necesitan secuencias adicionales del genoma de animales y plantas de importancia agrícola para estudiar la variación individual, diferentes razas domésticas y diversas variantes silvestres de cada especie. El análisis de secuencia de las comunidades microbianas, de las que no se pueden cultivar muchos miembros, proporcionará una fuente rica de información médica y ambientalmente útil. Y la secuenciación rápida y precisa, puede ser el mejor método para la monitorización microbiana de los alimentos y el medio ambiente, incluyendo la rápida detección y 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 Firma Clave. En la actualidad, las secuencias de firma de clave están dirigidas a regiones exónicas.

Una aplicación fundamental de la tecnología de ADN implica diversas estrategias de etiquetado para etiquetar un ADN que se produce por una polimerasa de ADN. Esto es útil en la tecnología de micromatrices: secuenciación de ADN, la detección de SNP, clonación, análisis por PCR, y muchas otras aplicaciones.

Ejemplo 1

Propósito: Uso de firmas claves y/o códigos de barras para permitir la identificación del genoma con tan sólo 8-18 nucleótidos y análisis de datos de secuencias muy cortas (lecturas), en tiempo real.

Algoritmos de construcción de matriz de sufijo de tiempo lineal se utilizaron para calcular el análisis de singularidad. El análisis determinó el porcentaje de todas las secuencias que eran únicas en varios genomas modelo. Se analizaron todas las longitudes de secuencia en un genoma. Se tienen en cuenta secuencias que ocurren sólo una vez en un genoma. El algoritmo de matriz sufijo funciona mediante el cálculo de un gráfico de puntuación de repetición que analiza la frecuencia de subsecuencias específicas dentro de una secuencia que se produce sobre la base de una ventana deslizante de dos pares bases. La información de genoma almacenada en el GenBank se utilizó para análisis in-silico. Se analizó un genoma viral, Lambda-fago, un genoma bacteriano, E. coli K12 MG1655, y el genoma humano. El porcentaje de lecturas única es una función de longitud de secuencia. Se hizo una suposición en relación con las secuencias que sólo producen emparejamientos sin ambigüedades y que producen solapamientos ambiguos para reconstruir el genoma. Lecturas únicas varían en tamaño de 7 a 100 nucleótidos. La mayoría de tamaños únicos fueron más cortos de 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 tal manera que las secuencias 400 nt son 99% únicas para el fago. Esto disminuye al 80% para las secuencias de fagos de 10 nt. Para las bacterias de E. coli (secuencias) de 18 nt del genoma son 97% únicas para E. coli. Para genomas humanos, las secuencias de 25 nt son 80% únicas para humanos y un aumento a 45 nt resulta en el 90% del genoma como único.

Reivindicaciones

1. Un método ex vivo de identificación de un material biológico en una muestra, que comprende:

(i) obtener una muestra que comprende dicho material biológico;

(ii) extraer una o más moléculas de ácido nucleico a partir de dicha muestra;

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

(iv) comparar dicha secuencia de un fragmento de nucleótidos con secuencias de ácidos nucleicos en una base de datos que utiliza emparejamiento probabilístico; y si dicha comparación de dicha secuencia de un fragmento de nucleótidos no resulta en una coincidencia que identifica el material biológico en dicha muestra, en virtud de la probabilidad de coincidencia del fragmento de nucleótidos de ser inferior a un umbral de una coincidencia objetivo, entonces el método comprende adicionalmente:

(v) generar información de secuencia adicional de dicha una o más molécula (s) de ácido nucleico, en el que dicha información de secuencia adicional comprende una secuencia de un fragmento de nucleótidos con uno o más nucleótidos adicionales;

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

(vii) repetir las etapas (v)-(vi) hasta que resulta una coincidencia en la identificación del material biológico en dicha muestra.

2. El método de la reivindicación 1, en el que dicha información de secuencia adicional comprende una secuencia de un fragmento de nucleótidos que consiste en un nucleótido adicional.

3. El método de la reivindicación 1, en el que dicha información de secuencia adicional comprende x nucleótidos adicionales, en el que x es menor de 50.

4. El método de la reivindicación 1, en el que dicha información de secuencia adicional comprende x nucleótidos adicionales, en el que x es mayor de 50.

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

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

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

8. Un sistema para la detección de material biológico, que comprende:

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

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

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

(iv) una base de datos que comprende secuencias de ácidos nucleicos de referencia; y

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

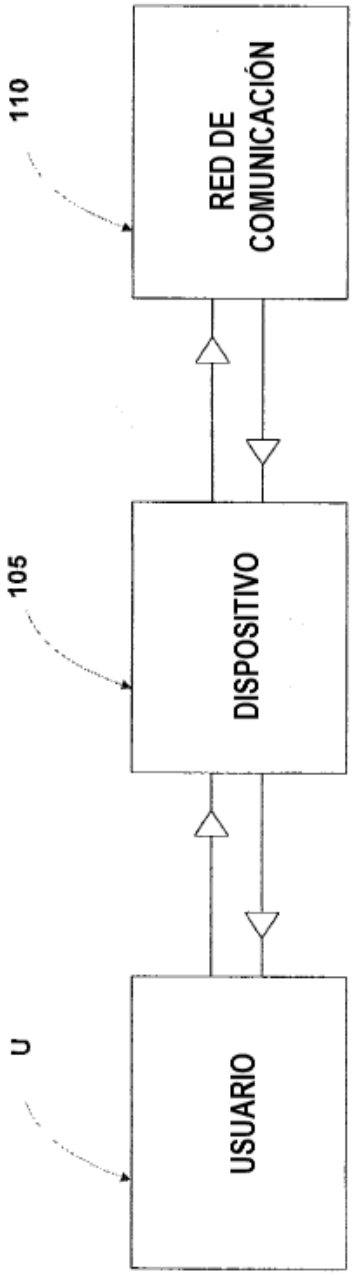

FIG. 1

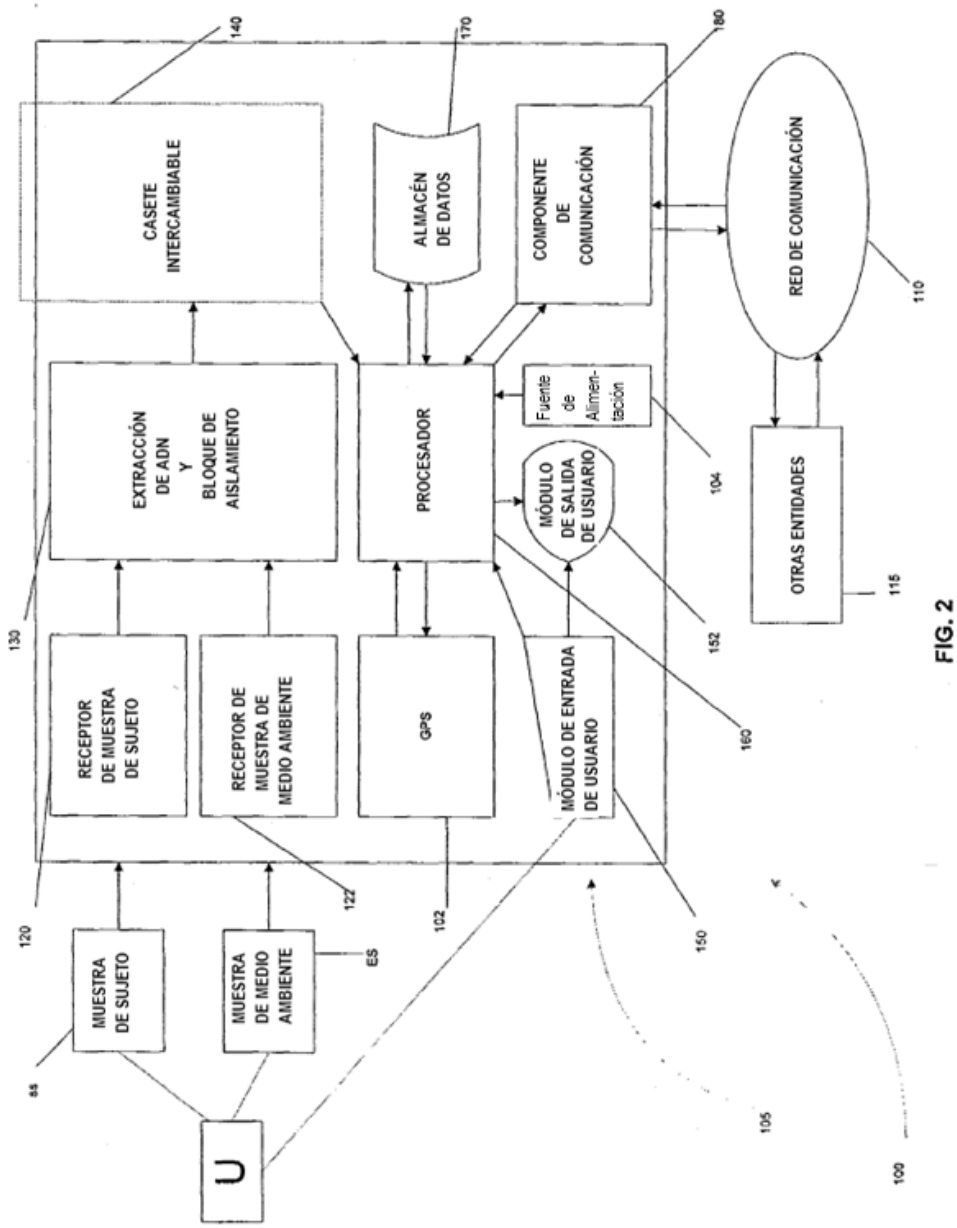
CASETE INTERCAMBIABLE
ALMACÉN DE DATOS
COMPONENTE DE COMUNICACIÓN
RED DE COMUNICACIÓN
EXTRACCIÓN DE ADN Y BLOQUE DE AISLAMIENTO
PROCESADOR
Fuente de Alimentación
MÓDULO DE SALIDA DE USUARIO
OTRAS ENTIDADES
RECEPTOR DE MUESTRA DE SUJETO
RECEPTOR DE MUESTRA DE MEDIO AMBIENTE
GPS
MÓDULO DE ENTRADA DE USUARIO
MUESTRA DE SUJETO
MUESTRA DE MEDIO AMBIENTE
U
140
170
180
110
130
104
152
115
160
120
122
102
150
105
100

FIG. 2

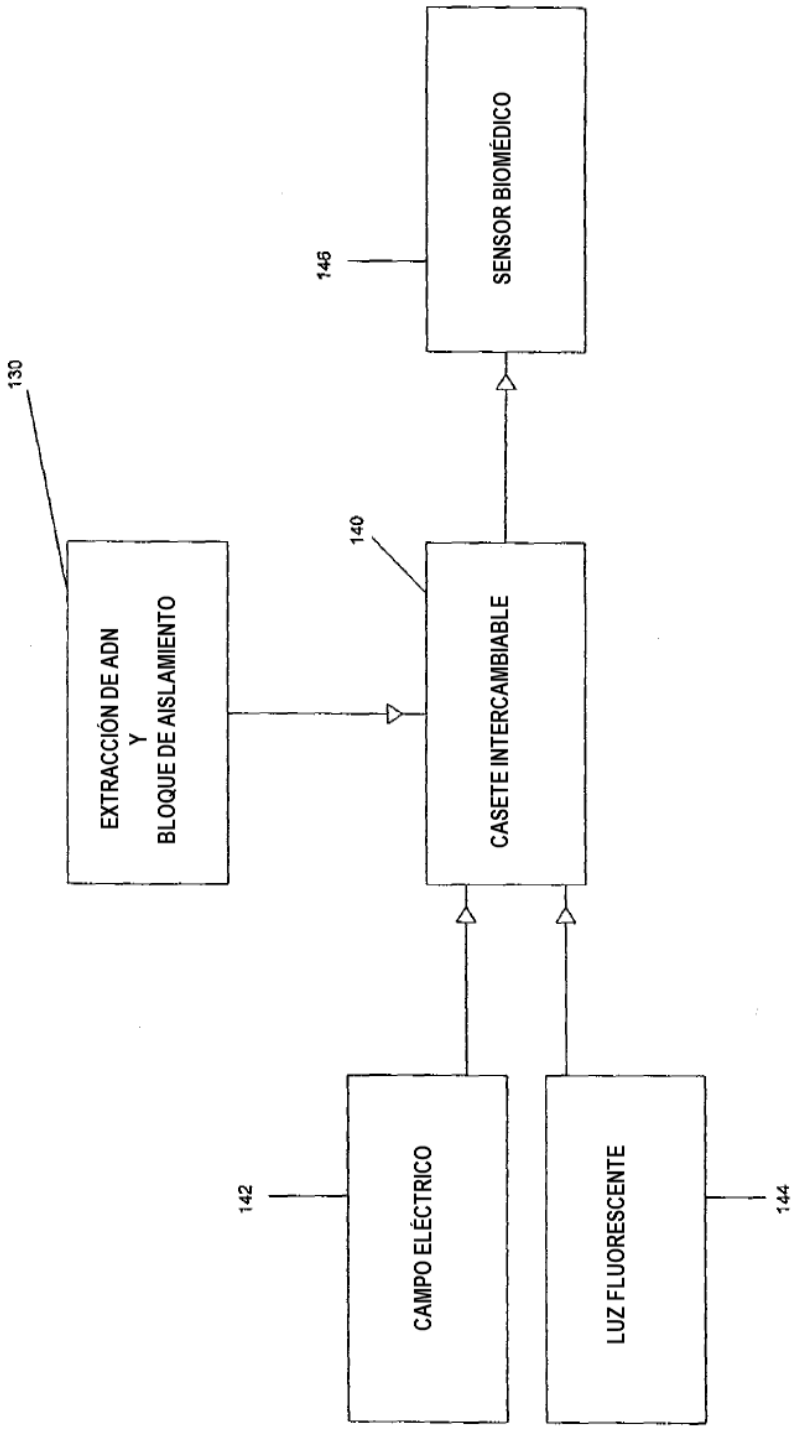
SENSOR BIOMÉDICO
146
130
140
EXTRACCIÓN DE ADN
Y
BLOQUE DE AISLAMIENTO
CASETE INTERCAMBIABLE
142
CAMPO ELÉCTRICO
LUZ FLUORESCENTE
144

FIG. 3

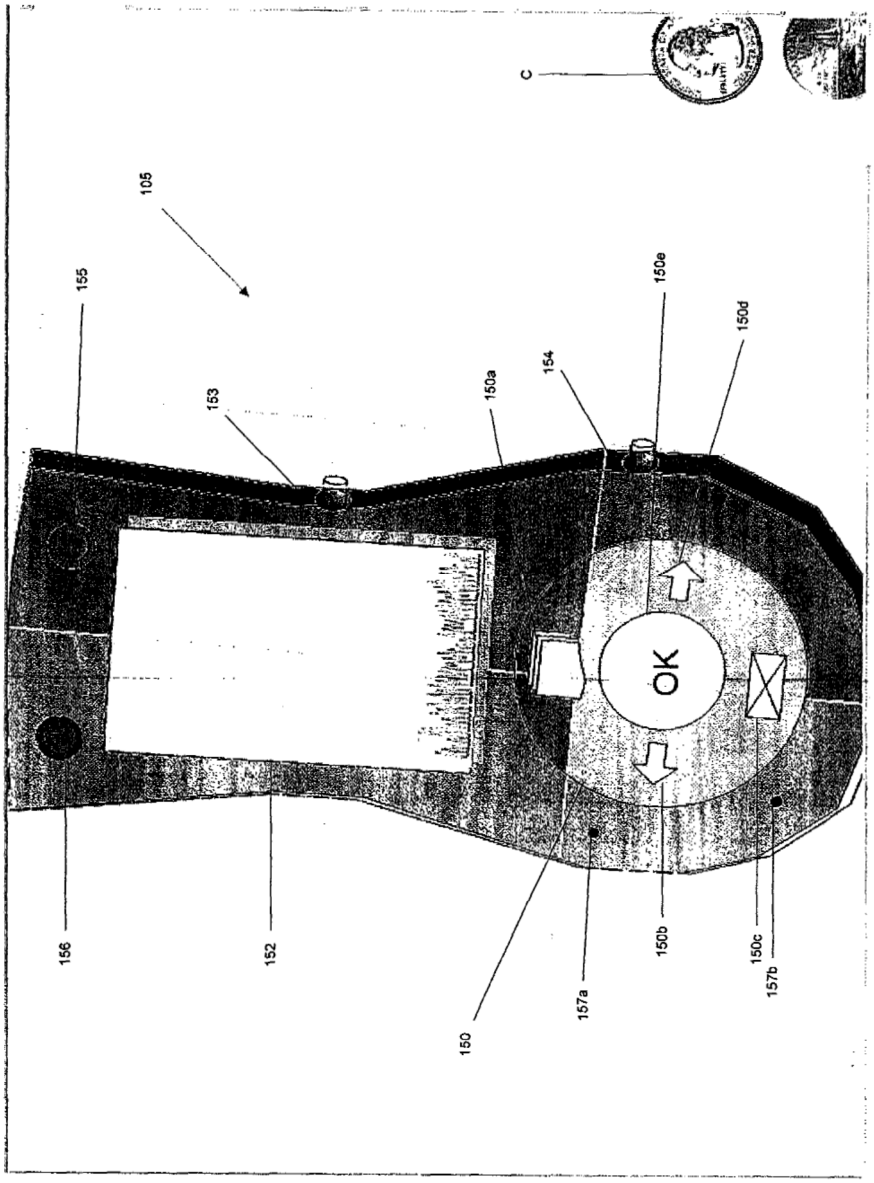
C
105
155
153
150a
154
150e
150d
OK
156
152
150
157a
150b
150c
157b

FIG. 4

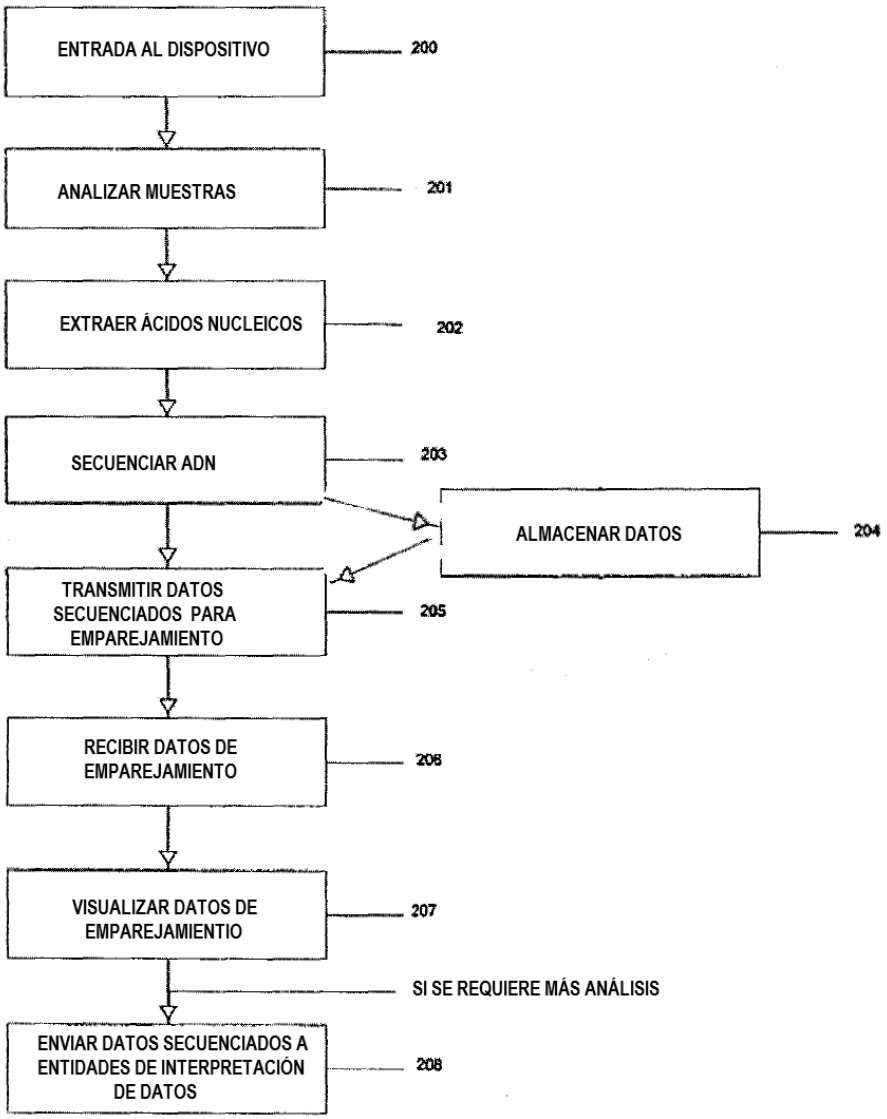
ENTRADA AL DISPOSITIVO
200
ANALIZAR MUESTRAS
201
EXTRAER ÁCIDOS NUCLEICOS
202
SECUENCIAR ADN
203
ALMACENAR DATOS
204
TRANSMITIR DATOS SECUENCIADOS PARA EMPAREJAMIENTO
205
RECIBIR DATOS DE EMPAREJAMIENTO
206
VISUALIZAR DATOS DE EMPAREJAMIENTIO
207
SI SE REQUIERE MÁS ANÁLISIS
ENVIAR DATOS SECUENCIADOS A ENTIDADES DE INTERPRETACIÓN DE DATOS
208

FIG. 5

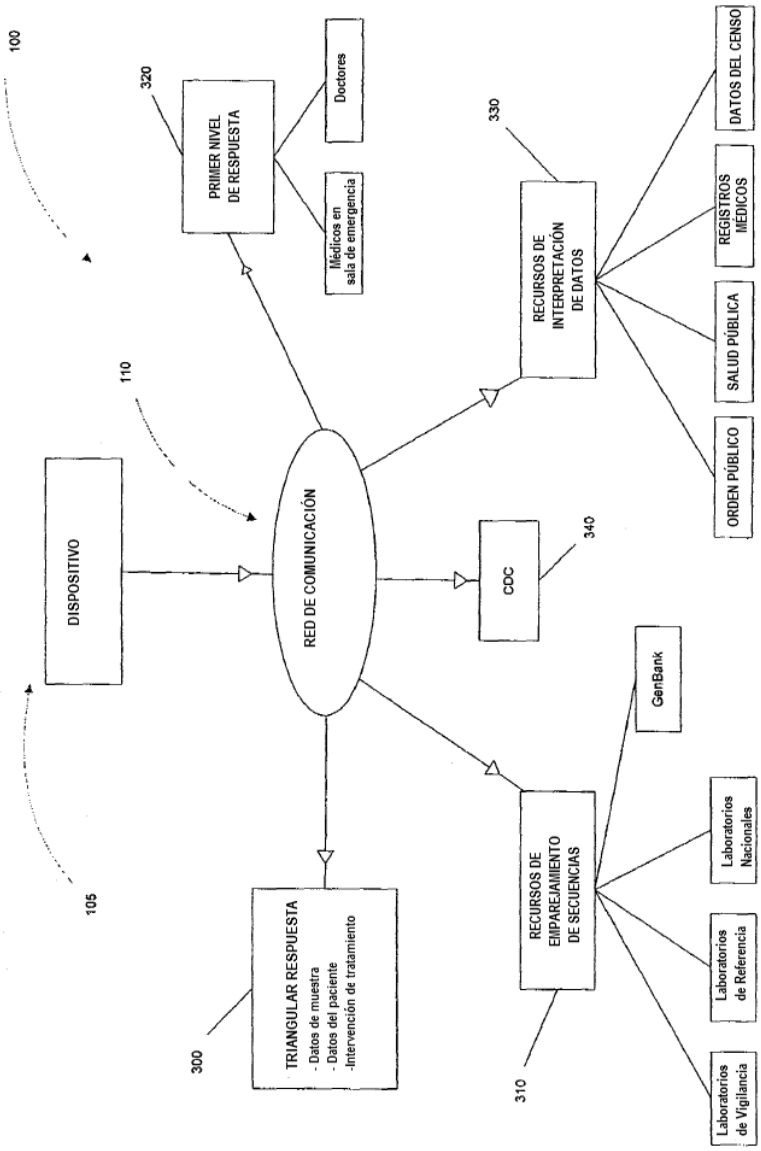
100
105
110
DISPOSITIVO
RED DE COMUNICACIÓN
TRIANGULAR RESPUESTA
- Datos de muestra
- Datos del paciente
-Intervención de tratamiento
300
PRIMER NIVEL DE RESPUESTA
320
Doctores
Médicos en sala de emergencia
CDC
340
RECURSOS DE INTERPRETACIÓN DE DATOS
330
DATOS DEL CENSO
REGISTROS MÉDICOS
SALUD PÚBLICA
ORDEN PÚBLICO
RECURSOS DE EMPAREJAMIENTO DE SECUENCIAS
310
GenBank
Laboratorios Nacionales
Laboratorios de Referencia
Laboratorios de Vigilancia

FIG. 6

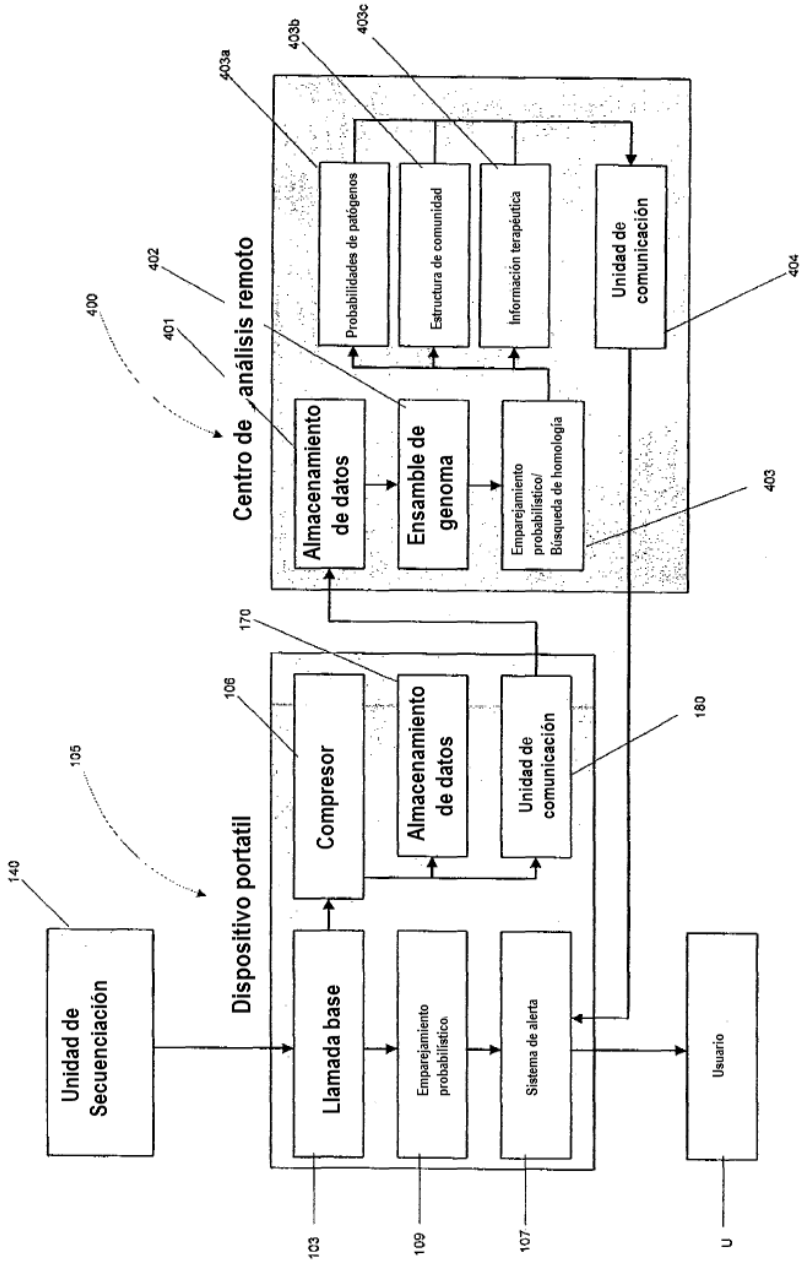
Unidad de Secuenciación
140
Dispositivo portatil
105
Llamada base
Compresor
106
Almacenamiento de datos
170
Unidad de comunicación
180
Emparejamiento probabilistico
Sistema de alerta
Usuario
103
109
107
U
Centro de análisis remoto
400
401
402
Almacenamiento de datos
Ensamble de genoma
Emparejamiento probabilistico/ Búsqueda de homologia
403
403a
403b
403c
Probabilidades de patógenos
Estructura de comunidad
Información terapéutica
Unidad de comunicación
404

FIG. 7

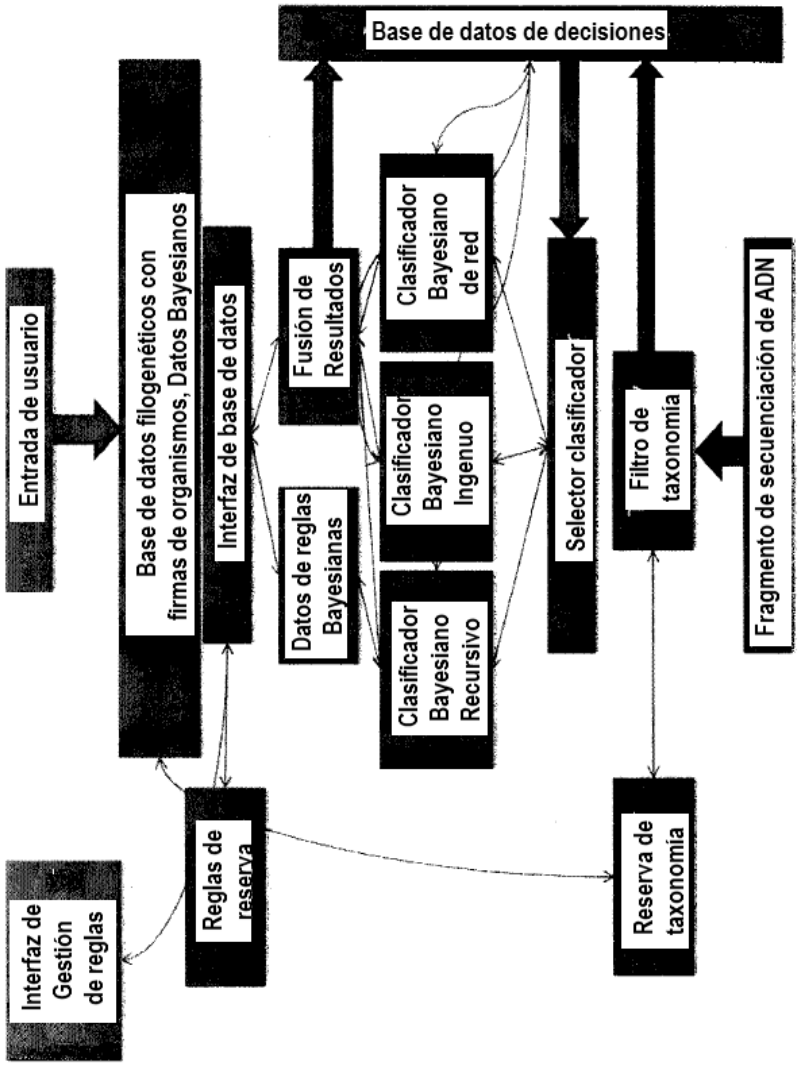
Base de datos de decisiones
Entrada de usuario
Base de datos filogenéticos con firmas de organismos, Datos Bayesianos
Interfaz de base de datos
Fusión de Resultados
Clasificador Bayesiano de red
Clasificador Bayesiano Ingenuo
Clasificador Bayesiano Recursivo
Datos de reglas Bayesianas
Selector clasificador
Filtro de taxonomía
Fragmento de secuenciación de ADN
Interfaz de Gestión de reglas
Reglas de reserva
Reserva de taxonomía

FIG. 8

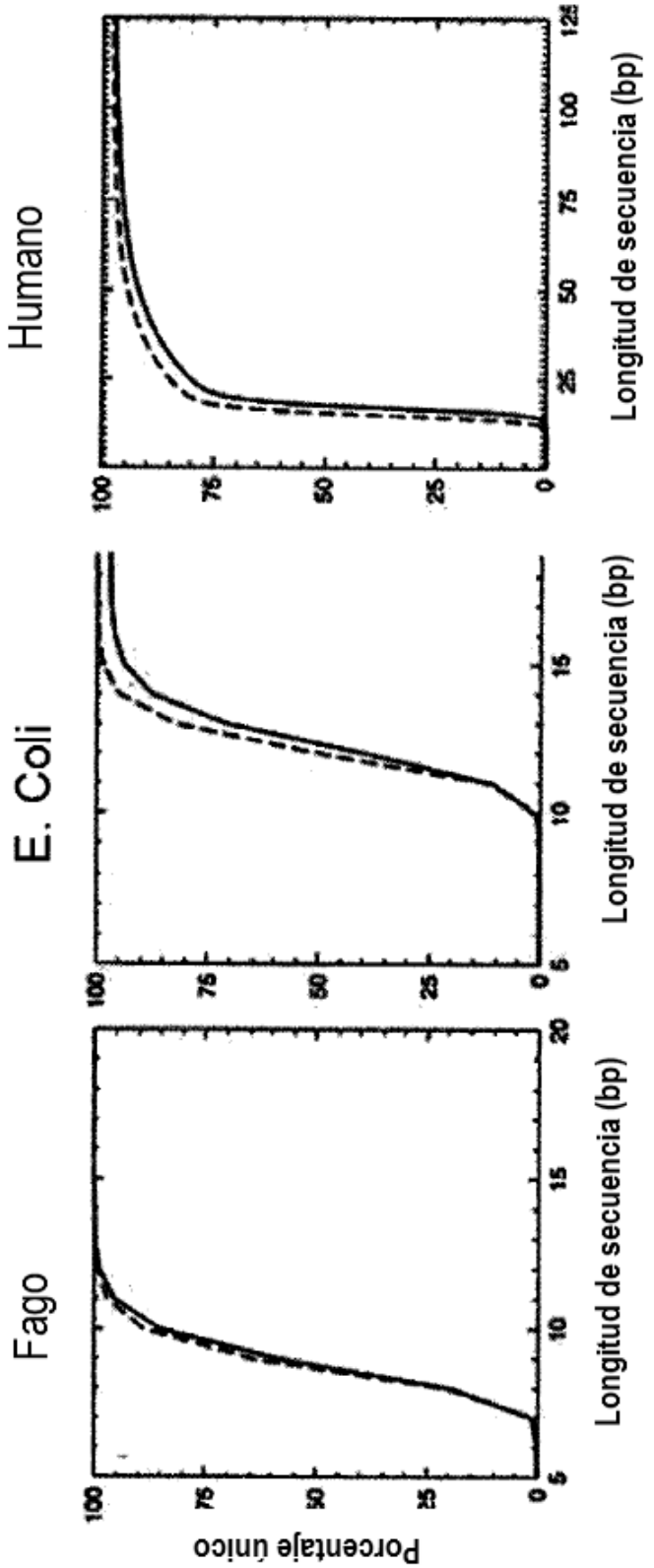
Fago
E. Coli
Humano
Porcentaje único
Longitud de secuencia (bp)
Longitud de secuencia (bp)
Longitud de secuencia (bp)
100
75
50
25
0
5
10
15
20
25
50
75
100
125

FIG. 9

FIG. 10